On the Empirical Content of the Formal-Informal Labor Market Segmentation Hypothesis

Ricardo Paes de Barros**
Gabriel Ulyssea***

Abstract

To date, the empirical literature on the formal-informal labor market segmentation hypothesis has largely focused on the analysis of wage determination in both sectors. The underlying premise in this literature is that, under segmentation, equally productive workers receive higher wages if located in the formal sector. Thus, identifying the existence of wage differentials between equally productive workers would imply that the labor market is segmented. The objective of this paper is to examine the extent to which the premise behind this literature actually has empirical content, and whether or not one can actually identify the existence of segmentation by analyzing wage data alone. For that, we develop a simple framework to describe the main determinants of the cross-sectional wage distributions in both sectors. We show that the evidence can be supported by completely different models of labor market functioning. It thus seems unlikely that one can actually identify the existence of labor market segmentation by analyzing data on wages alone.

Keywords: Segmentation, Informal Sector.

JEL Codes: O17.
1. Introduction

The existence of a large informal sector is a common feature of most Latin American labor markets [see Gasparini and Tornarolli (2007)]. In Brazil, for example, informal employment corresponds to nearly 40% of total employment, while gross wage differentials between formal and informal workers are also extremely high (above 80%). Even after controlling for observable characteristics, the formal-informal wage gap remains substantial and has increased over the past decade [see Ulyssea (2010b)]. These facts are not exclusive to the Brazilian labor market and have, in fact, long motivated an intense debate on whether or not labor markets in developing countries are segmented.

Despite the existence of a large empirical literature, there is still no consensus regarding the size, direction and the existence of a link between the observed formal-informal wage differentials and labor market segmentation. In an early study, Barros et al. (1990) analyze Brazilian construction workers’ wages in the formal and informal sectors and estimate substantial wage gaps, although the authors assume away selection on the unobservables. Carneiro and Henley (2001) use a standard, two-stage parametric selection correction method and find evidence that suggests that both formal and informal workers exploit comparative advantages, thus favoring the competitive view against the segmented one. Tannuri-Pianto and Pianto (2002) use a quantile regression model with a semi-parametric selection correction method, and find positive wage gaps in favor of formal workers throughout the entire distribution.\(^1\) Finally, Pratap and Quintin (2006) rely on the selection on the observables assumption to estimate the formal-informal wage gap in Argentina using a propensity score matching model. They find that the formal-informal wage gap vanishes once they match formal and informal workers with similar propensity scores.\(^2\)

In face of this mixed evidence, we take a step back and ask a more general question: Is it possible to identify the existence of segmentation only by analyzing the wage distributions in the formal and informal sectors and their different moments? In other words, the objective of this paper is to examine the extent to which the formal-informal labor market segmentation hypothesis has empirical content, and whether or not one can identify the existence of segmentation by analyzing wage data alone.

For that, we develop a simple framework to describe the main determinants of the cross-sectional wage distributions in both sectors. We consider a static environment populated by heterogeneous workers and jobs. In particular, workers differ with respect to their unobserved (to the econometrician) ability but are

---

1The authors find that these wage gaps are mostly explained by differences in workers’ attributes except for low income individuals, for whom differences in returns to their attributes explains around 30% of the gap.

2This is by no means an exhaustive list of the empirical studies on formal-informal segmentation. For a more comprehensive literature review, see Ulyssea (2006).
equal in all other dimensions. Wages in each sector are determined by the job’s quality and worker’s ability, where the relationship between ability and wages is described by a sector-specific hedonic function. As our main goal is to develop a simple statistical decomposition of the log-wage distribution, we do not analyze workers’ behavior directly. Instead, we assume the existence of a well-defined workers’ allocation process, but we remain agnostic as how it emerges.

Within this framework, we show that it is possible to decompose the log-wage distribution in each sector into three determinants:

(i) the workers’ selection process into the different sectors;

(ii) the reward functions in each sector, which determine the return to workers’ ability in each sector; and

(iii) the distribution of job quality in each sector.

Similarly, we also show that it is possible to decompose the average wage gap and the difference in log-wage variances into these three components, which provides a clear structure to interpret the empirical regularities actually observed in the data.

We specialize this general framework to three specific models of wage determination in the formal and informal sectors, and analyze how consistent they are with the stylized facts observed in the Brazilian data. The first model considers the case where jobs and the hedonic functions are the same in both sectors but workers’ allocation depends on ability. The second model assumes random allocation and homogeneous jobs, and it is thus very close to the competitive view of formal-informal labor markets. The third model assumes homogeneous workers but different job quality distributions and hedonic functions in both sectors, which is close to the underlying definition of segmentation used in most of the empirical literature.

Albeit the first model cannot account for the complete set of empirical regularities discussed in Section 2, both the second and third models are able to rationalize the existing empirical evidence. Indeed, they can support almost any given pattern regarding the moments of the log-wage distribution discussed here. Given that these models represent completely different views of wage determination in the formal and informal sectors, these results indicate that it is unlikely that one can actually identify the existence of labor market segmentation by analyzing data on wages alone.

The remaining of the paper is organized as follows. Section 2 presents some stylized facts regarding the log-wage distributions and wage differences between formal and informal workers. Section 3 presents the general model, while Section 4 specializes it to three different cases. Finally, Section 5 presents some final remarks.

Hence, our analysis can be thought of as being conditional on a vector of observable traits, such as schooling or gender, while there remains an (unobserved) worker’s ability term.
2. Facts

This section presents some stylized facts regarding the log-wage distributions in the formal and informal sectors, and the formal-informal wage gap. For that, we use the data from the 2008 National Household Survey (Pesquisa Nacional por Amostra de Domicílios – PNAD), which is an annual survey representative of the entire population and contains detailed information on socio-demographic characteristics as well as labor market outcomes. We restrict the sample to individuals who were either formal or informal employees at the period of the survey, who are between 18 and 65 years old, living in urban areas, and employed in the private sector (excluding domestic workers). Table 1 presents some basic descriptive statistics from the sample used.

