Gender wage gaps in formal and informal jobs, evidence from Brazil.

Sarra Ben Yahmed
Aix-Marseille University (Aix-Marseille School of Economics) and Sciences Po Paris

September 2013

Keywords: Informality, Gender Wage gaps, Selection.
JEL codes: J31, O17

Abstract

How does selection into formal vs. informal jobs shape gender wage inequality? This paper shows that the higher raw wage gap in the informal sector compared to the formal sector is an artefact of different male and female selection processes. First, women have better observable characteristics than men and the female advantage is stronger among formal employees. As a result, the adjusted wage gap is underestimated by the raw wage gap, especially in the formal sector. After controlling for observables, the formal and the informal wage gaps are not statistically different. Second, selection into work status differs between men and women. The difference in magnitude and direction of the selectivity bias affects the estimation of the gender wage gap. In the informal sector, observed wages overestimate wage offers for both men and women but the bias is higher for men. The selection-corrected gender wage gap is reduced and actually not significant in the informal sector. In the formal sector, however, observed wages underestimate wage offers for men, while for women observed wages overestimate wage offers. Therefore, the gender gap in observed wages understates the gender gap in wage offers. The selection-corrected gender wage gap is high and strongly significant in the formal sector.

1 Introduction

A striking characteristic of labour markets in developing countries is the existence of a large informal sector where labour regulations, including minimum wages and parental leaves, are inexistent. While labour protection and labour costs are lower in the informal segment of the labour market, informal jobs may offer other features valuable to workers such as flexibility. Those aspects may influence wages differently for men and women. It is thus important to distinguish the formal and the informal segments of the labour market when examining the gender wage gaps in developing countries. Moreover, policy makers may be concerned about the gender wage differences in the two segments separately as it can shed some light on how regulation affects women's prospects in the labour market. The aim of this paper is to investigate how informality shapes labour market outcomes for men and women, and in particular to examine whether there exist positive gender wage gaps in both formal and informal jobs and whether they are significantly different from one another. In doing so, we will raise the following questions: Do men and women sort differently across labour market statuses? How does the selection process affect the gender wage gaps in formal and informal jobs?

A vast literature has focused on earnings inequality due to informality in the labour market. Many papers have tried to understand whether the labour market is segmented or whether the existence of two different segments is the result of competitive allocation of workers. However, very few works have looked into gender differences within each segment. This paper investigates this issue and complements the limited empirical evidence on the gender wage gap in the informal labour market. Tansel (2001) estimates the gender pay differential among employees with social security coverage and workers without, in Turkey. She controls for self-selection into multiple work statuses and finds that the adjusted wage gap is substantial among covered (formal) workers but not significant among uncovered (informal) workers. Deininger et al. (2013) look at the gender wage gap in India and find that the share of the gap due to different returns to characteristics is higher among casual workers than among non-casual workers. They control for selection into labour market participation but they do not take into account the selection into multiple employment statuses conditional on being active. We depart from these papers in two ways. First in the definition of informality as we focus on employers' compliance with labour regulation rather than on social security coverage or temporary work. Second in the empirical methodology as we compare two approaches to deal with non-random selection into multiple employment outcomes.

This paper uses the Brazilian household survey, the Pesquisa Nacional por Amostra de Domicilios (PNAD), for the year 2009. The PNAD provides information on whether the worker's labour card is signed by the employer so that we are able to adopt a definition of informality based on employers' compliance with labour market regulation. A formal worker is an employee with a registered labour contract, hence entitled to labour rights and benefits, while an informal worker is employed without having a legal contract declared by his/her employer.

This paper raises the question of whether the differences in gender wage gaps across formal and informal jobs are due to labour regulation or to gendered selection into formal vs. informal employment. The endogeneity of work status is a major concern as failing to deal with non-random selection would lead to misleading estimates of gender wage gaps for two reasons. First, self-selection on unobservables would bias the coefficient estimates in the wage equation. Second, if selection is not random the observed wage gap does not reflect the gap in wage offers. It is thus important to recover the differences in wage *offers* to conduct the decomposition on the appropriate total wage difference. In the aim of controlling for self-selection, we first study the sorting of men and women into different employment statuses using a multinomial logit model. Some studies have focused on two alternatives, formal vs. informal. In our setting, it is relevant to describe potential work statuses more broadly as other situations are common alternatives to salaried work, especially for women, such as inactivity, self-employment and unemployment. Since the definition of the set of alternatives can affect the treatment of the selection bias, we choose to consider all the potential outcomes: inactive, unemployed, formal salaried work, informal salaried work, self-employment and employer.

We then investigate how selection into work status affects the estimation of the gender wage gaps. In the literature, the effect of selection on wage estimations is addressed with approaches similar to the well-known Heckman two-stage procedure. The control function consists of estimating a selection equation in a first stage and constructing the correction terms, the control function, to be used as a regressor in the main wage regression. The literature has proposed different methods for addressing selection into multinomial potential outcomes. We use the strategies of Lee (1983) and Dubin and McFadden (1984) where the selection model is specified as a multinomial logit model and we compare the results obtained under the different assumptions that those methods imply (see Bourguignon et al. (2007) for a discussion of the two approaches). Wage equations are estimated for formal salaried workers and informal salaried workers separately in order to compute the gender wage gap separately for formal workers and informal workers. We use a version of the Oaxaca-Blinder-Ransom decomposition that proposes a satisfactory solution to the choice of the non-discriminatory wage structure (Fortin, 2008).

Looking at the raw data, we find that women are more often unemployed than men and that the informality rate is higher among working women compared to working men. We also find that the size of the raw wage gap differs across groups of education, the formal and the informal gaps being significantly different only for certain education groups. This pattern is in line with recent evidence on the heterogeneity of informal labour markets (Gunther and Launov, 2012) and points to different labour market selection processes across formal and informal sectors. We show that men and women differ in the magnitude and direction of their selectivity bias in formal and informal jobs. Controlling for selection into work status affects the estimation of the gender wage gap, especially in the informal segment of the labour market where the gap is no longer significant. The gender wage gap remains significantly positive only among formal employees. Because labour market decisions and the gender wage gaps differ across the schooling distribution, we conduct the analysis for three different education groups.

This paper contributes to the small literature on labour market outcomes for men and women when a large share of employment is informal. This analysis is most closely related to the papers that study the gender wage gap among informal workers and formal workers separately. Tansel (2001) defines informality as the absence of social security protection and estimates the gender wage gaps among covered and uncovered wage earners in Turkey. In the wage equation, she controls for endogenous selection into multiple outcomes using the strategy developed by Lee (1983). The potential outcomes are divided into five categories: non-participation, private sector covered wage work, uncovered wage work, self-employment and other employment. She finds that, in 1994, the adjusted wage gap is strong and positive among covered workers but not significant among uncovered workers. Deininger et al. (2013) look at the gender wage gap in India in formal vs. informal jobs where informal work is defined as casual work in either agricultural or non-agricultural sectors. They control for selection into labour market participation using the Heckman's (1979) methodology. They find that the gender wage gap due to different returns to characteristics is particularly important for casual workers working in the agriculture; however, gender discrimination is much lower or inexistent in non-agricultural sectors. A few studies focus on the difference in the formal wage premium for men and women. Arabsheibani et al. (2003) study the evolution of wages for men and women in Brazil over the period 1988-1998. Their results seem to indicate that informality is more penalizing for men than for women, in other words that the formal wage premium is greater for men, however they do not test for the significance of the formal premium difference. Pagán and Ullibarri (2000) find that women tend to work more in unregistered (informal) firms In Mexico. Accordingly to the findings of Arabsheibani et al. (2003) for Brazil, they show that the informal wage penalty is lower for women than for men.

This paper is also related to the vast literature on the segmentation of the labour market and the formal wage premium. Maloney (1999) questions the dualistic view of the labour market and points out that the mobility of workers between the formal and the informal segments of the labour market suggests that the market is not segmented along this line. Carneiro and Henley (2002) explore how expected earnings differ in the informal and formal sectors controlling for selection of workers. Their selection correction approach consists in estimating the probability of being either a formal or an informal worker which has a significant impact on the estimation of earnings for both formal and informal workers. They find that some workers are actually better off choosing the informal segment of the labour market. Gunther and Launov (2012) also highlight the heterogeneous composition of the informal sector in Côte d'Ivoire. They do not reject the hypothesis that the labour market is dual as some workers are involuntarily employed without contract, even if some workers seem to choose informal jobs over formal employment. Magnac (1991) tests the hypothesis of segmented labour markets against the hypothesis of competitive markets in the urban areas of Colombia. He uses a sample of married women as he argues they represent the group that faces higher labour market entry costs because of domestic and familial responsibilities. He cannot reject the hypothesis of competitive labour markets for married women and concludes that unobserved characteristics, abilities or preferences are the drivers of the choice between formal and informal jobs. Pradhan and Van Soest (1995) study sector participation for both men and women in urban areas of Bolivia. They control for selection into formal jobs, informal jobs or non-participation and compare two selection models that make different assumptions on the underlying choice structure. In the ordered probit model, all workers want to enter the formal sector; informal jobs are a second best option. This model corresponds to a dualistic view of the labour market (Fields, 1975) where the formal sector offers better jobs that are rationed. The second approach uses an unordered model, the multinomial logit model and the selection term is constructed according to Lee's (1983) formulas that makes no assumptions about the ordering of sector preferences. They find that the ordered model describes best men outcomes. Male predicted wages are higher in the formal sector. For women, however, the opposite holds. According to the multinomial logit model, average expected earnings are higher in the informal sector for all females; according to the ordered probit, the predicted wage offers are higher in the formal sector for only 9 percent of women, those with a high level of education. They conclude that women's sector choice cannot be explained by restrictions to entry in the formal sector only. Women may choose to enter the informal sector to maximise their earnings. Voluntary employment in the informal sector can be explained by comparative advantage in informal jobs for workers who would not earn better wages in the formal sector. In addition to the articles mention above, this conclusion can also be found in Gindling (1991), Rosenzweig (1988), and Maloney (2004).

The present paper also contributes to the literature on gender wage gaps and selection into employment. Arabsheibani et al. (2003) study the gender wage differentials in Brazil over the period 1988-1998. They find that the gender wage gap, especially the part due to different returns of identical characteristics, has fallen over the period but remains positive. Madalozzo (2010) confirms the fall in the wage gap until the end of the 90s and finds no further decrease in the 2000s. Santos and Ribeiro (2006) find evidence of a glass ceiling in Brazil. These three papers study the gender wage gap in Brazil but they do not distinguish between formal and informal employees, nor do they investigate the impact of different selection biases between men and women on the gender wage gap.

A vast literature, starting in the late 1970s, have studied the effect of the selection bias on the gender wage gap, mostly focusing on the United States. Among recent papers, Blau and Kahn (2006) show that the decline in the gender wage gap during the 1980s was overstated as it is largely explained by sample selection. They also show that selection has also contributed to the slower reduction in the gender wage gap during the 1990s. Looking at European countries and the United States, Olivetti and Petrongolo (2008) point out that non-random selection explains why gender employment gaps are negatively correlated with gender wage gaps across countries. Women are on average positively selected into employment. Countries with particularly high gender employment gaps, such as southern Europe, are characterised by a strong female selection bias which in turn reduces the observed gender wage gap. This small observed gender wage gap is actually an artefact of the selection process: women who are employed have better abilities than non-employed women which overestimates female wage offers. Appleton et al. (1999) investigate how selection biases the gender wage gap in three African countries. She highlights that the observed wage gap is narrower than the gap in wage offers in Ethiopia and Uganda but not in the Côte d'Ivoire where female observed wages underestimate female wage offers. These studies find that correction for selection has important consequences for the assessment of gender wage gaps. We do not know any paper that has assessed the effect of the selectivity bias on gender wage gap estimations in labour markets where the co-existence of formal and informal jobs modifies the selection process.

The present paper is also linked to the empirical research on the heterogeneity of the wage gaps across groups with different skill levels. Albrecht et al. (2003) show that the gender wage gap is increasing along the wage distribution in Sweden. de la Rica et al. (2008) find that in Spain the gender wage gap is high and increases with wage (glass-ceiling effect) among highly educated workers while it is lower and decreases with wage among less educated workers (floor effect). The innovation of this paper is to explore how the wage gap differs by education groups for informal wage-earners and formal wage-earners separately.

The remainder of the paper is organized as follow. We start by discussing the impact of informality on gender employment and wage inequality while reviewing the related literature. In section 3 we describe the data and provide descriptive statistics on gender inequalities in the Brazilian labour market. Section 4 sets up the empirical model. In section 5 we discuss the results, looking at the selection into potential outcomes for men and women before moving onto the comparison of the gender wage gaps in the formal and informal sectors. The last section concludes.

2 Gender and Informality

Why would the gender wage gap differ across the formal and the informal segments of the labour market? As far as we know, the existing theoretical explanations have focused on the understanding of the formal wage premium, but they have not provided any explanation for gender differences in formal wage premium nor have they explained gender wage differences *within* each sector. Put differently, there are no theoretical models that explain why the formal gender wage gap may or may not differ from the informal gender wage gap. We use the existing literature to postulate hypotheses about the mechanisms behind the gender gaps *within* each segment and why those gaps may differ.

