Spring 2014

# Three essays on the labor market

Varun Kharbanda
*University of Iowa*

# THREE ESSAYS ON THE LABOR MARKET

by

Varun Kharbanda

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Economics
in the Graduate College of
The University of Iowa

May 2014

Thesis Supervisor: Professor George R. Neumann

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

---

PH.D. THESIS

---

This is to certify that the Ph.D. thesis of

Varun Kharbanda

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Economics at the May 2014 graduation.

Thesis Committee: _____
George R. Neumann, Thesis Supervisor

_____
Antonio F. Galvao

_____
N. Eugene Savin

_____
Forrest D. Nelson

_____
Douglas V. DeJong

I would like to dedicate my work to my parents Smt. Kamlesh Kharbanda and Shri. Subhash Chand Kharbanda. Your love and blessings kept me going.

## ACKNOWLEDGEMENTS

I would like to thank my advisor Prof. George R. Neumann along with my committee members, Prof. Antonio F. Galvao, Prof. N.Eugene Savin, Prof. Forrest D. Nelson, and Prof. Douglas V. DeJong. Their guidance and blessings helped me learn and do quality research. A special thanks to Alan Gelder for continuous feedback on my writing and work. My other friends and collegues including, but not limited to, Dohyoung Kwon, Michael Andrews, Huiyi Guo, Prof. Alexandra Nica, Michael Sposi, Husnain Ahmad, Adriana Gama, Chander Kochar, K. E. Carlson, and many more who helped me while working on my thesis.

## ABSTRACT

Using a three-essay approach, I focus on two issues related to the labor market: the effect of changes in regulatory costs on informal sector employment, and the role of endogeneity in the relationship between education and earnings.

In the first essay, I analyze the implications of regulatory costs on skill-based wage differences and informal sector employment. I use a two sector matching model with exogenous skill types for workers where firms have sector-specific costs and workers have sector-specific bargaining power. In general, there are multiple equilibria possible for this model. I focus on the equilibrium that best resembles the situation in the developing countries of sub-Saharan Africa and southern Asia. My results show that government policies which reduce regulatory costs decrease unemployment, earnings inequality, and the fraction of skilled workers in the informal sector. The different types of regulatory costs affect the skill premium differently and non-monotonically.

In the second essay, I test the hypothesis of linearity in returns to education in the Mincer regression with endogenous schooling and earnings. I estimate the marginal rate of return to education using a polynomial model and a semiparametric partial linear model based on the standard Mincer regression. To perform the analysis, I use a control function approach for IV estimation with spousal and parental education as instruments. Results suggest that estimates not accounting for endogeneity understate returns at the tails of the education spectrum and overstate returns for education levels between middle-school and college.

In the third essay, I empirically test the claim of Mookherjee and Ray (2010), based on a theoretical model of skill complexity, that "the return to human capital is endogenously nonconcave." I estimate the functional form of returns to education for India using a semiparametric partial linear model based on the standard Mincer regression. Marginal returns are estimated to test the nonconcavity of the functional form under both exogenous and endogenous schooling assumptions. My results show that the marginal rate of return declines during primary education and increases until high school, followed by stable returns for college and higher studies. However, the test of robustness of the functional form based on uniform confidence bands fails to reject the presence of nonconcavity in returns to education for India. This lends support to the claim of Mookherjee and Ray (2010).

## TABLE OF CONTENTS

## LIST OF TABLES

Table

## LIST OF FIGURES

Figure

# CHAPTER 1
# INTRODUCTION

Economists are interested in many issues related to the labor market such as, returns to education, unemployment, government regulations, marriage, and fertility. In my work, I focus on two of these: the skill-based effect of regulatory cost on the informal sector, and the relationship between education and wages under endogeneity. Through three independent essays, I contribute to the current knowledge on these two issues.

The rest of this chapter is divided into three sections. Each section briefly discusses the motivation, methodology, and contribution of my thesis.

## 1.1 Motivation

The two most important factors affecting the labor market are education and government policies. By observing statistics for the Indian labor market, one cannot ignore two facts. First, more than 70% of the workforce is engaged in the informal sector, which is comprised of legal economic activities that are unregulated by the government. Second, over 80% of the workforce does not have a high school degree. These facts motivate my thesis work on the labor market.

Like in India, informal sector employment is a significant part of the workforce in other developing countries. Governments use policies related to the labor market and the product market to improve the welfare of their people. But the complexity of the informal sector and its larger impact on social welfare makes designing effective

policies a challenge. By my work in the first essay, I try to simplify the observed complexity in the relation between wages and regulatory costs in the informal sector.

For my work on the latter part of my thesis, studies suggest that returns to education vary across geographic areas and demographic groups, and the information of perceived returns to education helps individuals optimize their schooling choice. Moreover, with the information on returns to education at different schooling levels, policymakers also get some direction in fund allocation across schooling levels. This increases my interest in the returns to education literature.

The linearity of returns to education makes the work easy for policymakers. However, the assumption of linearity in returns to education at different levels of schooling has been challenged by Hungerfor and Solon (1987). Card and Krueger (1992) claim that returns to education are approximately linear using 1980 US census data. Heckman et al. (1996) challenge this claim and Heckman et al. (2008) reject this assumption under exogenous schooling. The test under endogenous schooling is as yet undone. Through my work, I fill this gap.

In addition to non-linearity, the recent literature documents the convexity of the wage-education profile in the US and Latin American countries. Recently, Mookherjee and Ray (2010) claim that the returns to human capital are endogenously nonconcave. The empirical test of this claim may lead to a better understanding of challenges faced by individuals in the labor market and better theoretical models to account for these facts.

## 1.2 Methodology

The literature that studies the effect of regulations related to the informal sector in developing countries is limited. Existing theoretical models are designed for developing economies in Latin America and Eastern Europe; however, models for Africa and Asia are poorly designed. Part of the reason is that earnings in the informal sector in Asia and Africa can be higher than earnings in the formal sector due to high search costs for skilled workers.

To develop a general model and to simplify the structure of the informal sector with respect to productivity levels and wage formation under heterogeneous skill types, I construct a matching model with heterogeneous workers in both sectors. This model is the first part of my dissertation, and I use it to analyze the role of regulatory costs in the labor and product markets on the skill premium and skill composition of workers in the informal sector in different economic structures.

For my second and third essays, I use the polynomial and semiparametric approaches to estimate the relationship between education and the log of earnings, and test for the linearity assumption under endogenous earnings and schooling. I address the issue of endogeneity by using a control function approach for instrumental variable regression. Based on the availability of data and to be consistent with the prior literature, I use parents' and spouse's education as instrumental variables. The uniform confidence bands, based on a simple bootstrap method, are used to test for statistical significance.

Further, the third essay tests the theoretical claim of Mookherjee and Ray

(2010) on India. I assume education as an indicator of human capital and use the semiparametric approach to identify the functional form of returns to education.

### 1.3 Contribution

In my first essay, I specifically study the equilibrium under restrictions that are similar to the conditions observed in Asian and African economies. I contribute to the literature by extending the two sector matching model under sector-specific costs and bargaining power to a model with heterogeneously skilled workers.

Through the last two essays, I contribute in the identification of the functional form of returns to education by using a semiparametric approach. Further, I address the issue of endogenous schooling in the setting of partial linear models. This adds to the literature studying the changes in returns to education. The knowledge of returns to education in terms of money, health, and happiness for different schooling levels and training programs gives people guidance on how to appropriately invest resources.

The work documented in this thesis helps further the understanding of the informal sector and returns to education.

# CHAPTER 2
# WAGE DIFFERENCES, REGULATORY COSTS, AND THE INFORMAL EMPLOYMENT SECTOR

## 2.1 Introduction

The informal sector is comprised of business activities outside the purview of regulations due to limitations of government size. In many developing countries, especially in southern Asia and parts of Africa, this sector accounts for the majority of employment (see Figure 2.1). Given the size of the informal sector in these countries, business and labor policies that do not account for the informal sector may not serve their intended purpose. I study the effects of regulatory costs on the informal sector as well as the corresponding effects on inequality, unemployment, and the wage premium for high-skilled workers.[1] Regulatory costs can be separated into product market restrictions (PMR) and labor market restrictions (LMR). The main contribution of this essay is to develop the understanding of PMR and LMR in a model with heterogeneous workers. This will help policy makers understand how these restrictions impact wages and the relative proportions of skilled and unskilled workers in each sector. For developing countries with a large informal sector, as is the case in southern Asia, my analysis shows that increases in regulatory costs are coupled with increases in inequality, unemployment, the fraction of high-skilled workers in the informal sector, and the size of the informal sector. Output also decreases when

[1]Regulatory costs can be defined as the cost of starting and maintaining an organized firm.

regulatory costs increase. The skill premium in both sectors follows a 'U' shape for changes in PMR but a hump shape for changes in LMR. Reductions in LMR reduce wages in the formal sector, while informal sector wages follow a hump shape. Wages increase in both sectors with reductions in PMR.

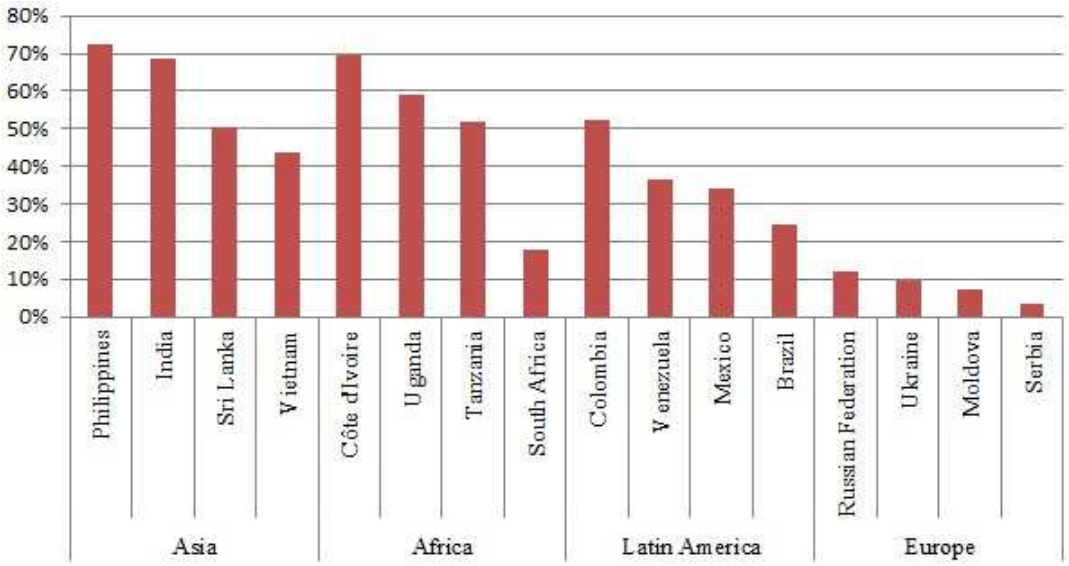


Figure 2.1: The percentage of non-agriculture workers in the informal sector. Source:- ILO, 2012 report.

The nature of the informal sector "differs significantly among countries with different structure" (Gërxhani, 2004). This makes the work for researchers more interesting, as they have to model the informal sector differently for different economic structures. For instance, Albrecht et al. (2009) analyze the effects of labor market policies, such as payroll taxes and severance taxes in the formal sector of an economy with a significant informal sector using a search and matching model. However, they

caution that their work focuses on Latin American economies and is not applicable to Africa without substantial modification. One possible reason for this is the high rate of literacy throughout Latin America, which is absent in many African and southern Asian economies.

High literacy rates make working in the informal sector largely a matter of choice in Latin America. For African and Asian economies, working in the informal sector is typically more a matter of survival. Thus, individuals try to find work in either sector to survive and the informal sector provides employment with higher probability (Gollin, 2008; Bernera et al., 2012; Xaba et al., 2002). These features of the informal sector in such economies require a model where a worker, irrespective of his skill, works in any job type.

The model presented in this essay is dynamic and has five different types of equilibrium. I focus on economies in which the formal sector offers different types of jobs and any worker can be in either sector. This essay attempts to answer the questions of how PMR and LMR affect each sector's wages and skilled worker participation.

The next section formalizes definitions of the informal sector and regulatory costs, highlighting the literature on these topics. Section 2.3 presents the basic structure of the model and describes the steady state equilibrium. Section 2.4 characterizes the other equilibria. In particular, it explains the pure cross-matching equilibrium characterizing developing countries in Asia and Africa. Different policies are analyzed through comparative statics in Section 2.5. Section 2.6 concludes.

## 2.2 Definitions and Previous Literature

The informal sector consists of small family-run businesses, street vendors, non-industrial farms, and other enterprises that are small enough to avoid extensive government documentation and therefore are allowed to be unregulated by the government.[2] Due to minimal documentation, researchers and governments commonly use firm size to distinguish between the formal and informal sectors.[3] The threshold size varies from country to country. In India, for example, a private enterprise is considered part of the informal sector if it employs less than ten workers and uses electricity, 20 if it does not. For Bolivia and Mexico, companies with fewer than six employees are included in the informal sector. Similarly, the threshold number is ten in Kenya, 20 in Sudan, and five in Central American economies. Recent research has used similar definitions for the informal sector (See Henley et al., 2006). These informal enterprises are not required to register with the government and are not under the scrutiny of government agencies (such as account controllers and labor ministries). Informal sector enterprises must, however, adhere to standard civic laws and pay taxes on their self-reported earnings.

Historically, in the absence of any regulations an economy is informal by definition. The formal sector began with the introduction of regulations and government

[2]The International Labor Office defines the informal sector as "private un-incorporated enterprises, which produce at least some of their goods or services for sale or barter, employ less than a specified number of employees, do not maintain complete books of accounts and are not registered" (Raveendran and Naik, 2006).

[3]The use of firm size to identify the informal sector means that ultimately we cannot empirically distinguish between the well-known firm-size effect and the formal/informal sector effect.

control over some aspects of the economy. Through regulations, governments try to provide a safe and stable environment for workers. Employment opportunities in the informal sector are therefore characterized by the absence of job security, flexible wages, and labor-intensive work. The majority of workers tend to be self-employed and derive their earnings from daily work. For example, poor hawkers and traders in India borrow money in the morning from a private lender to buy goods to sell at profit and repay the borrowed money by the evening. Workers in the informal sector may also work in a small enterprise, provide a service to earn tips, work as a household servant, provide informal transportation, or take care of cattle or machinery. Most of these tasks require no prior training, experience, or a significant amount of capital.

Unlike the formal sector, there is considerable variation in profits and earnings over time in the informal sector. Workers and firms tend to stay in business longer as the are willing to work at very low or negative economic profits.[4] The informal sector also responds to economic shocks much faster than the formal sector due to its flexibility and autonomy. These attributes make it difficult for economists to model.

Regulatory costs consist of product market restrictions (PMR) and labor market restrictions (LMR). PMR is defined as the cost that a firm has to spend in order to adhere to government policies while doing business. This includes a cost of registration, license fees and accounts maintenance. LMR are laws affecting only a worker's ability to bargain for wages through labor unions or otherwise. Since organized firms are regulated and can be scrutinized, they must obey these laws. PMR and LMR

[4]Social relationships are likely an important factor in this longevity.

within a country may change over time due to political and economic factors. A heavily regulated market may be created to benefit certain interest groups. Inefficiencies in government offices also add to regulatory costs. In general, any policy that makes it easier for a business to formally register and maintain its registration decreases PMR, and any labor law that provides more security to workers tends to increase the LMR.

High regulatory costs are often associated with low growth, high unemployment, high inequality, and large informal sector. Botero et al. (2004) study the labor regulations across 85 countries and conclude that stringent labor regulations "[are] associated with lower labor force participation and higher unemployment, especially of the young." Pagés-Serra (2000) uses data from the Latin American labor market and shows that job security regulations which tend to increase the bargaining power of workers in the formal sector reduce employment and increase inequality across workers. Besley and Burgess (2004), based on 1952–1992 data for Indian states, show that pro-worker regulations lower output, employment, investment, and productivity in the formal manufacturing sector, while output in the informal manufacturing sector increases. Ahsan and Pagés (2009) conduct empirical work on the Indian manufacturing sector and find that laws that increase the cost of dispute resolution between workers and firms substantially reduce formal sector employment and output. A cross-country regression analysis by Loayza et al. (2005) suggest an association between high levels of regulation and low growth. Djankov et al. (2006) find that better regulations positively affect annual growth rates. Antunes and de V. Cavalcanti (2007)

show that regulatory costs "account for differences in the size of the informal sector between United States and Mediterranean Europe."

My work is inspired by Charlot et al. (2012) and Albrecht et al. (2009). It extends the matching model of Albrecht and Vroman (2002) by adding an informal sector to an organized economy with sector-specific hiring costs and bargaining power. Charlot et al. (2012) study the implications of PMR and LMR in an economy with an informal sector and homogeneous workers. Boeri and Garibaldi (2007) and Albrecht et al. (2009) use a matching model with workers uniformly distributed over skill levels. Boeri and Garibaldi (2007) focus on explaining the "shadow puzzle," in which the size of the informal sector increases "in spite of improvements in technologies detecting tax and social security evasion." On the other hand, Albrecht et al. (2009) are more concerned with the implications of payroll taxes and severance taxes on the formal sector. Albrecht and Vroman (2002) use a one-sector matching model with two types of workers to study the effects of skill-based technological changes on wage differences.

On the other hand, Charlot et al. (2012), Zenou (2008), and Ulyssea (2010) study regulatory costs in an economy with an informal sector using a search friction model with homogeneous workers. Charlot et al. (2012) assume monopoly power in the goods market for both sectors, but firm size is endogenous. Zenou (2008) models the formal sector as a monopoly and the informal sector as a competitive market; in contrast, Ulyssea (2010) views both sectors as competitive. In any economy, we see many more skill types than are considered in these models. There are always highly skilled individuals who work only in the high-skilled jobs of the formal sector.

Similarly, the low-skilled individuals are highly likely to work in the informal sector. Additionally, individuals differ by discount rate, which accounts for the varying degrees of patience in searching for a job. These models cannot address issues such as skill based wage differences, participation rates, and unemployment.

My model follows Boeri and Garibaldi (2007) and Albrecht et al. (2009) by allowing for heterogeneous workers, and focuses only on the matching frictions in the labor market. Further, due to the intrinsic nature of the informal sector, I assume that a worker faces an exogenous matching rate in the informal sector where he produces at an exogenous output level independent of his skill type. To capture the fact that high-skill workers earn more than low-skill workers do, even in the informal sector, I allow for wage bargaining in both sectors. I also do not restrict firm entry and exit for the informal sector. I contribute to the existing literature by examining the effect of PMR and LMR on the informal sector in developing economies. My results support the empirical findings and confirm the theoretical results of Albrecht et al. (2009), as well as the complementing results of Charlot et al. (2012).[5] I also provide new theoretical results on the participation of high-skill workers in the informal sector.

## 2.3 Model

I use the matching model suggested by Mortensen and Pissarides (1994) in a two-sector framework to study the effects of government policies related to PMR and

[5]The empirical findings of Botero et al. (2004), Ahsan and Pagés (2009), Pagés-Serra (2000), Antunes and de V. Cavalcanti (2007), Besley and Burgess (2004), Djankov et al. (2006), and Loayza et al. (2005) are supported.

LMR. The two sectors are the formal and informal sectors. The model uses sector-specific cost to track the PMR and sector-specific bargaining power to track the LMR. The formal sector has a higher sector-specific cost than does the informal sector and may have higher sector-specific bargaining power in deciding wages. This reflects the fact that organized firms are bound by regulations; informal sector firms can bypass these regulations easily. Matching friction results from differences in workers' skill levels and the skills required by different jobs. This allows me to analyze the effects of changes in PMR and LMR on wages, unemployment, and vacancies offered in the economy.

I study a continuous time horizon, where workers are born with an exogenously given skill level and then live forever. I assume all workers are risk-neutral, and the mass of all workers is normalized to one. For simplicity, assume a fraction $P$ of workers has a low skill level, denoted by $s_1$, and a fraction $1-P$ has a high skill level, denoted by $s_2$. Each worker is endowed with one unit of time, during this time he works in either sector, or searches for a job. I use the notation $n_{sF}$, $n_{sI}$, and $u_s$ for worker's time spent in the formal sector, the informal sector, and unemployment, respectively. Here, the index $s \in \{L, H\}$ represents low and high skill levels, respectively. A job is described by its skill requirement, and is either filled by a worker or kept vacant. Filled jobs become vacant at an exogenous rate, $\delta$. Everybody in this economy faces common discount rate $r$, and the unemployment benefit is $b$.The technology in this

economy is given by:

$$x_j(s,y) = \begin{cases} y & \text{if } s \geq y \ \& \ j = F \\ y_0 & \text{if } j = I \\ 0 & \text{otherwise,} \end{cases}$$

where $j \in \{I, F\}$ identifies the sector: formal or informal, $s$ is the skill level of the worker, $y$ is the skill requirement of the job, and $y_0$ is the output produced by a worker in the informal sector. The informal sector technology produces output level $y_0$ irrespective of a worker's skill. I assume jobs in the informal sector are provided at the exogenous rate $\alpha$ and informal sector wages are bargained independently. Unlike informal sector firms, I assume that formal sector firms face competition through free entry and exit.

Firms bear a fixed cost for offering a job, which can be thought of as the cost of advertising a vacancy. Once the vacancy is filled, it represents the cost of resources provided to workers to fulfill their job. I assume this cost is constant and specific to each sector. The cost of a vacant job in the informal sector is $c_I$, and $c_F$ in the formal sector. The cost to a firm in sector $j$ for employing a worker of type $s$ for a job requiring skill level $y$ is $\omega_j(s,y) + c_j$, where $\omega_j(s,y)$ is the wage paid to the worker in sector $j$. The formal sector is regulated, and a large amount of resources is spent in posting vacancies and supporting current employees compared to the cost in the informal sector. Fixed costs differentiate the sectors and are higher in the formal sector. With a zero profit condition in the formal sector, a fixed job destruction rate, and the distribution of skills across workers, jobs in the formal sector will have one of two skill requirements, $y_1 = s_1$ or $y_2 = s_2$. A worker in sector $j$ has bargaining power $\beta_j$. Workers participate in a Nash bargaining game with an employer. The informal

sector offers one type of job, and workers bargain independently for wages.