Table 1
Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>Informal</th>
<th></th>
<th>Formal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Male (%)</td>
<td>0.483</td>
<td>0.497</td>
<td>0.603</td>
<td>0.489</td>
</tr>
<tr>
<td>Schooling (yrs)</td>
<td>7.7</td>
<td>4.2</td>
<td>9.4</td>
<td>3.8</td>
</tr>
<tr>
<td>Age (yrs)</td>
<td>33.7</td>
<td>12.9</td>
<td>34.5</td>
<td>11.0</td>
</tr>
<tr>
<td>Log-wage</td>
<td>6.0</td>
<td>0.81</td>
<td>6.7</td>
<td>0.62</td>
</tr>
<tr>
<td>Pctile 25</td>
<td>5.5</td>
<td>–</td>
<td>6.2</td>
<td>–</td>
</tr>
<tr>
<td>Pctile 50</td>
<td>6.0</td>
<td>–</td>
<td>6.6</td>
<td>–</td>
</tr>
<tr>
<td>Pctile 75</td>
<td>6.4</td>
<td>–</td>
<td>7.0</td>
<td>–</td>
</tr>
<tr>
<td>Pctile 99</td>
<td>8.3</td>
<td>–</td>
<td>8.7</td>
<td>–</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations from the 2008 PNAD.

Table 1 shows some basic regularities that are well-stablished in the informal sector literature: there is a larger share of women among informal workers, informal workers are on average less educated than formal ones, and formal workers have higher average wages. Moreover, the table also shows that the quantiles of the log-wage distribution in the formal sector are everywhere higher than those in the informal sector. Figure 1 presents similar evidence by depicting the cumulative density functions of the log-wage distributions in the informal and formal sectors. The graph shows that the log-wage distribution in the formal sector first order stochastically dominates the one in the informal sector.

Albeit interesting, these figures do not account for the substantial heterogeneity in workers’ observable characteristics both between and within sectors. Moreover, our analysis in the following sections considers observably equal workers and therefore we would like to spurge as much as possible of this observed heterogeneity. We thus move to a conditional analysis, using the following set of controls: dummies
for male, metropolitan area, age, years of schooling, race, activity sector, and state of residence.

Controlling for these covariates and running a standard OLS log-wage regression, one obtains an average formal-informal wage gap of 36.2% (the gross wage gap is 65.4%). As one could expect however, the average wage gap hides a great deal of variation. Figure 2 presents the log-wage quantile difference between formal and informal workers: Let \( \mu_q^1 \) and \( \mu_q^0 \) denote the \( q^{th} \) quantiles of the formal and informal log-wage distribution. We compute the differences \( \Delta \mu_q = \mu_q^1 - \mu_q^0 \) for each \( q = 1, \ldots, 99 \). The figure on the left panel shows the differences computed using the quantiles from the raw log-wage distributions. The panel on the right shows the differences computed using the quantiles from the log-residual distribution (the distribution of the residual computed from the wage regression mentioned above).

The figures show a strong declining pattern as one moves up in the quantiles of the wage distribution in both sectors. While the controlled wage gap remains high within the lower quantiles, it substantially decreases in the upper quantiles. Finally, the variance in the log-residual is greater among informal workers than formal ones (0.636 and 0.464, respectively). As we will argue in the following sections, these empirical regularities can be supported by very different models of wage determination and workers’ allocation and therefore cannot be used to infer if the labor market is actually segmented or not. Furthermore, this conclusion holds
Figure 2
Formal–informal log-wage quantile differences

(a) Log-wage

(b) Residual Log-wage
even if one uses more sophisticated econometric methods to estimate the formal-informal wage gap (such as in the studies briefly discussed in the introduction).

3. General framework

3.1 Set up

Let $P$ denote the population of workers. We assume that all workers $i \in P$ are homogeneous with respect to their observable attributes, but they may differ with respect to their unobserved abilities. We assume that the unobserved abilities can be summarized by a scalar, which we denote by $A_i$, for each $i \in P$.

A selection process is a mechanism to obtain a disjoint partition $(P_0, P_1)$ of $P$, such that if worker $i \in P_0$ she is an informal worker, and if $i \in P_1$ she is a formal worker. For each $i \in P$, let $d_i = 1 [i \in P_1]$ be an indicator function that equals one if the worker is employed in the formal sector and zero otherwise. For each worker $i \in P$, let $w_{i,0}$ and $w_{i,1}$ denote her potential log-wage in the informal and formal sectors, respectively. These log-wages are generated by allocating worker $i$ alternatively to the informal and formal sectors while holding constant the allocation of all other workers.

We assume that $w_{i,0}$ and $w_{i,1}$ have finite variances.

Throughout the paper, cumulative distribution functions (c.d.f.) are denoted by $F$, whereas densities are denoted by $f$. Subscripts will indicate to which random variable they refer to. For instance, $F_{w_0}$ denotes the c.d.f. for $w_0$ over the entire population $P$. Cumulative distributions constrained to the subpopulations $P_0$ and $P_1$ are identified by superscripts. For example, $F_{w_0}^0$ denotes the c.d.f. for $w_0$ constrained to $P_0$, that is, $F_{w_0}^0(\cdot) = F_{w_0}(\cdot | d = 0)$. Hence, if workers’ allocation is random, then $F_{w_0}^0 = F_{w_1}^1 = F_{w_0}$. It is worth noting that, in principle, the sectoral log-wages ($w_{i,k}$, $k = 0, 1$) could depend on the equilibrium allocation of workers across sectors. In other words, $F_{w_0}$ and $F_{w_1}$ would generally depend on which particular partition of $P$ is being considered. For example, a policy towards the formalization of labor relations would probably lead to a movement of workers from $P_0$ to $P_1$. This movement should be expected to depress wages in the formal sector while perhaps increasing them in the informal sector. Throughout this paper, however, the partition $(P_0, P_1)$ is taken as given and therefore no such general equilibrium effects are considered. Thus, it is not necessary to make explicit the dependence of the underlying log-wage distributions, $F_{w_0}$ and $F_{w_1}$, on the selection process.