According to the dualistic view of the labour market, the informal segment is characterized by lower wages. Empirical evidence, looking at salaried workers and not at self-employed, confirms that formal jobs offer on average higher wages than informal jobs (see Magnac (1991) who analyses female wages in Colombia, Gasparini and Tornarolli (2009) who focus on different Latin American countries and Almeida and Carneiro (2007) who find that the formal raw formal wage premium is positive in Brazil and decreases with regulation enforcement). Comparisons of raw average wage gaps are informative about the accepted wage offers but conceal heterogeneity in workers' observable and unobservable attributes. Empirical papers show that the formal-informal wage difference differs depending on workers' skill levels. Studying the urban labour market in Mexico, Gong and Van Soest (2002) find a significant wage premium in formal jobs for educated men but not for men with low-education who earn more on average in informal jobs. For women, the differences between formal and informal wages is small. It is thus informative to first look at gender wage differences for different education groups. It is also important to compute wage gaps adjusted for all characteristics in order to compare individuals with similar observable productivity within the two segments.

What would explain different gender wage gaps in formal and informal jobs once we have controlled for observable characteristics? From the labour supply side, if individuals have different preferences for the type of jobs, the theory of compensating wage differentials can give an explanation for formal-informal wage differences within groups and it can also help understand the gender wage gaps in formal jobs and in informal jobs. Formal jobs offer non-monetary benefits that are not available in informal jobs such as job severance contribution, maternity leave, unemployment benefits, social security. In a frictionless market, workers with identical productivity should earn a higher wage in the informal segment to compensate for the absence of non monetary benefits. If women value job protection more than men, for maternity reasons in particular, then women should be ready to accept lower wages compared to men in the formal sector but not in the informal sector. This would lead to a gender wage gap among formal employees only. However, if women value the flexibility of informal jobs more than men, we should also observe a gender wage gap among informal employees as well. As there are reasons to value the amenities of both sectors, workers' preferences over formal or informal employment will hinge on the balance of the advantages and disadvantages of both statuses depending on workers' characteristics. Gender differences in preferences are not a priori clear cut, which impedes us from drawing theoretical predictions on the overall effect of preferences on gender wage gaps in both sectors.

From the labour demand side, job offers stem from both registered and unregistered firms. Firms operating informally will not offer legal contracts to their employees. Firms operating formally might decide to hire workers formally or informally. Why would employers set different wages to a man and a woman with similar observable characteristics and employed under the same type of contract? Employers compare the costs and benefits of labour contract registration for both men and women and set their hiring decisions and wage setting rule accordingly. Employers may expect a higher quit rate among women because of, for example, permanent or temporary leave due to maternity. Lazear and Rosen (1990) provide a theoretical explanation where stronger domestic responsibilities generate higher female quite rates and lower female wages due to statistical discrimination. Bertrand et al. (2010) show that among high-skill employees small differences in labour market attachment in terms of working hours or short leave lead to enormous pay penalties for women. A higher quit rate generates higher costs because of vacancy and replacement costs; it can also generate forgone profits if no one can replace the employee on leave or if the time out of the job causes a loss of (general or specific) skills. Employers may want to compensate for the higher female quit rate by paying them lower wages. This argument applies especially to formal jobs where employers abide by the labour regulation such as the protection of the job during maternity leave. It should also be more stringent in high-skill jobs that require specific skills or training and less so in jobs that entail routine tasks only. de la Rica et al. (2008) analyses the gender wage gap in Spain using quantile techniques; they show that among highly educated workers, the wage gap increases along the wage distribution which is in line with the *glass ceiling* story. For these reasons, we would expect higher gender wage gaps among formal employees, especially for workers with high level of education. However, de la Rica et al. (2008) also show evidence of a *sticky floor*: among less educated worker, the wage gap is stronger for those at the bottom of the wage distribution. They explain this results by the much lower labour attachment of low-skilled women. Accordingly, the wage gap is expected to be significant among informal employees with low level of education, earning low wages.

The question of whether the gender wage gap is higher in the formal or the informal segment of the labour market has no straightforward answer and requires empirical investigations. Moreover, the empirical investigation need to account for the endogenous sorting of men and women into the different statuses as it can influence the wage equation estimates.

3 The Econometric model

To compare the gender wage gaps among formal and informal employees, we investigate how selection shapes the gender wage gaps in these two different segments of the labour market. We first compute the raw wage gaps and the wage gap adjusted for observable characteristics in both segments. Comparing the raw and the adjusted wage gaps enables us to say something about the role of observables characteristics on gender wage inequality. Next, we compute the wage gaps controlling for both observable characteristics and the selection into the different labour statuses.

3.1 The raw and the adjusted wage gaps

The raw wage gap in sector j is estimated from an equation where $\ln w_{ij}$ the hourly log wage is regressed on a constant and a female dummy only:

$$\ln w_{ij} = \beta_0 + \alpha_j F_{i(j)} + u_{ij} \tag{1}$$

where $F_{i(j)} = 1$ if employee i working in j is a woman. The raw wage gap is $E(\ln w|female) - E(\ln w|male) = \widehat{\alpha}_j$.

Different methods are used in the literature to compute the adjusted wage gap. One method is to estimate a mincerian wage equation on a pooled sample with a female dummy to capture the gender wage gap. The problem with this method is twofold. First, it might suffer from misspecification if the differences in returns to specific characteristics matter for the estimation of the wage gap. Second, we cannot estimate the selection rule for men and women separately using one wage equation on a pooled sample.

Instead, we use a version of the wage gap decomposition developed by Oaxaca (1973) and Blinder (1973) that avoids important methodological problems discussed in Oaxaca and Ransom (1994) and

Oaxaca and Ransom (1999). The decomposition methodology that follows has been presented in Fortin (2008) and is not sensitive to the choice of the reference wage structure. The reference wage structure is taken from the estimation of a common wage regression on the pooled sample of both men and women where the male advantage equals the female disadvantage with respect to the reference. We estimate three equations, two separate wage equations for men and women and a pooled wage equation with gender dummies and an identification restriction. Each equation is estimated separately for the formal and the informal segments denoted with the subscript $j = 2, 3$.

$$\ln w_{ipj} = \beta_{0pj} + \alpha_{pfj}F_i + \alpha_{pmj}M_i + \mathbf{X_i}\beta_{\mathbf{pj}} + u_{ij} \qquad \text{with } \alpha_{pfj} = -\alpha_{pmj} \tag{2a}$$

$$\ln w_{ifj} = \beta_{0fj} + \mathbf{X_i}\beta_{\mathbf{fj}} + u_{ifj} \tag{2b}$$

$$\ln w_{imj} = \beta_{0mj} + \mathbf{X_i}\beta_{\mathbf{mj}} + u_{imj} \tag{2c}$$

where X is a set of control variables that includes the number of years of education, the age and the age squared, the tenure and the tenure squared, whether the person is black, whether the person lives in an urban area, dummies for regions and sectors. To capture demand side effects, we use the regional unemployment rate that characterizes the state of the local labour market. We construct the regional unemployment for different education groups in order to identify the impact of lower labour demand even when controlling for regional dummies. The assumption is that labour markets are skill-specific, at least to some extent. Even if workers may accept a job for which they are overqualified, the unemployment rate among people of the same (generally defined) skill level will impact their decision to participate, their job finding rate and their wages.

The zero conditional mean assumption $E(u_m|x_m) = E(u_f|x_f) = 0$ ensures that the the error is uncorrelated with the regressors so that the OLS estimates are unbiased. The zero conditional mean assumption also ensures that the total average wage gap can be exactly decomposed into terms based on observables and their returns. For the wage decomposition to be exact though, only a weaker ignorability assumption is sufficient; what is needed is that the distribution of u given X is the same for the two groups. In other terms, the decomposition allows for selection on unobservables as long as they are the same for both men and women and yields identical selection biases. See Fortin et al. (2011) for a discussion of the assumptions required for identification in wage decompositions. Under the ignorability assumption, the total wage gap in each segment can be decomposed into three terms:

$$\overline{\ln W}_{mj} - \overline{\ln W}_{fj} = (\mathbf{\overline{X}'_m} - \mathbf{\overline{X}'_f})\widehat{\beta_{\mathbf{pj}}} + \mathbf{\overline{X}'_m}(\widehat{\beta}_{\mathbf{mj}} - \widehat{\beta_{\mathbf{pj}}}) + \mathbf{\overline{X}'_f}(\widehat{\beta_{\mathbf{pj}}} - \widehat{\beta}_{\mathbf{fj}})$$

The first term accounts for gender differences in characteristics, it is the endowment term. The last two term account for gender differences in the prices associated with given characteristics, it is also called the coefficient term and is here decomposed into the male advantage with respect to the reference prices and the female disadvantage with respect to the reference prices. The adjusted wage gap is the sum of the male advantage and the female disadvantage in the treatment of the characteristics :

$$\overline{WG}_j = \overline{\mathbf{X}}'_{\mathbf{m}}(\widehat{\beta}_{\mathbf{mj}} - \widehat{\beta_{\mathbf{pj}}}) + \overline{\mathbf{X}}'_{\mathbf{f}}(\widehat{\beta_{\mathbf{pj}}} - \widehat{\beta_{\mathbf{fj}}}) \tag{3}$$

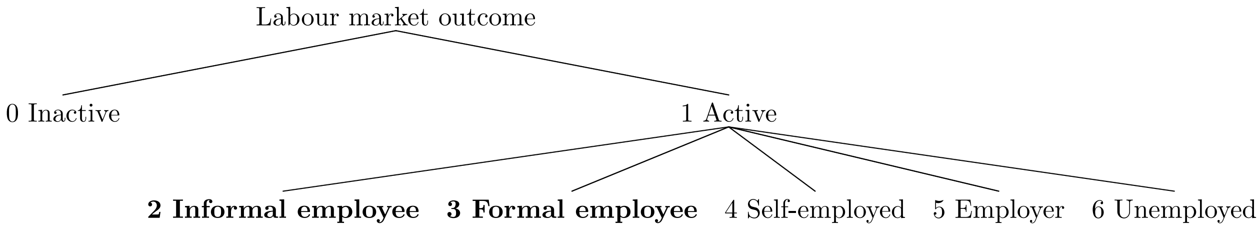
The adjusted wage gap takes into account the observable differences in characteristics between men and women, however it does not account for the selection of men and women into formal or informal jobs because of unobserved characteristics.

This can be problematic given that the conditional independence assumption is strong and that even the ignorability assumption may not hold in our case. Women have a much lower labour market participation rate than men and the selection of men and women into different types of jobs is certainly not random. What is more, selection into employment may follow different processes for men and women. The descriptive statistics (see below) show that the female unemployment rate is higher than the male unemployment rate and that the informality rate is higher among active women compared to active men (see table 3). If $E(u|X) \neq 0$ in equation (2), the coefficients of the wage equation are biased. If the ignorability assumption does not hold, men employed in a given type of job are different in observables and in unobservables from women who are employed in the same type of job. In that case, the selection biases differ for men and women and the estimations of the wage gaps are thus biased too. To eliminate the selection biases we adopt a control function approach that is presented in the next sub-section.

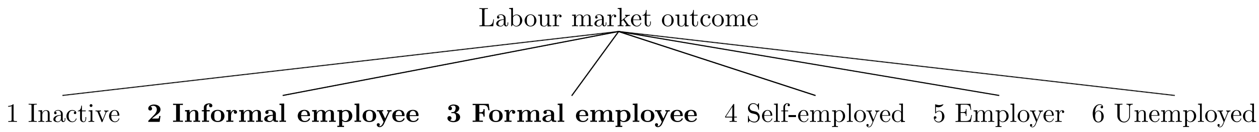
3.2 Treatment for selection into multiple employment statuses

Selection into formal salaried work vs. informal salaried work can be analysed using binary models but these models ignore potentially important differences among salaried workers and people in other situations such as inactivity, unemployment and self-employment. In this paper, we use a multinomial model to estimate the probability to be in formal employment and in informal employment taking into account that several relevant alternatives exist. In this setting, individuals have to choose between being inactive or entering the labour market where there are multiple potential outcomes. Individuals have different probabilities to be in a given work status depending on their preferences as well as on demand constraints and employers behaviours that may cause job rationing and segregation.

Different models of the individuals' (constrained or unconstrained) choices can be imagined. One possibility is to model a two-step sequential decision process where first the individuals decide to enter the labour market or not. In a second step, active individuals are selected into various employment categories or into unemployment.



Another alternative is to consider a one-step process where selection into inactivity, unemployment or the various employment categories happens simultaneously. In that case, the model has six mutually-exclusive outcomes denoted j: inactivity ($Y_i = 1$), informal employment ($Y_i = 2$), formal employment ($Y_i = 3$), self-employment ($Y_i = 4$), employer ($Y_i = 5$) and unemployment ($Y_i = 6$).



In the sequential process, individuals make the decision to participate or not to the labour market before knowing in which situation they will be in case of participation; they have a preference for being inactive vs. being active whatever the final situation. In our analysis, the first stage choice is simultaneous with the second stage chances of being in a given work status. We do not have determinants that may drive the first step choice and not the second step choice. In other words, the preference for being active is determined by the preferences over the different work statuses. For this reason, we prefer the simultaneous decision model.