Unemployed workers encounter open vacancies from each sector randomly. An unemployed worker comes across an informal sector vacancy at a rate $\alpha$. A formal sector job arrives according to the function $M(u, v)$, where $u$ is the unemployment rate and $v$ is the measure of vacancies. Assume $M(u, v)$ is characterized by constant returns to scale, and can be written as

$$M(u, v) = M(1, \frac{v}{u})u = m(\theta)u,$$

where $\theta = v/u$.

The arrival rate of open vacancies from the formal sector is therefore given by $m(\theta)$. Using standard assumptions, $m'(\theta) > 0$ implies that the matching rate is increasing in open vacancies per unemployed worker. In the limit, $\lim_{\theta \to 0} m(\theta) = 0$, implying that the matching rate converges to zero as the ratio of open vacancies to unemployed workers goes to zero. All workers find formal sector vacancies at the same rate, but low-skilled workers do not qualify for high-skilled job openings. Let $\phi$ be the fraction of formal sector vacancies that require a low skill level. The effective arrival rate of employment opportunities in the formal sector for low-skill workers is then $\phi m(\theta)$. The arrival rate of unemployed workers to a formal sector vacancy is given by $m(\theta)/\theta$. I assume that this rate is decreasing in $\theta$, which implies that as ratio of vacancies to unemployed workers increase, the arrival rate of unemployed workers to vacant jobs declines. In the limit, $\lim_{\theta \to 0}[\frac{m(\theta)}{\theta}] = \infty$, implying that the arrival rate of unemployed worker to a vacant job converges to infinity as the number of open vacancies per unemployed worker goes to zero. Since high-skilled vacancies cannot

be filled by low-skilled workers, let $\gamma = \frac{u_L P}{u_L P + u_H(1-P)}$ be the fraction of unemployed workers who are low-skilled. Then the effective arrival rate of high-skilled workers to high-skilled vacant jobs becomes $(1-\gamma)(\frac{m(\theta)}{\theta})$.

Given the values $\{P, s_1, s_2, \delta, r, c_I, c_F, y_0, \alpha, \beta_I, \beta_F\}$, together with the matching function $m(\theta)$, the steady-state equilibrium is defined as a collection of eight variables $\{\theta, \phi, u_L, u_H, n_{LF}, n_{LI}, n_{HF}, n_{HI}\}$ that satisfy the following conditions:

- Workers' and firms' choices are optimal given the actions of the other agents.
- In the long run, vacancy creation satisfies the zero-profit condition in the formal sector (given free entry and exit).
- The flow of low-skilled workers into and out of unemployment in each job type is equal, and likewise for high-skilled workers.
- The values $\{\phi, u_L, u_H, n_{LF}, n_{LI}, n_{HF}, n_{HI}\} \in [0,1]^7$ and $\theta > 0$.

To reiterate, the rates of job creation and matching are exogenous within the informal sector.

### 2.3.1 Match Formation and Wages

Workers' skill levels match job requirements when it is mutually beneficial for both the worker and the firm to enter into employment. The joint surplus when they match should be higher than the sum of their individual benefits when not matched. Formally, a match is formed if:

$$N_j(s,y) + J_j(s,y) \geq U(s) + V_j(y),$$

where $N_j(s,y)$ is the value of employment for a worker of type $s$ in a job of type $y$ for sector $j$; $J_j(s,y)$ is the value to the employer of filling a job of type $y$ with a worker of type $s$ for sector $j$; $U(s)$ is the value of unemployment for a worker of type $s$ and $V_j(y)$ is the value of a vacancy of type $y$ in sector $j$. When a match is formed, the wage $\omega_j(s,y)$ is such that the following Nash bargaining condition is satisfied:

$$N_j(s,y) - U(s) = \beta_j[N_j(s,y) + J_j(s,y) - U(s) - V_j(s)],$$

by which the benefit to a worker from employment is the same as the fraction $\beta_j$ of the total surplus. Here, $\beta_j$ is the exogenous share of the surplus that workers receive in sector $j$. Let $r$ be the common discount rate for workers and firms, and denote the unemployment benefit by $b$. In continuous time, a worker's benefit from employment is the sum of his instantaneous wage flow less the expected loss of becoming unemployed:

$$rN_j(s,y) = \omega_j(s,y) - \delta[N_j(s,y) - U(s)].$$

Simplified, this becomes:

$$N_j(s,y) = \frac{\omega_j(s,y) + \delta U(s)}{(r+\delta)}. \tag{2.1}$$

Similarly, the value of the job to the firm, conditional on $s > y$, is the sum of the instantaneous profit from employing a worker with skill level $s$ less the potential capital loss due to job dissolution:

$$J_j(s,y) = \frac{y - \omega_j(s,y) - c(y) + \delta V_j(y)}{(r+\delta)}. \tag{2.2}$$

Since, low-skilled workers in the formal sector can only fill low-skill jobs, and the effective arrival rate of low skill jobs is $m(\theta)\phi$, the value of unemployment for low-skilled workers is:

$$rU(s_1) = b + \alpha[\max\{N_I(s_1, s_1) - U(s_1), 0\}] + m(\theta)\phi[N_F(s_1, s_1) - U(s_1)]. \quad (2.3)$$

Hence, the value of unemployment for low-skilled workers is the sum of instantaneous unemployment benefits and the capital gains from acquiring a low-skill job in either sector. For high-skilled workers, the value of unemployment is given by:

$$rU(s_2) = b + \alpha[\max\{N_I(s_2, s_2) - U(s_2), 0\}] + m(\theta)[\phi \max\{N_F(s_2, s_1) - U(s_2), 0\} + (1 - \phi)(N_F(s_2, s_2) - U(s_2))]. \quad (2.4)$$

This is constructed in the same way as is the value of unemployment for low-skilled. However, since a high-skilled worker can work in either low or high-skill jobs, the capital gain is the expected value from being employed in each type of job. The condition, $\max\{N_F(s_2, s_1) - U(s_2), 0\}$, ensure that a high-skilled worker accepts a low-skill job only if it is more beneficial to him than is unemployment.

Finally, the respective values of low-skill and high-skill jobs in the formal sector are summarized as follows:

$$rV_F(s_1) = -c_F + \frac{m(\theta)}{\theta}[\gamma(J_F(s_1, s_1) - V_F(s_1)) + (1 - \gamma)\max\{J_F(s_2, s_1) - V_F(s_1), 0\}], \quad (2.5)$$

$$rV_F(s_2) = -c_F + \frac{m(\theta)}{\theta}(1 - \gamma)[J_F(s_2, s_2) - V_f(s_2)]. \quad (2.6)$$

A vacant job in the formal sector requiring a low skill level can be filled by either low-skill or high-skill workers. Therefore, the expected gain from filling a job in the the next period is the sum of the expected capital gains for the job filled by a low-skill or high-skill worker. This term, $\max[J_F(s_2, s_1) - V_F(s_1), 0]$, indicates whether or not a high-skill worker benefits by taking a low-skill job. Since a high-skill vacancy can only be filled by a high-skill worker, the gain from filling a high-skill job takes only high-skill workers into consideration.

For the informal sector, the value of a vacant job that is available to all workers can be specified similarly:

$$\begin{aligned} rV_I(y_0) = &- c_I + \tau[\gamma(J_I(s_1, y_0) - V_I(y_0)) \\ &+ (1-\gamma)\max\{J_I(s_2, y_0) - V_I(y_0), 0\}], \end{aligned}$$

where $y_0$ is the job type offered in the informal sector and $\tau$ is the exogenous arrival rate of unemployed workers to the vacancy. As mentioned previously, jobs in the informal sector can be filled by either low or high-skilled workers.

By the zero profit condition in the formal sector, $V_F(s_1) = V_F(s_2) = 0$, which simplifies Equation 2.1 to $-c(s) \geq rU(s)$. Conditional on $s \geq y$, Equation 2.2 gives us the wage of the worker:

$$\omega_j(s, y) = \beta_j(y_j - c_j) + (1 - \beta_j)rU(s).$$

Note that this is the weighted average of total gains from employment and the worker's flow value of unemployment. This leads to the three different wages to be paid in the

long run given the zero profit condition:

$$\omega_F(s_1, s_1) = \beta_F(s_1 - c_F) + (1 - \beta_F)rU(s_1),$$

$$\omega_F(s_2, s_1) = \beta_F(s_1 - c_F) + (1 - \beta_F)rU(s_2),$$

$$\omega_F(s_2, s_2) = \beta_F(s_2 - c_F) + (1 - \beta_F)rU(s_2).$$

Under parameter restrictions, the wage of a low-skill worker, $\omega_F(s_1, s_1)$, is the lowest in the formal sector. For the informal sector, I assume a similar wage specification:

$$\omega_I(s_1, y_0) = \beta_I(y_0 - c_I) + (1 - \beta_I)rU(s_1),$$

$$\omega_I(s_2, y_0) = \beta_I(y_0 - c_I) + (1 - \beta_I)rU(s_2).$$

Skilled workers in the informal sector earn more than low-skilled workers in the informal sector because their outside option is more valuable. The wage in the formal sector is highest for high-skilled workers. The long run zero profit condition of $V_F(s_1) = V_F(s_2) = 0$ in the formal sector, combined with Equation 2.2, gives us:

$$J_F(s, y) = \frac{(1 - \beta_F)[y - c_F - rU(s)]}{(r + \delta)}. \tag{2.7}$$

In other words, the value of a filled job to the employer is the discounted value of the employer's share of the surplus. Similarly, Equation 2.1 gives us:

$$N_j(s, y) = \frac{\beta_j[y - c_F - rU(s)]}{(r + \delta)} + U(s). \tag{2.8}$$

The value of a filled job to the employee is the discounted value of the employee's share of surplus over the value of unemployment.

## 2.4 Equilibrium

There are five different types of equilibria based on the respective parameter values. Each equilibrium restricts different worker types to different job types. Table 2.1 shows each equilibrium in terms of the restrictions on sector composition.

Table 2.1: Five equilibria of the model by presence of worker type in different job types.

| Equilibrium Type | Informal Job | Formal Job | |
|---|---|---|---|
| | | Low-skill | High-skill |
| (i) Pure cross-matching | $s_1$,$s_2$ | $s_1$,$s_2$ | $s_2$ |
| (ii) Cross-matching | $s_1$ | $s_1$,$s_2$ | $s_2$ |
| (iii) Weak ex-post segmentation | $s_1$,$s_2$ | $s_1$ | $s_2$ |
| (iv) Ex-post segmentation | $s_1$ | $s_1$ | $s_2$ |
| (v) Pure ex-post segmentation | $s_1$ | - | $s_2$ |

Note:- $s_1$ and $s_2$ represent the presence of low-skill or high-skill workers, respectively, in the given job column.

That is, the five equilibria are the following: (i) Pure cross-matching equilibria, defined as ones in which both types of workers are present in both sectors and high-skilled workers are in both types of job in the formal sector. (ii) Cross-matching equilibria in the formal sector with pure segmentation in the informal sector, in which the results are as in (i) without high-skill workers in the informal sector; low-skilled workers can work in either sector. (iii) Weak ex-post segmentation equilibria in the formal sector with cross-matching in the informal sector, in which a high-skilled worker is present in both sectors and not present in low-skill jobs in the formal sector. Again, low-skilled workers can work in either sector. (iv) Ex-post segmentation

equilibria, in which high-skilled workers prefer high skill jobs in the formal sector only and low-skilled workers can work in formal sector low skill jobs or in the informal sector. (v) Pure ex-post segmentation equilibria, in which the formal sector only offers high skill jobs, and thus high-skill workers are in the formal sector only and low-skill workers are in the informal sector.

Each equilibrium can also be seen as the combination of the following three assumptions. We allow (a) the high-skill worker to work in low-skill job in the formal sector: $N_F(s_2, s_1) - U(s_2) \geq 0$; (b) the high-skill worker to work in the informal sector: $N_I(s_2, y_0) - U(s_2) \geq 0$ ; (c) the low-skill worker to work in low-skill jobs in the formal sector: $N_F(s_1, s_1) - U(s_1) \geq 0$. These three conditions are further simplified as (a) $s_2 - c_F \geq rU(s_2)$, (b) $y_0 - c_I \geq rU(s_2)$, and (c) $s_1 - c_F \geq rU(s_1)$, respectively. The parameter value can be used to control these assumptions.

Since the data for developing countries show a large presence of both types of workers in both sectors, I focus on the first equilibrium type: pure cross-matching equilibria. Developed countries, on the other hand, typically have a large presence of high-skill workers in the high skill jobs of the formal sector, as depicted in equilibria (iv) and (v).

### 2.4.1 Pure Cross-matching Equilibria

Equilibrium in the pure cross-matching case consists of steady state values for eight endogenous variables $\theta, \phi, u_L, u_H$, $n_{LF}$, $n_{LI}$,$n_{HFL}$, and $n_{HFH}$. The last two variables represent the time spent by high-skill workers in the formal sector working

in low-skill and high-skill jobs, respectively. These variables should satisfy steady state conditions. Table 2.2 provides a summary of all notation for this equilibrium.

Table 2.2: Notation used to define this equilibrium

| Notation | Description |
|---|---|
| $\delta$ | Job destruction rate |
| $r$ | Interest rate |
| $\gamma$ | Fraction of low-skilled workers in unemployment pool |
| $\theta$ | Number of vacancies per unemployed |
| $\phi$ | Fraction of low-skilled jobs offered in the formal sector |
| $\beta_j$ | Bargaining power of a worker in sector $j$ |
| $\alpha$ | Matching rate in the informal sector |
| $y_0$ | Output level produced by a worker in the informal sector |
| $\omega_j(s,y)$ | Wage of a skill-type $s$ worker in a sector $j$ job that requires $y$ skill |
| $U(s)$ | Value of unemployment for a worker with skill level $s$ |
| $u_s$ | Unemployment rate among workers with skill level $s$ |
| $m(\theta)$ | Matching rate in the formal sector |
| $J_j(s,y)$ | Value to a sector $j$ employer of a $y$ skill job and $s$ skill-type worker |
| $N_j(s,y)$ | Value to the $s$ skill-type worker of a sector $j$ job that requires $y$ skill |
| $V_j(y)$ | Value of a vacancy requiring $y$ skill in sector $j$ to the employer |
| $n_{sj}$ | Fraction of workers, or time of a worker, with skill level $s$ in sector $j$ |
| $n_{sjy}$ | Fraction of workers with skill level $s$ in jobs that require skill $y$ in sector $j$ |
| $c_j$ | Cost of job in sector $j$ |

This equilibrium requires all three assumptions: (a) The high-skill worker's value of working in a low-skill job in the formal sector is higher than the value of unemployment: $N_F(s_2,s_1)-U(s_2)\geq 0$; (b) The high-skill worker's value of working in the informal sector job is higher than the value of unemployment: $N_I(s_2,y_0)-U(s_2)\geq 0$; (c) The low-skill worker's value of working in a low-skill job in the formal sector is higher than the value of unemployment: $N_F(s_1,s_1)-U(s_1)\geq 0$.

Since a low-skilled worker can either look for a job or work in one of the two

sectors, his time must satisfy the following constraint:

$$u_L + n_{LF} + n_{LI} = 1.$$

A high-skilled worker can spend time looking for a job or working in either sector. A high-skilled worker in the formal sector could be working in a low-skill or high-skill job, represented by the following equation:

$$u_H + n_{HFL} + n_{HFH} + n_{HI} = 1.$$

Using the condition that, in the formal sector, the flow of low-skilled workers out of unemployment equals the flow of low-skilled workers into unemployment, we get:

$$\phi m(\theta) u_L = \delta n_{LF}.$$

Here, $u_L$ shows the total mass of low-skilled unemployed workers, and the left-hand side represents the flow of low-skilled workers receiving employment in formal sector low-skill jobs. On the right-hand side, $n_{LF}$ represents the low-skilled workers employed in the formal sector and $\delta$ shows the fraction of workers who lose their jobs. Similarly for the informal sector, we get:

$$\alpha u_L = \delta n_{LI},$$

where $\alpha$ is the exogenous matching rate for jobs in the informal sector and $n_{LI}$ shows low-skilled workers in the informal sector. For high-skilled workers in the informal sector, the flow is matched by the following equation:

$$\alpha u_H = \delta n_{HI}.$$

High-skilled workers can work in either job in the formal sector. The steady state condition for the equal flow of workers out of and into unemployment gives the following condition for high-skilled workers in low-skill jobs:

$$\phi m(\theta) u_H = \delta n_{HFL}.$$

Since the probability of a high-skilled worker getting a low-skill job is the same as a low-skilled worker getting a low-skill job, the match rate is $\phi m(\theta)$. The term $u_H$ represents the high-skilled unemployed, so the left-hand side of this equation shows the flow of high-skilled workers becoming employed. The right-hand side constitutes the flow of high-skilled workers into unemployment. Since the match rate of high-skilled workers in high-skilled formal sector jobs is $(1-\phi)m(\theta)$, a condition for the equal flow of high-skilled workers into and out of high-skilled jobs is given by:

$$(1-\phi) m(\theta) u_H = \delta n_{HFH}.$$

Using these equations, we solve for the values of $u_L$, $n_{LF}$, $n_{LI}$, $u_H$, $n_{HFL}$, $n_{HFH}$, and $n_{HI}$ as functions of $\phi$ and $\theta$:

$$u_L = \frac{\delta}{\alpha + \delta + \phi m(\theta)}, \tag{2.9}$$

$$n_{LI} = \frac{\alpha}{\alpha + \delta + \phi m(\theta)}, \tag{2.10}$$

$$n_{LF} = \frac{\phi m \theta}{\alpha + \delta + \phi m(\theta)}, \tag{2.11}$$

$$u_H = \frac{\delta}{\alpha + \delta + m(\theta)}, \tag{2.12}$$

$$n_{HI} = \frac{\alpha}{\alpha + \delta + m(\theta)}, \tag{2.13}$$

$$n_{HFL} = \frac{\phi m\theta}{\alpha + \delta + m(\theta)}, \tag{2.14}$$

$$n_{HFH} = \frac{(1-\phi) m\theta}{\alpha + \delta + m(\theta)}. \tag{2.15}$$

For finding $\theta$ and $\phi$, I use the free entry and exit conditions of the formal sector: $V_F(s_1) = V_F(s_2) = 0$ . Using this condition, together with Equations 2.5, 2.6, and 2.7, I get:

$$\begin{aligned} -c_F + \frac{m(\theta)}{\theta}\frac{(1-\beta_F)}{(r+\delta)}[\gamma(s_1 - c_F - rU(s_1)) \\ + (1-\gamma)(s_1 - c_F - rU(s_2))] = 0, \end{aligned} \tag{2.16}$$

$$-c_F + \frac{m(\theta)}{\theta}\frac{(1-\beta_F)}{(r+\delta)}(1-\gamma)[s_2 - c_F - rU(s_2)] = 0. \tag{2.17}$$

I further solve Equations 2.3 and 2.4 using Equation 2.8 to get:

$$rU(s_1) = \frac{b(r+\delta) + \alpha\beta_I(y_0 - c_I) + m(\theta)\phi\beta_F(s_1 - c_F)}{r + \delta + \alpha\beta_I + m(\theta)\phi\beta_F}, \tag{2.18}$$

$$rU(s_2) = \frac{b(r+\delta) + \alpha\beta_I(y_0 - c_I) + m(\theta)\beta_F[(\phi s_1 + (1-\phi)s_2) - c_F]}{r + \delta + \alpha\beta_I + m(\theta)\beta_F}. \tag{2.19}$$

These equations show that the flow value of being unemployed is a weighted average of the benefit from unemployment and the gain from potential employment in either of the sectors. Since high-skilled workers can work in either job types in the formal sector, the gain from employment to high-skilled workers is the expected value from employment in the formal sector.

Substituting Equation 2.16 and 2.16 into the zero profit condition, $V_F(s_1) = V_F(s_2) = 0$, yields:

$$\gamma(s_1 - c_F - rU(s_1)) = (1-\gamma)(s_2 - s_1).$$

Using Equation 2.18, I get:

$$\gamma[(r+\delta)(s_1 - c_F - b) + \alpha\beta_I(s_1 - y_0 - c_F + c_I)] = (1-\gamma)[(s_2 - s_1)(r + \delta + \alpha\beta_I + m(\theta)\phi\beta_F)].$$

Adding $(1-\gamma)[(r+\delta)(s_1 - c_F - b) + \alpha\beta_I(s_1 - y_0 - c_F + c_I)]$ to both sides gives

$$[(r+\delta)(s_1 - c_F - b) + \alpha\beta_I(s_1 - y_0 - c_F + c_I)] = (1-\gamma)[(r+\delta)(s_2 - c_F - b) + \alpha\beta_I(s_2 - y_0 - c_F + c_I) + m(\theta)\phi\beta_F(s_2 - s_1)]. \quad (2.20)$$

Note that the left-hand side is constant and the right hand side is a function of $\gamma$, $\theta$, and $\phi$. Moreover, $\gamma$ is the fraction of unemployed workers who are low-skilled and

is given by $\gamma = \frac{Pu_L}{Pu_L+(1-P)u_H}$. With Equations 2.9 and 2.12, $\gamma$ can be rewritten as follows:

$$\gamma = \frac{P(\alpha+\delta+m(\theta))}{(\alpha+\delta+m(\theta)(P+(1-P)\phi))}. \tag{2.21}$$

This shows that $\gamma$ decreases in $\phi$ for a given value of $\theta$, which in turn implies that $(1-\gamma)$ increases in $\phi$. The right-hand side of Equation 2.20 is increasing in $\phi$ for a given value of $\theta$. Hence, given the value of $\theta$, there exists a unique $\phi$.