---

4If instead workers are characterized by a vector of skills but that cannot be unbundled and sold separately in the market, the assumption of scalar ability does not fundamentally change the present analysis.

5Unfortunately, however, this procedure may not completely describe how potential wages are generated. For instance, in a segmented labor market with queues, the order of arrival of a worker may affect her wage. Moreover, in the presence of quotas it may not make sense to reallocate one worker holding the allocation of the other workers fixed.
3.2 Wage determination

Throughout the analysis, we assume that log-wages are determined by two components: worker’s skills and the quality of the job she currently occupies. Since workers are observably equal, we omit the observable traits for notational convenience and only make explicit the dependence of wages on unobserved ability, $A_i$. Each sector can value ability differently; let $s_0(a)$ and $s_1(a)$ denote the average log-wage of workers with ability $A = a$ in the informal and formal sectors, respectively:

$$s_k(a) = E[w_k|A = a]$$  \hspace{1cm} (1)

for all $a \in \mathbb{R}$ and $k = 0, 1$.

These regression functions are henceforth referred to as hedonic functions [e.g. Rosen (1974)], and are assumed to be strictly increasing and differentiable.\(^6\) When $s_0(a) = s_1(a)$ for all $a$, we say the model has homogeneous hedonic functions. Equally able workers may also have different wages within a given sector. We assume that within-sector wage differences amongst equally skilled workers can emerge due to differences in the job quality component, which is denoted by $u_{i,k}$, where

$$u_{i,k} = w_{i,k} - s_k(A_i)$$  \hspace{1cm} (2)

and $k = 0, 1$.

Hence, the wage variance within a given sector that can be attributed to the components $u_{i,k}$ provides a measure of the degree of job heterogeneity in that sector. Accordingly, whenever $\text{Var}[u_1] > \text{Var}[u_0]$, we say that jobs in the formal sector are more heterogeneous than jobs in the informal sector. Using this notation, wages can be expressed as follows:

$$w_{i,k} = s_k(A_i) + u_{i,k}$$  \hspace{1cm} (3)

where $k = 0, 1$.

For simplicity, we assume that the job quality shocks, $u_{i,k}$, and worker’s ability, $A_i$, are independent. Note that, by construction, $E[w_0|A] = E[w_1|A] = 0$. So, without extra assumptions, $u_0$ and $u_1$ would only be mean-independent of $A$.

3.3 The selection process

Given a pair of hedonic functions, $s_k$, the potential log-wages, $w_{i,k}$ for each $i \in P$ and $k = 0, 1$, are completely determined by workers’ ability, $A_i$, and the quality of her potential job assignments, $u_{i,k}$, $k = 0, 1$. We restrict our analysis to the class of selection models where the assignment of workers to sectors may

---

\(^6\)Since both $s_0$ and $s_1$ are obtained from conditional expectations, they are well defined only almost surely. A more precise statement would thus be that there exist versions of $s_0$ and $s_1$ which are strictly increasing and differentiable.
depend on their ability levels but which is independent of the quality of their potential job assignments. More precisely, we assume:

\[ \Pr [d = 1 | A, u_0, u_1] = \Pr [d = 1 | A] \equiv \lambda(A) \]  

Expression (4) states that equally able workers have identical probabilities of being allocated to the formal sector, and it is independent of the quality of their potential job. This assumption excludes some interesting cases, such as the existence of labor market discrimination. Suppose that a demographic group is discriminated in the formal sector, such that whenever they are selected into that sector they are assigned to jobs of low quality. In this case, one should expect this group to be over-represented in the informal sector, even when holding ability constant. In this example, selection would be based on job prospects and, consequently, the above hypothesis of pure selection on ability would be violated.

The selection hypothesis in (4) has two important implications. First, it implies that the hedonic functions are not distorted by the selection process:

\[ E[w_k | A, d = 1] = E[w_k | A, d = 0] = E[w_k | A] = s_k(A) \]  

and second, it implies that the distribution of job quality in the two sectors is also not distorted by the selection process:

\[ F_{u_k}^0 = F_{u_k}^1 = F_{u_k} \]  

where \( k = 0, 1 \). In particular:

\[ E[u_k | d = 1] = E[u_k | d = 0] = E[u_k] = 0 \]  
\[ Var[u_k | d = 1] = Var[u_k | d = 0] = Var[u_k] \]  

3.4 Log-wage distributions by sector

Following our notation, let \( F_A, F_A^0 \) and \( F_A^1 \) denote the distributions of abilities over \( \mathbb{P}, \mathbb{P}_0 \) and \( \mathbb{P}_1 \), respectively. Using Bayes’ rule, it can be easily shown that these distributions are related to each other as follows:

\[ dF_A^0 = \frac{dF_A (1 - \lambda)}{1 - p} \]  
\[ dF_A^1 = \frac{dF_A \lambda}{p} \]  

where \( p = \int_{\mathbb{R}} \lambda(a) \times dF_A(a) \), which is the share of the population in the formal sector.

When \( \lambda = p \), the selection is said to be random and \( F_A = F_A^0 = F_A^1 \). In this special case, the distribution of abilities is the same in the formal and informal sectors. In the following section we provide a simple example to illustrate the case.
where the selection process depends on ability and how it can affect the distribution of abilities in each sector.