Let us denote V_{ij} the latent value (or utility) associated with being in state j. State j is observed $Y_i = j$ if the value associated with this state is higher than the value of the other states, or in other words, if status j is the best available option for individual i.

$$Y_i = j \text{ if } V_{ij} > \max_{k \neq j}(V_{ik})$$

We assume that the utility associated with work status j follows a linear function: $V_{ij} = \mathbf{Z_i}\alpha_\mathbf{j} + \mu_{ij}, \quad j = 1, ..., 6$. If we further assume that the errors are independent and identically distributed following a type I extreme value distribution, the probability of being in status j for individual i is defined by the multinomial logit model (McFadden, 1973):

$$P_{ij} = Pr(Y_i = j) = \frac{exp(Z_i \alpha_j)}{\sum_j^N exp(Z_i \alpha_j)} \tag{4}$$

The full model of selection and wage determination can be written as follows

$$\ln w_{ij} = \mathbf{X_{ij}}\lambda_{\mathbf{j}} + u_{ij}, \text{ if } V_{ij} > \max_{k\neq j}(V_{ik}) \text{ for } j = 2, 3$$

$$V_{ij} = \mathbf{Z_i}\alpha_{\mathbf{j}} + \mu_{ij}, \qquad j = 1, ..., 6$$

where individual i earns a wage w_{ij} if she is a formal worker $j = 2$ or an informal worker $j = 3$ and j is the observed outcome if the value associated with state j is the highest. A selection bias arises if the unobserved characteristics that influence wages u_{ij} are correlated with the unobserved determinants of the selection process μ_{ij}, if $E(u|x,)$.

The vector **X** includes the wage determinants, namely: years of education, age and age squared, tenure and tenure squared, whether the person is black, whether the person lives in an urban area, a macroeconomic demand side variable to capture rationing: regional unemployment rate by education group, regional and sector dummies. In the selection equation, the vector **Z** is composed of elements of **X** as potential earnings influence the choice of work status. We do not include tenure and sectoral dummies in **Z** as those characteristics are unknown before being employed. The vector **Z** additionally includes variables that are not in **X**. These excluded variables are important for addressing the selection bias and must meet two conditions. They should be orthogonal to the errors of the second-stage equation and also relevant to sectoral-choice determination in the outcome equation. We discuss the set of excluded variables in detail below.

To control for selection in the wage equation, we introduce a correction term that we denote $h(P_1, ..., P_6)$ where P_j denotes the probability to be in state j. The control function $h(.)$ is equal to the conditional mean of the residuals $E(u_j|X, Y = j)$. The methods available to compute $h(.)$ differ by their assumptions on the covariances between the error term of the wage equation and the error terms of the outcome equations.

Lee's (1983) approach assumes that the joint distribution of u_j and a transformation of μ_j does not depend on the other μ_k for $k \neq j$. Under this assumption and a additional linearity assumption the expected value of u_j, conditional on category j being observed is:

$$E(u_j|X, Y = j) = \sigma\rho_j \left(-\frac{\phi(\Phi^{-1}(P_j))}{P_j}\right)$$

where σ is the standard deviation of the wage errors and ρ is the correlation coefficient between the errors of the outcome equation and the errors of the wage equation. The control function is $h = -\frac{\phi(\Phi^{-1}(P_j))}{P_j}$ and $\sigma\rho_j$ are estimated by least squares. Only one correlation parameter ρ_j is estimated per wage equation under this method. Note that when $\sigma\rho_j$ is negative, workers are positively selected into work status j as $\sigma\rho_j \left(-\frac{\phi(\Phi^{-1}(P_j))}{P_j}\right)$ is strictly positive.

The distributional assumption might be too restrictive as the selection bias potentially originates in the correlation of u_j not only with μ_j but also with μ_k for $k \neq j$. Thus we also follow Dubin and McFadden (1984) who make less restrictive assumptions on the correlation between u_j and the $(\mu_k - \mu_j)$. The linearity assumption on the conditional mean of the wage equation residuals (Dubin and McFadden, 1984) is as follows:

$$E(u_j|X, Y = j) = \sigma \frac{\sqrt{6}}{\pi} \sum_k \rho_{jk}(\mu_k - E(\mu_k))$$

where j is the final outcome and $k = 1, ...6$ all the potential outcomes. ρ_{jk} is the correlation coefficient between u_j and μ_k and Dubin and McFadden (1984) make the restriction that the correlation coefficients sum up to zero $\sum_k \rho_{jk} = 0$. Given the multinomial logit formulas we have:

$$E(\mu_j - E(\mu_j)|V_j > \max_{s \neq j}(V_s), Z) = -\ln(P_j)$$

$$E(\mu_k - E(\mu_k)|V_j > \max_{s \neq j}(V_s), Z) = \frac{P_k \ln(P_k)}{1 - P_k}, \text{ for } k \neq j$$

The following wage equations corrected for selection are then estimated by least squares:

$$\ln w_{ipj} = \lambda_{pfj} F_i + \lambda_{pmj} M_i + \mathbf{X_i} \gamma_\mathbf{p} \mathbf{j} + \theta_\mathbf{pj} \mathbf{h_{pj}}(\mathbf{P_1}, ..., \mathbf{P_6}) + u_{ij} \text{ with } \lambda_{pfj} = -\lambda_{pmj} \tag{5a}$$

$$\log w_{ifj} = \mathbf{X_{ij}} \gamma_\mathbf{fj} + \theta_\mathbf{fj} \mathbf{h_{fj}}(\mathbf{P_1}, ..., \mathbf{P_6}) + \epsilon_{ifj} \tag{5b}$$

$$\log w_{imj} = \mathbf{X_{ij}} \gamma_\mathbf{mj} + \theta_\mathbf{mj} \mathbf{h_{mj}}(\mathbf{P_1}, ..., \mathbf{P_6}) + \epsilon_{imj} \tag{5c}$$

where $\theta_\mathbf{j} \mathbf{h_j}(\mathbf{P_1}, ..., \mathbf{P_6}) = E(u_j|X, Y = j)$ and depends on the model assumptions. The estimation of equations (5) allows us to recover ρ_j the correlation between u_j and μ_j when Lee's model is adopted and the correlation between u_j and all the μ_k for $k = \{1...j...6\}$ if the Durbin-Mac Fadden's approach is used.

We present the total decomposition with an additional term that captures the difference in average selection bias:

$$\overline{\ln W}_{mj} - \overline{\ln W}_{fj} = (\overline{\mathbf{X}}'_\mathbf{m} - \overline{\mathbf{X}}'_\mathbf{f})\widehat{\gamma_\mathbf{pj}} + \overline{\mathbf{X}}'_\mathbf{m}(\widehat{\lambda_\mathbf{mj}} - \widehat{\gamma_\mathbf{pj}}) + \overline{\mathbf{X}}'_\mathbf{f}(\widehat{\gamma_\mathbf{pj}} - \widehat{\gamma_\mathbf{fj}}) + \theta_\mathbf{mj} \mathbf{h_{mj}}(\mathbf{P_1}, ..., \mathbf{P_6}) - \theta_\mathbf{fj} \mathbf{h_{fj}}(\mathbf{P_1}, ..., \mathbf{P_6})$$

The last term capturing the selection effect has been treated in different ways in the literature on wage gap decomposition. Neuman and Oaxaca (2004) present different variations of the decomposition when selection is controlled for and show how the selection term can be included in the endowment term and/or in the coefficient term. We follow Yun (2007) who advocates treating selection as a separate term in the decomposition. In that way, the selection term provides a measure of the difference between the observed wage gap and the gap in wage offers[1].

The wage gap due to different returns to observable characteristics in sector j is:

$$\overline{WG}_{Sj} = \overline{\mathbf{X}}'_{\mathbf{m}}(\widehat{\gamma}_{\mathbf{mj}} - \widehat{\gamma_{\mathbf{pj}}}) + \overline{\mathbf{X}}'_{\mathbf{f}}(\widehat{\gamma_{\mathbf{pj}}} - \widehat{\gamma_{\mathbf{fj}}}) \tag{6}$$

The adjusted wage gap in equation (6) differs from the one in (3). First, the coefficients are now unbiased following the treatment for selection. Second, instead of explaining part of the total observed wage gap, the difference in returns now explains the gap in wage offers $\overline{\ln W}_{mj} - \overline{\ln W}_{fj} - (\theta_{mj} h_{mj}(P_1, ..., P_6) - \theta_{fj} h_{fj}(P_1, ..., P_6))$.

Equations (5) are also estimated for various education groups separately to explore how the selection rules and the gender wage gaps (6) differ across groups.

3.3 Identification

To identify the effect of selection and purge the wage estimates from the selection bias without relying on the difference in the functional forms, we need variables that determines the potential work status but do not affect directly wages. The validity of this method hinges on the exclusion restrictions. Given the data available, the excluded variables for this an analysis are various demographic characteristics: the presence of children, the presence of children under 14 years old, the marital status, a dummy for lone mothers and the number of family members holding formal jobs.

While it may be argued that children can affect the productivity of women on the job and thus may not be an appropriate excluded variable, the number of family members holding formal jobs has a priori no direct effect on wages. Moreover, it determines women sectoral-choice as the security brought by job protection and social security coverage of the household member makes labour participation less necessary and formal employment less valuable. In other words, having no family members in formal employment can make women more risk averse and hence willing to search more intensively for formal jobs than women with a household member in formal employment.

Our empirical approach will hence consist of computing the formal (informal) gender wage gaps adjusted for observable characteristics in a first step and in a second step controlling additionally for endogenous selection into formal (informal) employment. The latter step implies to estimate the probability to be in each outcome which we will do using a multinomial logit. This empirical strategy will be applied to the whole sample and to different education groups to capture potential heterogeneity in the selection patterns and wage gaps along the skill distribution.

[1] This approach has been adopted by **?** for the analysis of the ethnic wage gap in the U.S., by Wright and Ermisch (1991), Ogloblin (1999), Appleton et al. (1999) among others for gender wage gap decomposition, and by Ermisch and Wright (1993) for the estimation of wage offers in part-time and full-time jobs among women

4 Empirical results

4.1 The data

Individual information is taken from the 2009 Brazilian household survey, the Pesquisa Nacional por Amostras de Domicilio (PNAD), that covers both rural and urban areas. The PNAD provides information about the individuals of roughly 100,000 households. In 2009, around 252,000 working-age people (18-65) were interviewed, among whom 52% were women and 85% lived in urban areas. Sample weights ensure the representativeness of the survey.

The different employment categories are the following: employees (wage-earners of the public and private sectors) which include domestic workers employed by private households; self-employed; employer; unpaid and family workers The survey provides direct and reliable information that enables us to classify employees into formal and informal wage-earners. Individuals are asked if their labour card is signed by their employer; if it is not, they are not registered and are not entitled to any labour rights or benefits. The labour card is used in the private sector; workers in the public sector have other types of contracts and are considered as formal employees in this study. In this paper, we focus on gender differences among informal *wage-earners* only, including domestic workers but excluding self-employed, employers, unpaid and family workers.

Table 1 gives the demographic, household and educational characteristics of men and women holding formal and informal jobs. Informal employees are on average younger than formal employees. Men and women working formally are of the same age on average but in the shadow sector women are slightly older than men. Women who hold informal jobs are more often the head of the household and live less often in couples compared to women in formal jobs. A larger share of women have young children in the informal sector as 45% of women women working informally have children under 14 years of age against 40% for women in the formal sector.
The PNAD provides information on the composition of the household. A household can be made of several families, e.g. two families sharing a dwelling or one family hiring a domestic employee with or without his/her family. Women tend to live in households/families with a higher share of formal wage-earners; this differential can be explained by the higher male participation rate and lower male informality rate compared to the corresponding female rates, a difference that is discussed below. Both men and women in formal employment are better educated than those in informal employment and women are more educated than men in both segments of the labour market. Full-time work is less common among women and among informal workers. There are no major differences across gender or sector in the distribution of age at first job nor in the average tenure, which is somewhat surprising as we could have expected higher turnover and lower tenure in informal jobs.

Table 2 describes in more detail the educational attainment for different employment statuses. It reveals that the female distribution of school attainment dominates the male distribution. There are fewer low-educated women and more high-educated women participating to the labour market. The same applies for unemployed and shadow workers. The table also shows that the informal population is diverse. 37% of women, against 47% of men, have primary education or less, at the same time, 10% of unregistered women and 8% of unregistered men have tertiary education. This is consistent with a sorting of men and women where sex is a signal for labour market attachment or quit probability and a higher education level compensate for a higher average quit rate among women (see Lazear and Rosen (1990) for a theoretical model and de la Rica et al. (2008) for an empirical analysis where they explain the distribution of the wage gap in Spain with a similar rational.).

Table 3 highlights differences across gender and educational level in participation rates, unemployment rates and informality rates in 2009. The participation rate is lower for women and the participation gap decreases with education. The average participation is rate is 66% for women and 89% for men. Among people with primary education or less, only 53% of women decide to participate in the labour market while 85% of men do so, which corresponds to a gap of 31 precentage points. The participation rate increases with education and more rapidely for women. Among people with tertiary education, the participation is gap is of 10 percentage points.

The female unemployment rate is higher in all education-groups, the difference is larger for people with medium level of education. For workers with primary education, the unemployment gap is around 4 percentage points. For active people with secondary education, the unemployment rate is higher, especially for women at 13%, leading to a higher gender gap of 6 percentage points. The unemployment gap is lower among workers with tertiary education.

The informality rate measures the share of wage-earners without a labour contract; it is higher for women than for men, the difference being larger for people with secondary education or less. The informality rate decreases with education. Among female wage-earners, 30% of women with preimary education or less are employed without a contract, 22% among women with secondary education and 9% of women with tertiary education have no contract. The gender gaps in informality rates decreases with education as well, it is of 7 percentage points among workers with secondary education or less. It is lower of only 1 percentage point among workers with tertiray education.

We now turn to the distribution of formal and informal jobs across sectors. We can see in table 4 that 69% of female employees work in the service sector where the informality rate for women is 30%. Only 48% of male employees work in this sector and have a lower informality rate, 18%. The highest informality rate is in the construction and mining activities. Only 1% of working women are employed in the construction sector but 57% of them hold informal jobs. The manufacturing industry employs 14% of the labour force and the informality rates for men and women are similar, 16% and 15% respectively. However, in argiculture, which employs 22% of the labour force, the female informality rate is lower than the male informality rate by almost 20 percentage points.

