Decomposing the Gender Wage Gap with Sample Selection Adjustment: Evidence from Colombia

Decomposing the Gender Wage Gap with Sample Selection Adjustment: Evidence from Colombia[1]

Alejandro Badel
badel@stls.frb.org
Federal Reserve Bank of St. Louis

Ximena Peña
xpena@uniandes.edu.co
Universidad de Los Andes

October, 2010

Abstract

Despite the remarkable improvement of female labor market characteristics, a sizeable gender wage gap exists in Colombia. We employ quantile regression techniques to examine the degree to which current small differences in the distribution of observable characteristics can explain the gender gap. We find that the gap is largely explained by gender differences in the rewards to labor market characteristics and not by differences in the distribution of characteristics. We claim that Colombian women experience both a "glass ceiling effect'' and also (what we call) a "quicksand floor effect" because gender differences in returns to characteristics primarily affect women at the top and the bottom of the distribution. Also, self selection into the labor force is crucial for gender gaps: if all women participated in the labor force, the observed gap would be roughly 50% larger at all quantiles.

Keywords: Gender gap, Mincer, wage, semiparametric, quantile regression, selection.
JEL classification numbers: C21, J22, J31.

[1]The authors thank James Albrecht, Susan Vroman and Frank Vella for helpful conversations, this volume's editors Hugo Ñopo and Marcela Perticará for their suggestions, seminar participants at the Universidad de Los Andes, Universidad del Rosario, NIP-Colombia, LACEA 2009 and BIARI 2009 for their comments, and Christopher Martinek, Liliana Olarte, Natalia Perdomo and Daniel Wills for able research assistance. All remaining errors and omissions are our own.

I. Introduction

Between 1976 and 2006, Colombian women increased their labor market participation from 30% to 60%. Women also improved the quality of their observable labor market characteristics and penetrated occupations and sectors previously reserved to men. For example, women surpassed men in college attainment (Peña, 2006) and the fraction of women in the industry sector increased from 29% to 37%. In spite of these improvements, male and female hourly wages are still dramatically different – the mean gender wage gap was 14% in 2006.[2]

In this paper we employ quantile regression techniques to study the gender gap. Our aim is to analyze gender gaps defined not just as the differences between the means of male and female wages but across all quantiles of these wage distributions. The Machado Mata (MM hereafter) decomposition technique is used to determine, at every percentile of the wage distribution, the portion of the gap due to gender differences in labor market characteristics such as education and age -composition effect- and the portion due to gender differences in the rewards to these characteristics -price effect.

Our data comes from the Colombian Household Survey (CHS), a repeated cross-section carried out by the Statistics Department. CHS collects information on demographic and socioeconomic characteristics. We use the June 2006 wave.

Our main finding says that men get paid significantly more than women after controlling for observable factors and the gap displays a U-shape: women's wages lie further below men's at the extremes of the distribution than around the middle of the distribution. This result complements the international literature on gender gap distributions. De la Rica, Dolado and Llorens (2006) also find a U shaped gap in data from Spain, while Albrecht Van Vuuren and Vroman (2009) find that, in the Netherlands, the gap is larger at the top of the distribution. A substantial gender gap at the top quantiles of the wage distribution is commonly referred to as a "glass ceiling effect". We refer to the sizable gap observed at the bottom of the distribution in Spain and Colombia as a "quicksand floor effect".

Several papers decompose the gender wage gap across the distribution for different countries. [3] However, only Albrecht, Van Vuuren and Vroman (2009) control for sample selection bias in the estimation of female wage regressions. Since female participation is far from universal in Colombia this is an important issue. We follow the extension of MM proposed by Albrecht et al. (2009) in order to account for sample selection. Their approach uses the two-stage sample selection correction procedure introduced by Buchinsky (1998). This procedure combines a semiparametric binary model for the participation equation with a linear quantile regression model for the wage equation.

[2] The conditional gap, that is, the gender mean wage gap after controlling for labor market characteristics, was 11,4% in 2006.