From the assumption that $A$ and $(u_0, u_1)$ are independent, it follows that

$$F_{w_0}^0(\omega) = \int_{-\infty}^{+\infty} F_{u_0}^0(\omega - x)dF_{u_0}^0(x) \tag{11}$$

$$F_{w_1}^1(\omega) = \int_{-\infty}^{+\infty} F_{u_1}^1(\omega - x)dF_{u_1}^1(x) \tag{12}$$

and if the distribution of abilities is absolutely continuous with density $f_A$ then one can write

$$F_{w_0}^0(\omega) = \int_{-\infty}^{s_0^{-1}(\omega)} f_A(a)da = \int_{-\infty}^{s_0^{-1}(\omega)} \frac{1 - \lambda(a)}{1 - p} f_A(a)da \tag{13}$$

$$F_{w_1}^1(\omega) = \int_{-\infty}^{s_1^{-1}(\omega)} f_A(a)da = \int_{-\infty}^{s_1^{-1}(\omega)} \frac{\lambda(a)}{p} f_A(a)da \tag{14}$$

where we use the fact that the hedonic functions are assumed to be strictly increasing.

Using expressions (13) and (14), one can re-write (11) and (12) as follows:

$$F_{w_0}^0(\omega) = \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{s_0^{-1}(\omega - x)} \frac{1 - \lambda(a)}{1 - p} f_A(a)da \right] dF_{u_0}^0(x) \tag{15}$$

$$F_{w_1}^1(\omega) = \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{s_1^{-1}(\omega - x)} \frac{\lambda(a)}{p} f_A(a)da \right] dF_{u_1}^1(x) \tag{16}$$

Expressions (15) and (16) show that the observed log-wage distributions can potentially differ due to three factors:

(i) biased selection, $\lambda \neq p$;

(ii) differences in the hedonic functions $s_0 \neq s_1$; and

(iii) differences in the distributions of job quality, $F_{u_0} \neq F_{u_1}$.

In Section 4 we introduce three particular models that highlight the role of each one of these factors in generating observably distinct sectoral log-wage distributions.

3.5 Decomposing and interpreting differences in means, variances and quantiles

In this section we investigate how the observed differences in means, variances, and quantiles of the log-wage distribution can be decomposed such that the components are in a one-to-one correspondence with the three potential sources of differences among the observed distributions:
(i) non-random selection;
(ii) differences in the hedonic functions; and
(iii) differences in the distributions of job quality.

Differences in means

The observed wage gap between formal and informal workers can be written as follows:

\[ \Delta \mu = E \left[ w_1 | d = 1 \right] - E \left[ w_0 | d = 0 \right] \]

\[ = E \left[ s_1(A) - s_0(A) | d = 1 \right] + E \left[ s_0(A) | d = 1 \right] - E \left[ s_0(A) | d = 0 \right] \]

\[ + E \left[ u_1 | d = 1 \right] - E \left[ u_0 | d = 0 \right] \]

\[ = \Delta \mu_s + \Delta \mu_b \] (17)

where \( \Delta \mu_s \equiv E \left[ s_1(A) - s_0(A) | d = 1 \right] \) and \( \Delta \mu_b \equiv E \left[ s_0(A) | d = 1 \right] - E \left[ s_0(A) | d = 0 \right] \).

Note that \( E \left[ u_k | d = 1 \right] = E \left[ u_k | d = 0 \right] = E \left[ u_k \right] = 0 \).

The term \( \Delta \mu_s \) corresponds to the treatment on the treated parameter (TOT), which is entirely driven by differences in the hedonic functions. Additionally, because selection does not depend on the \( u_k \), the formal-informal wage gap is not affected by differences in the distribution of job quality across sectors. The term \( \Delta \mu_b \) captures the selection bias, which creates a difference between the (average) causal impact of holding a formal contract, \( \Delta \mu_s \), and the observed gap, \( \Delta \mu \).

The selection bias term can be re-written in a more informative way:

\[ \Delta \mu_b = E \left[ s_0(A) | d = 1 \right] - E \left[ s_0(A) | d = 0 \right] = \int s_0(a) \left[ dF_A^1(a) - dF_A^0(a) \right] \]

\[ = \left[ \frac{1}{p(1 - p)} \right] \int s_0(a) \left[ \lambda(a) - p \right] dF_A(a) \]

\[ = \left[ \frac{1}{p(1 - p)} \right] Cov \left( s_0(A), \lambda(A) \right) \]

By assumption, average log-wages increase with ability, \( \frac{\partial s_0(a)}{\partial a} > 0 \) for all \( a \). Thus, if more skilled workers are more likely to self-select into the formal sector, \( \frac{\partial \lambda(a)}{\partial a} > 0 \), for all \( a \), then \( \Delta \mu_b > 0 \). In this case, the observable difference, \( \Delta \mu \), would overstate the true average effect of being in the formal sector, \( \Delta \mu_s \). Of course, the term \( \Delta \mu_b \) vanishes if there is no self-selection based on ability, so that \( \lambda = p \). However, it is also possible that \( \Delta \mu_b = 0 \) even if the selection process is non-random.\(^7\)

\(^7\)As an example, let \( A \sim U(0,1) \), \( s_0(a) = a \), for all \( a \), and \( \lambda \) be defined as follows:

\[ \lambda(a) = \begin{cases} 
1 & \text{if } a \in (1/4, 3/4) \\
0 & \text{if } a \in (0, 1/4) \cup (3/4, 1) 
\end{cases} \]
Differences in variances