4.2 Wage distributions across genders

To complete the preliminary description of the gender differences in the formal and the informal segments of the labour market, we compute raw wage differences. Table 1 shows that average raw hourly wages are higher in formal jobs and that in both formal and informal jobs men earn on average more than women. Figure 1 displays the wage distributions for men and women in both the formal and informal sectors. Among formal workers, the female wage distribution is shifted farther to the left compared to the male wage distribution which indicates that the raw difference between male and female wages is positive especially in the middle of the wage distribution. On the other hand, in the informal sector, the male and female wage distributions almost overlap except at the bottom where the lower tail of the female distribution is fatter. This description is valid for wage-earners working in the service sector and in the manufacturing industries. However in agriculture the two wage distributions almost overlap in the formal sector except for a fatter lower tail of the female distribution, while in the informal sector the female wage distribution is to the left of the male wage distribution. This pattern holds for urban workers but not for rural workers as figure 2 shows. In rural areas, the female wage distribution dominates the male distribution in the formal sector. However, in the informal sector, the female distribution has larger tails both at the bottom and at the top of the wage distribution. For this reason, we separate rural and urban workers in the following analysis of gender wage gaps in informal and in informal jobs.

4.3 Selection into multiple potential employment statuses

We start our empirical analysis by estimating the multinomial logit equation (4) to understand the impact of supply side and demand side variables on the probability of being in each outcome. We estimate the multinomial logit model for men and women separately. The marginal effects are reported in table 5 for urban workers and in table 10 in the appendix for rural workers. The tables provide an estimate of the effect of a marginal change in each variable, for an individual with average characteristics in the male sample and in the female sample. The relative risk ratios of the multinomial logit estimation are provided in the appendix.

Education and age determine men's and women's outcomes in the same direction though the magnitudes of the effects differ. The number of years of education reduces the probability of being out of the labour force much more for women; it also reduces the probability to be informally employed while increasing the chances to be formally employed, the latter effect being stronger for women again. The probability of formal and informal salaried work decreases with age for both men and women, as does the probability of being unemployed.

Other variables such as the family structure have opposite effects on men and women. The presence of young children and living in couples reduce the probability of inactivity for men while it increases it for women. A woman with young children has a lower probability to be formally or informally employed and will choose self-employment more often. This does not hold for lone mothers who have a greater probability to be working in a salaried job. Contrary to women living in couples, men with young children have a lower probability to be inactive or self-employed but a higher probability to hold a formal job. Those results are consistent with the traditional division of roles within the household.

We find that higher regional unemployment rates increase non-participation for women although the marginal effect is not significant. Regional unemployment rate reduces the probability to find a formal job and increases the probability to hold an informal job for women. The opposite holds for men. Higher unemployment rates increase labour participation. There is no discouragement effect in the Brazilian urban labour market. In addition, higher unemployment rates increase the probability that men hold formal jobs while it reduces their probability to be self-employed or employers. This may reveal an insurance effect: as it becomes tougher to find a job, men tend to search more intensively for formal jobs that are more secured and provide unemployment benefits in case of lay off.

4.4 Wages

Tables 6 and 7 present the estimates of the female wage equations and the male wage equations in urban areas. In the appendix, table 11 gives the reference wage structure computed on the pooled sample and used in the wage decomposition as suggested by Fortin (2008). Tables 12 to 17 provide the wage equation estimates for the three education groups by gender. Tables 18 and 19 show the results for rural areas.

We can see in tables 6 and 7 that the return to education is stronger in the formal sector for both men and women, this pattern is robust to the introduction of the selction control function.
Age has also a significant positive impact on wages, the effect is of the same magnitude in both segments but it is stronger for men than for women. We do not see here evidence of a concave effect of age on wages.

Tenure in current firm increases male wages in formal jobs but not in informal jobs. As for women, the effect is not significant in formal jobs while it has a negative effect in informal jobs. As we control for age, this result does not mean that actual wages are declining for women. Negative returns to tenure in informal jobs means that women that keep on working informally for the same employer have lower wages compared to women who have changed job more recently. The negative returns of tenure with the same employer may be due to low female job mobility along with monopsony power of employers. Kambourov and Manovskii (2009) and Sullivan (2009) argue that specific skills are occupation rather than firm specific. They find that tenure with the same employer has zero or negative effects on wages when tenure in occupation and in the industry are controlled for. Schmiedern (2007) also finds negative returns to tenure with the employer for women in Germany.

The regional unemployment rate affects negatively wages; this effect is also robust to selection treatment. It has a stronger negative impact on female wages. For both women and men, unemployment reduces more the formal employees' wages; this result indicates that the formal segment of the labour market is competitive.

In columns (3) to (6) of tables 6 and 7, the control function is included as an additional regressor. The selection bias is significant in both the formal and the informal sectors, for both men and women. Tables 6 and 7 report the correlations between the errors of the wage equation and the errors of the selection equation when all education groups are pooled together. The correlation coefficient gives us the direction of the average selection rules for men and for women. Note that when Lee's approach is adopted, a negative $\sigma_j \rho_j$ implies a positive selection bias as $\sigma_j \rho_j \left(-\frac{\phi(\Phi^{-1}(P_j))}{P_j}\right)$ is strictly positive. Men are positively selected in informal employment and negatively selected in formal employment according to Lee's method. For given values of observable characteristics, men holding informal jobs have unobserved characteristics that are most valued in this sector. Consequently, observed wages overestimate male wage offers in informal jobs. On the other hand, men are negatively selected into the formal sector. Those with the highest wage potentials in formal jobs do not self-select into those jobs and choose other work statuses. Negative selection occurs when the reservation wage is increasing with the wage offer. This selection pattern remains the same when we estimate the selection term for different education groups. As for women, they are positively selected in both informal and formal jobs. The average selection rule hides heterogeneity across education groups. The positive selectivity bias in formal jobs holds for women with secondary education or less but not for women with tertiary education who are negatively selected in formal jobs as men are. Highly educated women working in the formal sector are those with lower wage potential compared to highly educated women in other work statuses.

Results for rural workers are shown in the appendix in tables 18 and 19. The return to education are smaller than in urban areas and tenure has no significant impact on formal wages. For women, tenure is negatively associated with wages in the informal segment as we observe in urban areas. The unemployment rate has a much lower downward effect for women and is not significant for men.

4.5 The gender wage gap in informal and formal jobs

Table 8 displays the estimated gender wage gaps among urban salaried workers, for the whole population as well as for different groups of education. Table 20 in the appendix shows the results for rural workers.

The total raw wage gaps are positive and significant in both formal and informal jobs. The average raw gap is significantly higher among informal employees when it is estimated on the whole population. However, this conceals different composition of the male and female labour force. When we estimate the gaps for different education groups, we see that the gap in informal jobs is higher than the gap in formal jobs only for the most educated employees but not significantly so. The small sample size for workers with tertiary education might be responsible of the lower precision in the estimates. For workers with primary education or no education, the wage gap is stronger among employees with a legal contract which is at odds with the intuition that labour market regulation, in particular minimum wages, should reduce the scope for wage gaps at the bottom of the wage distribution. Another striking pattern is the increase in the wage gap with the education level which can be interpreted as a form of "glass ceiling" in both the formal and the informal segments. The raw wage gaps does not account for the labour force heterogeneity. As men and women might have different characteristics in both types of jobs, a more detailed analysis is needed. Does the gender pay gap differ systematically between the formal and informal sectors once we control for the workforce characteristics? Does it depend on the education group the workers belong to?

Controlling for observable characteristics such as the exact number of years of education, age, tenure, sector of activity and location, increases the wage gap in both formal and informal sectors. This is an expected result as women are more educated than men and working women present overall better characteristics on average than working men in this sample. Since, the female advantage in observables is stronger in the formal sector, the formal adjusted wage gap increases more than the informal adjusted wage gap (from panel 1 to panel 2 in table8). As a result, the wage gaps in formal jobs and in informal jobs are not statically different from one another. Although skills receive lower returns in informal jobs, the gender differences in returns is about the same in percentage terms in formal and informal jobs: the average wage gap is about 0.2 log points which amounts to a difference of 22 percent. If we look at the gaps for different education groups, adjusting women's returns to the returns obtained by men would increase women's wages by 17 to 21 percent among workers with primary education and by 22 to 25 percent among workers with secondary education.

For high-skilled workers however, the adjusted wage gaps are statistically different in the two segments. It is higher among formal employees at 26 percent compared to 19 percent among informal employees. The so-called "glass ceiling effect" is stronger in jobs with legal labour contracts which goes in the direction of one hypothesis formulated in section 2.

We now turn to the effect of the selection bias on the gender wage gaps. The data give information on observed wages only, for workers working in a given sector. To infer the magnitude of the wage gap correctly though, we want to compare wage *offers* (that would be) made to all men and women. If selection into sector is non-random, observed wages either overstate or understate wage offers. If the selection bias differs by gender, the observed raw wage gap, given in the first panel of table 8, will not reflect the raw difference in wage offers. Controlling for sector participation enables us to recover the average wage offers within each sector, providing our control function captures properly the selection bias. Panel 3 shows how selection changes the average gender wage gaps differently in the informal and formal sectors.

In the informal sector, the observed wage gap overstates the wage gap in wage offers. This is because observed informal wages overestimate informal wage *offers* for both men and women but male wage offers are more strongly overestimated than female wage offers. Controlling for inclusion in the informal salaried worker sample reduces the average informal wage offer more for men than for women, and thus reduces the wage gap that can be explained by differences in characteristics and returns. Among informal employees, differences in returns have no role in explaining the gap anymore. Put differently, after purging the estimates from the selection effect, we cannot reject the hypothesis that men and women receive equal treatment for their skills in the informal sector.

In the formal sector, on the other hand, the gender gap in wage offers is offset by the selection bias and is underestimated by the observed gender wage gap. This is because male observed wages underestimate male wage *offers* while female observed wage overestimate female wage offers. Observable characteristics are better among working women which makes the part of the gender wage gap due to different returns even bigger than the total wage difference adjusted for the selection bias. The increase in the wage gap with education is robust to the treatment of selection.

These results highlight that labour regulation may impact gender wage inequality in the urban labour market in Brazil. The finding that wage gaps are positive and significant only in formal jobs is consistent with the following explanation. If employers believe that women have a higher quite probability, statistical discrimination induces employers to pay lower wages to women because they expect higher average female labour cost. We argue that the gap in gender expected labour cost because of gender differences in labour market attachment is higher in jobs where employment protection is binding. When an employee takes a temporary leave, his/her job must remain available to him/her, generating costs due to vacancy and replacement. This effect is expected to be weaker in informal job because the job of the employee on leave can be allocated to another worker permanently. We also find that the wage gap is higher among high-skilled workers in formal jobs, a result that is commonly found in the literature on gender wage gaps. This finding is often explained by statistical discrimination that produces higher gender gaps in high-wage jobs. Higher gender differences in pay among formal workers and the increase in the pay gap with the education level are thus consistent.

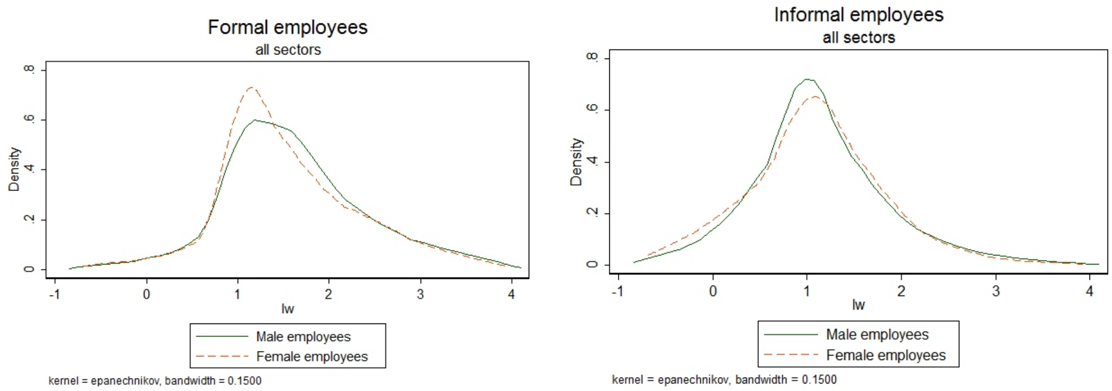
5 Conclusion

This paper investigates gender wage inequality in formal and informal jobs in Brazil. The data shows that the total average gender wage gap is positive and significant in both sectors. The informal sector features the highest total average gender wage gap but this conceals differences in male and female characteristics. When we ignore the selection bias, the differences in returns are the same in formal and informal jobs and are responsible for about 22% of the pay gap. This paper aditionnaly shows that the similitude in the formal and informal wage gaps is artificially generated by different selection of men and women in formal and informal jobs. In the informal sector, both male and female observed average wages overestimate their respective average wage offers but not by the same magnitude. The stronger male selection bias in informal jobs displaces male observed wage distribution further to the right compared to the female observed wage distribution. As a result the observed wage gap overestimates the gap in wage offers. We find that the difference in average wage offers faced by men and women is actually completely explain by differences in selection bias. The gender gap due to different returns is not significant in the informal sector. The opposite happens in the formal sector. The gender difference in selectivity bias narrows the gender gap in observed wages. This is because female observed wages overestimate female wage offers while male observed wages underestimate male wage offers. As a result, even controlling for selection, the gender wage gap due to different returns is strongly positive in formal jobs. Moreover, the gender wage gap increases with education in the formal sector.