[3] See Albrecht et al. (2003) for Sweden, de la Rica et al. (2007) for Spain, Hoyos, Ñopo and Peña (2010) for Colombia, Ganguli and Terrell (2005) for Ukraine, and Ñopo (2006) and Fernández, (2006) for Chile.

Our second finding says that despite having one of the highest female labor participation rates in Latin America, self selection of women into work is important in the Colombian case. If all women participated in the labor force, the observed gender wage gap would be larger by roughly 50% at all quantiles. Finally, we find that the bulk of the gender gap (both before and after controlling for selection) is explained by differences in the returns to characteristics. This is similar to what has been found in other countries (see, for example AVV).

The next section describes the data used and presents descriptive statistics. Section III describes the metholodogy used, and results are presented in Section IV, both before and after controlling for sample selection. Sections V concludes, discusses policy implications of our results and directions for future work.

II. Descriptive Statistics and Data

We use the Colombian Household Survey (CHS), a repeated cross-section carried out by the Statistics Department. CHS collects information on demographic and socioeconomic characteristics such as gender, age, marital status and educational attainment, as well as labor market variables for the population aged 12 or more including occupation, job type, income and sector of employment. We use the June 2006 wave to analyze the evolution of the raw gap and then focus on the latter wave to perform the selection correction and decomposition exercises.

Our analysis focuses on the seven main cities which account for 60% of the urban population, and according to 2005 Census data 78% of Colombians live in urban areas[4]. In the 7 main cities 93% of men between 25 and 55 years of age work, while only 69% of women do. When we compare Bogotá and the other cities, we find that even though the levels of male participation are comparable, women participation is significantly higher in Bogotá: 75% vs. 65%.

We use only observations with a complete set of covariates and restrict our sample to prime-aged individuals (between 25 and 55 years of age) who report working between 16 and 84 hours per week[5] and earn more than one dollar per day. Table 1 shows the sample selection for 2006, which retains 15,423 observations, equivalent to nearly 4 million using weights, 47% of which are female. Our sample selection criteria seek to minimize measurement error in the log hourly wage.

---

[4] Bogotá accounts for 45% of the population in the 7 main cities but given the design of the CHS, the sample size corresponds to only 15%. Sample weights are used to get representative results. Instead of extending the MM methodology to include sample weights we perform calculations for Bogotá and Elsewhere separately, and then we build the weighted distribution as follows:
a) Let q_i be the percetiles of the log wage distributions for $i=\{Bogotá, Elsewhere\}$.
b) Calculate at the j distribution the percentile levels at which q_i lies and call these P_i. E.g. $P_{Bog} = F_{bog}(q_{else})$.
c) The percentiles q_{else} correspond to the $Pr(z=bog)*(P_{Bog})+(1-Pr(z=bog))*(0.01,0.02,0.03...0.99)$ percentile levels of the country distribution.
d) Obtain the country percentiles by linear interpolation.

[5] The legally defined full time work is 48 hours per week in Colombia.

Table 1: Sample Selection, April-June 2006

	No. Observations	Weighted	% Men
7 main cities, 12+ years	46.439	14.200.850	0,44
Ages 25 to 55 years…	23.915	6.047.089	0,43
who work,	16.513	4.302.923	0,51
report 16-84 hours per week	15.563	4.012.872	0,52
and earn more than US$1 per day.	15.423	3.978.580	0,52

In addition to the differences in participation rates, men and women also display differences in hours worked per month. Even though in our sample both have median hours of 208, men work on average 220 hours per month while women work 197 hours.

The dependent variable is log hourly wage. The explanatory variables included in the estimations are: age and its square[6], 4 education categories[7], and dummies for marital status[8] and head of household.

A fundamental right hand side variable in Mincerian wage regressions is experience. However, most datasets employed in the empirical labor literature do not contain direct measures of experience. Our study is not exempt from this problem. Many studies employ "potential experience" which is a proxy for experience constructed from education and age data.[9] It is well known that "potential experience" typically overstates actual experience for women (see Altonji and Blank (1999) for a complete discussion).