The observed difference in the formal and informal log-wage variances, $\Delta \sigma^2$, can be written as follows:

$$\Delta \sigma^2 = \text{Var}[w_1 | d = 1] - \text{Var}[w_0 | d = 0] = \text{Var}[s_1(A) | d = 1] - \text{Var}[s_0(A) | d = 1] + (\text{Var}[s_0(A) | d = 1] - \text{Var}[s_0(A) | d = 0]) + \text{Var}[u_1] - \text{Var}[u_0] = \Delta \sigma^2_s + \Delta \sigma^2_u + \Delta \sigma^2_b$$

where

$$\Delta \sigma^2_s \equiv \text{Var}[s_1(A) | d = 1] - \text{Var}[s_0(A) | d = 1]$$
$$\Delta \sigma^2_b \equiv \text{Var}[s_0(A) | d = 1] - \text{Var}[s_0(A) | d = 0]$$
$$\Delta \sigma^2_u \equiv \text{Var}[u_1] - \text{Var}[u_0]$$

Similarly to the difference in means decomposition, one can immediately identify the three distinct factors that account for the observed sectoral difference in log-wage variances:

(i) differences between sectors with respect to the value they attach to ability, $\Delta \sigma^2_s$;

(ii) differences between sectors with respect to the degree of job heterogeneity, $\Delta \sigma^2_u$; and

(iii) the effect of nonrandom selection, $\Delta \sigma^2_b$.

Similarly to the difference in means, it is possible to show that the term $\Delta \sigma^2_s$ can be expressed as the following covariance:

$$\Delta \sigma^2_s = \text{Var}[s_1(A) | d = 1] - \text{Var}[s_0(A) | d = 1] = E[s_1(A) - s_0(A) | d = 1]^2 - E[s_1(A) | d = 1]^2$$
$$\Delta \sigma^2_u = \text{Var}[u_1] - \text{Var}[u_0] = E[s_1(A) - s_0(A) | d = 1]^2$$

By assumption, $s_1$ and $s_0$ are strictly increasing and thus $\Delta \sigma^2_s > 0$ whenever $s_1 - s_0$ is also an increasing function. But, $s_1 - s_0$ is increasing if and only if $\frac{\partial s_1(a)}{\partial a} \geq \frac{\partial s_0(a)}{\partial a}$ for all $a$. In words, if the marginal value of ability is higher in the
formal sector, over the entire support of the ability space, then $\Delta \sigma^2 > 0$. As for
the term $\Delta \sigma^2$, there is no simple relationship between its sign and the behavior of $\lambda$.8

**Differences in quantiles**

Up to now we have been successful in decomposing $\Delta \mu$ and $\Delta \sigma^2$ in additive components which are in a one-to-one correspondence with the three factors shown to generate differences between the observed log-wage distributions in the formal and informal sectors. Thus, as far as differences in means and variances are concerned, these three factors do not interact with each other. We now consider the observed differences in quantiles. The $\alpha$-th quantile of the wage distribution in sector $k$ is given by:

$$ \mu^k_\alpha = (F^k_{w_k})^{-1}(\alpha) $$  \hspace{1cm} (20)

for $k = 0, 1$ and $0 < \alpha < 1$.

Then, the quantile differences can be simply written as

$$ \Delta \mu_\alpha = \mu^1_\alpha - \mu^0_\alpha $$  \hspace{1cm} (21)

Opposed to the mean and variance differences, it does not seem possible to obtain a similar simple analytical decomposition for the differences in quantiles. One thus have to consider special cases in order to make progress in comparing the different quantiles of the log-wage distribution in the formal and informal sectors. In the next section, we consider different models which constitute special cases of the general framework developed in this section.

4. **Specific models of wage determination**

4.1 **Homogeneous Jobs and Hedonic Functions**

In this model all jobs are equally good, $u_0 = u_1 = 0$, and ability is equally valuable in both sectors, $s_0 = s_1 = s$. Consequently, workers are completely indifferent to which sector they should belong. In this case, the formal and informal sector are completely symmetric and there is no labor market segmentation.9 Nonetheless, we assume that selection could still depend on ability, in which case the log-wage distributions in both sectors simplify to:

$$ F^0_{w_k}(\omega) = \int_{-\infty}^{s_0^{-1}(\omega)} \frac{1 - \lambda(a)}{1 - p} f_A(a) da $$  \hspace{1cm} (22)

$$ F^1_{w_k}(\omega) = \int_{-\infty}^{s_1^{-1}(\omega)} \frac{\lambda(a)}{p} f_A(a) da $$  \hspace{1cm} (23)

---

8Heckman and Honore (1990) characterize $\Delta \sigma^2$ for some particular selection models.

9We do not explicitly define segmented versus integrated markets. For this discussion in a slightly more general context, see Ulyssea (2010a) in this volume.
As the above expressions show, the observed differences in the log-wage distributions only reflect differences in the distribution of abilities across sectors, which differ only due to the nonrandom selection process. The formal-informal wage gap (expression 17) is also completely determined by the selection mechanism, as the hedonic functions are the same and therefore $\Delta \mu_s = 0$. In this case, the wage gap has no causal interpretation regarding the impact of holding a formal contract, as $\Delta \mu = \Delta \mu_b$. The same holds for the difference in variances (expression 18): jobs are homogeneous, and thus we have that $\Delta \sigma^2_s = 0$; and because the hedonic functions are the same, $\Delta \sigma^2 = \Delta \sigma^2_b$. Consequently, in this particular model $\Delta \sigma^2 = \Delta \sigma^2_b$.

As for the quantile differences, expressions 22 and 23 already suggest that it is not possible to obtain a closed-form expression for the quantiles of the log-wage distribution in each sector. Indeed, it is necessary to impose further structure to the model in order to obtain closed-form expressions for the quantiles and the quantile differences. In what follows we consider a very simple case to illustrate a model with positive selection into the formal sector ($\frac{\partial \lambda(a)}{\partial a} > 0$).