Using $V_F(s_2) = 0$, I get

$$c_F = \frac{m(\theta)}{\theta}\frac{(1-\beta_F)}{(r+\delta)}(1-\gamma)(s_2 - c_F - rU(s_2).$$

Inserting the expression for $rU(s_2)$ from Equation 2.19 gives

$$c_F = \frac{m(\theta)}{\theta}\frac{(1-\beta_F)}{r+\delta}(1-\gamma) \frac{[(r+\delta)(s_2-c_F-b)+\alpha\beta_I(s_2-y_0-c_F+c_I)+m(\theta)\phi\beta_F(s_2-s_1)]}{r+\delta+\alpha\beta_I+m(\theta)\beta_F}.$$

Finally, using Equation 2.20 yields

$$c_F = \frac{m(\theta)}{\theta}\frac{(1-\beta_F)}{r+\delta}\frac{[(r+\delta)(s_1-c_F-b)+\alpha\beta_I(s_1-y_0-c_F+c_I)]}{r+\delta+\alpha\beta_I+m(\theta)\beta_F}. \tag{2.22}$$

The right-hand side of Equation 2.22 is decreasing in $\theta$ for all positive values of $\theta$ and the left-hand side is constant, which gives a unique solution for $\theta$. Now, using

Equations 2.20 and 2.21, it is possible to solve for the values of $\phi$ and $\gamma$. The model can be solved using Equations 2.9 to 2.15. To further analyze the effects of government policies on the market, I present comparative statics results.

## 2.5 Comparative Statics

Product market restrictions (PMR) are controlled by the sector-specific cost of jobs, $c_F$, and labor market restrictions (LMR) by using the sector-specific bargaining power of workers, $\beta_F$ and $\beta_I$. First, I analyze the effect of PMR on the size of the informal sector and wages by changing the value of $c_F$ and maintaining the same LMR in two sectors, $\beta_F = \beta_I$. The equilibrium values of $\{\theta,\ \phi,\ u_L,\ u_H,\ n_{LF},\ n_{LI},\ n_{HF},\ n_{HI}\}$ are solved by assuming the functional value of $m(\theta) = 2\theta^{1/2}$ and the following parameter values: $s_1 = 1$, $s_2 = 1.2$, $P = 2/3$, $b = 0.01$, $\beta_F = \beta_I = 0.5$, $\delta = 0.2$, $r = 0.1$, $\alpha = 0.8$, and $y_0 = 0.9$. These parameters satisfy the conditions for a pure cross-matching equilibrium and most of them are used by Albrecht and Vroman (2002).[6] The value of $y_0$, being less than one, ensures that the total output produced in the informal sector is lower than the value of output produced by low-skilled or high-skilled workers. This reflects the formal sector wage premium documented in the empirical studies.[7]

Since PMR is substantially lower in the informal sector, I fix the cost of creating

[6] Asian and African countries provide few or no unemployment benefits, and generally have high interest rates. To account for this and to satisfy the equilibrium conditions, I change the values of $b$ and $r$ in Albrecht and Vroman (2002) from 0.1 to 0.01 and from 0.05 to 0.1, respectively. The values of $\alpha$ and $y_0$ are determined by a rule of thumb to satisfy the equilibrium conditions.

[7] See Maloney (1999) and Pratap and Quintin (2006)

and maintaining a job in the two sectors: $c_I = 0.05$ in the informal sector and $c_F = 0.4$ in the formal sector. The difference can be thought of as the regulatory costs to offer a vacancy in the formal sector. Then, I decrease the value of $c_F$ from 0.4 to 0.1 to show the effect of reducing PMR. Table 2.3 shows the results for this exercise for an economy where the low-skilled individuals are in the majority with $P = 2/3$. The equilibrium values for $\theta$, $\gamma$, $\phi$, and $u$ as well as wages, skill premiums, the percentage of high-skilled workers in the informal sector, the Gini coefficient for earnings inequality, and the size of the informal sector in terms of workers are presented in Table 2.3.[8]

Table 2.3 also shows that reductions in PMR increase total vacancies per unemployed, $\theta$, indicating more jobs are offered by firms in the formal sector. Overall, this increases the value of unemployment, $U(s)$, and reduces the size of the informal sector. Wages for both types of workers in the formal sector rise because of the increase in output net of sector-specific cost, $(y_j - c_j)$, and the increase in $U(s)$. The later also increases wages for workers in the informal sector. Overall, due to the decrease in unemployment, $u$, and the increase in wages, $\omega_j(s, y)$, the reduction in PMR reduces income inequality and boosts output.

The fraction of vacancies requiring a low skill level in the formal sector, $\phi$, follows a 'hump' shape with the reduction in PMR. It results in an opposite response in the fraction of low-skilled workers in the unemployment pool, $\gamma$, which follows a

[8]The skill premium is calculated by taking the difference between the expected wages of a high-skilled worker and a low-skilled worker. The expected wage for high-skilled workers is a weighted average of wages in each job type. Weights are assigned by the fraction of high-skilled workers in each job type.

Table 2.3: Effect of reducing product market restrictions, $c_F$.

| | $c_F$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0.4 | 0.35 | 0.3 | 0.25 | 0.2 | 0.15 | 0.1 |
| $\theta$ | 0.274 | 0.548 | 0.951 | 1.554 | 2.508 | 4.170 | 7.632 |
| $\phi$ | 0.560 | 0.787 | 0.898 | 0.951 | 0.965 | 0.946 | 0.891 |
| $u$ | 11.6% | 8.84% | 7.1% | 5.86% | 4.89% | 4.05% | 3.27% |
| $\gamma$ | 0.721 | 0.696 | 0.682 | 0.675 | 0.673 | 0.676 | 0.688 |
| $U(s_1)$ | 5.22 | 5.63 | 6.07 | 6.53 | 7.03 | 7.54 | 8.09 |
| $U(s_2)$ | 5.75 | 5.94 | 6.24 | 6.63 | 7.10 | 7.66 | 8.35 |
| Gini coefficient | 0.124 | 0.059 | 0.050 | 0.043 | 0.041 | 0.037 | 0.029 |
| Output | 1.79 | 1.88 | 1.94 | 1.98 | 2.01 | 2.05 | 2.08 |
| Wages | | | | | | | |
| $\omega_F(s_1, s_1)$ | 0.561 | 0.606 | 0.653 | 0.702 | 0.751 | 0.802 | 0.855 |
| $\omega_F(s_2, s_1)$ | 0.587 | 0.622 | 0.662 | 0.706 | 0.755 | 0.808 | 0.867 |
| $\omega_F(s_2, s_2)$ | 0.687 | 0.722 | 0.762 | 0.806 | 0.855 | 0.908 | 0.967 |
| $\omega_I(s_1, y_0)$ | 0.686 | 0.706 | 0.728 | 0.751 | 0.776 | 0.802 | 0.830 |
| $\omega_I(s_2, y_0)$ | 0.712 | 0.722 | 0.737 | 0.756 | 0.779 | 0.808 | 0.842 |
| Skill Premium | | | | | | | |
| Formal | 7.0% | 3.7% | 1.9% | 0.9% | 0.7% | 1.1% | 2.4% |
| Informal | 2.6% | 1.6% | 0.9% | 0.5% | 0.4% | 0.6% | 1.3% |
| Informal Sector | | | | | | | |
| % of workers | 47% | 35% | 28% | 23% | 20% | 16% | 13% |
| % of skilled workers | 39% | 32% | 27% | 23% | 19% | 16% | 12% |

Note:-The other parameters are fixed as follows: $s_1 = 1$, $s_2 = 1.2$, $b = 0.01$, $\beta_F = \beta_I = 0.5$, $\delta = 0.2$, $r = 0.1$ $P = 2/3$, $c_I = 0.05$, $\alpha = 0.8$, $y_0 = 0.9$, and $m(\theta) = 2\theta^{1/2}$.

'U' shape. The rates of change for $U(s_1)$ and $U(s_2)$ are dependent on the respective values of $\theta$, $\phi$, and $\gamma$. Since the skill premium is derived from the difference in wages, which in turn depend on $U(s_1)$ and $U(s_2)$, the skill-premium follows the 'U' shape. The fraction of high-skilled workers in the informal sector declines continuously with reductions in PMR.

To analyze the impact of LMR related policies, I set the bargaining power of workers in the informal sector at $\beta_I = 0.3$ and $\beta_F = 0.7$ for the formal sector. The PMR cost remains fixed at $c_F = 0.4$. Table 2.4 shows the equilibrium effects of decreases in LMR by reducing formal sector bargaining power until $\beta_F = 0.3$. As

LMR decreases, wages in the formal sector, unemployment, the size of the informal sector, and the number of high-skilled workers in the informal sector decline. Total output increases. Wages in the informal sector and the skill premium in both sectors follow a hump shape.[9]

The reduction in LMR directly affects the formal sector wage rate. Given that the flow value of being unemployed ($rU(s)$) is lower than the net output per worker, a decline in a worker's bargaining power due to a reduction in LMR decreases the wage. A reduction in LMR makes the formal sector more profitable for employers, attracting new firms. This leads to an increase in the number of vacancies per unemployed, $\theta$, and attracts workers from outside the formal sector. Unemployment and the size of the informal sector decline due to additional vacancies offered in the formal sector. Given that the productivity in the informal sector is lower than that of the formal sector, overall output increases.

To understand the hump shaped response of informal sector wages and the skill premium to reductions in LMR, the relationship between LMR and the value of unemployment needs to be explored. In contrast to the effect of reductions in PMR, the fraction of low-skill vacancies offered in the formal sector, $\phi$, follows a U shape. The number of low-skilled workers in the unemployment pool, $\gamma$, varies in the opposite direction from $\phi$. Initially, an increase in $\theta$ leads to an increase in the value of being unemployed for both types of workers. However, due to the reductions

[9]The hump in wages for low-skilled workers is very mild and might not be observed empirically.

in bargaining power and wages, the opportunity cost of being unemployed declines for both workers, as does the value of unemployment. Because there are no changes in the sector-specific cost or the worker's bargaining power in the informal sector, a hump shaped response of $U(s)$ to changes in LMR leads to a similar response in wages in the informal sector. Increases in employment decrease earnings inequality.

Table 2.4: Effects of reducing labor market restrictions, $\beta_F$.

| | $\beta_F$ | | | | |
|---|---|---|---|---|---|
| | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 |
| $\theta$ | 0.232 | 0.374 | 0.568 | 0.843 | 1.26 |
| $\phi$ | 0.964 | 0.925 | 0.917 | 0.932 | 0.970 |
| $u$ | 10.31% | 9.25% | 8.25% | 7.27% | 6.24% |
| $\gamma$ | 0.671 | 0.676 | 0.678 | 0.677 | 0.671 |
| $U(s_1)$ | 5.02 | 5.04 | 5.05 | 5.04 | 5.02 |
| $U(s_2)$ | 5.08 | 5.17 | 5.19 | 5.16 | 5.07 |
| Gini coefficient | 0.091 | 0.076 | 0.064 | 0.052 | 0.035 |
| Output | 1.81 | 1.85 | 1.89 | 1.93 | 1.96 |
| Wages | | | | | |
| $\omega_F(s_1, s_1)$ | 0.571 | 0.562 | 0.552 | 0.542 | 0.531 |
| $\omega_F(s_2, s_1)$ | 0.572 | 0.567 | 0.560 | 0.550 | 0.535 |
| $\omega_F(s_2, s_2)$ | 0.712 | 0.687 | 0.660 | 0.630 | 0.595 |
| $\omega_I(s_1, y_0)$ | 0.606 | 0.608 | 0.608 | 0.608 | 0.606 |
| $\omega_I(s_2, y_0)$ | 0.610 | 0.617 | 0.618 | 0.616 | 0.610 |
| Skill Premium | | | | | |
| Formal | 0.67% | 1.40% | 1.54% | 1.25% | 0.52% |
| Informal | 0.41% | 0.89% | 1.00% | 0.81% | 0.34% |
| Informal Sector | | | | | |
| % of workers | 41% | 37% | 33% | 29% | 25% |
| % of skilled workers | 41% | 36% | 32% | 28% | 25% |

Note:- The other parameters are fixed as follows: $s_1 = 1$, $s_2 = 1.2$, $b = 0.01$, $\beta_I = 0.3$, $\delta = 0.2$, $r = 0.1$ $P = 2/3$, $c_I = 0.05$, $c_F = 0.4$ $\alpha = 0.8$, $y_0 = 0.9$, and $m(\theta) = 2\theta^{1/2}$.

The skill premium follows a hump shape due to the skill-based difference in the response rates of the value of being unemployed. The rate of change of $U(s_1)$ is smaller than that of $U(s_2)$ because net output is high for the high-skilled workers compared to the low-skilled workers in the formal sector.

In summary, reductions in PMR and LMR increase output and employment; they reduce earnings inequality, the size of the informal sector, and the fraction of high-skilled workers in the informal sector. The effect on wages and the skill premium differs for reductions in PMR and LMR. A reduction in PMR increases wages for all workers, and the skill premium in each sector follows a U shape. A reduction in LMR reduces wages in the formal sector, and wages in the informal sector and the skill premium in both sectors follows a hump shape.

Many of these results are supported empirically in previous literature. Djankov et al. (2006), Antunes and de V. Cavalcanti (2007), and Loayza et al. (2005) show that the lower PMR positively affects annual growth rates and reduces the size of the informal sector. Botero et al. (2004), Ahsan and Pagés (2009), Besley and Burgess (2004), and Pagés-Serra (2000) show that higher LMR is associated with higher unemployment, a larger informal sector, more poverty and inequality, and lower output and investment. Based on the availability of data, further empirical studies can be done on the relationship between: (a) regulation costs and the fraction of high-skill workers in the informal sector; (b) PMR and earnings inequality; (c) PMR and unemployment; (d) PMR and wages across sectors; and (e) LMR and wages in the formal sector.

## 2.6 Conclusion

This chapter studies the effect of government policies related to the product and labor markets on unemployment, wages, and the size of the informal sector. Supporting the previous literature, my results show that government policies reducing PMR or LMR decrease the size of the informal sector, earnings inequality, and unemployment. Further, I find that a reduction in PMR or LMR also decreases the fraction of skilled workers in the informal sector. With a reduction in PMR, the skill premium follows a U shape, whereas with a decline in LMR, it follows a hump shape.

Many of my results support the empirical findings of Ahsan and Pagés (2009), Djankov et al. (2006), Djankov et al. (2002), Botero et al. (2004), and Loayza et al. (2005). The result on skilled worker participation in the informal sector can be empirically tested using cross-country data.

It would be beneficial to extend the model to study the impact of state policies in the regional economics framework and empirically test the model using state data for India. Data for India are particularly suitable, as the World Bank provides an indicator for the ease of doing business (a proxy for regulatory cost) for 17 Indian states. Further, National Sample Survey Organisation (NSSO) data can be utilized to measure the degree of unionization (a proxy for labor law tightness) and skilled labor participation for each state. Moreover, the fraction of individuals in each sector varies based on the type of economy and the given parameters. Further research can be done using the different equilibrium cases of this model to study diverse economies and regions.

# CHAPTER 3
# THE ROLE OF ENDOGENEITY IN THE RELATIONSHIP BETWEEN EDUCATION AND EARNINGS: A SEMIPARAMETRIC APPROACH

## 3.1 Introduction

The linearity of returns to education has been debated since early 1970's. Some labor economists provide evidence against the linearity assumption but ignore endogeneity.[1] They use polynomial and dummy variable approaches to test the linearity assumption. However, dummy variable approach gives crude estimates and estimates from a polynomial approach may behave erratically at the end points. Furthermore, these studies ignore endogeneity in education. To ensure these issues are addressed, I test the linearity assumption under endogeneity using a semiparametric approach. An application of the semiparametric approach to 1980 and 2000 census data shows that under an assumption of exogeneity, the marginal rate of return to education (MRRE) varies non-monotonically across years of schooling in the US.[2] Under endogenous schooling and earnings, the non-monotonicity is reduced to some extent.[3] Further comparison of estimates suggests that the failure to account for endogeneity leads to overestimation of returns to education between middle school and college,

[1]See Hungerford and Solon (1987), Heckman et al. (1996), Jaeger and Pagé (1996), Heckman (2008)

[2]I mainly discuss the 1980 census data in this chapter. The application of this methodology to the 1990, 2000 census and 1992, 2005, and 2013 CPS March Supplement data results in an even stronger conclusion against linearity.

[3]I use a control function approach for IV estimation in a semiparametric partial linear model. Spousal and parental education are used as instruments.

and underestimation of returns at other levels of the education spectrum.[4]

The question of linearity is studied because it contributes to the human capital accumulation problem. The knowledge of returns to education influences individuals decisions on schooling and can also be useful to policy makers in allocating funds for education.[5] In the theory of human capital accumulation, Becker (1967, 1975) attributes the difference in MRRE across schooling levels to heterogeneity in ability and educational opportunities. This theory can be used to explain regional differences in returns to education and may prove helpful in policy making.

The log-linear regression model of Mincer (1974) is the standard model to estimate returns to education in labor economics. Here, log earnings are linear in schooling and quadratic in potential experience.[6] To model the non-linearity in this relationship, researchers generally use dummy variables in a parametric model to estimate returns at different schooling levels. While studying the impact of school quality on returns to education for cohorts born before 1960, Card and Krueger (1992) use the dummy variable approach and claim that returns to education for different states in the US are approximately log-linear above some state-specific threshold level of schooling. Heckman et al. (1996) challenge this claim by noting that regional labor market variables affect the returns for low-skilled workers much more than those for

[4]This result is applicable for standard Mincer regression in comparison to the estimates based on spouse's education as instrumental variable.

[5]See Jensen (2010)

[6]A quadratic function of potential experience (defined as the difference between age and years of education plus six in most studies) is often used in Mincer's log-linear model to capture the effect of on-the-job training on earnings

high skilled workers.[7]

Estimating the internal rate of return to education for a limited number of schooling levels by nonparametric regression, Heckman et al. (2008) show the importance of relaxing Mincer's assumptions of linearity in schooling and of separability between schooling and potential experience. They formally reject the hypothesis of linearity in returns to education in the Mincer regression using the US national level census data for all census years between 1940 and 1990, inclusive.[8] I confirm the non-linearity in MRRE while incorporating data from all schooling levels to give a clearer picture of the function of returns to education.

To identify this function, I generalize the standard Mincer regression by adding higher order polynomials for years of schooling and potential experience.[9] To further ensure that the non-linear relationship between years of education and log of earnings is fully captured, an alternative approach is to use a non-parametric regression assuming no pre-specified functional form and relying on the data to estimate the

[7]Heckman et al. (2003) systematically analyze Mincer's assumptions using CPS and census data from the US and find that the estimates overstate the returns to education if taxes and tuition are not accounted for. Moreover, the uncertainty and changing nature of the economic environment also has important effects on estimates of the returns to education. In this chapter, I ignore taxes and tuition, and expect that the qualitative results of my chapter remain unchanged due to the equal effect of the economic environment and uncertainty on all individuals.

[8]In Heckman et al. (2008), the log earnings model is tested against an alternative where the coefficient on education is allowed to differ for each schooling level to test the linearity assumption.

[9]Murphy and Welch (1990) find the use of a quartic, rather than quadratic, function for experience more appropriate for fitting a model on CPS data from 1964 to 1987. Lemieux (2006) verified for 1979-81, 1989-91 and 1999-2001 CPS data. Lemieux (2006) further gives evidence of a need to include the higher order polynomials in potential experience to "fine-tune" the standard Mincer equation.

functional form (Härdle, 1990).[10] Since this approach suffers from the 'curse of dimensionality', Heckman et al. (2003) use an alternative non-parametric approach to estimate returns to schooling. They regress experience on the log of earnings separately for each schooling level. However, their estimates are confined to schooling levels of between six to 16 years. To estimate MRRE at all given schooling levels, I use a semiparametric method known as the partial linear model.

This semiparametric method contains a non-parametric component of years of education along with a parametric component of the remaining variables in the standard Mincer regression to explain the log of wages. It estimates the total return at all schooling levels. I use spline functions over these estimates to approximate the non-linearity between the log of earnings and schooling. The first derivative of this function at each schooling level estimates the level-specific marginal rate of return to education (MRRE). The semiparametric partial linear model captures the non-linear relationship between the log of earnings and schooling while accounting for the useful linear effects of other variables.

Schooling is implicitly assumed to be exogenous in the Mincer model. This assumption has been challenged by many studies and the debate is not yet settled.[11] To identify the appropriate approach for this chapter, I follow Mincer by assuming that individual schooling levels are exogenous. After identifying the approach, I

[10]The census data are large enough to estimate a fully non-parametric model using a dummy for each value of the variables, though identification under endogeneity would be challenging.

[11]See Griliches, (1977); Willis and Rosen, (1979); Card, (1995); (1999); Heckman and Vytlacil, (2001); (2003).

address the issue of endogeneity.

The instrumental variable method is generally used to overcome the issue of endogeneity.[12] In estimating returns to education, endogeneity can arise from omitted variables, self-selection, or measurement error. Similar to self-selection, some omitted variables like ability tend to show as an upward bias, whereas omitted variables like financial constraints tend to show a downward bias in OLS estimates. Measurement error biases OLS estimates towards zero. However, Card (1999) shows that the OLS estimates are biased downwards because individuals with high discount rates choose low levels of schooling, which have much higher marginal rates of return. Given the results of this chapter, Card's claim may not apply. Dearden (1999) suggests little or no selection bias in returns to education.

Researchers use different variables as instruments for addressing the issue of endogeneity. Examples include presence of a college nearby (Card, 1993; Cameron and Taber, 2004), local labor market earnings for low skill workers (Cameron and Taber, 2004), local unemployment rate (Cameron and Taber, 2004), educational reforms (Devereux and Fan, 2011; Dickson and Smith, 2011), and education level of parents or spouse (Trostel et al., 2002).[13] Given census data for the US, the educational levels of both parents and spouses can be identified by merging the information of

[12]See Imbens (2014) for more on those assumptions that should be satisfied while deciding on an intrumental variable.