Bigger gender differences in returns in the formal sector can certainly not lead to the conclusion that employment protection legislation is detrimental to women. First, the formal segment of the labour market provides higher wages to women, even if the formal wage premium is lower for women than for men. Additionally, given that women face a higher unemployment rate and need to take maternity leave, the flow of earnings of women relative to men can be higher in the formal sector because unemployment benefits and maternity leave benefits compensate for wage losses in the formal sector while wage losses are not compensated for in the informal sector. Further work is needed, first to really identify the impact of labour regulation on discriminatory behaviours, second to investigate how participating to the informal sector affects gender differences in earnings over the life cycle.

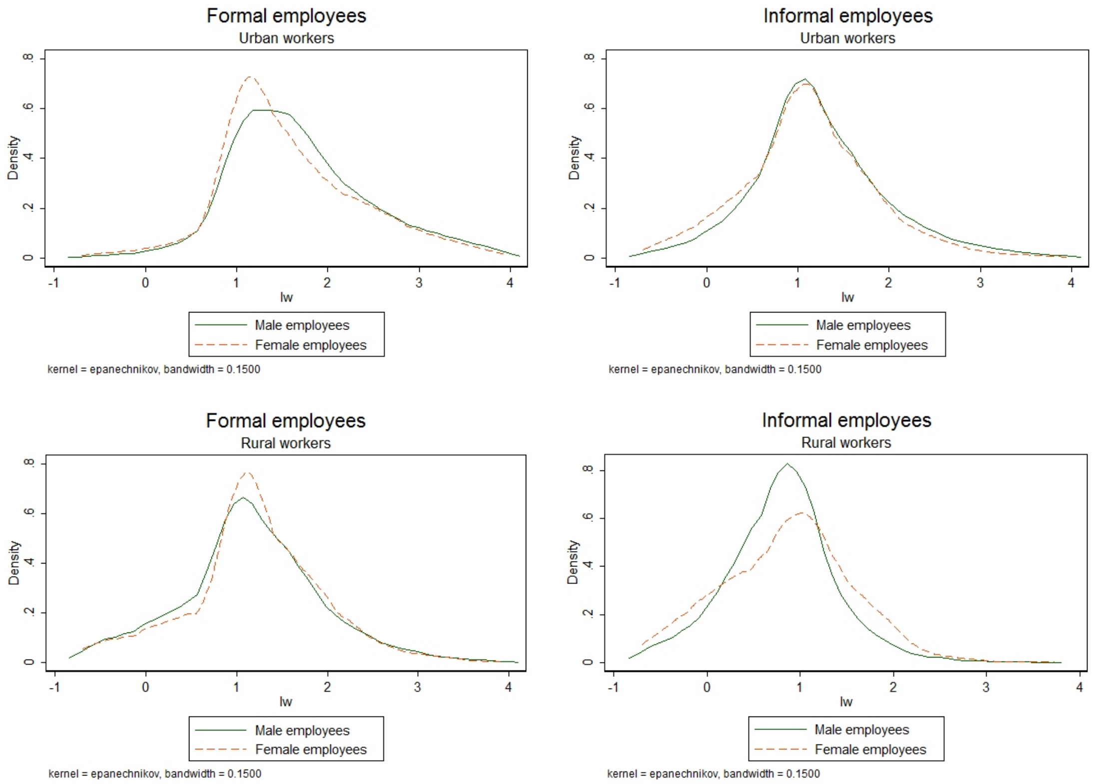
6 Tables

Figure 1: Wage Distributions by sex



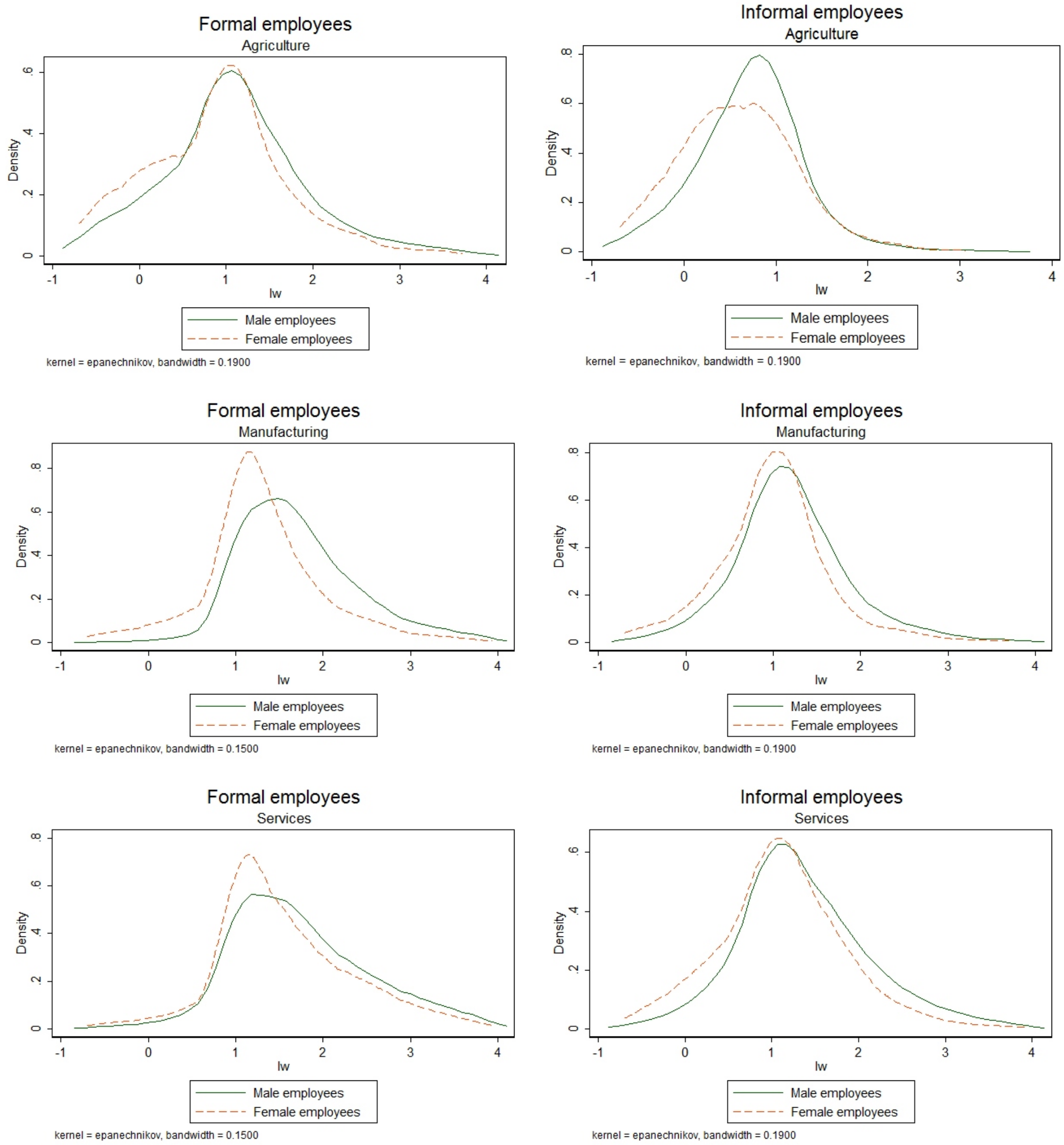
Source: PNAD, 2009, IBGE, Brazil.

Figure 2: Wage Distributions by sex,urban and rural wage-earners



Source: PNAD, 2009, IBGE, Brazil.

Figure 3: Wage Distributions by sex and sector



Source: PNAD, 2009, IBGE, Brazil.

Table 1: Descriptive statistics by sex, 2009

	Formal				Informal			
	(1)		(2)		(3)		(4)	
	Men		Women		Men		Women	
Demographics								
Age (mean)	38.15	(12.04)	38.06	(11.66)	33.98	(12.17)	35.31	(11.64)
Head of household	0.64	(0.48)	0.27	(0.45)	0.51	(0.50)	0.30	(0.46)
Living in couple	0.81	(0.39)	0.70	(0.46)	0.76	(0.43)	0.62	(0.48)
Children under 14	0.40	(0.49)	0.40	(0.49)	0.40	(0.49)	0.45	(0.50)
Household composition								
Number of people	3.8	(1.6)	3.7	(1.6)	4.1	(1.9)	4	(1.9)
Family members in the household	3.5	(1.3)	3.3	(1.3)	3.8	(1.7)	3.5	(1.5)
among the working-age members								
Share of the household members with a formal job[a]	0.29	(0.40)	0.40	(0.44)	0.15	(0.30)	0.32	(0.41)
Share of the family members with a formal job[a]	0.29	(0.39)	0.41	(0.42)	0.15	(0.30)	0.33	(0.42)
Mother lives in the household	0.30	(0.46)	0.32	(0.46)	0.40	(0.49)	0.29	(0.45)
Education								
Illiterate	0.07	(0.26)	0.05	(0.22)	0.12	(0.33)	0.07	(0.25)
Years of schooling (mean)	8.16	(4.34)	9.30	(4.35)	6.73	(4.36)	7.93	(4.21)
Job related variables								
Hourly Wage	7.99	(19.27)	6.46	(18.26)	4.13	(14.85)	3.60	(5.70)
Hours of Work	43.5	(11.7)	36.2	(14)	42.7	(12.5)	35	(15.3)
Full time	0.88	(0.34)	0.66	(0.48)	0.84	(0.40)	0.59	(0.50)
Several jobs	0.05	(0.22)	0.05	(0.22)	0.05	(0.22)	0.05	(0.21)
Union membership	0.22	(0.41)	0.21	(0.41)	0.07	(0.25)	0.04	(0.20)
Public sector	0.12	(0.33)	0.21	(0.41)	.	.	.	.
Civil servant	0.07	(0.25)	0.13	(0.34)	.	.	.	.
Age at first job								
Under 10	0.13	(0.34)	0.10	(0.30)	0.13	(0.34)	0.09	(0.29)
10-14	0.39	(0.49)	0.31	(0.46)	0.42	(0.49)	0.34	(0.47)
15-17	0.26	(0.44)	0.26	(0.44)	0.26	(0.44)	0.27	(0.44)
17-19	0.14	(0.35)	0.18	(0.39)	0.12	(0.33)	0.15	(0.36)
20-24	0.06	(0.24)	0.12	(0.32)	0.05	(0.22)	0.10	(0.30)
25-29	0.01	(0.09)	0.02	(0.16)	0.01	(0.09)	0.02	(0.15)
More than 30	0.00	(0.03)	0.02	(0.12)	0.00	(0.04)	0.02	(0.14)
Tenure (mean number of years)	2.80	(3.00)	2.82	(3.06)	2.77	(2.90)	2.96	(3.03)
Night work	0.02	(0.14)	0.01	(0.08)	0.02	(0.13)	0.01	(0.08)
N	81027		59015		17060		16549	

Source: Author's calculation based on the PNAD 2009, IBGE, Brazil. The columns give the shares among male formal wage-earners (1), female formal wage-earners (2), male informal wage-earners (3) and female informal wage-earners (4). Standard deviations in parentheses.

[a] The share of working-age household/family members holding formal jobs excludes the respondent.

We will use this variable to explain the sorting of individuals across job types.

Table 2: Share of educated people among active, unemployed and informal workers

	All		Active		Unemployed		Informal workers	
	Men	Women	Men	Women	Men	Women	Men	Women
Primary or less	35	32	33	26	23	19	47	37
Secondary	50	49	51	51	63	65	45	53
Tertiary	15	19	16	23	14	16	8	10
	100	100	100	100	100	100	100	100

Source: Author's calculation based on the PNAD 2009, IBGE, Brazil.

Table 3: Descriptive statistics by education and sex groups, 2009

Level of education	Participation rate		Unemployment rate		Informality rate among working individuals	
	Men	Women	Men	Women	Men	Women
Total	89	66	6	11	19	25
Primary or less	85	53	4	8	23	30
Secondary	91	68	7	13	15	22
Tertiary	90	80	5	7	8	9

Source: Author's calculation based on the PNAD 2009, IBGE, Brazil.

Table 4: Employment shares and informality rate by sex and sectorss

	Employment share			Informality rate		
Sector	Overall	Men	Women	Overall	Men	Women
Agriculture	22	24	18	20	27	8
Industry	14	16	12	16	16	15
Construction Mining	7	12	1	29	28	57
Services	57	48	69	24	18	30

Source: Author's calculation based on the PNAD, 2009, IBGE, Brazil.

Table 5: Labour market status, arginal effects for men and women separately. Urban workers.