**Example: Positive selection into the formal sector**

Let $\lambda(a) = 0$ for $a < 0$ and $\lambda(a) = 1 - e^{-a}$ for $a \geq 0$, so there is positive selection into the formal sector, $\frac{\partial \lambda(a)}{\partial a} \geq 0$, and thus higher ability implies a higher probability of being assigned to the formal sector. Additionally, let us assume that abilities are exponentially distributed with unit mean: $f_A(a) = 0$ for $a < 0$ and $f_A(a) = \exp\{-a\}$ for $a \geq 0$. In this case we obtain $p = \frac{1}{2}$, $f^0_A(a) = f^1_A(a) = 0$ for $a < 0$ and

\[
\begin{align*}
    f^0_A(a) &= 2 \exp\{-a\} \\
    f^1_A(a) &= 2 \exp\{-a\} \left(1 - \exp\{-a\}\right)
\end{align*}
\]

for $a \geq 0$. In this case, half of the population is in each sector, the distribution of abilities in the informal sector is exponential with mean equal to $\frac{1}{2}$, while the average ability in the formal sector is $\frac{3}{2}$; however, the ability distribution in the formal sector is not exponential. The variance of the distribution of abilities over $\mathbb{P}$ is 1, while the conditional variances are $\frac{1}{4}$ and $\frac{5}{4}$ for the informal and formal sectors, respectively.

Moreover, assume that $s_0(a) = s_1(a) = \beta a$ for all $a \in \mathbb{R}$ (and $u_0 = u_1 = 0$). In this case, the observed distribution of log-wages can then be expressed as follows:

\[
\begin{align*}
    F^0_{w_0}(\omega) &= 1 - e^{-\frac{\omega}{2}} \\
    F^1_{w_1}(\omega) &= \left(1 - e^{-\frac{\omega}{2}}\right)^2
\end{align*}
\]

for all $\omega > 0$.

First, note that in this case the log-wage distribution in the formal sector first order stochastically dominates the one in the informal sector, as $F^1_{w_1}(x) \leq F^0_{w_0}(x)$.
for all $x$. As for the observed means, they are given by $\mu_0 = \frac{2}{2}$ and $\mu_1 = \frac{3\beta}{3}$ for the informal and formal sectors, respectively. Hence, $\Delta \mu = \Delta \mu_b = \beta > 0$. With respect to the variances, we obtain $\sigma_0^2 = \frac{\beta}{4}$ and $\sigma_1^2 = \frac{5\beta^2}{4}$ for the informal and formal sectors, respectively. Hence, $\Delta \sigma^2 = \Delta \sigma_b^2 = \beta^2 > 0$. In summary, although there exist no differences among jobs and ability is equally valuable in both sectors, the observed log-wage distributions are very different. In fact, the process of selection on ability was able to generate a pair of observable distributions were both the level and the inequality are higher in the formal sector.

Finally, the quantiles of the log-wage distribution are given by

$$
\mu_0^{\alpha} = -\frac{\beta}{2} \log (1 - \alpha) \\
\mu_1^{\alpha} = -\beta \log (1 - \sqrt{\alpha})
$$

Hence, for all $\alpha \in [0, 1]$, $\Delta \mu_\alpha > 0$ and $\frac{\partial \Delta \mu_\alpha}{\partial \alpha} > 0$ (see Figure 3). In this example, positive and increasing quantile differences were obtained by specifying a selection process in which the probability of being allocated to the formal sector increases with ability. This is clearly at odds with the evidence discussed in Section 2. Although this simple model is able to match the sign of the wage gap, it does not mimic the behavior of the quantile differences nor the sign of $\Delta \sigma^2$. Hence, this very simple model is not adequate to represent the basic log-wage distribution facts.

### 4.2 Random selection with homogeneous jobs

In this model, jobs are assumed to be homogeneous and, due to random selection, abilities are identically distributed across sectors. This is therefore also very close to the idea of a non-segmented labor market. Since jobs are assumed to be homogeneous, equally able workers must face the same wage within each sector, but wages may vary across sectors.

The cross-sector wage variation only reflects differences regarding their hedonic functions. Formally, the log-wages of workers with ability $a \in \mathbb{R}$ are given by

$$
w_k = s_k(a)
$$

(24)

where $k = 0, 1$. Similarly, the observed log-wage distributions in the formal and informal sectors differ only because of differences in the hedonic functions:

$$
F_{w_0}^k(\omega) = F_A(s_k^{-1}(\omega)) = \int_{-\infty}^{s_k^{-1}(\omega)} f_A(a)da
$$

(25)

This is given by

$$
F_{w_0}^0(x) - F_{w_1}^1(x) = 2 \exp \left( -\frac{x}{2} \right) \left[ 1 - \exp \left( -\frac{x}{2} \right) \right] \geq 0
$$

where the above inequality holds strictly except for the point $x = 0$. 

---

Brazilian Review of Econometrics  30(2) November 2010  303
As an additional consequence of random selection, the observed distributions equal the underlying unconditional log-wage distributions, that is, $F_{w_0}^0 = F_{w_0}$ and $F_{w_1}^1 = F_{w_1}$. This fact permits the observed differences to have simple causal interpretations. Take the case of the wage gap: due to random selection, $\Delta \mu_b = 0$, and thus the observed wage gap represents the average causal effect of formal labor contracts on wages, $\Delta \mu = \Delta \mu_s$. Similarly, the difference in variances is entirely determined by the term that captures the differences between sectors with respect to the value they attach to ability, $\Delta \sigma^2$.