[13]Dickson and Harmon (2011) cite Heckman and Urza (2010) and summarize some potential problems with the IV approach as follows: "IV estimates rest on strong, a priori data assumptions; in a heterogeneous model, different instruments will give different estimates; and finally, the IV estimate, depending on the instrument used and assumptions made, will give different estimates of the return to education, which are often incorrectly interpreted."

individuals with their parents and spouses.[14] I check the estimates with mother's, father's, and spouse's education as instrumental variable. The sample selection of individuals sharing a house with their parents or spouse may give us an inconsistent IV estimate. Wang (2013), using data for China and the US, shows that the impact of sample selection on IV estimates for returns to education is either statistically insignificant or modest.

In the next section, I discuss the different model specifications for estimating returns to education.[15] Section 3.3 sheds light on the issue of endogeneity in semiparametric models and explains the search for an instrumental variable approach that is identifiable and provides consistent estimates. Section 3.4 introduces the data and presents preliminary findings. Results for the polynomial and semiparametric specification are presented and analyzed in subsection 3.5.1, followed by a comparison with the results from other chapters in subsection 3.5.2. Section 3.6 presents results for IV models that tackle the endogeneity issue. Section 3.7 summarizes the findings with concluding remarks.

## 3.2 Models with Exogenous Schooling

Most models for estimating returns to education use the standard Mincerian earnings equation. Heckman et al. (2003) summarize the work done using Mincer regression. They also examine the importance of relaxing functional form assump-

[14]I use the household identifier and family relation variable in census data.

[15]The estimation of marginal rates of return and the estimation procedure for the semiparametric model are discussed in Appendices A and B, respectively.

tions in the Mincer regression and analyze modifications to the standard Mincer by accounting for taxes and tuition with exogenous schooling and earnings. Their work is useful in understanding the fundamental basis of the Mincer regression. I contribute to this literature by generalizing the Mincer regression to obtain robust estimates for the MRRE and test the linearity assumption under endogeneity.[16] Since the Mincer specification assumes exogenous schooling, I follow it initially. Later, I test my findings under endogenous earnings using instrumental variables.

The standard log-linear model suggested by Mincer for estimating returns to education is shown below:

$$log(w_i) = \beta_{10} + \beta_{11} YED_i + \beta_{21} PEX_i + \beta_{22} PEX_i^2 + \beta_{31} X_i + \varepsilon_i, \tag{3.1}$$

where for individual $i$, $w_i$ is the wage rate, $YED_i$ is years of education, $PEX_i$ represents potential experience, $X_i$ includes all other factors to be discussed in Section 3.4, and $\varepsilon_i$ is the unobserved error term. $\beta$ are the model parameters. I drop the subscript $i$ until required.

To capture the non-linearity in the Mincer regression, economists often use the standard dummy variable method. The following specification of the Mincer regression without intercept can be used to estimate returns to education at each

[16]Heckman et al. (2003) use an alternative non-parametric approach independent of the Mincer model to estimate returns to schooling by regressing experience on the log of earnings for each schooling level separately. Gorodnichenko and Peter (2005) use the semiparametric method to construct a counterfactual wage distribution for a cross-country comparison over the returns to education between Russia and the Ukraine. Carneiro et al. (2010) also use semiparametric regression to study the marginal policy relevant treatment effect (MPRTE) in terms of the returns to education.

schooling level:

$$log(w) = \sum_{l=1}^{L} \beta_{1l} D_{YED=l} + \beta_{21} PEX + \beta_{22} PEX^2 + \beta_3 X + \varepsilon, \tag{3.2}$$

where $L$ are total levels of schooling in the data, $l$ takes the value of different schooling levels, and $D_{YED=l}$ are dummy variables that take the value 1 for an individual with schooling level $l$. The coefficients of the dummy variables capture the non-linear relationship between schooling and the log of earnings.

Higher degree polynomials can also be used to capture the non-linear effect of years of education. Some researchers suggest using this method for potential experience as well (Heckman et al., 2003). I use the first, second, and third degree polynomial terms for schooling and potential experience.[17] With these specifications, I have the following model:

$$\begin{aligned} log(w) = \beta_0 + \beta_{11} YED + \beta_{12} YED^2 + \beta_{13} YED^3 \\ + \beta_{21} PEX + \beta_{22} PEX^2 + \beta_{23} PEX^3 + \beta_3 X + \varepsilon. \end{aligned} \tag{3.3}$$

The standard Mincer model given in Equation 3.1 gives only one estimate of the rate of return for all levels of education. The dummy variable approach in Equation 3.2 allows for a different estimate at each schooling level. We will see that the marginal returns have greater variance than do those of other methods. In Equation 3.3, the interdependence between the higher polynomial orders of YED makes the returns to education less volatile across schooling levels. However, the

[17]The model with a third degree polynomial is sufficient to test linearity. The estimates from a model with additional polynomials do not add much to the fit of the data.

polynomial model may behave erratically at the tails and thus is biased towards falsely rejecting linearity assumptions. To ensure the estimates do not behave erratically at the tails, I use a partial linear model which generalizes the Mincer regression further. This is shown below:

$$log(w) = \beta_0 + f_1(YED) + \beta_{21}PEX + \beta_{22}PEX^2 + \beta_{23}PEX^3 + \beta_3 X + \varepsilon. \quad (3.4)$$

The structure of the function $f_1(.)$, along with the other parameters of the model, are estimated using the data. The estimation of this semiparametric model is discussed in more detail in Appendix B.[18]

Because the endogeneity between earnings and education may affect the results, I use parents' and spouse's education as instrumental variables. In the next section, an approach to estimate the semiparametric model using instrumental variables is discussed.

### 3.3 A Model with Endogenous Schooling and Earnings

I estimate marginal rates of return for all schooling levels using polynomial and semiparametric models. To estimate with endogenous schooling and earnings, I need IV estimation method for both models. IV estimation method for linear models is well-accepted and documented. In this section, I discuss IV estimation method for partial linear models.

Blundell and Powell (2003) (hereafter BP) mention three approaches to instrument variables in semiparametric models: a standard instrumental variable ap-

[18]The qualitative results based on estimates of marginal returns are the same whether I use splines or higher order polynomials on potential experience in a partial linear model.

proach, a fitted value approach, and a control function approach. They show that the identification requirements on an unspecified function are simpler to interpret for the control function approach than for the other two. BP further claim that the control function approach gives consistent estimates even under nonadditive unspecified functions.[19] To ensure the consistency of estimates and the identification of the unspecified function, I use a control function approach.[20]

The control function approach treats the endogeneity problem as one of omitted variables. In this approach, the effect of any omitted variables is assumed to be captured in the estimated residual of the first stage model. This model regresses the endogenous variable on the instrumental variable along with the other independent variables of the second stage model. The estimated residual is then included as an independent variable in the second stage to control for endogeneity. To implement the control function approach in the semiparametric model, the first-stage residuals, $\hat{\nu}_i$, are calculated for each individual by estimating the model as follows:

$$YED_i = \pi_1(Z_i) + \pi_2(PEX_i) + \gamma_3 X_i + \nu_i, \tag{3.5}$$

where $Z_i$ is an instrumental variable, $\nu_i$ is the unobserved error term, $X_i$ and $PEX_i$ are the independent variables from the second-stage equation (as in Equation 3.6), $\gamma_3$ is the coefficient parameter for $X_i$, and $\pi_1(.)$ and $\pi_2(.)$ are the unspecified functions

[19]Das (2005) claims that the continuously distributed endogenous regressors are "ill-posed" and cannot be identified. The discreteness ensures that the error independence condition is met. The semiparametric estimation model I use is an additive model, and the independent and instrumental variables have finite support.

[20]Yatchew and No (2001) also use the control function approach for partial linear models to estimate gasoline demand in Canada with regions as the IV for prices.

on $Z_i$ and $PEX_i$, respectively.[21] The functional form on potential experience is left unspecified in the first stage model, as we want to control for the non-linear effect of potential experience on the endogenous variables following BP.

The estimated residuals from the first stage model are used in the second-stage equation to control for omitted variables.[22] The second stage semiparametric model is given as follows:

$$log(w_i) = \beta_0 + f_1(YED_i) + g_1(\hat{\nu}_i) + f_2(PEX_i) + \beta_3 X_i + \varepsilon_i, \tag{3.6}$$

where $g_1(.)$ is an unspecified function on estimated residuals, $\hat{\nu}_i$. This ensures that the effect of omitted variables account for the endogeneity in the model. After the estimation, the restriction for applying the control function approach as suggested in BP is as follows:

$$\begin{aligned} E(\varepsilon|YED, PEX, X, Z) &= E(\varepsilon|YED, PEX, X, \nu) \\ &= E(\varepsilon|\nu). \end{aligned} \tag{3.7}$$

These conditions can be analyzed for the models used in this chapter. However, the estimation technique use in practice gives results that always satisfy these restrictions.

To estimate the semiparametric model with IV and its derivative for calculat-

[21] I use the unspecified function for PEX in the partial linear model instead of including the higher order polynomial as a matter of convenience.

[22] Instead of residuals from the first-stage equation, one can use the predicted value of years of education in the second-stage equation. This will give us the fitted value approach. However, this approach suffers from identification problem as discussed in Blundell and Powell (2003).

ing marginal rate of return, I use the same approach as used for the semiparametric model under exogenous schooling.

The next section describes the data for the US, followed by a section presenting estimation results to show the application of the procedure discussed above.

## 3.4 Data

Following the literature on returns to education in the US, I use a five percent sample for 1980 and 2000 census data.[23] Motivated by Heckman et al. (2003) and Gelbach (2009), I exclude members of the armed forces from the data. I keep individuals who are in the full time workforce, ages 16 to 64, and focus on males who describe themselves as "White Only" or "Black Only." To further clean the data, only individuals whose total wage and salary is greater than $1, have more than one hour of work per week, and worked for at least one week during the year are kept. After these exclusions, I am left with a smaller sample for each dataset which is used in the analysis. The 1980 dataset has 2,387,770 observations and the 2000 dataset has 2,634,230.

Table 3.1 shows the distribution of the employed male population in the US by education level attained. 1980 census data have more reliable information on the years of schooling completed rather than degree completed. 2000 census data use the highest grade completed with information on the type of degree. To maintain comparability across datasets in Table 3.1, I impute the grade or degree level from

[23]The census data are downloaded from the IPUMS (Integrated Public Use Microdata Series) website.

years of schooling in 1980 census data.

Table 3.1: Distribution of the employed (male) population in the US by level of educational attainment (based on 1980 and 2000 public use census data samples).

| Education Levels | % of Employed Population | |
|---|---|---|
| | 1980 | 2000 |
| <1st Grade | 0.36 | 0.43 |
| 1st to 4th | 1.18 | 0.24 |
| 5th or 6th | 1.91 | 0.71 |
| 7th or 8th | 5.56 | 1.31 |
| 9th | 3.96 | 1.72 |
| 10th | 5.65 | 2.85 |
| 11th | 5.95 | 3.27 |
| 12th, no diploma | – | 3.14 |
| High school with diploma | 35.86[a] | 29.12 |
| Some college but no degree | 6.9[a] | 23.53[b] |
| Associate degree | 10.67 | 6.82 |
| Bachelor's degree | 13.82[a] | 17.35 |
| Master's degree | 2.7[a] | 6.05 |
| Professional degree | 1.65[a] | 2.22 |
| Doctorate | 2.43[a] | 1.23 |

[a] Individuals with 'No Schooling Completed', 'Nursery School, Preschool' and 'Kindergarten' are assigned 0 years of schooling, which is the same as '<1st Grade'. Similarly, 'Grade 12' and 'Some college, but less than 1 year' are assigned 12 years of schooling or as 'High School graduate, with diploma'.

[b] In the 2000 census, the education variable reports two values for 'some college but no degree': (a) <1 year of some college, 7.51%, and (b) 1 or more years of college, 16.02%.

Before discussing the results, it is important to keep in mind the changes in the labor market and the schooling system over time in the US. Schooling through high

school is funded mainly by the government and laws related to compulsory education are strictly enforced. This is evident from the distribution of employed individuals based on their highest level of educational attainment, as over 75% and 85% of the employed population are high school graduates in 1980 and 2000, respectively. Regarding employment, the minimum working age differs by state from 12 years to 16 years, and the number of hours a young person can work is often limited. I take 14 years to be the minimum age for employment.[24]

On the variable "other factors," $X$, I have used different factors to control for social and geographic influences on wages. As suggested in Psacharopoulos and Patrinos (2004), factors related to work profile, such as occupation type or industry, are ignored as they obscure the effect of education. In the literature, Gabriel and Rosenthal (1999) show the importance of a location variable in a returns to education regression and suggest the use of Standard Metropolitan Statistical Area as a dummy variable to ensure that the estimates do not suffer from omitted variable bias. Other than location, most chapters on returns to education use race as a variable

Dummy variables are created to identify race and location for each individual. The variable "Black" is created with value 1 for individuals with race given as "Black Only" and 0 for individuals with "White Only;" other races are not included in the analysis. The location variable "Non-Metropolitan" identifies if the individual is from a non-metropolitan or undefined metropolitan area (1) or a metropolitan area (0).

[24]Federal employment rules regulating child labor in the US are set under the Fair Labor Standards Act. Each state also enacts their own laws regarding child labor. The more rigorous standard between federal and state law is applied.

Table 3.2: Mean or percentages of key variables used in the study across datasets for the US.

| | Census | |
|---|---|---|
| | 1980 | 2000 |
| Potential Exp. | 18.2 | 19.7 |
| Years of Education | 12.6 | 13.3 |
| Hourly wages | $ 9 | $ 22 |
| Weekly wages | $344 | $928 |
| Annual wages | $16,422 | $43,730 |
| % Rural/Non-metro | 28.2 | 23.5 |
| % Black | 8.7 | 10.0 |

Note:-The dollar values are not adjusted for inflation and are in current prices of the given year.

Table 3.2 presents the summary of key variables used in this study for all datasets. Since Card and Krueger (1992) use weekly wages from the 1980 census data, I use the same for comparison. Heckman et al. (2008) use annual business income, including wages and salary, for 1980 and 2000 census data; I use the same for in my analysis. The mean of potential experience, years of education, and all types of wages are increasing over time for the male workforce.[25] The percentage living in rural or non metropolitan areas is declining. The percentage of the workforce that is Black has increased over the period.

[25]Wages are not adjusted for inflation.

## 3.5 Results for Exogenous Models

To test for the linearity of returns to education, I estimate all models discussed in Section 3.2.[26] The non-zero coefficients for the higher order polynomial on years of education in the polynomial model (Equation 3.3) imply non-linearity in returns to education. In the next section, I show that semiparametric estimates are more conservative in showing non-linearity than are the polynomial model estimates.[27]

### 3.5.1 Results based on 1980 Census Data

Each model found in Section 3.2 is estimated using the 1980 census dataset. Table 3.3 shows parameter estimates and key measures for each model with standard errors in parentheses. The more generalized specifications adds only 0.03 points in R-squared value in comparison to the standard Mincer regression. The small incremental change in explanatory power helps explain the long standing popularity of the standard Mincer regression. The comparison of coefficients other than PEX and YED across different models have the expected effects only: location and race affect wages negatively, and the value of these coefficients do not show much variation across models.[28]

Comparing coefficients across the different models shows that potential experience and years of schooling have a non-linear relationship with the log of annual

[26]Due to memory limitations, all semiparametric estimates are based on a simple random sample of 350,000 observations after exclusions. It covers around 13% of the cleaned data.

[27]The semiparametric estimates use cubic smoothing splines to penalize curvature.

[28]The estimates of dummy variables over YED are shown in the Appendix C Table C.1.

Table 3.3: Parameter estimates and key goodness-of-fit indicators for the different models discussed in Section 3.2 based on the 1980 US census dataset.

| Variables | Standard Mincer | Dummy Mincer | Polynomial Mincer | Semi-parametric |
|---|---|---|---|---|
| Intercept | 7.352 | | 7.6565 | 7.0137 |
| | (0.0024) | | (0.0060) | (0.0015) |
| YED | 0.0930 | | -0.0941 | 0.1004 |
| | (0.0002) | | (0.0017) | (0.0006) |
| $(YED)^2$ | | | 0.1757 | |
| | | | (0.0016) | |
| $(YED)^3$ | | | -0.0495 | |
| | | | (0.0005) | |
| Smoothing Parameter (YED) | | | | 0.9914 |
| PEX | 0.1033 | 0.0997 | 0.1875 | 0.1775 |
| | (0.0001) | (0.0001) | (0.0003) | (0.0002) |
| $(PEX)^2$ | -0.171 | -0.0164 | -0.0674 | -0.0611 |
| | (0.0003) | (0.003†) | (0.0002) | (0.0078†) |
| $(PEX)^3$ | | | 0.0077 | 0.0067 |
| | | | (0.0022†) | (0.0011†) |
| Black | -0.3073 | -0.2956 | -0.3199 | -0.3266 |
| | (0.0017) | (0.0017) | (0.0016) | (0.0009) |
| Non-Metropolitan | -0.0961 | -0.0961 | -0.0973 | -0.1004 |
| | (0.0011) | (0.0010) | (0.0010) | (0.0006) |
| $R^2$ | 0.374 | 0.3880 | 0.4060 | 0.4074 |
| F-Value | 285311 | 16570000 | 203988 | |

Note:- Each $j^{th}$ degree polynomial variable is multiplied by $10^{-(j-1)}$ to get the standardized coefficient value. All estimates are significant at the 1% level. Number of observations across models are same at 2,387,770 except for semiparametric model where I use 350,000 observations based on simple random sample due to memory limitations.

†Multiplied by $10^{-2}$.

earnings. The presence of high degree polynomials for YED in the polynomial model suggests a considerable amount of non-linearity between schooling and the log of earnings. The linear estimate of YED in the semiparametric model mirrors the lin-

ear effect of YED in the standard Mincer regression. However, the non-parametric portion is also significant in explaining the log of wages, as the smoothing parameter given in Table 3.3 is significant at the 1% level, which implies non-linearity in returns to education. The linear estimate for return to education remains around 10% under both standard and semiparametric estimation.

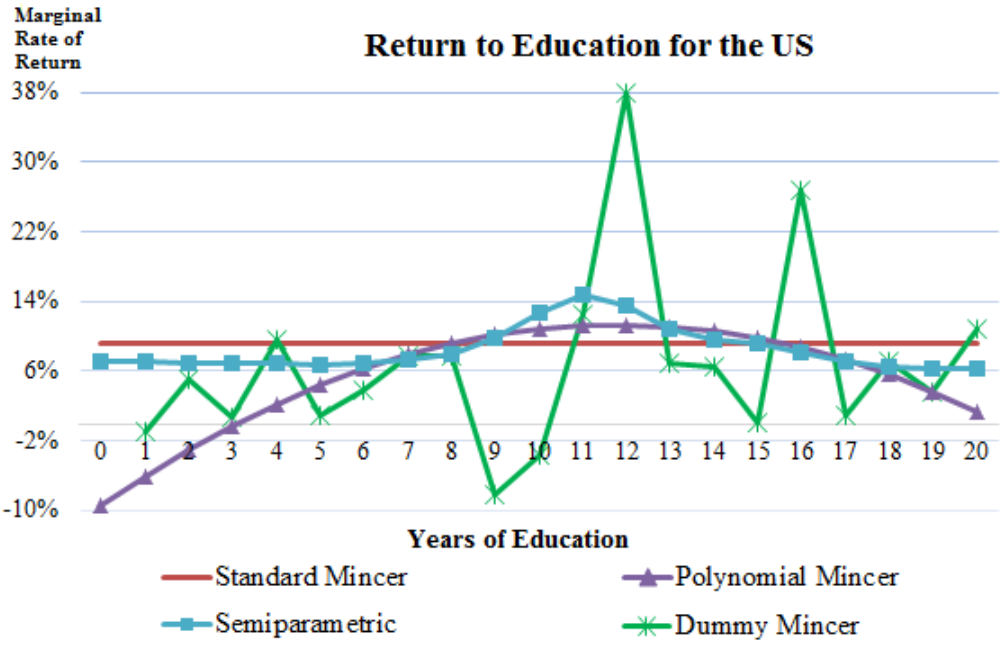


Figure 3.1: Estimates of the marginal rate of returns to education by year of schooling for 1980 US census data across different models.

Figure 3.1 shows the estimates for the MRRE across years of schooling in the US for all the models. The estimated points can be interpreted as the marginal return to education from attending school for an additional year. For instance, based on semiparametric estimates, an individual with nine years of schooling will get a

return of around 11% by attending school for an extra year. This increases to 13% when this individual completes ten years of schooling.

The standard Mincer regression estimates the marginal rate of return as a constant 9.7% per year of schooling. Conversely, the more generalized versions of the Mincer regression show changes in the estimates of MRRE across years of schooling. The estimates from the dummy variable model are highest for 12, 16, and 20 years of schooling, representing the high-school, college, and doctoral degree levels. The lowest estimate of MRRE for the dummy variable approach is at -8% at nine years of education. Although this shows the presence of the sheepskin effect, it rejects linearity in returns to education. Smoothing hides these effects and estimates a pattern with less variance. Estimates for the polynomial and semiparametric models show a similar pattern except at the tails.

As mentioned by Green and Silverman (1994), the polynomial regression has various drawbacks. One of the drawbacks is overfitting, which can be seen in Figure 3.1 at the tails. The semiparametric model solves this issue. Based on the estimates of the semiparametric model, the MRRE at each schooling level remains stable for the first eight years of schooling at around 7% and increases to 15% during high school. The marginal returns decline to 7% after a high school degree.[29]

[29]The functional form estimated using semiparametric regression is found to be significant with 95% uniform confidence bands. These bands are created using a bootstrap method.

### 3.5.2 Results in Comparison to Other Studies

In this subsection, I use the polynomial and dummy variable models to analyze the linearity assumption in other datasets and compare them with the results of Card and Krueger (1992) (hereafter CK) and Heckman et al. (2008) (hereafter HLT). The estimates in CK are based on the 1980 census, whereas HLT use the sample from 1940–2000 census data and 1964–2006 CPS March supplement data. To compare their results with estimates from the semiparametric approach, I use the public use sample of the 1980 and 2000 census data.[30]

Using the dummy variable approach to estimate total returns to education[31], CK suggest that the relationship between earnings and education "is approximately log-linear for the levels of education above a minimum threshold."[32] They present the results for three cohorts of white men based on birth year, namely, 1920, 1930, and 1940 for white men born between 1920-29, 1930-39, and 1940-49, respectively. Using the same sample data and methodology, I try to replicate their estimates and compare the results between the dummy variable and semiparametric approaches. Due to the paucity of space and similarity between the different cohorts, I present

[30]CK use weekly wages as the earnings for each individual whereas HLT use annual earnings, total of business income, wages and salary for all census years, and annual wages and salary earnings in CPS data. I use both methods for comparison and proceed to follow HLT.