Women	Inactive	Informal employee	Formal employee	Self-employed	Employer	Unemployed
Age	0.005***	-0.003***	-0.001***	0.002***	0.001***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	-0.021***	-0.012***	0.033***	0.000	0.003***	-0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Having Children	0.009*	-0.000	-0.007	-0.004	-0.002*	0.004
	(0.004)	(0.003)	(0.003)	(0.002)	(0.001)	(0.003)
... under 14	0.010**	-0.008**	-0.013***	0.020***	0.003**	-0.012***
	(0.004)	(0.003)	(0.004)	(0.003)	(0.001)	(0.002)
Living in couple	0.066***	-0.043***	-0.040***	-0.002	0.007***	0.012***
	(0.004)	(0.003)	(0.003)	(0.002)	(0.001)	(0.002)
Lone mother	-0.086***	0.037***	0.031***	-0.004	0.000	0.022***
	(0.006)	(0.005)	(0.006)	(0.004)	(0.002)	(0.005)
Formal workers	0.023***	-0.003	-0.005*	-0.018***	-0.007***	0.010***
in the household	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Unemployment rate	0.127	0.218***	-0.976***	0.289*	-0.020	0.368***
(regional, education specific)	(0.104)	(0.058)	(0.066)	(0.114)	(0.022)	(0.048)
Men	Inactive	Informal employee	Formal employee	Self-employed	Employer	Unemployed
Age	0.002***	-0.004***	-0.002***	0.005***	0.002***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	-0.007***	-0.011***	0.024***	-0.006***	0.006***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Having children	0.035***	0.001	-0.047***	-0.008*	-0.005**	0.028***
	(0.002)	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)
... under 14	-0.097***	0.000	0.090***	0.026***	0.015***	-0.034***
	(0.002)	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)
Living in couples	-0.034***	-0.015***	0.055***	-0.012**	0.017***	-0.010***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.002)	(0.002)
Formal workers	0.014***	0.004**	0.007**	0.009**	-0.022***	0.005***
in the household	(0.001)	(0.001)	(0.002)	(0.003)	(0.002)	(0.001)
Unemployment rate	-0.482***	-0.040	0.418***	-0.287***	-0.156***	0.196***
(region, education specific)	(0.056)	(0.060)	(0.081)	(0.066)	(0.035)	(0.043)

Notes: Marginal effects, standard errors in parenthesis. The marginal effects of each explanatory variables on the probability to be in the six different outcomes are computed based on the multinomial logit estimation.

Table 6: Hourly wages in the informal and formal sectors. Women in urban areas

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Years of education	0.046**	0.083**	0.036**	0.018	0.137**	0.237**
	(0.002)	(0.002)	(0.002)	(0.023)	(0.009)	(0.016)
Age	0.037**	0.033**	0.036**	0.024**	0.035**	0.011
	(0.002)	(0.003)	(0.003)	(0.005)	(0.002)	(0.006)
Age^2	-0.000**	-0.000**	-0.000**	-0.000**	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	-0.022**	0.004	-0.022**	-0.022**	0.005	0.003**
	(0.004)	(0.003)	(0.003)	(0.001)	(0.004)	(0.000)
Tenure^2	0.002**	-0.001**	0.002**	0.002**	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Black	-0.041**	-0.092**	-0.039**	-0.038**	-0.090**	-0.086**
	(0.012)	(0.008)	(0.009)	(0.001)	(0.004)	(0.004)
Unemployment rate	-4.095**	-6.805**	-3.694**	-1.939**	-6.466**	-3.578**
(region, education specific)	(0.466)	(0.755)	(0.309)	(0.504)	(0.095)	(0.226)
Constant	-0.080	0.482**	-0.288**	0.659	-0.760**	-2.106**
	(0.077)	(0.079)	(0.092)	(0.529)	(0.267)	(0.036)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			0.478**	1.811	0.857**	12.931*
			(0.059)	(1.000)	(0.253)	(5.389)
ρ_1				0.454**		0.001
				(0.060)		(0.010)
ρ_2			-0.288**			-1.141**
			(0.032)			(0.046)
ρ_3				0.183	-0.702**	
				(0.503)	(0.019)	
ρ_4				0.274		0.401**
				(0.289)		(0.067)
ρ_5				-1.799**		-0.572**
				(0.195)		(0.176)
ρ_6				0.821**		0.983**
				(0.147)		(0.035)
R^2	0.28	0.52				
N	14,511	32,133	14,511	14,511	32,133	32,133

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

Table 7: Hourly wages in the informal and formal sectors. Men in urban areas

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Years of education	0.058**	0.086**	0.041**	0.044*	0.080**	0.116**
	(0.002)	(0.002)	(0.004)	(0.017)	(0.000)	(0.007)
Age	0.052**	0.053**	0.048**	0.037**	0.050**	0.046**
	(0.003)	(0.003)	(0.001)	(0.000)	(0.000)	(0.001)
Age^2	-0.001**	-0.000**	-0.001**	-0.001**	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	-0.004	0.004	-0.004**	-0.003	0.004**	0.002**
	(0.005)	(0.003)	(0.000)	(0.009)	(0.001)	(0.001)
Tenure^2	0.000	-0.001*	0.000	0.000	-0.001**	-0.000*
	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
Black	-0.106**	-0.087**	-0.105**	-0.104**	-0.087**	-0.085**
	(0.013)	(0.008)	(0.008)	(0.003)	(0.007)	(0.007)
Unemployment rate	-4.122**	-6.491**	-4.047**	-1.867*	-6.603**	-5.355**
(region, education specific)	(0.322)	(0.392)	(0.274)	(0.904)	(0.031)	(0.097)
Constant	-0.079	0.107	-0.345**	0.878**	0.323**	-0.720**
	(0.064)	(0.064)	(0.020)	(0.067)	(0.018)	(0.066)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			0.852	3.673**	0.266**	9.481**
			(0.460)	(0.153)	(0.001)	(1.891)
ρ_1				-0.172**		0.001
				(0.005)		(0.018)
ρ_2			-0.396**			-1.032**
			(0.014)			(0.024)
ρ_3				0.517**	0.270**	
				(0.178)	(0.011)	
ρ_4				0.378*		0.519**
				(0.189)		(0.093)
ρ_5				-1.469**		-0.677**
				(0.073)		(0.064)
ρ_6				0.742**		0.977**
R^2	0.34	0.48				
N	12,594	41,679	12,594	12,594	41,679	41,679

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

Table 8: Gender wage gap decomposition. Informal and formal sectors, urban areas.

Level of Education		**All**	**Primary or less**	**Secondary**	**Tertiary**
1-Raw wage gap					
$\overline{\ln W}_{mj} - \overline{\ln W}_{fj}$					
	Informal	**0.133****	**0.058***	0.193**	0.244**
		(0.014)	(0.026)	(0.015)	(0.030)
	Formal	**0.075****	**0.176****	0.197**	0.215**
		(0.005)	(0.020)	(0.016)	(0.012)
	Welch's t-statistics	-2.51	3.94	0.18	-1.26
2-Controlling for observables only					
Part due to differences in returns					
$\overline{WG}_j = \overline{\mathbf{X}}'_{\mathbf{m}}(\widehat{\beta}_{\mathbf{mj}} - \widehat{\beta_{\mathbf{pj}}}) + \overline{\mathbf{X}}'_{\mathbf{f}}(\widehat{\beta_{\mathbf{pj}}} - \widehat{\beta_{\mathbf{fj}}})$					
	Informal	0.200**	0.191**	0.223**	**0.173****
		(0.019)	(0.030)	(0.018)	(0.026)
	Formal	0.214**	0.151**	0.197**	**0.232****
		(0.010)	(0.014)	(0.012)	(0.013)
	Welch's t-statistics	0.08	-1.60	-1.37	2.66
3-Controlling for observables and self-selection					
3.1-Wage gap after subtracting the selection bias					
$\overline{\ln W}_{mj} - \overline{\ln W}_{fj} - (\theta_{mj}h_{mj} - \theta_{fj}h_{fj})$					
Lee	Informal	-0.125	-0.289	-0.159	0.455
		(0.237)	(0.279)	(0.312)	(0.853)
	Formal	0.612**	0.622*	0.500	0.639
		(0.219)	(0.263)	(0.295)	(0.516)
DMF	Informal	0.219	0.146	0.451	-0.934
		(0.308)	(0.347)	(0.445)	(1.053)
	Formal	0.325**	0.082	0.303*	0.572**
		(0.115)	(0.227)	(0.121)	(0.121)
3.2-Part due to difference in returns					
$\overline{WG}_{Sj} = \overline{\mathbf{X}}'_{\mathbf{m}}(\widehat{\gamma}_{\mathbf{mj}} - \widehat{\gamma_{\mathbf{pj}}}) + \overline{\mathbf{X}}'_{\mathbf{f}}(\widehat{\gamma_{\mathbf{pj}}} - \widehat{\gamma_{\mathbf{fj}}})$					
Lee	Informal	**-0.063**	**-0.147**	**-0.139**	0.389
		(0.237)	(0.284)	(0.310)	(0.851)
	Formal	**0.974****	**0.388****	**0.553****	0.329**
		(0.050)	(0.019)	(0.021)	(0.015)
	Welch's t-statistics	4.28	1.70	2.21	-0.07
DMF	Informal	0.271	0.295	0.478	-0.990
		(0.310)	(0.353)	(0.445)	(1.053)
	Formal	0.455**	0.049	0.297*	0.565**
		(0.114)	(0.226)	(0.121)	(0.120)
	Welch's t-statistics	0.56	-0.59	-0.39	1.47
Number in Informal		27,105	9,856	14,443	2,806
Share of women		53%	51%	54%	57%
Number in Formal		78,378	11,586	39,717	17,705
Share of women		44%	33%	41%	60%

Notes: * $p < 0.05$; ** $p < 0.01$ s.e. in parenthesis. Panel 1: equation (1). Panel 2: equation (3). Panel 3.2: equation (6). The results are expressed on the logarithmic scale. To obtain the difference in percentage points: $(exp(WG) - 1) \times 100$. The Welch's test is applied to test the difference between the formal and the informal gaps with different population sizes and variances. Bold characters indicate that the difference between the formal and the informal wage gaps is significant at 10% when $|t| > 1.64$, the difference is significant at 5% if $|t| > 1.96$

References

Albrecht, James, Anders Bjorklund, and Susan Vroman, "Is There a Glass Ceiling in Sweden?," *Journal of Labor Economics*, 2003, *21* (1), pp. 145–177.

Almeida, Rita and Pedro Carneiro, "Inequality and employment in a dual economy: Enforcement of labor regulation in Brazil," 2007.

Appleton, Simon, John Hoddinott, and Pramila Krishnan, "The gender wage gap in three African countries," *Economic Development and Cultural Change*, 1999, *47* (2), 289–312.

Arabsheibani, G Reza, Francisco Galrão Carneiro, and Andrew Henley, *Gender wage differentials in Brazil: trends over a turbulent era*, Vol. 3148, World Bank, 2003.

Bertrand, Marianne, Claudia Goldin, and Lawrence F Katz, "Dynamics of the gender gap for young professionals in the financial and corporate sectors," *American Economic Journal: Applied Economics*, 2010, pp. 228–255.

Blau, Francine D and Lawrence M Kahn, "The US Gender Pay Gap in the 1990s: Slowing Convergence," *Industrial and Labor Relations Review*, 2006, pp. 45–66.

Blinder, Alan S., "Wage Discrimination: Reduced Form and Structural Estimates," *The Journal of Human Resources*, 1973, *8* (4), pp. 436–455.

Bourguignon, François, Martin Fournier, and Marc Gurgand, "Selection bias corrections based on the multinomial logit model: Monte Carlo comparisons," *Journal of Economic Surveys*, 2007, *21* (1), 174–205.

Carneiro, Francisco Galrão and Andrew Henley, "Modelling Formal vs. Informal Employment and Earnings: Micro-econometric Evidence for Brazil," 2002.

de la Rica, Sara, Juan Dolado, and Vanesa Llorens, "Ceilings or floors? Gender wage gaps by education in Spain," *Journal of Population Economics*, 2008, *21*, 751–776.

Deininger, Klaus, Songqing Jin, and Hari Nagarajan, "Wage Discrimination in India's Informal Labor Markets: Exploring the Impact of Caste and Gender," *Review of Development Economics*, 2013, *17* (1), 130–147.

Dubin, Jeffrey A and Daniel L McFadden, "An econometric analysis of residential electric appliance holdings and consumption," *Econometrica: Journal of the Econometric Society*, 1984, pp. 345–362.

Ermisch, John F. and Robert E. Wright, "Wage Offers and Full-Time and Part-Time Employment by British Women," *The Journal of Human Resources*, 1993, *28* (1), pp. 111–133.

Fortin, Nicole M, "The Gender Wage Gap among Young Adults in the United States The Importance of Money versus People," *Journal of Human Resources*, 2008, *43* (4), 884–918.

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo, "Decomposition methods in economics," *Handbook of labor economics*, 2011, *4*, 1–102.

Gasparini, Leonardo and Leopoldo Tornarolli, "Labor informality in Latin America and the Caribbean: patterns and trends from household survey microdata," *Desarrollo y Sociedad*, 2009, (63), 13–80.

Gong, Xiaodong and Arthur Van Soest, "Wage differentials and mobility in the urban labour market: a panel data analysis for Mexico," *Labour Economics*, 2002, *9* (4), 513–529.

Gunther, Isabel and Andrey Launov, "Informal employment in developing countries: opportunity or last resort?," *Journal of Development Economics*, 2012, *97* (1), 88–98.

Kambourov, Gueorgui and Iourii Manovskii, "Occupational mobility and wage inequality," *The Review of Economic Studies*, 2009, *76* (2), 731–759.

Lazear, Edward P. and Sherwin Rosen, "Male-Female Wage Differentials in Job Ladders," *Journal of Labor Economics*, 1990, *8* (1), pp. S106–S123.

Lee, Lung-Fei, "Generalized econometric models with selectivity," *Econometrica*, 1983, pp. 507–512.

Madalozzo, Regina, "Occupational segregation and the gender wage gap in Brazil: an empirical analysis," *Economia Aplicada*, 2010, *14* (2), 147–168.

Magnac, Th, "Segmented or competitive labor markets," *Econometrica: journal of the Econometric Society*, 1991, pp. 165–187.