Since the allocation of workers is random, the distribution of abilities in the two sectors is identical, $F_A^1 = F_A^0 = F_A$. Because all jobs are assumed to be equally good, the log-wage distribution in each sector is given by this common underlying distribution of abilities transformed by the sector specific hedonic function. Since the hedonic functions are assumed to be strictly increasing, one can write $\mu^k = s_k(q_A(\alpha))$, for $k = 0, 1$ and $0 < \alpha < 1$; $q_A(\alpha)$ denotes the $\alpha$ quantile of the distribution of abilities in the overall population. Therefore, for $0 < \alpha < 1$:

$$\Delta \mu_\alpha = \mu^1_\alpha - \mu^0_\alpha = s_1(q_A(\alpha)) - s_0(q_A(\alpha))$$

$$\frac{\partial \Delta \mu_\alpha}{\partial \alpha} = \left[ \frac{\partial s_1(q_A(\alpha))}{\partial \alpha} - \frac{\partial s_0(q_A(\alpha))}{\partial \alpha} \right] \frac{\partial q_A(\alpha)}{\partial \alpha}$$
Notice that $\Delta \mu_\alpha > 0$ for all $0 < \alpha < 1$ if and only if $s_1(a) > s_0(a)$ for all $a > 0$. Moreover, $\frac{\partial \Delta \mu_\alpha}{\partial s_\alpha} > 0$ for all $0 < \alpha < 1$ if and only if $\frac{\partial s_1(q\alpha)}{\partial a} \geq \frac{\partial s_0(q\alpha)}{\partial a}$ for all $a$, since $\frac{\partial \Delta \mu_\alpha}{\partial s_\alpha} \geq 0$ for all $0 < \alpha < 1$. In summary, the observed quantile differences are positive and increasing if and only if the differences between the hedonic functions in the formal and informal sectors are also everywhere positive and increasing, while the reverse is also true.

4.2.1 Discussion

As shown above, in this particular model the observed differences can be expressed as follows:

$$
\Delta \mu = \Delta \mu_s \equiv E[s_1(A) - s_0(A)]
$$

$$
\Delta \sigma^2 = \Delta \sigma^2_s \equiv Var[s_1(A)] - Var[s_0(A)]
$$

$$
\Delta \mu_\alpha = \mu_1^\alpha - \mu_0^\alpha \equiv s_1(q\alpha) - s_0(q\alpha)
$$

$$
\frac{\partial \Delta \mu_\alpha}{\partial \alpha} = \left[ \frac{\partial s_1(q\alpha)}{\partial \alpha} - \frac{\partial s_0(q\alpha)}{\partial \alpha} \right] \frac{\partial q\alpha}{\partial \alpha}
$$

In Section 2 we presented evidence on these objects and we can thus interpret them in the realm of this specific model. Notice that $\Delta \mu_\alpha > 0$ for all $0 < \alpha < 1$ if and only if $s_1(a) > s_0(a)$ for all $a \in \mathbb{R}$. Hence, everywhere positive quantile differences (which we observe in the data) imply that ability is more valuable in the formal than in the informal sector for any ability level. This immediately implies that the average relative wage gain of formal labor contracts ($\Delta \mu_s$) must be positive, which is also observed in the data.

As just discussed, the quantile differences are decreasing if and only if the marginal value of ability is lower in the formal sector, i.e., $\frac{\partial s_1(a)}{\partial a} \leq \frac{\partial s_0(a)}{\partial a}$ for all $a \in \mathbb{R}$. Moreover, as discussed in Section 3, we have that

$$
\frac{\partial s_1(a)}{\partial a} \leq \frac{\partial s_0(a)}{\partial a} \Rightarrow \Delta \sigma^2_s < 0
$$

which in this model means that $\Delta \sigma^2_s < 0$, as observed in the data.

In sum, the empirical regularities discussed in Section 2 are compatible with a model in which the selection process is random and jobs are homogeneous. In the context of this model, this implies that $s_1(a) > s_0(a)$ and $\frac{\partial s_1(a)}{\partial a} \leq \frac{\partial s_0(a)}{\partial a}$, for all $a \in \mathbb{R}$. Note that the pattern observed in the quantile differences is crucial in determining how the findings should be interpreted in the context of this model. To further characterize this model, consider the following example illustrated in Figure 4.

Here again we use the structure of the example discussed in subsection 4.1, and we assume that workers’ skills have an exponential distribution with unit.
Hedonic functions: $s_0(x) = b_0 x$ and $s_1(x) = c + b_1 x$.

Model 1: $b_0 = 2$, $b_1 = 1$, $c = 2.5$; model 2: $b_0 = 1$, $b_1 = 2$, $c = 0$.

mean. In the particular parameterization used in Figure 4, in Model 1 (see the Figure’s footnote) we have that $\mu_0 = b_0$ and $\mu_1 = c + b_1$ (remember that ability is exponentially distributed with unit mean). Thus, for any parameterization such that $c > b_0 - b_1$, we have that $\Delta \mu > 0$, although the quantile differences might be decreasing if $b_0 > b_1$. Additionally, if $b_0 > b_1 \geq 1$ then $\sigma^2_0 = b_0^2 > b_1 = \sigma^2_1$.

More broadly, the model with random selection and homogeneous jobs is based on three primitives: $f_A$, $s_0$, and $s_1$. However, only two objects can be observed, $f_{w_0}$ and $f_{w_1}$. Therefore, as it was the case in the previous model, this model cannot be refuted, since that for any given pair of observed log wage densities $(f_{w_0}^0, f_{w_1}^1)$, one can always find a triple – $f_A$, $s_0$, and $s_1$ – which perfectly rationalizes the observed densities. Indeed, by normalizing one of the primitives at any given value, we can recover the other two uniquely from the observed distributions. For example, given $s_0$, the other primitives can be uniquely recovered as follows:

$$f_A(a) = \left( \frac{\partial s_0(a)}{\partial a} \right) f_{w_0}^0 (s_0 (a))$$

$$F_A (a) = F_{w_1}^1 (s_1 (a))$$
4.3 Homogeneous Workers

In this model all workers are equally able and therefore selection is necessarily random. As a consequence, the shape of the observed log-wage distributions is solely determined by the distribution of job quality in each sector, while the hedonic functions determine the location of the distributions. This can be regarded as an extreme version of segmentation, since all workers are equal and therefore any wage difference is a consequence of differences across sectors in the returns to ability or job quality.