The problem is clearly illustrated by Perticará (2007) and Perticará and Bueno (2009) in the literature that studies Latin American labor markets. These papers analyze gender gaps in Chile controlling for experience. Chilean data displays substantial differences between potential and effective experience. Specifically, "potential experience'' overestimates effective experience for women, due to lower participation of women over their life cycle or, as these papers put it, due to gender differences in the "timing" of experience acquisition. These papers show that overestimating female experience implies a downward bias in estimates of the part of the raw gender gap that can be accounted for by returns to experience.

A favored alternative when the preferred dataset does not contain measured of experience is to use "predicted experience" calculated using predictions from a regression model of experience. The model is estimated using an auxiliary dataset. We cannot use this alternative since, to our knowledge, there are no auxiliary data sets with "actual experience" available in Colombia.

[6] There is no available information in the survey regarding work experience, nor information about the number of births per woman -this is only identifiable for the head of household or spouse.

[7] The education groups are: no completed education, completed primary, completed secondary and completed tertiary.

[8] We summarize the marital status information into two categories: 'together' including individuals married or cohabiting which we refer to as married, and 'alone' which includes the categories single, divorced, separated and widowed.

[9] The typical proxy is constructed as *potential experience=age-years of education-6*.

We include education and a polynomial in age in order to proxy for experience. Our specification is designed to mitigate the downward bias in female "potential experience" by allowing the effect of age and education on experience to vary by gender. This flexibility comes at a cost: we cannot separately identify the effect of age and education from the effect of experience in labor market returns. The cost turns to be low here since this distinction is not important for our analysis.

The descriptive Statistics are summarized in Table (2). First, men earn higher mean hourly wages than women: the average log wage for men is 7.86 and 7.72 for women. There are sizeable differences between the traditional labor market characteristics of working and non-working women, which is suggestive of non-random selection into work. The distribution of age and schooling is very similar between men and working women. Working men and women have similar average age, whereas non-working women are nearly 2 years older. Working women are the most educated, followed by men and finally non-working women; the education distribution of working women first-order stochastically dominates that of working men which in turn first order stochastically dominates that of non-working women. Working men and non-working women display similar proportions of married individuals, 69% and 67% respectively, whereas only 48% of working women report being married. Males are more often head of household than females: 69% of men are head of household, while only 30% of working women and 17% of non-working women are.

Table 2. Descriptive Statistics, Wage Equation

	Men	Women	
	Working	**Working**	**Not Working**
Log Wage	7,86	7,72	
	(0,76)	(0,82)	
Age	38,33	38,01	39,93
	(8,57)	(8,34)	(9,17)
Education			
< Primary	0,07	0,07	0,10
Primary +	0,34	0,31	0,39
Secondary +	0,41	0,40	0,40
University	0,18	0,22	0,10
Married	0,69	0,48	0,67
Head of Household	0,69	0,30	0,17
Bogotá	0,43	0,47	0,33
Home Ownership	0,49	0,52	0,57
# Children 2-6yrs			
2	0,18	0,13	0,15
1	0,03	0,02	0,02
# Children <1yr	0,04	0,02	0,02
Log Non-Earned Income	12,28	11,95	12,33
	(1,35)	(1,31)	(1,40)
Log Other Family Income	13,43	13,77	13,67
	(1,18)	(1,12)	(1,00)
No. Obs	8.368	7.055	5.670

Note: Standard errors in parentheses.

To gain identification, some variables included in the selection equation must be excluded from the Mincer wage equation. We exclude home ownership, number of children between 2 and 6 years of age, presence of children under 1, personal non-earned income (NEI) and other family income (OFI). The selection equation is calculated only for women. Again, working and non-working women have different sets of characteristics regarding these variables. Home Ownership is a dichotomous variable indicating whether the person owns the house they inhabit. A higher proportion of women not working tend to be home-owners as compared to those who work: 57% vs. 52%, respectively. A smaller percentage of working women has children: twice as many women not working have children under 1 as compared to women working, and there is a slightly higher fraction of non-working women with children between 2 and 6 years of age.