[31]The total returns to education in dummy variable approach is the value of the coefficient at each schooling level. It is the sum of the base marginal return plus plus the base return.

[32]In the model estimating returns to education, CK control for the state of birth and state of residence effects by including dummies for each. Other than controlling for race and location, they also include a dummy to identify married and unmarried individuals.

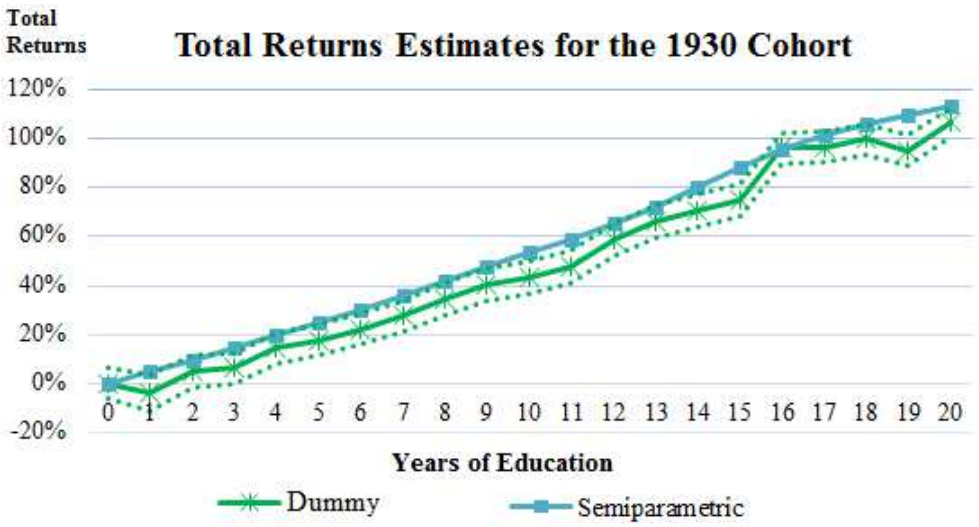


Figure 3.2: Dummy variable and semiparametric model estimates of the total returns to education by year of schooling for 1930 cohort based on the1980 US census data. Comparable to Figure 2 in Card and Krueger (1992). Dotted lines show the pointwise 95% confidence interval for dummy variable estimates.

and discuss results for the 1930 cohort only.[33]

Figure 3.2 gives the estimates for total returns based on the dummy variable approach and the semiparametric approach for cohorts of white men born in the 1930's. Since the estimate in CK (Figure 2 in Card and Krueger (1992)) starts with zero, I normalize the estimates of both models.[34] Figure 3.2 supports the claim of CK. The total returns estimates of the semiparametric regression are closer to linear than are the estimates of the dummy variable approach.

Although the earnings-education relation appears to be better approximated

[33]All regression output on estimation for this cohort and results from other cohorts is available on request.

[34]The normalization for estimates is done by subtracting the first estimate from all estimates, thus making the first estimate equal to zero. The pattern in estimates for the dummy variable model matches the estimates from CK (Figure 2 in Card and Krueger (1992)) after three years of education. The reason for the mismatch is unknown.

by a log-linear function, the marginal returns estimates suggest otherwise. In Figure 3.3, the estimates of MRRE using the same approaches show non-linearity in returns to education. The dummy variable estimates show greater variance than do the semiparametric estimates, as seen earlier. A confidence interval for the dummy estimates fails to reject the null hypothesis that the returns are equal to zero at all schooling levels except for high-school, college, and doctoral degrees.[35] This suggests the presence of a sheepskin effect (Table 3.4 shows the estimates based on the dummy variable approach in numbers). Within the 95% uniform confidence band, the semiparametric estimates show that the marginal rate of return is constant for the first 10 years of schooling and increases until 15 years of schooling and declines thereafter to stabilize at 19 and 20 years of schooling. The non-constant estimates of MRRE suggest the presence of non-linearity in returns to education for the 1930 cohort based on 1980 census data.

HLT formally reject the linearity assumption and give estimates of internal rate of return for education assuming a work life of 47 years in a schooling model. They estimate the standard Mincer regression and progressively relax the assumptions on schooling and potential experience to observe changes in estimates.[36] Although the

[35]The confidence interval for the estimate of marginal returns based on dummy variables is calculated by adding the standard error of each coefficient estimate used in calculating marginal return. I assume zero covariance between the coefficients of two adjacent dummies.

[36]The estimates are given for census data 1940–2000 for schooling years 6-8, 8-10, 10-12, 12-14, 14-16, and 12-16. They estimate four types of models: Mincer specification, Mincer with relaxed linearity in schooling, relaxed linearity in schooling and quadratic in potential experience, and relaxed assumptions of linearity and of parallelism between education and potential experience.

Table 3.4: Estimates of Total Rates of Return and Marginal Rates of Return with Standard Error based on a dummy variable approach for the 1930 cohort from 1980 US census data.

| Years of Schooling | Total Returns | | Marginal Returns | |
|---|---|---|---|---|
| | Estimate | Standard Error | Estimate | Standard Error |
| 0 | 0 | 0.0321 | | |
| 1 | -0.0401 | 0.0366 | -0.0401 | 0.0487 |
| 2 | 0.0477 | 0.0335 | 0.0878 | 0.0496 |
| 3 | 0.0613 | 0.0323 | 0.0136 | 0.0465 |
| 4 | 0.1416 | 0.0325 | 0.0803 | 0.0458 |
| 5 | 0.1766 | 0.0323 | 0.035 | 0.0459 |
| 6 | 0.2206 | 0.0321 | 0.044 | 0.0455 |
| 7 | 0.2738 | 0.0324 | 0.0533 | 0.0456 |
| 8 | 0.3442 | 0.0323 | 0.0703 | 0.0457 |
| 9 | 0.4012 | 0.0327 | 0.0570 | 0.0459 |
| 10 | 0.4317 | 0.0328 | 0.0305 | 0.0463 |
| 11 | 0.4754 | 0.033 | 0.0438 | 0.0465 |
| 12 | 0.5853 | 0.0326 | 0.1099 | 0.0463 |
| 13 | 0.6575 | 0.0327 | 0.0723 | 0.0462 |
| 14 | 0.7001 | 0.0325 | 0.0425 | 0.0461 |
| 15 | 0.7489 | 0.0325 | 0.0489 | 0.0459 |
| 16 | 0.9574 | 0.0319 | 0.2084 | 0.0455 |
| 17 | 0.9612 | 0.0317 | 0.0038 | 0.0449 |
| 18 | 0.9937 | 0.0312 | 0.0325 | 0.0444 |
| 19 | 0.94752 | 0.0309 | -0.0462 | 0.0439 |
| 20 | 1.0605 | 0.0298 | 0.1129 | 0.0429 |

semiparametric model discussed in this chapter relaxes the assumptions on linearity and being quadratic on potential experience, it uses a different estimation procedure and thus the two are not comparable.

To check for non-linearity, Figure 3.4 shows the estimates from the semiparametric approach for the 1980 and 2000 census datasets using the conditions of HLT.

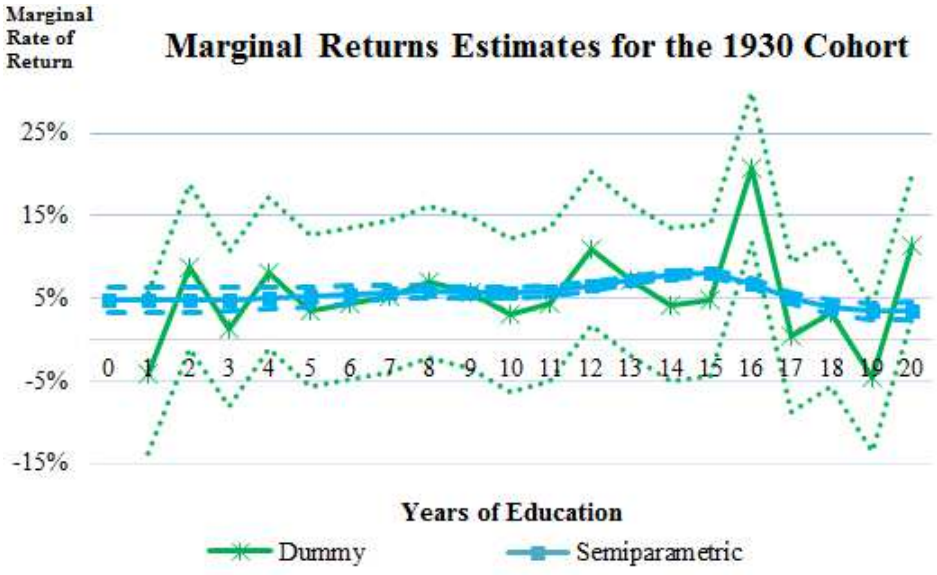


Figure 3.3: Dummy variable and semiparametric model estimates of the marginal rate of returns to education by year of schooling for the 1930 cohort based on 1980 US census data. Dashed lines show the 95% confidence band for the semiparametric estimate and dotted lines show the pointwise 95% confidence interval for the dummy variable estimates.

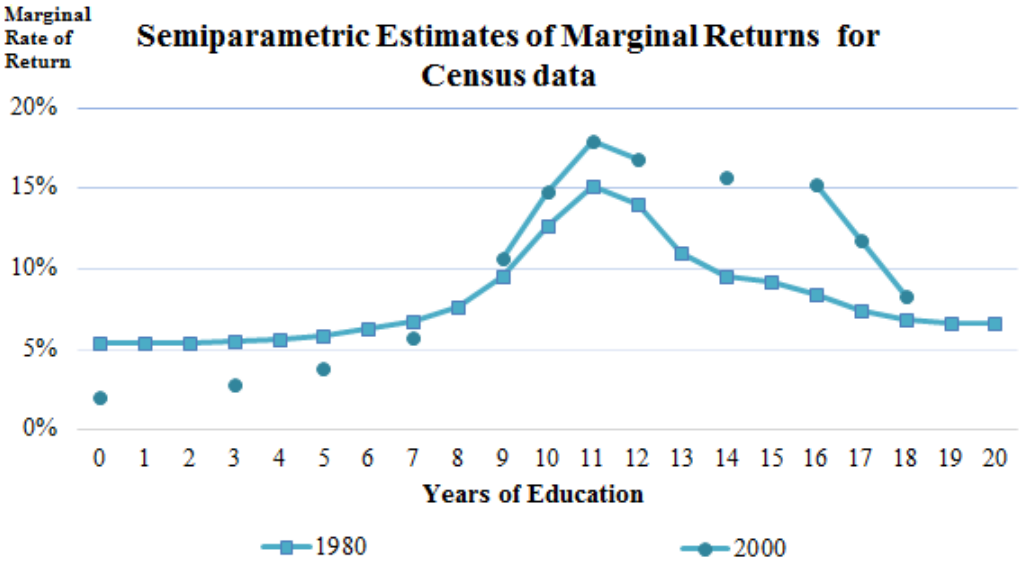


Figure 3.4: Estimates of the marginal rates of return to education by year of schooling for different census years.

The missing line between different estimates of 2000 census data is due to the absence of information for those years of education in the census data.[37] As shown in the graph for both census years, the peak is attained between the 10th and 12th years of schooling. This matches the high estimates for the 10-12 category of the 1980 and 2000 census data from HLT (the only exception is the 'Relax linearity in schooling and parallelism' estimates for 1980 census data).[38] The nonlinearity is evident from the marginal returns curve, supporting the results of HLT.

To further ensure that the non-linearity is observed even after controlling for endogeneity, in the next section I analyze the results of the available instrumental variables on the 1980 census data.

### 3.6 Results for Endogenous Model

To tackle the endogeneity issue in this chapter, I use parental and spousal education as an instrumental variable from the 1980 census and further test the restrictions suggested by Blundell and Powell (2003). Given the US social norm that an independently earning son is not supposed to live with his parents, the observations matching father and son or mother and son in the same household are found to be

[37] As discussed, the 1980 census dataset has information on years of schooling. For the 2000 census data, the mapping from education level to schooling years is redefined as suggested in HLT. Most of the mapping between degree level and years of schooling for the 2000 census data matches the mapping done by HLT. The exceptions are that HLT give (1) 14 years of schooling to everyone with some college but no degree; (2) 17 years to professional degree holders (3) 18 years for masters degree or doctoral degree holders.

[38] I also check for the estimates using the 1990 census dataset and find that the estimates for the 1990 and 2000 census datasets are close to each other. One can observe the same in estimates from HLT (Table 2a).

less than one percent of the observations that match between a working male and his spouse. Observations that have education for parents and spouse are rare. This restricts us to use one instrument at a time.

The non-constant MRRE curve based on the semiparametric regression is sufficient to show the presence of non-linearity in returns to education. In this section, I present results for the IV approach in the semiparametric model only.[39] The estimates of the IV model are given in Table C.2 and C.3 in the Appendix C. By the construction of the regression solution, the residuals of the first stage equation are independent of the IV and potential experience. Therefore, there is no way to ensure that the restriction for applying the control function approach as suggested by Blundell and Powell (2003) in Equation 3.7 are satisfied. However, numerically the restrictions for the IV model (Equation 3.7) as mentioned by Blundell and Powell (2003) are tested for each model and found to be satisfied.

As shown in Table 3.5, the estimates from the spouse's IV models are lower and within 0.25% of the estimates of the model under exogenous schooling for the first three years of education. In contrast, estimates of other IVs remain higher by 0.5 to 1% for first 7 years of education. Between eight and fourteen years of schooling the estimates of IV model are 2.5% lower than the estimates from exogenous model. The model under exogeneity give estimates for above fourteen years of school which

[39]I also estimate the polynomial model using a control function approach for IV. The nonzero coefficients of the higher order polynomial for years of education are statistically significant at the 5% level and thus reject the linearity assumption for all instrumental variables except father's education. However, the number of observations in the IV model with father's education is small. Results from the polynomial IV model are available on request.

Table 3.5: Estimates of marginal returns based on semiparametric IV model for each instrumental variable and the semiparametric estimates under exogenous schooling using 1980 US census data.

| Years of | Endogenous | Instrumental Variable | | |
|---|---|---|---|---|
| Schooling | Semiparametric | Spouse's Education | Father's Education | Mother's Education |
| 0 | 7.06 | 6.83 | 7.75 | 7.53 |
| 1 | 7.06 | 6.87 | 7.77 | 7.65 |
| 2 | 7.01 | 6.97 | 7.82 | 7.74 |
| 3 | 6.94 | 7.13 | 7.89 | 7.81 |
| 4 | 6.85 | 7.30 | 7.98 | 7.92 |
| 5 | 6.82 | 7.50 | 8.06 | 8.01 |
| 6 | 6.95 | 7.78 | 8.03 | 7.95 |
| 7 | 7.30 | 8.12 | 7.98 | 7.97 |
| 8 | 8.05 | 8.50 | 8.07 | 8.29 |
| 9 | 9.83 | 9.05 | 8.24 | 8.88 |
| 10 | 12.66 | 9.88 | 8.47 | 9.65 |
| 11 | 14.80 | 10.65 | 8.53 | 10.28 |
| 12 | 13.62 | 10.64 | 8.08 | 10.31 |
| 13 | 10.87 | 10.02 | 7.32 | 10.08 |
| 14 | 9.59 | 9.60 | 6.75 | 10.03 |
| 15 | 9.15 | 9.43 | 6.60 | 10.11 |
| 16 | 8.24 | 8.90 | 6.68 | 10.07 |
| 17 | 7.07 | 8.28 | 6.68 | 9.87 |
| 18 | 6.47 | 8.20 | 6.56 | 9.80 |
| 19 | 6.24 | 8.37 | 6.47 | 9.82 |
| 20 | 6.27 | 8.62 | 6.47 | 9.94 |

are lower by 1.85% of the estimates from spouse's and mother's IV model. Overall, the spousal education estimates are up to 3% lower for the first three years, and 7 to 30% lower for schooling between 8 and 14 years, and 6 to 10% and 10 to 40 % higher for schooling years between 4 to 7 and above 15, respectively, in comparison to the model under exogenous schooling.

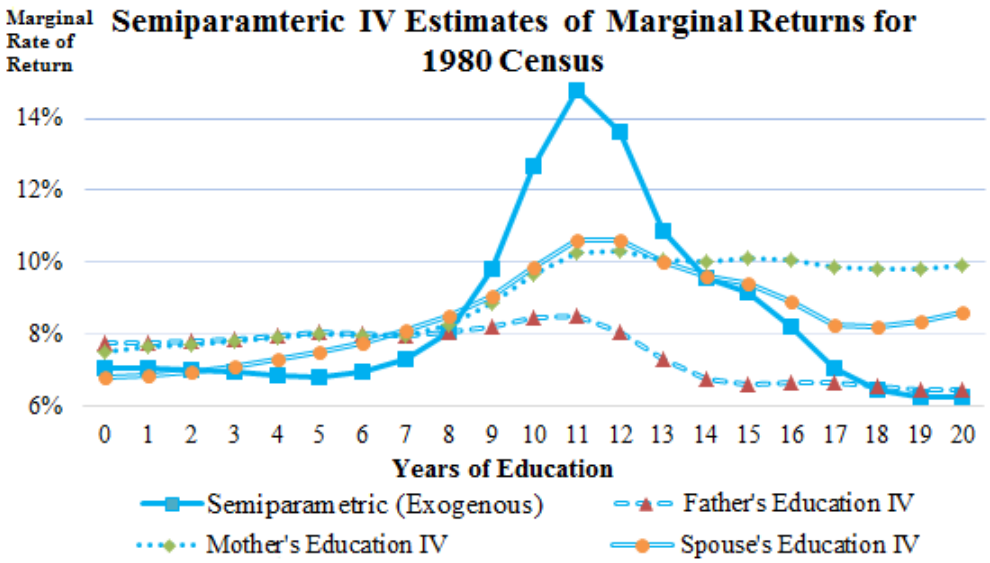


Figure 3.5: Estimates of the marginal rate of returns to education by years of schooling for different instrumental variables using 1980 US census data. I use a control function approach on semiparametric models to address the issue of endogeneity in the estimation of returns to education. Uniform confidence bands based on a simple bootstrap method show that the functional form is significant at the the 95% level for the spouses' education IV only.

Figure 3.5 shows the affect on the curvature of estimates with different IV's. Compared to the exogenous schooling case, after controlling for endogeneity, marginal returns to education are higher in the schooling levels before tenth grade and after college. This suggests the Mincer regression underestimates returns to education for both ends of the schooling spectrum and overestimates for schooling levels around high school and college.

## 3.7 Conclusion

Semiparametric and polynomial specifications are used under a Mincer framework to estimate returns to education for each schooling level in the US. The gener-

alized specification based on the semiparametric model provide smoother estimates at each given year of schooling compared to the standard dummy variable approach and the polynomial specification based on the Mincer regression. However, these specifications only add marginally to the explanatory powers of the standard Mincer specification, supporting the popularity of the Mincer regression over the last 40 years.

The estimates from the semiparametric model show that the MRRE for the 1980 census data are largely constant except for schooling levels between middle school and college. The dummy variable estimates show the presence of the sheepskin effect at the high school completion, college graduation and doctoral degree level. The nonlinearity of returns to education is statistically significant in both the polynomial and the semiparametric models.

# CHAPTER 4
# ARE RETURNS TO EDUCATION CONCAVE? AN APPLICATION TO INDIA

## 4.1 Introduction

The curvature of the schooling function in the log earnings model differs across countries (Colclough et al., 2010). However, the theory based on skill complexity presented by Mookherjee and Ray (2010) claims that "the return to human capital is endogenously nonconcave." To test this theory with respect to India, I use a semiparametric approach based on the standard Mincer regression to identify the functional form of returns to schooling in the log earnings model.[1] This approach relaxes the assumptions of linearity in schooling and potential experience, and penalizes curvature to give a 'smoother' functional form. Under exogenous schooling, my results show that returns to education for India are nonconcave above the primary schooling level.[2] To control for endogeneity between earnings and schooling, I use parents' and spouse's education as instrumental variables in a control function approach. The quantitative result does not change after controlling for endogeneity. However, the uniform confidence bands for the years prior to primary school fails to establish the statistical significance of concavity in returns to education. Therefore,

[1]I assume human capital is best approximated by years of education. Returns to education are nonconcave when the marginal returns to education are non-decreasing, which implies weak convexity in returns to education.

[2]I use India's National Sample Survey Organisation's Employment and Unemployment Survey 2004-05 data for males. The result is true at the aggregate level but may not hold at the state or regional levels.

I fail to reject the claim of Mookherjee and Ray (2010) for India.

The direct contribution of this essay is to estimate marginal returns to education for all given schooling levels in India using a semiparametric approach after controlling for endogeneity. Indirectly, the essay provides an empirical base for the theory of human capital accumulation of Becker (1967, 1975). This theory attributes differences in marginal rate of return to education across schooling levels to heterogeneity in ability and educational opportunities. Asadullah and Yalonetzky (2012) document inequality in educational opportunities in India for 1983-2004. My work may be used to complement the work by Asadullah and Yalonetzky (2012) to support Becker's claim.

Mincer (1974) suggests a log-linear regression model, in which log earnings are linear in schooling and quadratic in potential experience, to estimate returns to education.[3] The rate of return to education is represented by $\beta_{11}$ in the following equation:

$$log(w) = \beta_{10} + \beta_{11}YED + \beta_{21}PEX + \beta_{22}PEX^2 + \beta_{31}X + \varepsilon, \tag{4.1}$$

where $w$ is the wage rate, $YED$ is years of education, $PEX$ represents potential experience, and $X$ includes all other factors.