Maloney, William F, "Does informality imply segmentation in urban labor markets? Evidence from sectoral transitions in Mexico," *The World Bank Economic Review*, 1999, *13* (2), 275–302.

McFadden, D, "Conditional logit analysis of qualitative choice behavior," *Frontiers in Econometrics*, 1973, pp. 105–142.

Neuman, Shoshana and Ronald L Oaxaca, "Wage decompositions with selectivity-corrected wage equations: A methodological note," *The Journal of Economic Inequality*, 2004, *2* (1), 3–10.

Oaxaca, Ronald, "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 1973, *14* (3), pp. 693–709.

Oaxaca, Ronald L and Michael R Ransom, "On discrimination and the decomposition of wage differentials," *Journal of Econometrics*, 1994, *61* (1), 5 – 21.

Oaxaca, Ronald L. and Michael R. Ransom, "Identification in Detailed Wage Decompositions," *The Review of Economics and Statistics*, 1999, *81* (1), pp. 154–157.

Ogloblin, Constantin G., "The Gender Earnings Differential in the Russian Transition Economy," *Industrial and Labor Relations Review*, 1999, *52* (4), pp. 602–627.

Olivetti, C. and B. Petrongolo, "Unequal Pay or Unequal Employment? A Cross-Country Analysis of Gender Gaps," *Journal of Labor Economics*, 2008, *26* (4), 621–654.

Pagán, José A and Miren Ullibarri, "Group heterogeneity and the gender earnings gap in Mexico," *Economía Mexicana, Nueva Época*, 2000, *9* (1), 23–40.

Pradhan, Menno and Arthur Van Soest, "Formal and informal sector employment in urban areas of Bolivia," *Labour Economics*, 1995, *2* (3), 275–297.

Santos, Renato Vale and Eduardo Pontual Ribeiro, "Diferenciais de Rendimentos entre Homens e Mulheres no Brasil Revisitado: Explorando o Teto de Vidro," *Centro Universitario Newton Paiva/MG e PPGE/UFRGS*, 2006.

Schmiedern, Johannes, "Returns to Tenure: Is Specific Human Capital Acquired on the Job?," mimeo 2007.

Sullivan, Paul, "Estimation of an occupational choice model when occupations are misclassified," *Journal of Human Resources*, 2009, *44* (2), 495–535.

Tansel, Aysit, "Wage Earners, Self Employed and Gender in the Informal Sector in Turkey," in "in" Economic Research Forum 2001.

Wright, Robert E and John F Ermisch, "Gender discrimination in the British labour market: a reassessment," *The Economic Journal*, 1991, *101* (406), 508–522.

Yun, Myeong-Su, "An extension of the Oaxaca decomposition using generalized residuals," *Journal of Economic and Social Measurement*, 2007, *32* (1), 15–22.

A Multinomial logit estimates

Table 9: Labour market status. Urban workers

Relative risk ratios from the multinomial logit estimation. Formal employee is the base outcome

Women	Inactive	Informal employee	Self-employed	Employer	Unemployed
Age	0.020***	-0.018***	0.031***	0.040***	-0.027***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
Years of education	-0.206***	-0.226***	-0.132***	0.075***	-0.175***
	(0.002)	(0.003)	(0.003)	(0.007)	(0.003)
Children	0.057**	0.028	-0.010	-0.103	0.066*
... under 14	0.085***	-0.017	0.249***	0.215***	-0.076*
	(0.022)	(0.029)	(0.032)	(0.063)	(0.030)
Living in couple	0.368***	-0.148***	0.146***	0.576***	0.256***
	(0.022)	(0.027)	(0.030)	(0.065)	(0.031)
Lone mother	-0.426***	0.122**	-0.182***	-0.086	0.086
	(0.039)	(0.044)	(0.053)	(0.127)	(0.050)
Formal workers	0.098***	0.011	-0.161***	-0.376***	0.121***
in the household	(0.015)	(0.019)	(0.025)	(0.068)	(0.019)
Unemployment rate	4.116***	5.580***	7.536***	2.039	7.393***
(regional, by education group)	(0.425)	(0.566)	(0.600)	(1.279)	(0.580)
Constant	0.579***	1.564***	-1.821***	-5.372***	0.621***
	(0.111)	(0.138)	(0.172)	(0.357)	(0.150)
N	110918				
Men	Inactive	Informal employee	Self-employed	Employer	Unemployed
Age	0.022***	-0.025***	0.031***	0.046***	-0.030***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Years of education	-0.127***	-0.154***	-0.113***	0.063***	-0.102***
	(0.003)	(0.003)	(0.002)	(0.004)	(0.004)
Children	0.432***	0.134***	0.059*	0.000	0.567***
	(0.026)	(0.028)	(0.025)	(0.042)	(0.035)
... under 14	-1.181***	-0.246***	-0.114***	0.129***	-0.735***
	(0.029)	(0.025)	(0.023)	(0.037)	(0.032)
Living in couple	-0.425***	-0.271***	-0.229***	0.271***	-0.294***
	(0.027)	(0.027)	(0.026)	(0.049)	(0.032)
Formal workers	0.103***	0.023	-0.094***	-0.362***	0.069***
in the household	(0.012)	(0.014)	(0.017)	(0.041)	(0.014)
Unemployment rate	-5.140***	-1.324*	-0.799	-4.334***	1.710*
(regional, by education group)	(0.573)	(0.583)	(0.502)	(0.792)	(0.682)
Constant	-0.424**	1.468***	-0.499***	-4.090***	-0.111
	(0.145)	(0.127)	(0.119)	(0.219)	(0.176)
N	99079				

Notes: In the multinomial logit model, the risk of $y = j$ is measured as the risk of the outcome relative to the base outcome, $Pr(y = j)/Pr(y = 3) = exp^{X\beta_j}$ and the relative risk ratios for a one-unit change in X is the exponentiated value of the coefficient

Table 10: Labour market status, rural workers.
Marginal effects for men and women separately

Women	Inactive	Informal employee	Formal employee	Self-employed	Employer	Unemployed
Age	-0.002***	-0.002***	0.001***	0.005***	0.000***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	-0.014***	-0.003***	0.028***	-0.012***	0.001***	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Children	-0.030**	-0.004	0.020**	0.017	-0.002	-0.000
	(0.010)	(0.007)	(0.006)	(0.010)	(0.002)	(0.004)
... under 14	-0.019*	-0.004	-0.010	0.032**	0.005*	-0.003
	(0.010)	(0.006)	(0.006)	(0.010)	(0.002)	(0.004)
Living in couple	0.006	-0.043***	-0.049***	0.103***	-0.007*	-0.010
	(0.012)	(0.009)	(0.009)	(0.012)	(0.003)	(0.005)
Lone mother	-0.042*	0.062***	-0.013	-0.022	-0.003*	0.018*
	(0.019)	(0.015)	(0.012)	(0.021)	(0.001)	(0.009)
Formal workers	0.030***	0.000	-0.015***	-0.018*	-0.001	0.003
in the household	(0.007)	(0.004)	(0.005)	(0.007)	(0.001)	(0.002)
Unemployment rate	0.998***	0.486***	-0.808***	-0.858***	-0.002	0.184*
(region, education specific)	(0.236)	(0.143)	(0.127)	(0.257)	(0.029)	(0.072)
Men	Inactive	Informal employee	Formal employee	Self-employed	Employer	Unemployed
Age	0.001***	-0.006***	-0.002***	0.007***	0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	-0.002**	-0.020***	0.019***	-0.002	0.005***	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Children	0.023***	-0.025**	-0.037***	0.040***	-0.007*	0.007**
	(0.005)	(0.009)	(0.009)	(0.010)	(0.004)	(0.002)
... under 14	-0.053***	0.038***	0.076***	-0.062***	0.006	-0.004
	(0.004)	(0.008)	(0.007)	(0.009)	(0.003)	(0.002)
Living in couples	-0.036***	-0.050***	0.037***	0.054***	0.005	-0.011**
	(0.006)	(0.009)	(0.008)	(0.010)	(0.003)	(0.003)
Formal workers	0.017***	-0.010*	-0.008	0.016**	-0.016***	0.000
in the household	(0.002)	(0.004)	(0.005)	(0.006)	(0.004)	(0.001)
Unemployment rate	0.373**	0.680**	-0.708***	-0.494*	-0.072	0.221***
(region, education specific)	(0.127)	(0.216)	(0.177)	(0.251)	(0.075)	(0.058)

Notes: Marginal effects, standard errors in parenthesis. The marginal effects of each explanatory variables on the probability to be in the six different outcomes are computed based on the multinomial logit estimation.

B Wage equations on the pooled sample, the reference wage structure for the Oaxaca-Blinder decomposition

Table 11: Hourly wages in the informal and formal sectors. Pooled sample, urban areas.

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Female	-0.100**	-0.107**	-0.097**	-0.076**	-0.109**	-0.040**
	(0.005)	(0.004)	(0.005)	(0.006)	(0.005)	(0.006)
Male	0.100**	0.107**	0.097**	0.076**	0.109**	0.040**
	(0.005)	(0.004)	(0.005)	(0.006)	(0.005)	(0.006)
Years of education	0.053**	0.085**	0.047**	0.063**	0.082**	0.090**
	(0.002)	(0.002)	(0.002)	(0.005)	(0.003)	(0.006)
Age	0.044**	0.045**	0.043**	0.037**	0.044**	0.031**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
Age^2	-0.000**	-0.000**	-0.000**	-0.000**	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	-0.014**	0.005*	-0.014**	-0.013**	0.002	0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Tenure^2	0.001**	-0.001**	0.001**	0.001**	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Black	-0.072**	-0.090**	-0.072**	-0.069**	-0.091**	-0.088**
	(0.009)	(0.007)	(0.009)	(0.009)	(0.007)	(0.007)
Unemployment rate	-4.202**	-6.614**	-4.057**	-2.866**	-6.780**	-6.150**
(region, education specific)	(0.342)	(0.548)	(0.339)	(0.327)	(0.512)	(0.469)
Constant	-0.070	0.240**	-0.177*	1.223**	0.222**	0.272**
	(0.060)	(0.059)	(0.075)	(0.176)	(0.066)	(0.097)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			0.456**	5.585**	0.784**	4.690**
			(0.061)	(1.838)	(0.117)	(0.563)
ρ_1				0.290**		0.590**
				(0.041)		(0.059)
ρ_2			-0.129**			0.024
			(0.038)			(0.080)
ρ_3				0.485**	-0.051*	
				(0.128)	(0.022)	(0.087)
ρ_4				0.635**		0.255
				(0.092)		(0.137)
ρ_5				-1.444**		-1.139**
				(0.339)		(0.234)
ρ_6				1.656**		1.376**
				(0.229)		(0.166)
N	27,105	78,367	27,105	27,105	69,009	69,009

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

C Wage equations by education groups

C.1 Workers with primary education or less

Table 12: Hourly wages. Women with primary education or less, urban areas.

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Years of education	0.026**	0.009**	0.026**	0.012	0.013	0.000
	(0.004)	(0.003)	(0.008)	(0.013)	(0.013)	(0.026)
age	0.024**	-0.007	0.024**	0.020	-0.007**	-0.002**
	(0.005)	(0.005)	(0.001)	(0.015)	(0.002)	(0.000)
Age^2	-0.000**	0.000*	-0.000**	-0.000	0.000**	0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	-0.017	-0.012*	-0.017	-0.018	-0.011**	-0.012**
	(0.010)	(0.005)	(0.011)	(0.011)	(0.001)	(0.003)
Tenure^2	0.001	0.001	0.001	0.001	0.001**	0.001**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Black	0.005	-0.027*	0.004	0.004	-0.027**	-0.027**
	(0.018)	(0.012)	(0.020)	(0.013)	(0.005)	(0.007)
Constant	-0.084	1.338**	-0.080	0.170	1.206**	1.521**
	(0.106)	(0.097)	(0.167)	(0.481)	(0.078)	(0.581)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			0.298**	0.593	0.121**	0.581
			(0.002)	(0.354)	(0.024)	(1.473)
ρ_1				0.562		0.266
				(0.374)		(0.796)
ρ_2			0.008			0.951
			(0.047)			(0.495)
ρ_3				-0.325	-0.219**	
				(0.592)	(0.053)	
ρ_4				-0.108		-0.498
				(0.534)		(0.934)
ρ_5				-0.983		0.258
				(0.811)		(0.396)
ρ_6				0.760		-0.916**
				(0.762)		(0.113)
R^2	0.21	0.17				
N	5,032	4,043	.	.	.	.

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

Table 13: Hourly wages. Men with primary education or less, urban areas.

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Years of education	0.020**	0.027**	0.014**	0.010	0.022**	0.031**
	(0.004)	(0.003)	(0.005)	(0.018)	(0.001)	(0.001)
age	0.031**	0.022**	0.028**	0.023**	0.020*	0.018**
	(0.004)	(0.003)	(0.001)	(0.007)	(0.008)	(0.005)
Age^2	-0.000**	-0.000**	-0.000**	-0.000**	-0.000	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	0.023**	0.013**	0.023	0.023**	0.013	0.013**
	(0.008)	(0.004)	(0.018)	(0.005)	(0.008)	(0.001)
Tenure^2	-0.003**	-0.001**	-0.003	-0.003**	-0.001	-0.001**
	(0.001)	(0.000)	(0.002)	(0.001)	(0.001)	(0.000)
Black	-0.089**	-0.042**	-0.089**	-0.088**	-0.041**	-0.040**
	(0.019)	(0.008)	(0.005)	(0.020)	(0.000)	(0.011)
Constant	0.161*	0.542**	-0.029	1.047*	0.758**	0.581
	(0.073)	(0.046)	(0.161)	(0.490)	(0.193)	(0.354)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			0.446	2.440	0.190**	3.150
			(0.544)	(1.268)	(0.007)	(1.814)
ρ_1				0.086		-0.117
				(0.193)		(0.148)
ρ_2			-0.362**			-0.678**
			(0.116)			(0.146)
ρ_3				0.644**	0.322**	
				(0.187)	(0.003)	
ρ_4				0.105		0.356*
				(0.451)		(0.149)
ρ_5				-1.520**		-0.924**
				(0.299)		(0.083)
ρ_6				0.746		1.289**
				(0.478)		(0.145)
R^2	0.22	0.19				
N	4,824	8,467	.	.	.	.