In principle, even the small rate of male non participation observed in our sample could imply that Mincer regressions for men are also subject to sample selection bias. In fact, it is well known that labor force participation rates complicate analyzing wage differences between black and white men in the US literature. Butler and Heckman (1977) are the first to point out that relative wages amongst blacks are overstated because the generosity of transfer programs leads to selective withdrawal of the least-skilled blacks from the labor force. However, apart from this example, the potential sample selection bias for men has been ignored in most of the literature. We follow the literature for two reasons. First, we want to be comparable with Albrecht, Van Vuuren and Vroman (2009) which is the only study employing the same methodology as ours. Second, we want to avoid the extreme lack of precision in wage gap estimates that would likely arise if we tried to control for selection by estimating our semi-parametric selection equation from a sample where only a tiny fraction of workers opts out of the labor force.

The next variable, NEI, is defined as income not related to labor market activities: accrued interest rates, rentals, pensions, remittances and other concepts. A low percentage of women report positive NEI: 19% of women who do not work report positive NEI, while 14% of working women do. Of those who report strictly positive NEI, women not working report higher levels on average than those working. Finally, OFI is defined as the total household income minus the individual's total income. Not surprisingly, since a higher proportion is married, a higher percentage of non-working women report strictly positive levels vis-à-vis working ones, 87% vs. 80%. However, the average OFI for working women is higher than for non-working women.

III. Methodology

Our aim is to analyze gender gaps defined not just as the differences between the means of male and female wages but across all quantiles of these wage distributions. Concretely, let random variables w^M and w^F denote male and female wages, respectively. Let θ be a number between 0 and 1 that indicates the quantile (e.g. $\theta = 0.5$ tells us we are looking at the median of the distribution). Finally, let $Q_\theta(x)$ denote the $\theta - th$ quantile of the

distribution of random variable x. We can thus express the raw gender gap between the $\theta - th$ quantiles of w^M and w^F as $Q_\theta(w^M) - Q_\theta(w^F)$.

Roughly speaking, the Machado Mata (MM) technique, described below, allows us to produce counterfactual wage distributions, say w^C which we can compare to actual distributions. For example, we can compute, in a way explained below, the wage distribution that would prevail if women had the same characteristics of men, but were still "paid as women". Denote this hypothetical distribution by w^C. Then we can decompose the raw gap into two terms:

$$Q_\theta(w^M) - Q_\theta(w^F) = [Q_\theta(w^M) - Q_\theta(w^C)] + [Q_\theta(w^C) - Q_\theta(w^F)]$$

The first term in brackets is interpreted as the contribution of differences in how men and women are paid (or differences in returns). Recall that in w^M and w^C male and female workers have the same distribution of characteristics. The second term in brackets is interpreted as the contribution of differences in characteristics across men and women. Recall that in w^C and w^F women are "paid like women". The next section describes how the Machado Mata technique allows us to produce counterfactual distributions w^C.

Machado-Mata Technique

The MM technique is based on linear Quantile Regressions. These can be described as follows. Suppose that for each individual *i* in either population of females, *F*, or males, *M*, we observe the log wage w_i and a vector of covariates x_i. Further, assume that for each population *j=M,F*, the conditional θ-quantile of w_i^j, conditional on the set of covariates x_i^j, is given by $Q_\theta(w_i^j) = x_i^j \beta_\theta^j$. Then we can define an error term as $e_{\theta_i}^j = y_i^j - x_i^j \beta_{\theta_i}^j$, where $e_{\theta_i}^j$ is a random disturbance that satisfies $Q(e_{\theta_i}^j) = 0$ by construction. The linear QR model for population *j* is thus $w_i^j = x_i^j \beta_\theta^j + e_{\theta_i}^j$. The standard techniques to estimate β_θ^j can be found in Greene (2002). Our next section reviews an extension to these techniques that allows consistent estimation in the presence of nonrandom sample selection issues.