Murphy and Welch (1990) and Lemieux (2006) use US data to show the need for a generalized Mincer model. I use a semiparametric partial linear model with cubic splines to generalize the standard Mincer model. This method has a nonparametric

[3]Potential experience is defined as the difference between age and years of education minus six in most of the studies. It captures the effect of on-the-job training.

component in which the functional forms of years of education and of potential experience are unspecified. Other variables adhere to parametric restrictions. To visualize the curvature of the functional form, I estimate the first derivative using splines, giving us marginal returns to education. A negative slope of the marginal returns to education curve indicates the presence of concavity in the returns to education function. I also use dummy variable and polynomial models to compare the results.

In these models, schooling is implicitly assumed to be exogenous. To ensure that the issue of endogeneity is addressed before empirically testing the claim of Mookherjee and Ray (2010), I use instrumental variables in a semiparametric model.[4] The control function approach is advocated by Blundell and Powell (2003) to give consistent and identifiable estimates for an unspecified function in the partial linear model.

Given the data available for India and my focus on the male workforce, the educational level of both parents and spouse can be identified by merging the information of individuals with their parents and spouse. I test my results with mother's, father's and spouse's education as instrumental variables.[5]

The purpose of this essay is to estimate the functional form of returns to education. There are infinitely many functions that can estimate the returns to

[4]Trostel et al. (2002) use these instruments to study returns to education. See Imbens (2014) for detailed discussion on the use of instrumental variable for controling endogeneity.

[5]The selective sub-sample of individuals who live in extended families with their parents or wife may cause sample selection bias in the IV estimate based on these IVs. Wang (2013) estimates the returns to education with spousal and parental education as IV for US and China. He claims that the sample selection bias has a modest to statistically insignificant impact for the spousal and parental education IV.

education. To test the statistical significance of the estimated function, I use uniform confidence bands. The uniform confidence bands ensure that the function has a uniform level of confidence at each point irrespective of function value. A simple bootstrap method is used to construct the uniform confidence bands.[6]

These models are applied on micro level labor data on males from the Employment and Unemployment Survey (EUS) for India in 2004–05. Results show that the semiparametric estimates have less variance than do the estimates from the dummy variable model. Furthermore, they do not exhibit irregular behavior at the tails as do the polynomial model estimates.

My research on India follows Duraisamy (2002) and Dutta (2006), who present the estimates for marginal rate of return to education for India using NSSO data for the periods of 1983–1994 and 1983–1999, respectively. Other studies estimating returns to education are based on different data sources (Kingdon, 1998). Agarwal (2012) uses the India Human Development Survey (IHDS) 2005 data with quantile regression to examine the effect of education across the wage distribution. However, these studies use parametric Mincer regressions, using dummy variables for different schooling levels to capture the non-linear effect of schooling on education. These studies ignore people who drop out during primary school and focus to estimate only for certain schooling levels.[7] These studies also include other variables like occupation, union status of workers, and personal attributes, which serve the purpose of modeling

[6]I could not find information on constructing uniform confidence bands for the partial linear model in the literature.

[7]See Table D.2 in Appendix D for schooling levels evaluated by earlier studies.

earnings instead of evaluating the rate of return on schooling (Becker, 1964).

My work, on the other hand, estimates the marginal rate of return to education at each schooling level for more years of schooling than are commonly reported. Selection bias in the workforce and measurement error in reported schooling are ignored as many studies find the overall bias in the estimates to be negligible and statistically insignificant (Psacharopoulos and Patrinos, 2004; Ashenfelter et al., 2000).

Given a lack of data on taxes, the dominance of public schools, and the equal effects of the economic environment and uncertainty at aggregate level across individuals, I expect that the qualitative results of my essay remain unaffected in the absence of these factors.[8] My estimates for returns to education for India are qualitatively comparable to the estimates of Agarwal (2012).[9]

In the next section, I briefly discuss different model specifications, the estimation of the marginal rate of return, and estimation procedures for the semiparametric model. Section 4.3 introduces the data and presents preliminary findings. Results are presented and analyzed in Section 4.4. Section 4.5 summarizes the findings with concluding remarks.

## 4.2 Model Estimation

To estimate the functional form of the marginal rate of return to education, I start with a general model and discuss issues related to the estimation and identification of the functional form. This general model is further developed with restrictions

[8]See Kingdon (1996) for the role of public schools in India

[9]Check Table D.2 in Appendix D

to address these issues. The variables from the standard Mincer regression are used as a base in these specifications.

To estimate returns to education, one needs to model the relationship between wages and years of schooling. The basic structure of the model is given by:

$$log(w_i) = g(YED_i, PEX_i, X_i) + \varepsilon_i,$$

where $w_i$ is the wage rate, $YED_i$ is years of education, $PEX_i$ represents potential experience, and $X_i$ includes all other factors for individual $i$. I will ignore the subscript $i$ from here. The function $g(.)$, if left unspecified along with the error distribution, represents a nonparametric regression.

A well known problem with nonparametric regression is the 'curse of dimensionality.' Ramsay and Silverman (2005) describe this as a combination of an increase in computation cost, a decline in the optimal rate of convergence, and potential problems with model identification. To overcome the curse of dimensionality, some researchers limit the number of independent variables to one. Another suggestion is to assume that the $g(.)$ function is separately additive in each independent variable. In this case, the function $g(.)$ is specified as:

$$g(YED, PEX, X) = f_1(YED) + f_2(PEX) + f_3(X),$$

where $f_q(.)$, for $q = 1, 2, 3$, represents the unspecified functional forms for $YED$, $PEX$, and $X$, respectively.[10] With the addition of structure to the model and no

[10]For more on generalized additive models, see Hastie and Tibshirani (1986). Delgado and Robinson (1992) survey nonparametric and semiparametric methods for economics.

pre-specified functional form for any independent variable, we move into the area of semiparametric models. This generalization helps us to preserve the non-linear relationship of each variable with the dependent variable and to avoid pitfalls of nonparametric estimation. Local linear regression and spline functions are some of the methods used in applied econometrics to estimate the unspecified function over each variable.

Horowitz (1998) mentions other semiparametric methods to resolve the curse of dimensionality. Single index models use a parametric specification to reduce the dimension of the model to one. He then applies the nonparametric regression using that index. In terms of my model, this is given by the following specification of $g(.)$:

$$g(YED, PEX, X) = f(\beta_1 YED + \beta_2 PEX + \beta_3(X)),$$

where $f(.)$ is a function specified by the nonparametric regression and $\beta$ are the parameters. The parametric specification aggregates the effect of all independent variables. The nonparametric model captures the non-linearity between the dependent variable and the aggregate effect. Another method frequently used in empirical studies is a partial linear model, which uses an additive model with parametric restrictions on a limited number of independent variables. In my study, the partial linear model can be specified as:

$$log(w) = f_1(YED) + f_2(PEX) + \beta_3(X) + \varepsilon, \tag{4.2}$$

where $f_1(.)$ and $f_2(.)$ are unspecified functions and $\beta$ are the parameters for the model. Given the focus of this chapter, I will use the partial linear model to capture the non-

linearity between years of education and the log of earnings while accounting for the non-linearity between potential experience and the log of earnings.

Previous studies use the dummy variable approach to capture nonlinearity. However, this approach gives estimates with high variance and thus is not preferred to test the claim of Mookherjee and Ray (2010). The polynomial approach could be used to test said claim, but the estimates could behave erratically at the tails and may not give correct information. To compare all methods and show these issues, I estimate these models as well.

I use the dummy variable model given by:

$$log(w) = \sum_{l=1}^{L} \beta_{1l} D_{YED=l} + \beta_{21} PEX + \beta_{22} PEX^2 + \beta_3 X + \varepsilon. \tag{4.3}$$

In this equation, $L$ represents the total levels of schooling, $l$ represents different schooling levels, and a dummy variable, $D_{YED=l}$, is 1 for an individual with schooling level $l$, else 0. The estimated values of $\beta_{1l}$ show the total returns to education.

The polynomial model used is:

$$\begin{aligned} log(w) = \beta_0 &+ \beta_{11} YED^1 + \beta_{12} YED^2 + \beta_{13} YED^3 \\ &+ \beta_{21} PEX + \beta_{22} PEX^2 + \beta_{23} PEX^3 \\ &+ \beta_3 X + \varepsilon. \end{aligned} \tag{4.4}$$

I use polynomials up to the third degree for potential experience and years of education.

An IV estimation method for the partial linear model is required to addressing the issue of endogeneity. Blundell and Powell (2003) discuss different approaches to

address endogeneity for nonparametric and semiparametric models. They suggest the use of a control function approach to address endogeneity for partial linear models to ensure identification and consistency of estimates.

In theory, the control function approach addresses endogeneity by including the effect of omitted variables in the main equation to account for the variables creating endogeneity. In practice, a two stage regression is used in which the first stage estimates the endogenous variable using an instrumental variable and other independent variables. The residual estimated from the first stage regression is used as an independent variable in the second stage regression.

In the first stage, I estimate the following equation:

$$YED = \pi_{1k}(Z_k) + \pi_2(PEX) + \gamma_3 X + \nu. \tag{4.5}$$

In this equation, $Z_k$ represents the instrumental variables: parental and spousal education. $\pi_{1k}(.)$ is the unspecified function on the instrumental variable, $Z_k$. The independent variables $X$ and $PEX$ are included to account for effect in the main equation. As in the main model, the independent variable $X$ has a parametric restriction and $PEX_i$ is left unspecified. The function $\pi_2(.)$ and parameter $\gamma_3$ represent this unspecified function and parametric restriction, respectively. The unobserved error $\nu$ is estimated for each individual using this regression.

The estimated error term, $\hat{\nu}$, which accounts for omitted variables and noise in the first stage model, is included as an independent variable in second stage equation. Since the model is for semiparametric regression, the instrumental variable should be

left unspecified. This model is given as follows:

$$log(w_i) = \beta_0 + f_1(YED_i) + g_1(\hat{\nu}_i) + f_2(PEX_i) + \beta_3 X_i + \varepsilon_i. \tag{4.6}$$

Here, $g_1(.)$ represents an unspecified functional form for $\hat{\nu}$.

Further, Blundell and Powell (2003) give the following restrictions to ensure the endogeneity issue is addressed:

$$\begin{aligned} E(\varepsilon|YED, PEX, X, Z) &= E(\varepsilon|YED, PEX, X, \nu) \\ &= E(\varepsilon|\nu). \end{aligned} \tag{4.7}$$

These restrictions ensure that the second stage error term, $\varepsilon$, has zero covariance with all independent variables of both the equations.[11] For the purpose of this study, I assume that these restrictions are satisfied.

To estimate the partial linear model, I use the SAS procedure Proc GAM.[12] This procedure estimates the following model:

$$\begin{aligned} log(w) = \beta_0 + \beta_{11} YED + s_{12}(YED) + \beta_{21} PEX \\ + s_{22} PEX + \beta_3 X + \varepsilon. \end{aligned}$$

The $s_{12}$ and $s_{22}$ are the unknown spline functions which captures the non-linearity of YED and PEX, respectively, over the log of wages. Ordinary least squares are used to estimate the other models.

[11] While estimating a model, the error terms are forced to have zero covariance with independent variables. This forced zero covariance implies that all error terms would satisfy these restrictions. Therefore, there is no way to ensure these restrictions are satisfied.

[12] The robustness check of my results confirm that the qualitative results do not change with method of estimation. Appendix B further discusses the estimation procedure for the partial linear model. For more details on the Proc GAM procedure, see Xiang (2001).

Using the estimates of $\beta_{11}$ and $s_{12}(.)$ form the Proc GAM procedure, I calculate the total returns to education by:

$$\text{Total Return to Education} = \beta_1(YED) = \hat{\beta}_{11} YED + \hat{s}_{12}(YED).$$

The marginal returns to education are derived by two steps. In the first step, I use the piecewise polynomial form of the cubic spline to estimate the function of the total returns over schooling levels. The second step estimates the first derivative of this spline function.[13] This gives an estimate of marginal returns to education for each schooling level. To show the application of the procedure discussed and estimate the models, the next section explores the data for India. Later sections include the estimation results.

### 4.3 Data

In India, the National Sample Survey Organization (NSSO) conducts surveys on employment and unemployment. I use data from the 2004–2005 survey. This survey encompasses rural and urban areas and covers 124,680 households and 602,833 persons. Weights are used to replicate national level figures. For the application of the Mincer equation, I focus on individuals aged 16–64 who have positive wage income at the time of the survey.[14] Because literacy rates are low, governmental and non-governmental organizations run programs to educate individuals. Since most illiterates are adults, some programs provide informal education and NSSO data record

[13]MATLAB functions are used to run these steps. See Appendix A for more on estimation of returns to education.

[14]Agarwal (2012), Duraisamy (2002), and Dutta (2006) focus on the 16 to 64 age group.

the source of education for these individuals. For the formally educated, the highest grade completed is recorded. Table 1 shows the distribution of employed male individuals by educational status. Only 16.5% have passed high school, with 8.9% of the total male working population holding a college degree. The informally literate make up 2.6% of the total male working population.

Table 4.1: Distribution of the employed (male) population in India by level of educational attainment (based on NSSO Employment and Unemployment Survey 2004–05).

| Education Levels | % of Employed Population |
|---|---|
| Non-Literate | 29.3 |
| Literate without Formal Schooling | 2.6 |
| Below Primary School | 9.6 |
| Completed Primary Education | 15.1 |
| Completed Middle School | 17.4 |
| Completed Secondary School | 9.5 |
| Completed High School | 5.2 |
| Diploma/Certificate Course | 2.4 |
| College Graduate | 6.5 |
| Postgraduate and Above | 2.4 |

The NSSO records all economic and non-economic activities an individual is engaged in over the week prior to the survey. Because a person can be engaged in more than one job, I focus on his total earnings for the week and use it to calculate hourly wages. The survey also identifies the primary job of each respondent, which I use to identify job type. To make the study comparable to studies on other countries, I base my study on the wage earning male population and exclude individuals with

an informal education, pensioners, the disabled, the unemployed, and those who are engaged in 'non-economic' activities like prostitution, begging, and own-household workers.[15] Since the study is based on years of education and the measure we have in the NSSO survey is given by category, I assign a number of years required to complete each given category of education. I assign 0 years of school to illiterate individuals, three for below primary, five years if primary education has been completed, and so on until 18 years of school for post graduate and above.[16] After all of the exclusions, I am left with 61,473 observations that represent individuals across India.

I define potential experience as age minus years of schooling minus six.[17] Given that 37.5% of the sample has no formal schooling and many individuals drop out of school early, I calculate the potential experience for individuals with less than seven

[15]Jaeger and Pagé (1996), Card and Krueger (1992), Cameron and Taber (2004), Keane and Wolpin (2001), and Johnson (2010) use male only data. I could not find any study that includes informal education in estimating the returns to education (informal education represents roughly 2.5% of India's working population). Duraisamy (2002) only looks at individuals who are wage earners.

[16]In India, the masters degree in non-engineering and non-medical subjects requires 17 years of schooling, whereas for engineering and medical courses the requirements are 18 and 19 years, respectively. The M.Phil. or M.D. requires two more years of schooling, with an option of attending 3 more years of schooling to obtain a Ph.D. The total years of schooling at each level depends on the course or field of study. As the NSSO data do not record coursework information and combine all studies after a college degree into one category of 'post graduate and above,' I assign 18 years of schooling for this category. Because the proportion of students studying for Ph.D, M.Phil. or M.D. degrees is generally quite low, the average number of years required for 'post graduate degree or above' is expected to be close to 18.

[17]In India, a child should be at least five years old before being admitted to primary school. Duraisamy (2002) uses five years as the minimum age to calculate potential experience. Previous literature on the US uses six years as a minimum age to calculate the same (Jaeger and Pagé, 1996). Since five is a minimum age by law in India, I use six years as the age to enter primary school, assuming that the average age would be greater than five and close to six.

years of schooling as age minus 14, because the Constitution of India prohibits children below the age of 14 years from working in any factory or mine or engaging in any other hazardous employment. By the year 2000, laws were enacted that made employing children or facilitating child labor a criminal act, punishable with a prison term. Given the poor enforceability of the law in India, children from poor families usually start working at an age younger than 14. However, the experience accumulated before the age of 14 is not considered legal and on-the-job learning is not high. Therefore, I use 14 years of age as the age when one can start accumulating experience.

Following Becker (1964) and Psacharopoulos and Patrinos (2004), I do not control for household characteristics and job profile. This ensures that the model estimates only a Mincerian rate of return to education. However, in the literature, location as the external factor is used in Duraisamy (2000) to capture differences between the rural and urban economies in India. Further, Madheswaran and Attewell (2009) document that the social class has an effect on wage differences in India.[18]

Table 4.2: Mean or percentages of key variables used in the study for India.

| Variable | India |
|---|---|
| Potential Experience | 19.1 |
| Years of Education | 5.8 |
| Average Weekly Wage (in Indian Rupee) | 708 |
| % Rural Area | 67.1 |
| % Underprivileged | 71.5 |

[18]Bhaumika and Chakrabarty (2009) cite that in 2005, religion is not significant in explaining wage differences in India. For a detailed discussion check Abraham (2012).

I use the dummy variable "Underprivileged" and "Rural Area" to control for social and geographic factors, respectively, and include them in 'Other Factors'. Underprivileged takes value 1 for backward and 0 for upper caste individuals and Rural Area takes the value 1 if the individual is from a rural area and 0 if the individual is from an urban area.

Table 4.2 presents the summary of key variables used in this study. The population covered in the study has average potential experience of 19 years and average schooling of 6 years. The average weekly wage is Rs. 708. Urbanization is under 35% in India and the backward social class accounts for over 70% of the workforce.

## 4.4 Results

Using the NSSO data discussed in the previous section, I estimate models in Section 4.2. Subsection 4.4.1 discuss the results from models assuming exogeneity and subsection 4.4.2 studies the results of endogenous models.

### 4.4.1 Results for Models assuming Exogeneity

The measures of model fit and parameter estimates of variables for each model are given in Table 4.3 with standard errors in parentheses. The explanatory power of a model estimated by R-squared value shows that the more generalized specifications adds only 0.02 points to the standard Mincer regression. The Rural Area and Underprivileged variables have negative parameter estimates and do not change much across models. This seems in accordance to the previous literature on India and

Table 4.3: Parameter estimates and key goodness-of-fit indicators for the different models discussed in Section 4.2 using NSSO Data for India 2004–05.

| Variables | Standard Mincer | Semiparametric | Dummy | Polynomial Mincer |
|---|---|---|---|---|
| Intercept | 1.3892 | 1.6262 | | 1.5177 |
| | (0.0118) | (0.0002) | | (0.0148) |
| YED | 0.0917 | 0.0916 | | 0.0276 |
| | (0.0006) | (0.0014†) | | (0.0033) |
| $(YED)^2$ | | | | 0.0453 |
| | | | | (0.0055) |
| $(YED)^3$ | | | | -0.0014ˆ |
| | | | | (0.0023) |
| PEX | 0.047 | 0.0159 | 0.0457 | 0.0432 |
| | (0.0009) | (0.0006) | (0.0009†) | (0.0021) |
| $(PEX)^2$ | -0.0070 | | -0.0069 | -0.0056 |
| | (0.0002) | | (0.0002) | (0.0011) |
| $(PEX)^3$ | | | | -0.0002ˆ |
| | | | | (0.0002) |
| Underprivileged | -0.1585 | -0.1404 | -0.1406 | -0.1404 |
| | (0.0064) | (0.0002) | (0.0063) | (0.0063) |
| Rural Area | -0.4614 | -0.4635 | -0.4638 | -0.4641 |
| | (0.00063) | (0.0002) | (0.0062) | (0.0062) |
| R-Square | 0.4645 | 0.4794 | 0.4816 | 0.4807 |
| F-Value | 10661.5 | | 49378.6 | 7114.5 |

Note:- Each $j^{th}$ degree polynomial variable is multiplied by $10^{-(j-1)}$ to get standardized coefficient value. Estimates are significant at the 1% level unless specified otherwise.Number of observation for each model is 61,473. The smoothing parameters for YED and PEX in the semiparametric model are 0.5192 and 0.9999, respectively.

ˆInsignificant at 95% confidence level.

†Multiplied by $10^{-2}$.

similar countries.

The parametric estimates of the coefficient for years of education for the standard Mincer and semiparametric models are close indicating that the linear effect is

caputred equally well in both the models. The same is true for the coefficient estimates of potential experience in all models except the semiparametric. The polynomial model has a statistically significant positive estimate for the second order polynomial for years of education. This suggests that years of education is convex of the log of hourly wages.[19] To see the marginal estimates of each model in comparison to other models, refer to Figure 4.1.

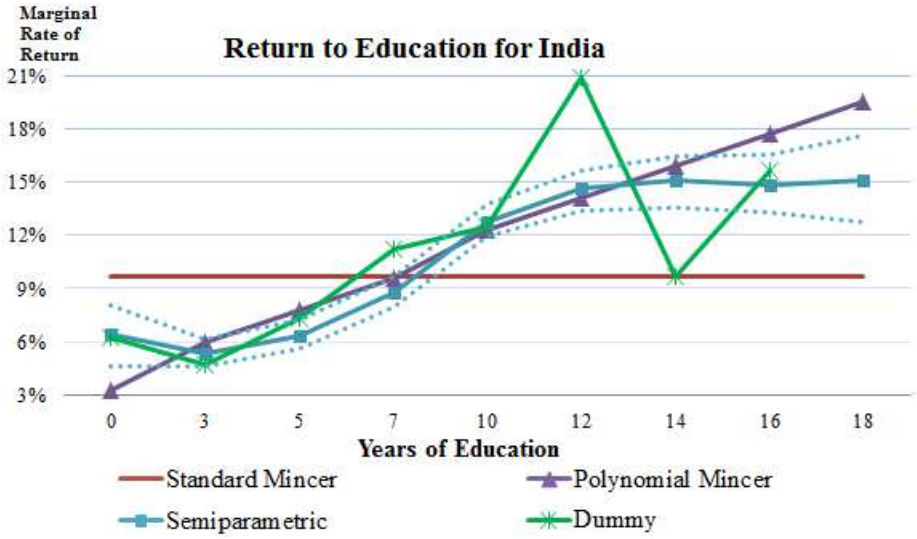


Figure 4.1: Marginal returns to education by years of education for India based on the NSSO Employment and Unemployment Survey 2004-05. Dotted lines show the uniform confidence band for the semiparametric estimates based on a simple bootstrap method.