In this context, observed log-wage distributions truly reveal the underlying job heterogeneity, as they can be written as follows:

\[ F_{0w} (\omega) = F_{0u} (\omega - s_0 (\pi)) \] (26)
\[ F_{1w} (\omega) = F_{1u} (\omega - s_1 (\pi)) \] (27)

The wage gap only depends on the difference in the sectoral hedonic functions: \( \Delta \mu = \Delta \mu_s \), while the same is not true for within-sector differences in wages, which are uniquely determined by the within-sector job heterogeneity. Similarly, the difference in variances only depends on the difference in the job quality heterogeneity in both sectors: \( \Delta \sigma^2 = \Delta \sigma^2_u \), where the last equality follows from the fact that workers are homogeneous and therefore \( \Delta \sigma^2_s = 0 \). Analogously, the quantiles can be written as:

\[ \mu^k_{\alpha} = s_k (\pi) + q_{u_k} (\alpha) \] (28)

where \( q_{u_k} (\alpha) \) denotes the \( \alpha \)-th quantile of the job quality distribution in sector \( k \); for \( k = 0, 1 \) and for all \( 0 < \alpha < 1 \). One can thus write:

\[ \Delta \mu_{\alpha} = \mu^1_{\alpha} - \mu^0_{\alpha} = s_1 (\pi) - s_0 (\pi) + \mu_{\alpha} (u_1) - \mu_{\alpha} (u_0) \]

\[ \frac{\partial \Delta \mu_{\alpha}}{\partial \alpha} = \frac{\partial \mu_{\alpha} (u_1)}{\partial \alpha} - \mu_{\alpha} (u_0) \frac{\partial u_0}{\partial \alpha} \]

In the data we observe that \( \Delta \mu > 0 \), which under this model can be interpreted as evidence of segmentation, since on average equally productive workers get higher wages simply for holding a formal job. Similarly, if one observes \( \Delta \sigma^2 < 0 \) then it must be the case that jobs are more heterogeneous in the informal sector, which seems to be corroborated in the data [e.g. Maloney (2004)]. The interpretation and implications for the pattern of the quantile differences are not clear in this model without further assumptions.

4.3.1 Discussion

This model has four primitives, \( F_{u0}, F_{u1}, s_1 (u), \) and \( s_0 (u) \), but only only two objects are observed, \( f_{0w} \) and \( f_{1w} \). This implies that any pair of log-wage
distributions $F^k_w$, $k = 0, 1$, can be rationalized in this model by letting

\[
\begin{align*}
    s_0 (\bar{a}) &= \int \omega dF^0_w (\omega) = E \left[ w_0 | d = 0 \right] \\
    s_1 (\bar{a}) &= \int \omega dF^1_w (\omega) = E \left[ w_1 | d = 1 \right] \\
    F_{w_0} (x) &= F^0_w (x + s_0 (\bar{a})) \\
    F_{w_1} (x) &= F^1_w (x + s_1 (\bar{a}))
\end{align*}
\]

Both the model discussed here and the one in the previous section can rationalize the empirical regularities discussed in Section 2. However, the interpretations under the different models differ sharply. In the present model, all workers are identical, so that the wage gap and quantile differences emerge solely due to differences in the job quality distributions across sectors and differences in the hedonic functions. Hence, in this model the wage gap and the other empirical regularities can be indicative of the existence of labor market segmentation, as equally productive workers have a higher compensation simply for holding a formal job.

It is worth highlighting, however, that a complete characterization of labor market segmentation typically involves some form of barrier to workers’ mobility. In a dynamic setting, these differences in wages could not be sustained as an equilibrium outcome without barriers to mobility, otherwise workers would keep entering the formal sector until the wage gap was eliminated (remember that worker are homogeneous). However, here we abstract from such dynamic and general equilibrium effects.

5. Final Remarks

Most of the empirical literature that aims at identifying the existence of formal-informal labor market segmentation relies on the analysis of earnings functions in both sectors. In particular, it focuses on the identification of the wage gap between equally productive formal and informal workers. The premise behind this literature is that if the labor market is segmented, then equally productive formal workers would receive higher wages than their informal counterpart. Hence, identifying the existence of such wage gaps would imply identifying the existence of segmentation.

As discussed in the introduction, the evidence in this literature is, at best, mixed. We therefore take a step back and try to assess to which extent the premise that guides this literature actually has empirical content. More specifically, we ask if the existence of formal-informal labor market segmentation can be identified by analyzing wages alone. For that, we develop a simple but general statistical framework that provides a structure to interpret the observed empirical regularities. We specialize this framework to three different models that reflect very different data generating processes and ask whether or not they can rationalize the facts observed in the data.
The first model is one that assumes homogeneous jobs and hedonic functions. The second model (subsection 4.2) is close to the competitive view of formal-informal labor markets, as workers’ selection is random and jobs are homogeneous. In this model the hedonic functions differ but, as discussed in Heckman and Hotz (1986), this does not imply the existence of segmentation. Finally, the third model discussed (subsection 4.3) illustrates the premise that seems to guide most empirical examination of the formal-informal segmentation hypothesis: Workers are homogeneous but have higher wages solely because they hold a formal job. Hence, within the framework considered here, the second and third models represent two extreme views regarding wage determination in the formal and informal sectors.

While the first model is not able to match the observed empirical regularities, both the second and third models are able to rationalize the existing evidence. In fact, these two models can rationalize any pair of observed sectoral log-wage distributions. Thus, our analysis indicates that is not possible to identify the existence of segmentation only by examining wage distributions and differentials.

References