The distribution of w conditional on x is fully described by $Q_\theta(w \mid x)$ in the way ordinary sample quantiles fully characterize any given distribution. Hence, realizations of w_i given x_i can be interpreted as independent draws from $Q_\theta(w_i \mid x_i)$ where θ_i is a uniform random variable in [0,1]. This property of Q_θ suggests a way to obtain a random sample of the (estimated) distribution of wages conditional on x, which is described below.

First, one draws a large number (say K=1000) of values of θ from a uniform distribution in [0,1]. Second, for each draw θ_k, one estimates the corresponding QR coefficient vector in population F, denoted $\hat{\beta}_k^F$. Third, one obtains a random sample of size 1000 from the sample of covariates of population M. Denote each of these 1000 samples by x_k^M The samples are indexed by k since there is a random draw of x^F per each draw θ_k. Finally the counterfactual distribution is constructed as $w_k^{FM} = x_k^M \beta_k^F$.[10]

The distribution w^{FM} can, given our notation, be interpreted as the counterfactual distribution of female log wages that would prevail if we maintain the returns to observable characteristics of women β_θ^F, but endow women with the male distribution of labor market characteristics x^F. Theoretically w^{FM} is given by:

$$w^{FM} = \beta_\theta^F x^M + e^{FM} \quad \text{with} \quad Q_\theta\left(e^{FM}\right) = 0. \qquad (1)$$

Machado and Mata (2005) proves that the four-step sampling procedure described above gives a consistent estimator of the counterfactual distribution w^{FM}. This follows because the quantiles of the empirical distribution w_k^{FM} converge in probability to the theoretical quantiles of w^{FM}.

Estimating a quantile regression of Female Wages

A consistent estimator of $\hat{\beta}_\theta^F$ is vital to build counterfactual distributions. However, given that women self-select into work, the usual problem of sample selection bias applies to its estimation. If, for example, the fraction of women actually participating is higher at the top of the potential work distribution, observed data under-samples the low potential earners and oversamples the high potential ones. Therefore, we need to correct for selection in a QR framework.

We estimate the Mincer equation for working women correcting for selection where each quantile is given by:

$$Q_\theta\left(w \mid \cdot\right) = \mu_\theta + x^F \beta_\theta^F + P_\theta\left(z\gamma\right) + e_\theta \quad \text{for} \quad \theta \in (0,1). \qquad (2)$$

Where x^F is a vector of labor market characteristics, z^F is the set of observables that influence the participation decision[11], and P is the probability of participating. The term

[10] The probability integral transformation states that if U is a uniform random variable on [0,1], the $F^{-1}(U)$ has the density F.

[11] An important assumption is that x_i is a subvector of z_i, and that z_i includes at least one continuous variable not present in x_i. The particular exclusion restrictions in our application are made explicit in the data section.

$P_\theta(z\gamma)$ adjusts for selection at each quantile $\theta \in [0,1]$. Once β_θ^F has been consistently estimated, the MM procedure is conducted as described above.

We follow Buchinsky (1998) to account for selection in a QR framework[12]. This procedure shares the spirit of the popular Heckman (1978) two-step selection correction model but differs from Heckman in two important ways. First, quantiles, as opposed to mean regressions, are considered. Second, normality and homoskedasticity in the selection model are not assumed. Therefore, while in Heckman's the selection bias term takes the usual `inverse Mills ratio' form, in Buchinsky (1998) the form of the selection bias term is unknown.