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

C.2 Workers with secondary education

Table 14: Hourly wages. Women with secondary education, urban areas.

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Years of education	0.018**	0.056**	0.001	0.060	0.068**	0.198**
	(0.003)	(0.003)	(0.009)	(0.031)	(0.006)	(0.000)
age	0.041**	0.016**	0.039**	0.024**	0.016**	-0.003
	(0.003)	(0.002)	(0.000)	(0.008)	(0.000)	(0.003)
Age^2	-0.000**	-0.000	-0.000**	-0.000**	-0.000**	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	-0.023**	0.002	-0.023**	-0.022**	0.002	0.002
	(0.006)	(0.003)	(0.007)	(0.006)	(0.001)	(0.003)
Tenure^2	0.002**	-0.001*	0.002**	0.002**	-0.001**	-0.001*
	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)
Black	-0.037*	-0.060**	-0.035**	-0.034**	-0.060**	-0.058**
	(0.015)	(0.007)	(0.011)	(0.011)	(0.002)	(0.000)
Constant	-0.176	0.283**	-0.311**	0.263	0.046	-1.506**
	(0.117)	(0.085)	(0.017)	(0.266)	(0.163)	(0.136)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			0.489**	1.214	0.178**	6.029**
			(0.048)	(0.716)	(0.009)	(1.070)
ρ_1				-0.007		-0.149**
				(0.202)		(0.022)
ρ_2			-0.319**			-1.094**
			(0.045)			(0.022)
ρ_3				0.606*	-0.248*	
				(0.298)	(0.102)	
ρ_4				-0.700*		0.374**
				(0.334)		(0.092)
ρ_5				-0.687		-0.349**
				(0.426)		(0.093)
ρ_6				0.895*		1.001**
				(0.357)		(0.010)
R^2	0.22	0.26				
N	7,878	17,924	.	.	.	.

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

Table 15: Hourly wages. Men with secondary education, urban areas.

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Years of education	0.045**	0.063**	0.013	0.085**	0.045**	0.091**
	(0.004)	(0.002)	(0.015)	(0.023)	(0.004)	(0.008)
age	0.058**	0.043**	0.052**	0.053**	0.036**	0.041**
	(0.004)	(0.004)	(0.006)	(0.008)	(0.001)	(0.002)
Age^2	-0.001**	-0.000**	-0.001**	-0.001**	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	-0.013	0.000	-0.013**	-0.014*	0.000	-0.000
	(0.009)	(0.004)	(0.001)	(0.006)	(0.002)	(0.007)
$Tenure^2$	0.002	-0.000	0.002**	0.002*	-0.000	-0.000
	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)
Black	-0.093**	-0.072**	-0.093**	-0.092**	-0.072**	-0.070**
	(0.015)	(0.007)	(0.012)	(0.011)	(0.002)	(0.003)
Constant	-0.514**	-0.173	-0.662**	0.457	0.343**	-0.821**
	(0.068)	(0.095)	(0.132)	(0.266)	(0.081)	(0.147)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			1.130**	4.544**	0.262**	3.911**
			(0.380)	(1.472)	(0.019)	(0.758)
ρ_1				-0.184		-0.019
				(0.205)		(0.025)
ρ_2			-0.416			-0.906**
			(0.316)			(0.022)
ρ_3				0.464**	0.559**	
				(0.106)	(0.067)	
ρ_4				0.521*		0.579**
				(0.248)		(0.066)
ρ_5				-1.389**		-0.804**
				(0.091)		(0.065)
ρ_6				0.809**		0.950**
				(0.181)		(0.070)
R^2	0.26	0.32				
N	6,565	26,205	.	.	.	.

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

C.3 Workers with secondary education

Table 16: Hourly wages. Women with tertiary education, urban areas.

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Years of education	0.127**	0.152**	0.096**	0.174**	0.121**	0.148**
	(0.014)	(0.004)	(0.001)	(0.039)	(0.012)	(0.018)
age	0.032**	0.048**	0.014	0.014	0.044**	0.029**
	(0.012)	(0.004)	(0.010)	(0.011)	(0.009)	(0.001)
Age^2	-0.000	-0.000**	-0.000	-0.000*	-0.000*	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	-0.017	0.009	-0.019	-0.022	0.009**	0.008
	(0.027)	(0.005)	(0.017)	(0.027)	(0.003)	(0.010)
Tenure^2	0.001	-0.001*	0.001	0.001	-0.001**	-0.001
	(0.003)	(0.000)	(0.003)	(0.003)	(0.000)	(0.001)
Black	-0.119*	-0.138**	-0.119	-0.121**	-0.138**	-0.134**
	(0.049)	(0.016)	(0.075)	(0.031)	(0.024)	(0.001)
Constant	-1.167**	-0.997**	-1.810	1.789	-0.318	-0.553*
	(0.410)	(0.275)	(0.960)	(1.586)	(0.459)	(0.223)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			4.358**	24.638	0.377**	7.519**
			(1.546)	(15.131)	(0.039)	(1.074)
ρ_1				0.158		0.420*
				(0.174)		(0.199)
ρ_2			-0.387**			-0.654**
			(0.019)			(0.009)
ρ_3				0.892**	0.590**	
				(0.280)	(0.178)	
ρ_4				-1.196*		-0.929**
				(0.573)		(0.002)
ρ_5				-0.209		-0.423**
				(0.474)		(0.040)
ρ_6				0.366		1.151**
				(0.688)		(0.165)
R^2	0.31	0.35				
N	1,601	12,513	.	.	.	.

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

Table 17: Hourly wages. Men with tertiary education, urban areas.

	Informal	Formal	Informal Lee	DMF	Formal Lee	DMF
Years of education	0.121**	0.150**	0.110**	0.097**	0.131**	0.128**
	(0.011)	(0.006)	(0.002)	(0.033)	(0.004)	(0.008)
age	0.043**	0.065**	0.034**	0.038*	0.054**	0.025*
	(0.012)	(0.004)	(0.006)	(0.019)	(0.005)	(0.011)
Age^2	-0.000*	-0.001**	-0.000	-0.000	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	-0.024	0.002	-0.028**	-0.030	0.003**	0.001
	(0.018)	(0.007)	(0.009)	(0.025)	(0.001)	(0.006)
Tenure^2	0.002	-0.001	0.002**	0.002	-0.001	-0.000
	(0.002)	(0.001)	(0.000)	(0.002)	(0.000)	(0.001)
Black	-0.192**	-0.153**	-0.191**	-0.188**	-0.153**	-0.148**
	(0.039)	(0.022)	(0.044)	(0.044)	(0.009)	(0.014)
Constant	-0.290	-1.554**	-1.126	0.584	-0.852**	-0.194
	(0.283)	(0.115)	(0.788)	(0.938)	(0.030)	(0.572)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			3.353	7.439	0.464**	4.528*
			(2.675)	(7.806)	(0.032)	(2.300)
ρ_1				-0.222		-0.610**
				(0.446)		(0.155)
ρ_2			-0.379**			0.371
			(0.038)			(0.563)
ρ_3				-0.200	0.761**	
				(0.386)	(0.165)	
ρ_4				1.194**		-0.197
				(0.432)		(0.538)
ρ_5				-1.016**		-0.869**
				(0.186)		(0.198)
ρ_6				0.364		1.302**
				(0.450)		(0.034)
R^2	0.27	0.36				
N	1,205	9,215	1,205	1,205	9,215	9,215

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

D Results for rural workers

Table 18: Hourly wages. Women in rural areas.

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Years of education	0.037**	0.062**	0.033**	0.062**	0.070**	0.053*
	(0.007)	(0.004)	(0.006)	(0.019)	(0.012)	(0.025)
age	0.033**	0.033**	0.029**	0.020	0.034**	0.021**
	(0.010)	(0.005)	(0.005)	(0.011)	(0.002)	(0.007)
Age^2	-0.000**	-0.000**	-0.000**	-0.000*	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	-0.036	-0.005	-0.034*	-0.032*	-0.005	-0.009
	(0.023)	(0.013)	(0.017)	(0.016)	(0.008)	(0.009)
Tenure^2	0.003	0.000	0.002	0.002	0.000	0.001
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
Black	-0.049	-0.036	-0.045	-0.050	-0.036*	-0.032
	(0.028)	(0.022)	(0.036)	(0.040)	(0.016)	(0.022)
Unemployment rate	-2.228*	-2.448**	-1.674	-2.424**	-2.455**	-1.551*
(region, education specific)	(0.946)	(0.769)	(1.337)	(0.933)	(0.509)	(0.763)
Constant	0.121	0.428*	-0.238	0.046	0.233	1.360*
	(0.191)	(0.166)	(0.327)	(0.556)	(0.372)	(0.640)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			0.687**	2.109	0.172**	1.257*
			(0.225)	(1.545)	(0.059)	(0.536)
ρ_1				0.817		1.207**
				(0.465)		(0.223)
ρ_2			-0.309**			0.897
			(0.117)			(0.619)
ρ_3				1.146	-0.170	
				(0.636)	(0.222)	
ρ_4				-1.292*		0.547
				(0.515)		(0.570)
ρ_5				0.198		-1.692
				(1.638)		(0.917)
ρ_6				0.209		0.639*
				(0.547)		(0.316)
R^2	0.30	0.32				
N	1,571	1,822	1,571	1,571	1,822	1,822

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

Table 19: Hourly wages. Men in rural areas.

	OLS		Selection			
	Informal	Formal	Informal		Formal	
Control function			Lee	DMF	Lee	DMF
Years of education	0.017**	0.044**	0.016*	-0.009	0.046**	0.053**
	(0.004)	(0.004)	(0.008)	(0.016)	(0.002)	(0.014)
Age	0.037**	0.036**	0.037**	0.029**	0.037**	0.041**
	(0.004)	(0.005)	(0.005)	(0.010)	(0.003)	(0.006)
Age^2	-0.000**	-0.000**	-0.000**	-0.000**	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Tenure in years	0.003	0.009	0.003	0.002	0.009	0.009
	(0.015)	(0.006)	(0.010)	(0.013)	(0.010)	(0.008)
Tenure^2	0.000	-0.001	0.000	0.000	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black	-0.029	-0.049**	-0.029	-0.028	-0.049**	-0.049**
	(0.023)	(0.015)	(0.016)	(0.026)	(0.011)	(0.014)
Unemployment rate	0.077	-0.956	0.122	0.968	-0.943*	-0.601
(region, education specific)	(0.576)	(0.665)	(0.484)	(0.909)	(0.475)	(0.515)
Constant	0.184*	0.484**	0.150	0.163	0.374**	0.280
	(0.083)	(0.132)	(0.166)	(0.233)	(0.069)	(0.428)
Sector dummies	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes
σ^2			0.254**	0.450	0.168**	0.777**
			(0.047)	(0.315)	(0.008)	(0.241)
ρ_1				-0.490		1.544**
				(0.917)		(0.425)
ρ_2			-0.067			-0.099
			(0.213)			(0.378)
ρ_3				-0.577	-0.149**	
				(0.342)	(0.055)	
ρ_4				-0.044		0.890**
				(0.682)		(0.268)
ρ_5				-1.907**		-0.678
				(0.633)		(0.682)
ρ_6				0.318		0.246
				(0.315)		(0.275)
R^2	0.25	0.28				
N	4,196	3,737	.	.	.	.

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis. Clustered s.e. in OLS regressions. Bootstrap estimates of the s.e. when controlling for selection to account for the two-step procedure.

Table 20: Gender wage gap in the informal and formal sectors, rural areas

Level of Education	**All**	**Primary**	**Secondary**	**Tertiary**
Raw wage gap from equation (1)				
Informal	0.095**	0.120*	0.156*	0.059
	(0.018)	(0.023)	(0.028)	(0.165)
Formal	-0.064*	0.027	0.073*	0.015
	(0.028)	(0.030)	(0.031)	(0.039)
Controlling for observables from equation (3)				
Informal	0.268**	0.283**	0.229**	0.227
	(0.044)	(0.038)	(0.058)	(0.168)
Formal	0.099**	0.037	0.115**	0.098*
	(0.020)	(0.026)	(0.024)	(0.045)
Controlling for observables and self-selection from equation (6)				
Lee				
Informal	0.408*	0.589	0.531	.
	(0.208)	(0.309)	(0.566)	.
Formal	0.114	0.100	-0.121	.
	(0.154)	(0.210)	(0.160)	.
DMF				
Informal	0.840*	1.292**	0.901	.
	(0.335)	(0.484)	(0.572)	.
Formal	-0.149	0.116	-0.348	.
	(0.251)	(0.508)	(0.226)	.
Number of men, Informal	4,198	3,025	1,135	38
Number of women, Informal	1,571	831	688	52
Number of men, Formal	3,983	1,875	1,525	231
Number of women, Formal	2,191	505	958	489

Notes: * $p < 0.05$; ** $p < 0.01$ Standard errors in parenthesis.