The standard Mincer shows a constant return of 9.2% across all schooling levels. As suspected, the estimates of marginal returns from a polynomial Mincer

[19]The estimates of years of education for the dummy variable model is given in the Table D.1 in the Appendix D.

deviate more at the tails. The estimates from dummy variable model are close to semiparametric model estimates for the first ten years of schooling, but fluctuate around the semiparametric estimates for high school and beyond. The dummy variable estimates for high school completion and diploma or certificate courses fall outside of the uniform confidence bands of the semiparametric estimates. This suggests the presence of sheepskin effect.

Based on semiparametric model, the marginal rate of return to education is around 6% when an individual first enrolls in school. This declines to 5% in the next couple of years before going back up to 6% after finishing primary education. The marginal rate of return increases to 8% after finishing middle school, 11% after finishing secondary school, and 14% after finishing high school (or higher secondary). After 12 years of education, the marginal return varies around 14% – 15% during college and higher studies.[20]

The decline in marginal returns during primary education indicates the presence of concavity in returns to education. This decline may reflect financial constraints. Children from poor families with no adequate source of income drop out from primary schools to work as laborers in the informal sector. Due to the increased supply of workers and high competition from workers with completed primary schooling or no schooling, wages for primary school dropouts are depressed and the marginal returns are lower than those at the no schooling level. Since the marginal returns are

[20]Below Primary, Primary, Middle, Secondary, Higher Secondary, and diploma/Certificate Course are mapped to 3, 5, 8, 10, 12, and 14 years of schooling, respectively.

positive, workers with an incomplete primary education still make more money than those with no education. However, the uniform confidence band shows the decline is statistically insignificant at the 5% level. To ensure the functional form of returns to education is nonconcave inder endogeneity, I check for results using instrumental variables in the next subsection.

### 4.4.2 Results on Endogenous Models

In my search for an instrumental variable to estimate the endogenous model, I use the fact that the extended family culture is quite prevalent in rural India. Thus, the household level survey can be used to extract information on parental and spousal education. Using the household identifier and variable for relation with the head of the family, I match different individuals with their parents and spouse.[21] This gives me a large sample of individuals with their parental and spousal education. I use these variables as instruments to control for endogeneity in schooling and estimate equations 4.5 and 4.6 using the Proc GAM procedure in SAS.

Table 4.4 shows the estimate for first stage regression on $YED$ for each instrumental variable (IV) as an independent variable and with all three IVs together, I call it all-IV model. Across all IVs, spouse's education level has the highest explanatory power for explaining variation in individual education level, suggesting assortative mating is present in the Indian marriage market. The all-IV suggests the spouse's

[21]Some families have more than two married sons and some families have more than one spouse of the head of the family. These observations are less than 5% of the selected data. The younger spouses' data are deleted for the purpose of this study. Separate results with bigger a sample shows that the qualitative results remain unchanged.

Table 4.4: Parameter estimates and key goodness-of-fit indicators from the first stage model of IV with YED as the dependent variable (Equation 4.6) for each instrumental variable using NSSO Data for India 2004–05.

| | Relatives Education as Instrumental Variable | | | |
|---|---|---|---|---|
| | Spouse's | Father's | Mother's | All-IV |
| Obs. Used | 45,565 | 17,918 | 13,714 | 7,311 |
| Intercept | 5.1799 | 6.4124 | 7.6439 | 4.2461 |
| | (0.0016) | (0.0022) | (0.0025) | (0.0037) |
| Spouse's Education | 0.7162 | | | 0.5103 |
| | (0.0001) | | | (0.0003) |
| Father's Education | | 0.5630 | | 0.3039 |
| | | (0.0002) | | (0.0003) |
| Mother's Education | | | 0.6272 | 0.0071 |
| | | | (0.0003) | (0.0005) |
| PEX | -0.0333 | -0.0189 | -0.0385 | -0.0519 |
| | (0.0040†) | (0.0001) | (0.0001) | (0.0002) |
| Underprivileged | -0.6524 | -0.7689 | -0.8903 | -0.2492 |
| | (0.0010) | (0.0016) | (0.0019) | (0.0023) |
| Rural Area | -1.4042 | -1.2016 | -1.4295 | -0.1977 |
| | (0.0010) | (0.0016) | (0.0019) | (0.0025) |
| Smoothing Parameter | | | | |
| (Spouse's Education) | 0.7076 | | | 0.7214 |
| (Father's Education) | | 0.7606 | | 0.7431 |
| (Mother's Education) | | | 0.8485 | 0.8239 |
| (PEX) | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| R-Square | 0.5424 | 0.342 | 0.2903 | 0.5432 |
| Sum of Squares | | | | |
| Total | 1,429,013 | 422,429 | 323,517 | 195,261 |
| Explained | 775,080 | 144,516 | 93,903 | 106,069 |

†Multiplied by $10^{-2}$.

education level has the highest coefficient value among three IVs, followed by father and then mother. Potential experience is negatively associated with years of education. A male from a rural location attends 1.2–1.9 fewer years of schooling and a male from backward class attends 0.24–0.89 fewer years of schooling. The residual,

$\nu$, is estimated for each model and I use it as an independent variable in the second stage regression.

Table 4.5: Parameter estimates and key goodness-of-fit indicators from second stage model of IV with log of earnings as dependent variable (Equation 4.5) for each instrumental variable using NSSO Data for India 2004–05.

| | Relative's Education as Instrumental Variable | | | |
|---|---|---|---|---|
| | Spouse's | Father's | Mother's | All-IV |
| Intercept | 1.3207 | 1.2039 | 1.0435 | 1.132 |
| | (0.0004) | (0.0006) | (0.0008) | (0.0008) |
| YED | 0.1222 | 0.0943 | 0.1074 | 0.1063 |
| | (0.0025†) | (0.0051†) | (0.0068†) | (0.0056†) |
| 1st Stage Residual ($\hat{\nu}$) | -0.0419 | -0.0322 | -0.0465 | -0.0575 |
| | (0.0032†) | (0.0059†) | (0.0075†) | (0.0075†) |
| PEX | 0.0176 | 0.0291 | 0.0305 | 0.0282 |
| | (0.0007†) | (0.0017†) | (0.0022†) | (0.0029†) |
| Underprivileged | -0.0501 | -0.0998 | -0.0759 | -0.0686 |
| | (0.0002) | (0.0003) | (0.0003) | (0.0004) |
| Rural Area | -0.386 | -0.233 | -0.1806 | -0.2446 |
| | (0.0002) | (0.0003) | (0.0003) | (0.0005) |
| Smoothing Parameter | | | | |
| (YED) | 0.5052 | 0.5440 | 0.5397 | 0.5099 |
| (1st Stage Residuals($\nu$)) | 1 | 1 | 1 | 1 |
| (PEX) | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| R-Square | 0.5368 | 0.3803 | 0.3663 | 0.4317 |
| Sum of Squares | | | | |
| Total | 43,749 | 11,951 | 9,030 | 5,366 |
| Explained | 23,483 | 4,545 | 3,307 | 2,317 |

†Multiplied by $10^{-2}$.

Number of observations remains the same as in Table 4.4.

Table 4.5 provides the estimates of the second stage regression for different instrumental variables. The estimated first stage residual, $\hat{\nu}$, is assumed to include

all omitted variables. Across all models, $\hat{\nu}$ negatively affects the log of earnings after controlling for the effect of education, experience, geographical location, and social class. All other variables have same effects as in the model without IV, which suggests that the omitted variables are now being accounted for. A negative coefficient for $\hat{\nu}$ suggests the omitted variables are negatively correlated with log earnings. This suggests that the omitted variables constrain earnings and schooling and therefore show a downward bias in estimates. Further research on the movement in returns to education could shed some light on these variables.

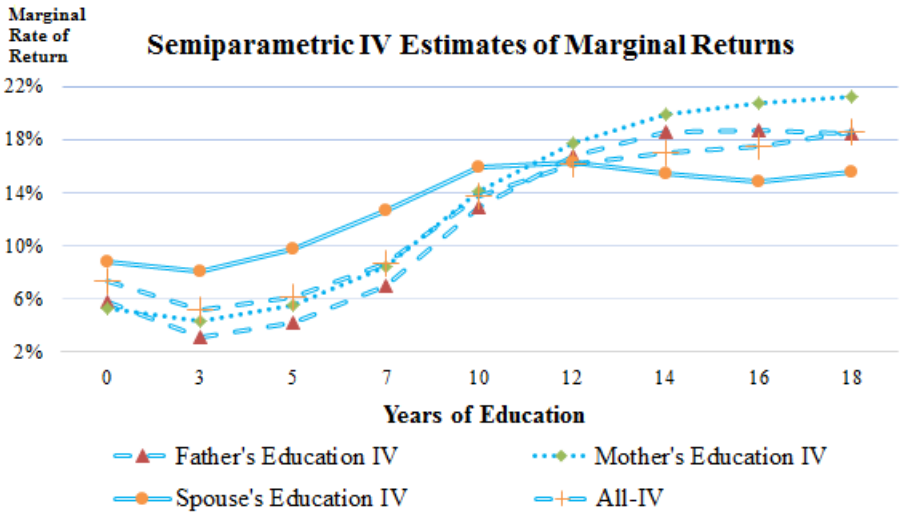


Figure 4.2: Semiparametric model estimates of returns to education by year of schooling in India for different IVs and the all-IV model.

The estimates of marginal returns for the models presented in Table 4.5 and 4.4 are plotted on Figure 4.2.[22] The IV estimates based on spouse's education are rel-

[22]The estimates summarized by Figure 4.2 are presented in Table D.3 in the Appendix D.

atively higher than are estimates based on other IV models for the first ten years of schooling and lower for high school and beyond. The mother's education IV shows higher marginal returns for schooling years 12 and above in comparison to other IVs. Except for spousal education, all other estimates show non-decreasing marginal returns after high-school. Similar to the estimates under exogenous schooling, all estimates suggest a decline in marginal returns to education during primary schooling.

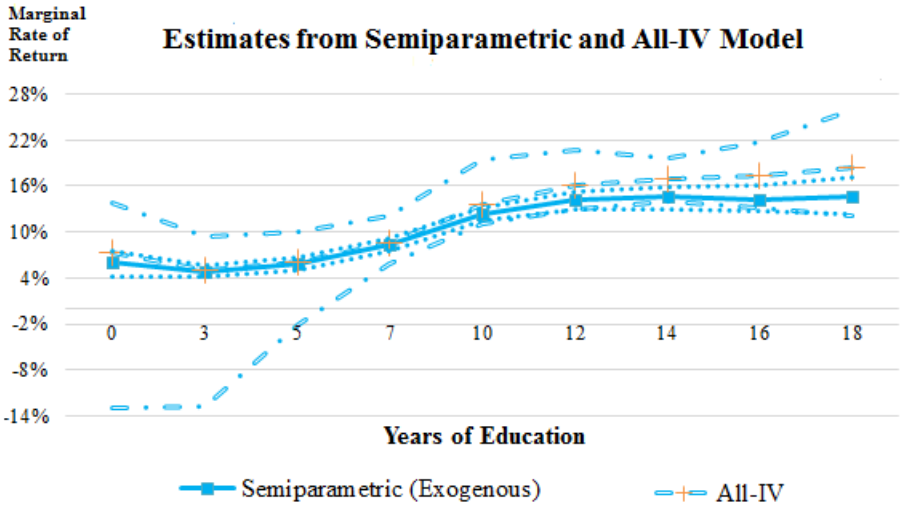


Figure 4.3: Uniform confidence bands for the semiparametric model and all-IV model estimates of the returns to education by year of schooling for India.

However, the uniform confidence bands constructed at a 5% level suggest statistical insignificance of this decline for each IV.[23] Figure 4.3 presents the uniform confidence bands for semiparametric estimates under an exogeneity assumption and

[23]See Figure D.1, D.2, and D.3 in Appendix D for Graphs on individual IV

under endogenous schooling and earnings. This shows that the decline in marginal returns during primary education is not statistically significant for either model. The wider band for IV estimates is possibly due to the smaller sample size. In all, we fail to reject the presence of nonconcavity in returns to education for India based on given sample.

## 4.5 Conclusion

The generalized cubic polynomial specification increases the adjusted R-squared value by 0.02 points over the standard Mincer regression but fails to estimate the returns to education function appropriately. The semiparametric specification does a better job in estimating the returns to education function. Like other studies on India, my results show that returns to education increase at a higher rate after the first seven years of schooling. Moreover, I find that the marginal rate of return on education varies non-monotonically with respect to years of schooling in India. The estimates for the marginal rate of return to education for individuals who drop out during the first five years of schooling are lower than those for individuals with no schooling. The marginal rate of return increases with schooling level after primary school until college and does not vary much after high school. However, the uniform confidence bands suggest the decline is statistically insignificant at a 5% level.

Under endogenous education and earnings, estimates based on parental and spousal education as IV suggest a similar pattern. Comparing different IV estimates with semiparametric estimates under exogeneity suggests that controlling for endo-

geneity gives higher estimates of marginal returns for schooling level after high-school. The uniform confidence bands over the estimated function for all IVs also fail to reject nonconcavity in returns to education. All of these results suggest the presence of concavity in returns to education during primary schooling for India, but this concavity is statistically insignificant. Thus, we fail to reject the claim of nonconcavity in returns to human capital made by Mookherjee and Ray (2010).

# APPENDIX A
# DERIVING THE RETURNS TO EDUCATION

The marginal return is derived by taking the derivative of $log(w)$ with respect to $YED$. In the standard Mincer regression (Equation 3.1 and 4.1), it is represented by the value of the coefficient for years of education:

$$\frac{d(log(w))}{d(YED)} \mid_{YED} = \beta_1.$$

This is a single estimate for all education levels and is interpreted as the return of an additional year of schooling.

For the dummy variable model, the marginal return at a given schooling level is estimated by taking the change in total returns from the previous schooling level divided by the years of schooling for the given level:

$$\text{Marginal Return to } l \text{ level of education} = \frac{\beta_{1(l)} - \beta_{1(l-1)}}{YED_l - YED_{l-1}}.$$

This method leads to the loss of the estimate at the first level.

The MRRE at each schooling level for the polynomial model can be calculated by using the following equation:

$$\frac{d(log(w))}{d(YED)} \mid_{YED} = \beta_{11} + \beta_{12} 2YED + \beta_{13} 3YED^2.$$

Based on the estimated structure for the non-parametric function of $\beta_1(.)$, the total returns to education for the semiparametric specification is given by:

$$\text{Total Return to education} = \beta_1(YED) = \beta_{11} YED + s_{12}(YED),$$

where Proc GAM output in SAS provides a linear estimate for YED, $\hat{\beta}_{11}$, as well as a non-parametric estimate of YED, the value of the function $\hat{s}_{12}(.)$ for each schooling level.

To derive the marginal returns to education, I use MATLAB on the estimated values of the total returns to education across schooling levels. First, I use the "spline" function to get the piecewise polynomial form of the cubic spline fitting total returns to education for each schooling level. Then, a first derivative function of the piecewise polynomial function is found by using the "fnder" function. The value of this first derivative, an estimate of the marginal returns to education at each schooling level, can then be found by using the "ppval" function.

# APPENDIX B
# ESTIMATING THE SEMIPARAMETRIC REGRESSION

The estimation of a partial linear model includes an estimation of parameters for the parametric portion and the estimation of parameters and the functional form of the non-parametric portion. In the literature, a similar model is used by Tobias (2003) to estimate returns to ability. He uses a two step method which separates the estimation for the parametric and the non-parametric model. In the first step, OLS is used to estimate the parametric model by cancelling out the nonparametric part. In my model, the first step of this process can be done by subtracting observation $(i+1)$ from $i$ when the schooling level in these two observations are equal. This can also be shown as two individuals with equal schooling levels, $YED_i = YED_{i+1}$. The first order differencing gives us following:

$$\begin{aligned} log(w_i) - log(w_{i+1}) = \beta_{21}(PEX_i - PEX_{i+1}) + \beta_{22}(PEX_i^2 - PEX_{i+1}^2) \\ + \beta_{23}(PEX_i^3 - PEX_{i+1}^3) + (f_1(YED_i) - f_1(YED_{i+1})) \\ + \beta_3(X_i - X_{i+1}) + \varepsilon_i - \varepsilon_{i+1}. \end{aligned}$$

Assuming the continuity of the function $f_1$, the nonparametric part will cancel out from the model and OLS can be used to estimate the $\beta$ values. These estimates are used in the second step to remove the effects of variables specified in the parametric model from the dependent variable. The residual is then used as the dependent variable in a local linear regression (a nonparametric regression technique) where $YED$ is the independent variable.

I use different methods to estimate the partial linear models and find that all methods, including the one used by Tobias (2003), produce similar qualitative results for the purpose of this study. Here, I discuss the results of Generalized Additive Models (GAM) because the SAS procedure used for GAM is readily available to researchers and it shows the linear and non-linear parts of the non-parametric model separately. The SAS procedure Proc GAM is used to estimate the partial linear model. The model I estimate has the following form:

$$\begin{aligned} log(w) = \beta_0 + \beta_{11}YED + s_{12}(YED) + \beta_{21}PEX + \beta_{22}PEX^2 \\ + \beta_{23}PEX^3 + \beta_3 X + \varepsilon. \end{aligned}$$

Here, $s_{12}$ is an unknown function that captures the non-linearity of YED over the log of wages.

The Proc GAM procedure uses a two loop iteration process referred to as a Back-fitting (call it inner loop) and Local Scoring algorithm (call it outer loop) (Xiang, 2001). This procedure starts with a loop using the Local Scoring algorithm by initializing an unknown function ($s_{12}$) and estimates some other values (such as the weights and the adjusted dependent variable). This loop has a nested loop, a

back-fitting algorithm, that uses values from the outer loop to get estimates of $\beta$ and $s_{12}$. For my model, the inner loop runs until it fails to decrease or satisfies the convergence criterion over the following equation:

$$RSS = \frac{1}{n}\|log(w) - \hat{\beta}_0^{(m)} - \hat{\beta}_{11}^{(m)} YED - \hat{s}_{12}^{(m)}(YED) - \hat{\beta}_{21}^{(m)} PEX - \hat{\beta}_{22}^{(m)} PEX^2 - \hat{\beta}_{23}^{(m)} PEX^3 - \hat{\beta}_3^{(m)} X - \hat{\beta}_3^{(m)} X\|^2,$$

where $m$ is the number of the iteration. Based on the estimated values of $\beta$ from the weighted back-fitting algorithm, the outer loop calculates a new set of weights and an adjusted dependent variable. The outer loop runs until the estimates satisfy the convergence criterion (which is $10^{-8}$). The literature on Proc GAM is well established and any discussion of the estimation is outside of the scope of this paper. For more information on the estimation of $\beta$ and $s_{12}$, please see SAS documentation or Xiang (2001).

For estimating the unknown function, I use cubic smoothing splines as smoothers to find the function with two continuous derivatives. Cubic smoothing splines are the unique minimizer of the penalized least squares, which is a a method to measure the fit of a function on the data using least squares with a penalty for curvature. The smoothing parameter for each nonparametric part of the model is selected by the Generalized Cross Validation (GCV) method. A larger value of the smoothing parameter produces smoother curves. The statistical significance of each smoothing parameter is also given in the SAS output. Proc GAM does the entire process and gives the final estimates for the following model as the output:

$$log(w) = \beta_0 + \beta_{11} YED + s_{12}(YED) + \beta_{21} PEX + \beta_{22} PEX^2 + \beta_{23} PEX^3 + \beta_3 X + \varepsilon.$$

Data are robust to the choice of method for nonparametric estimation. For more details on Proc GAM, refer to the SAS support documentation on Proc GAM.

# APPENDIX C
# CHAPTER 3 SUPPORTING TABLES

Table C.1: Parameter estimates for YED in the Dummy Mincer Model with Standard Error for the 1980 US census data.

| Dummy for YED | Parameter Estimate | Standard Error |
|---|---|---|
| 0 | 7.8132 | 0.0087 |
| 1 | 7.8037 | 0.0120 |
| 2 | 7.8554 | 0.0098 |
| 3 | 7.8632 | 0.0076 |
| 4 | 7.9601 | 0.0070 |
| 5 | 7.9691 | 0.0059 |
| 6 | 8.0073 | 0.0044 |
| 7 | 8.0858 | 0.0040 |
| 8 | 8.1633 | 0.0027 |
| 9 | 8.0809 | 0.0026 |
| 10 | 8.0448 | 0.0022 |
| 11 | 8.1692 | 0.0021 |
| 12 | 8.5487 | 0.0013 |
| 13 | 8.6180 | 0.0020 |
| 14 | 8.6836 | 0.0019 |
| 15 | 8.6854 | 0.0026 |
| 16 | 8.9538 | 0.0018 |
| 17 | 8.9639 | 0.0028 |
| 18 | 9.0353 | 0.0030 |
| 19 | 9.0796 | 0.0038 |
| 20 | 9.1798 | 0.0032 |

Table C.2: Key statistics from First Stage Model of IV with YED as the dependent variable (Equation 3.6) for each instrumental variable using 1980 US census data.

| | Instrumental Variable | | |
|---|---|---|---|
| | Spouse's Education | Father's Education | Mother's Education |
| Obs. Used | 1,542,650 | 6,981 | 23,298 |
| Intercept | 5.3461 | 10.514 | 11.0792 |
| | (0.0059) | (0.0278) | (0.0165) |
| Z | 0.7034 | 0.3092 | 0.3018 |
| | (0.0004) | (0.0021) | (0.0012) |
| PEX | -0.0492 | -0.0359 | -0.0496 |
| | (0.0001) | (0.0007) | (0.0004) |
| Black | -1.0450 | -0.3899 | 0.8356 |
| | (0.0039) | (0.0308) | (0.0149) |
| Non-Metropolitan | -0.5528 | -0.7704 | -0.7973 |
| | (0.0022) | (0.0192) | (0.0105) |
| Smoothing Parameter | | | |
| (Z) | 0.9926 | 0.9868 | 0.9933 |
| (PEX) | 0.9992 | 0.9998 | 0.9993 |
| R-Square | 0.430 | 0.192 | 0.191 |
| Sum of Squares | | | |
| Total | 64,253,704 | 1,673,361 | 5,270,370 |
| Explained | 27,645,869 | 321,231 | 1,005,746 |

The estimates for the semiparametric model for spousal education IV is based on simple random sample of 300,000 observations due to memory limitations.

Table C.3: Parameter estimates and key goodness-of-fit indicators from Second Stage Model of IV with log of earnings as dependent variable (Equation 3.5) for each instrumental variable using 1980 US census data.

| | Instrumental Variable | | |
|---|---|---|---|
| | Spouse's Education | Father's Education | Mother's Education |
| Intercept | 8.4506 | 8.5597 | 8.3181 |
| | (0.0020) | (0.0135) | (0.0118) |
| YED | 0.0935 | 0.0761 | 0.0957 |
| | (0.0001) | (0.001) | (0.0009) |
| 1st Stage Residual ($\hat{\nu}$) | -0.0268 | -0.0094 | -0.0273 |
| | (0.0002) | (0.0011) | (0.0009) |
| PEX | 0.0003 | 0.0003 | 0.0003 |
| | (0.0049‡) | (0.0003†) | (0.0002†) |
| Black | -0.2785 | -0.3255 | -0.2502 |
| | (0.0010) | (0.0049) | (0.0034) |
| Non-Metropolitan | -0.1245 | -0.1072 | -0.1358 |
| | (0.0006) | (0.0032) | (0.0024) |
| Smoothing Parameter | | | |
| (YED) | 0.9923 | 0.9899 | 0.9893 |
| (1st Stage Residuals($\nu$)) | 1 | 0.0099 | 1 |
| (PEX) | 0.9992 | 0.9978 | 0.9993 |
| R-Square | 0.199 | 0.489 | 0.190 |
| Sum of Squares | | | |
| Total | 2,752,649 | 77,117 | 251,643 |
| Explained | 546,402 | 37,739 | 47,712 |

Number of observations remain the same as in Table C.2

†Multiplied by $10^{-2}$.

‡Multiplied by $10^{-4}$.

# APPENDIX D
# CHAPTER 4 SUPPORTING TABLES AND GRAPHS

Table D.1: Parameter estimates for YED in the Dummy Mincer Model with Standard Error using NSSO Data for India 2004–05.

| Dummy for YED | Parameter Estimate | Standard Error |
|---|---|---|
| 0 | 1.5006 | 0.0121 |
| 3 | 1.6737 | 0.0133 |
| 5 | 1.7581 | 0.0119 |
| 7 | 1.8943 | 0.0114 |
| 10 | 2.2158 | 0.0125 |
| 12 | 2.4553 | 0.0146 |
| 14 | 2.8625 | 0.0190 |
| 16 | 3.0473 | 0.0132 |
| 18 | 3.3512 | 0.0189 |

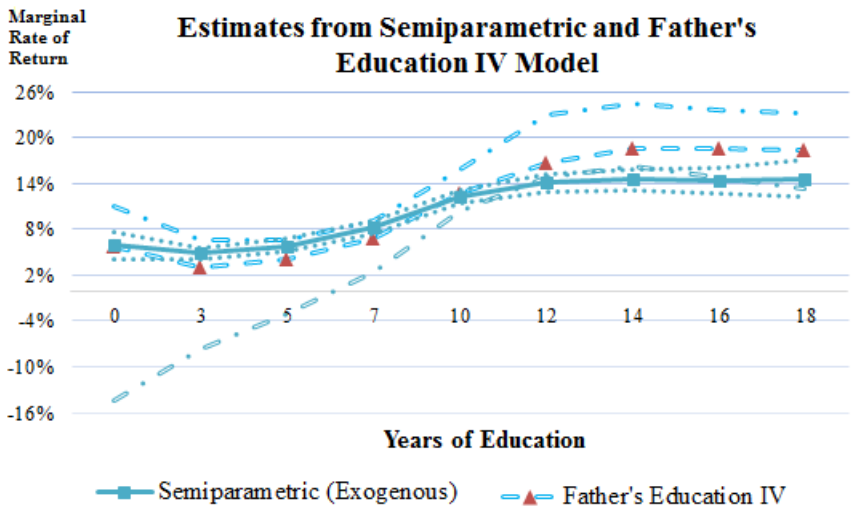


Figure D.1: Uniform confidence bands for father's education IV model estimates of returns to education by year of schooling for India.

Table D.2: Estimates of marginal rate of return to education for males in this study and some of the previous studies for India under exogeneity.

| Schooling Level | This Paper | Agarwal(2012)* | Dutta (2006)** |
|---|---|---|---|
| Non-Literate | 5.98 | | |
| Below Primary | 4.86 | | |
| Primary | 5.87 | 5.47 | 5.6 |
| Middle | 8.32 | 6.15 | 3.5 |
| Secondary | 12.30 | 11.38 | 6.1 |
| Higher Secondary | 14.16 | 12.21 | |
| Diploma/Certificate Course | 14.63 | | |
| College Graduate | 14.30 | 15.87 | 12.3 |
| Postgraduate and Above | 14.62 | | |

* Males and females both included. Based on India Human Development Survey (IHDS) 2005

** For years 1999-2000. Males regular workers only.

Note:-The estimates of my study are directionally similar to other studies. This paper estimates returns for all given levels of schooling. The difference in the estimates can be attributed to differences in data source, sample exclusions, and time frame of the study.

Table D.3: Estimates of marginal returns based on the semiparametric IV model for each instrumental variable and the semiparametric estimates under exogenous schooling using NSSO Data for India 2004–05y.

| Years of Schooling | Relative's Education as Instrumental Variable | | | | Exogenous Semiparametric |
|---|---|---|---|---|---|
| | All | Spouse's | Father's | Mother's | |
| 0 | 7.23 | 8.78 | 5.74 | 5.26 | 5.98 |
| 3 | 5.10 | 7.98 | 3.06 | 4.24 | 4.86 |
| 5 | 6.05 | 9.66 | 4.12 | 5.49 | 5.87 |
| 7 | 8.65 | 12.64 | 6.93 | 8.33 | 8.32 |
| 10 | 13.63 | 15.85 | 12.79 | 13.98 | 12.30 |
| 12 | 16.08 | 16.16 | 16.69 | 17.67 | 14.16 |
| 14 | 16.93 | 15.36 | 18.54 | 19.86 | 14.63 |
| 16 | 17.38 | 14.70 | 18.66 | 20.64 | 14.30 |
| 18 | 18.47 | 15.45 | 18.35 | 21.19 | 14.62 |

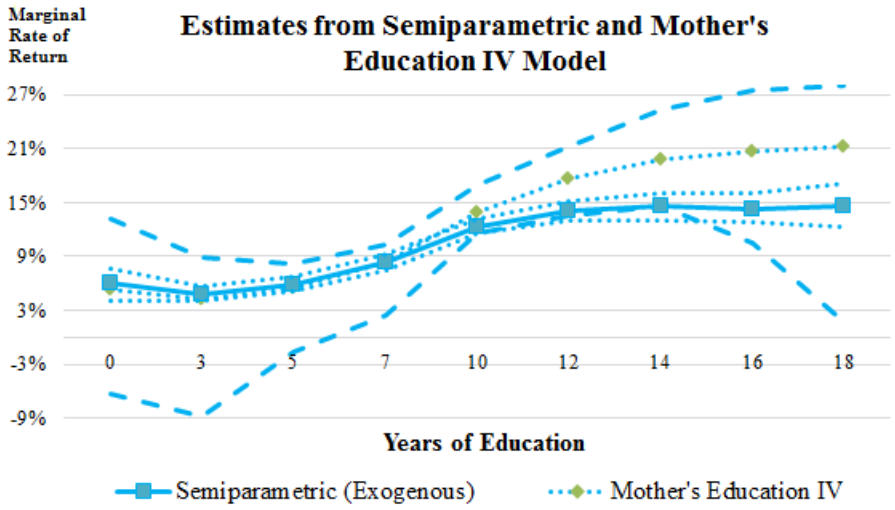


Figure D.2: Uniform confidence bands for mother's education IV model estimates of returns to education by year of schooling for India.

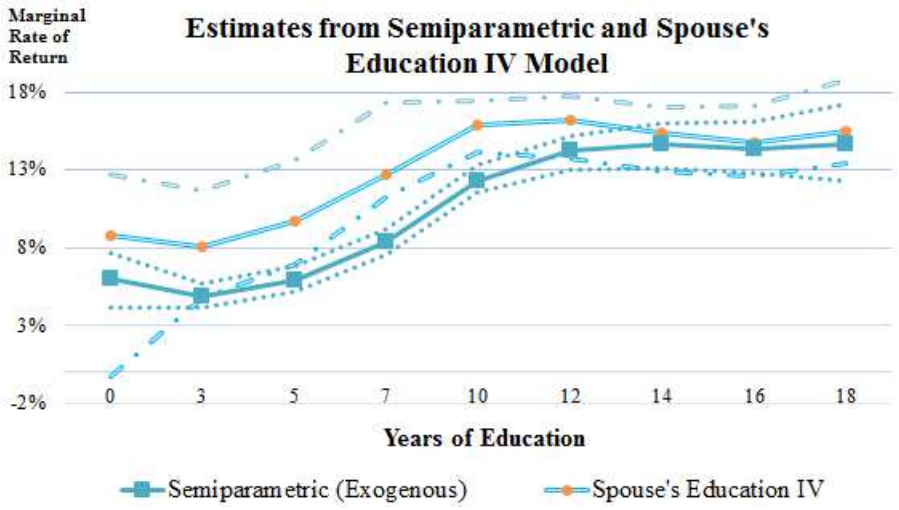


Figure D.3: Uniform confidence bands for Spouse's education IV model estimates of returns to education by year of schooling for India.

# REFERENCES

Abraham, V. (2012). Wages and earnings of marginalized social and religious groups in india: Data sources, scope, limitations and suggestions.

Agarwal, T. (2012). Returns to education in india: Some recent evidence. *Journal of Quantitative Economics* 10(2):131–51.

Ahsan, A. and Pagés, C. (2009). Are all labor regulations equal? evidence from indian manufacturing. *Journal of Comparative Economics* 37(1):62–75.

Albrecht, J., Navarro, L. and Vroman, S. (2009). The effects of labor market policies in an economy with an informal sector. *The Economic Journal* 119:1105–1129.

Albrecht, J. Navarro, L. and Vroman, S. (2002). A matching model with endogenous skill requirement. *International Economic Review* 43:283–305.

Antunes, A.R. and de V. Cavalcanti, T.V. (2007). Start up costs, limited enforcement, and the hidden economy. *European Economic Review* 51(1):203 – 224.

Asadullah, M.N. and Yalonetzky, G. (2012). Inequality of educational opportunity in india: Changes over time and across states. *World Development* 40:1151–1163.

Ashenfelter, O., Harmon, C. and Oosterbeek, H. (2000). A review of estimates of the schooling/earnings relationship, with tests for publication bias. NBER Working Paper 7457, January, 2000.

Becker, G.S. (1964). Human capital: A theoretical and empirical analysis. National Bureau of Economic Research.

Becker, G.S. (1967). *Human Capital and the Personal Distribution of Income: An Analytical Approach*. Woytinsky Lecture no.1. Ann Arbor: University of Michigan, Institute of Public Administration.

Becker, G.S. (1975). Investment in human capital: Rates of return. In: *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. NBER.

Bernera, E., Gomeza, G. and Knorringa, P. (2012). 'helping a large number of people become a little less poor': The logic of survival entrepreneurs. *European Journal of Development Research* 24:382 – 396.

Besley, T. and Burgess, R. (2004). Can labor regulation hinder economic performance? evidence from india. *The Quarterly Journal of Economics* 119(1):91–134.

Bhaumika, S.K. and Chakrabarty, M. (2009). Is education the panacea for economic deprivation of muslims?: Evidence from wage earners in india, 19872005. *Journal of Asian Economics* 20:137–149.

Blundell, R. and Powell, J.L. (2003). Endogeneity in nonparametric and semiparametric regression models. In: *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Volume II,*, chap. 8, pp. 312–357. Cambridge University Press.

Boeri, T. and Garibaldi, P. (2007). Shadow sorting. NBER International Seminar on Macroeconomics 2005, MIT Press.

Botero, J., Djankov, S., La Porta, R., de Silanes F., L. and Shleifer, A. (2004). The regulation of labor. *The Quaterly Journal of Economics* 119:1339–1382.

Cameron, S. and Taber, C. (2004). Estimation of educational borrowing constraints using returns to schooling. *Journal of Political Economy* 112:132–82.

Card, D.E. (1993). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. National Bureau of Economic Research 4483.

Card, D.E. (1995). Earnings, schooling, and ability revisited. In: *Research in Labor Economics*, vol. 14, pp. 23–48. JAI Press, Greenwich, CT.

Card, D.E. (1999). The causal effect of education on earnings. In: *Handbook of Labor Economics,*, vol. 3, chap. 30, pp. 1801–1863. Elsevier, Amsterdam.

Card, D. and Krueger, A.B. (1992). Does school quality matter? returns to education and the characteristics of public schools in the united states. *The Journal of Political Economy* 100:1–40.

Carneiro, P., Heckman, J.J. and Vytlacil, E.J. (2010). Estimating marginal returns to education. Working Paper 16474, National Bureau of Economic Research.

Charlot, O., Malherbet, F. and Terra, C. (2012). Informality in developing economies: Regulation and fiscal policies. Revised version of IZA Discussion Paper No. 5519, under revision.

Colclough, C., Kingdon, G. and Patrinos, H. (2010). The changing pattern of wage returns to education and its implications. *Development Policy Review* 28:733–747.

Das, M. (2005). Instrumental variables estimators of nonparametric models with discrete endogenous regressors. *Journal of Econometrics* 124(2):335 – 361.

Dearden, L. (1999). The effects of families and ability on men's education and earnings in britain. *Labour Economics* 6(4):551 – 567.

Delgado, M.A. and Robinson, P.M. (1992). Nonparametric and semiparameetric methods for economic research. *Journal of Economic Surveys* 6:201–249.

Devereux, P.J. and Fan, W. (2011). Earnings returns to the british education expansion. *Economics of Education Review* 30(6):1153 – 1166. Special Issue: Economic Returns to Education.

Dickson, M. and Harmon, C. (2011). Economic returns to education: What we know, what we dont know, and where we are goingsome brief pointers. *Economics of Education Review* 30(6):1118 – 1122. Special Issue: Economic Returns to Education.

Dickson, M. and Smith, S. (2011). What determines the return to education: An extra year or a hurdle cleared? *Economics of Education Review* 30(6):1167 – 1176. Special Issue: Economic Returns to Education.

Djankov, S., La Porta, R., de Silanes F., L. and Shleifer, A. (2002). The regulation of entry. *The Quaterly Journal of Economics* 117:1–37.

Djankov, S., McLiesh, C. and Ramalho, R.M. (2006). Regulation and growth. *Economic Letters* 92:395–401.

Duraisamy, P. (2002). Changes in the returns to education in india, 198394: by gender, age-cohort and location. *Economics of Education Review* 21(6):609–622.

Dutta, P. (2006). Returns to education: New evidence for india, 19831999. *Education Economics* 14(4):431–451.

Gabriel, S.A. and Rosenthal, S.S. (1999). Location and the effect of demographic traits on earnings. *Regional Science and Urban Economics* 29:445–461.

Gelbach, J.B. (2009). When do covariates matter? and which ones, and how much? Mimeo, Eller College of Management, University of Arizona.

Gërxhani, K. (2004). The informal sector in developed and less developed countries: A literature survey. *Public Choice* 120(3-4):267–300.

Gollin, D. (2008). Nobody's business but my own: Self-employment and small enterprise in economic development. *Journal of Monetary Economics* 55(2):219 – 233.

Gorodnichenko, A. and Peter, K.S. (2005). Returns to schooling in russia and ukraine:a semiparametric approach to cross-country comparative analysis. *Journal of Comparative Economics* 33:324–350.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models:- A roughnesss panalty approach.* 1st ed. Chapman and Hall.

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica* 45:1–22.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* 1(3):297–318.

Heckman, J.J., Lochner, L.J. and Todd, P. (2003). Fifty years of mincer earnings regressions. NBER Working Paper 9732.

Heckman, J.J. and Vytlacil, E. (2001). Identifying the role of cognitive ability in explaining the rising return to education. *Review of Economics and Statistics* 83(1):1–12.

Heckman, J.J. and Vytlacil, E. (2003). Econometric evaluation of social programs. In: *Handbook of Econometrics*. Amsterdam: North-Holland.

Heckman, J., Layne-Farrar, A. and Todd, P. (1996). Human capital pricing equations with an application to estimating the effect of schooling quality on earnings. *The Review of Economics and Statistics* 78(4):pp. 562–610.

Heckman, J.J. (2008). Schools, skills and synapses. *Economic Inquiry* 46:289–324.

Heckman, J.J., Lochner, L.J. and Todd, P.E. (2008). Earnings functions and rates of return. *Journal of Human Capital* 2(1):pp. 1–31.

Heckman, J.J. and Urza, S. (2010). Comparing {IV} with structural models: What simple {IV} can and cannot identify. *Journal of Econometrics* 156(1):27 – 37. Structural Models of Optimization Behavior in Labor, Aging, and Health.

Henley, A., Arabsheibani, G.R. and Carneiro, F. (2006). On defining and measuring the informal sector. World Bank Policy Research Working Paper 3866.

Horowitz, J.L. (1998). *Semiparametric Methods in Econometrics*. Springer.

Hungerford, T. and Solon, G. (1987). Sheepskin effects in the returns to education. *The Review of Economics and Statistics* 69(1):pp. 175–177.

ILO, D.O.S. (2012). Statistical update on employment in the informal economy.

Imbens, G.W. (2014). Instrumental variables: An econometrician's perspective. Working Paper 19983, National Bureau of Economic Research.

Jaeger, D.A. and Pagé, E.A. (1996). Degree matter: New evidence on sheepskin effects in the returns to education. *The Review of Economics and Statistics* 78:733–740.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *Quaterly Journal of Economics* 125(2):515–548.

Johnson, M. (2010). Borrowing constraints, college enrolment, and delayed entry. Mathematica Policy Research, Working Paper.

Keane, M. and Wolpin, K. (2001). The effect of parental transfers and borrowing constraints on educational attainment. *International Economic Review* 42:1051–1103.

Kingdon, G. (1996). The quality and efficiency of private and public education: a case-study of urban india. *Oxford Bulletin of Economics & Statistics* 58(1):57–82.

Kingdon, G. (1998). Does the labour market explain lower female schooling in india? *Journal of Development Studies* 35(1):39–65.

Lemieux, T. (2006). The mincer equation thirty years after schooling, experience, and earnings. In: *Jacob Mincer A Pioneer of Modern Labor Economics*. Springer Science+Business Media Inc.

Loayza, N., Oviedo, A. and Servn, L. (2005). The impact of regulation on growth and informality- cross-country evidence. World Bank Policy Research Working Paper No. 3623.

Madheswaran, S. and Attewell, P. (2009). Caste discrimination in the indian urban labour market: Evidence from the national sample survey. *Economic and Political Weekly* 42:4146–4153.

Maloney, W.F. (1999). Does informality imply segmentation in urban labor markets? evidence from sectoral transitions in mexico. *The World Bank Economic Review* 13:275–302.

Mookherjee, D. and Ray, D. (2010). Inequality and markets: Some implications of occupational diversity. *American Economic Journal: Microeconomics* 2(4):pp. 38–76.

Mortensen, D.T. and Pissarides, C.A. (1994). Job creation and job destruction in the theory of unemployment. *Review of Economic Studies* 61:397–415.

Murphy, K.M. and Welch, F. (1990). Empirical age-earnings profiles. *Journal of Labor Economics* 8:202–229.

Pagés-Serra, C. (2000). The cost of job security regulation: Evidence from latin american labor markets. *Journal of LACEA Economia* .

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

Psacharopoulos, G. and Patrinos, H.A. (2004). Returns to investment in education: A further update. *Education Economics* 12(2):111–134.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. 2nd ed. Springer.

Raveendran, G. Murthy, S.V.R. and Naik, A.K. (2006). Expert group on informal sector statistics (delhi group). In: *Outline and Progress Report on the Manual on Surveys of Informal Employment and Informal Sector*.

Tobias, J.L. (2003). Are returns to schooling concentrated among the most able? a semiparametric analysis of the ability-earnings relationships. *Oxford Bulletin of Economics and Statistics* 65:1–29.

Trostel, P., Walker, I. and Woolleyc, P. (2002). Estimates of the economic return to schooling for 28 countries. *Labor Economics* 9:1–16.

Ulyssea, G. (2010). Regulation of entry, labor market institutions and the informal sector. *Journal of Development Economics* 91:87–99.

Wang, L. (2013). Estimating returns to education when the {IV} sample is selective. *Labour Economics* 21(0):74 – 85.

Willis, R.J. and Rosen, S. (1979). Education and self-selection. *Journal of Political Economy* 87(5):S7–S36.

Xaba, J., Horn, P., Motala, S. and Singh, A. (2002). The Informal Sector in Sub-Saharan Africa. Employment Sector 2002/10, Working Paper on the Informal Economy, International Labour Office Geneva.

Xiang, D. (2001). Fitting generalized additive models with the gam procedure. SAS Institute Inc.

Yatchew, A. and No, J.A. (2001). Household gasoline demand in canada. *Econometrica* 69(6):1697–1709.

Zenou, Y. (2008). Job search and mobility in developing countries. theory and policy implications. *Journal of Development Economics* 86:336–355.