The model is summarized as follows. Let y_i be a participation dummy and *G* an unknown function of the single index $z_i\gamma$. The probability of participating is given by

$$P(y_i = 1) = G(z_i\gamma) \quad \text{for} \quad i = 1,...,N. \qquad (3)$$

We then construct the flexible selection correction term, *P(·)*, as a polynomial of the index,

$$P(z_i\gamma) = \lambda_{\theta,0} + \lambda_{\theta,1}\, r(a + b(z_i\gamma)) + \lambda_{\theta,2}\, r(a + b(z_i\gamma))^2 + ... + \lambda_{\theta,q}\, r(a + b(z_i\gamma))^q, \qquad (4)$$

where *a* and *b* are location and scale parameters, and *r(·)* denotes the inverse mills ratio $r(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$ evaluated at $a + b(z_i\gamma)$. A key point here is that the λ's vary with θ. We separate the location and scale parameters from the index since these are not identified in the semiparametric single-index framework[13]. Following Buchinsky, a Hausman specification test is used. We test the null hypothesis of normal errors, given the existence of the single index estimator which is consistent under both null and alternative hypotheses[14]. Probit should be used in the first step of the selection correction when errors are normally distributed; the single-index estimator should be used otherwise[15].

---

[13] To see this, note that for any pair *(a,b)* and a function $G(a + b(z_i\gamma))$ there is a function $\hat{G}(z_i\gamma)$ such that $G(a + b(z_i\gamma)) = \hat{G}(z_i\gamma)$ for all z_i. Following Buchinsky, we estimate a and b by running a probit regression of y_i on the semiparametrically estimated index $z_i\gamma$.

[14] The Hausman Test is perfomed using Klein and Spady's (1993) estimator. Under the null hypothesis of normally distributed errors, $(d_I - d_p)'(V_I - V_p)^{-1}(d_I - d_p) \sim \chi^2(d_f)$ where for *i={single index, probit}*, d_i are the estimates, V_i the covariance matrices and $d_f = \dim(d_i)$. The delta method is used to compute the covariance matrix of the probit estimates.

[15] While Buchinsky (1998) and AVV (2007) use the Ichimura single-index estimator, we employ the quasi-maximum likelihood estimator of Klein Spady (1993). The latter is superior since it achieves the semiparametric efficiency bound of Chamberlain and Cosslet.

Last, note that μ_θ and $\lambda_{\theta,0}$ are not separately identified in the quantile regression model above. We follow Buchinsky and estimate them by the method proposed by Andrews and Schafgans (1998) of identification at infinity. The intuition is as follows: if we choose a subsample of women with labor market characteristics such that the probability of working given those characteristics is arbitrarily close to 1, we can use this subsample to estimate the intercept in the Mincer equation, μ_θ, without adjusting for selection.

Clearly, de variance-covariance matrix for the MM procedure has to account for the variability of the selection correction estimates. AVV prove asymptotic normality of the MM quantiles in this context, and extend the covariance matrix estimator in Buchinsky (1998) for quantile regression with selection correction to the MM quantiles. We employ their formulation.

IV. Results

Raw Gap Decompositions, without Selection Correction

As a starting point, we estimate Quantile Regression (QR) wage equations using standard off-the-shelf methods.[16] Log hourly wage is regressed on the specified set of covariates. Results are reported in Tables (3) and (4) for women and men, respectively. Included variables have the expected sign. Education has a monotonic effect on wages; since the left out category is Completed Primary, lower levels negatively affect wages whereas higher ones have a positive effect. Age is sometimes significant while Age squared appears not be. Being married or head of household positively affects wages.

Table 3: Mincer Equation, Women

	Bogota				Elsewhere			
	20%	**40%**	**60%**	**80%**	**20%**	**40%**	**60%**	**80%**
Constant	6,62***	7,18***	7,33***	7,2***	5,82***	6,64***	7,13***	7,12***
	(0,95)	(0,38)	(0,32)	(0,43)	(0,34)	(0,19)	(0,18)	(0,23)
Age	0,01	0	0	0,01	0,04***	0,02***	0	0,01
	(0,05)	(0,02)	(0,02)	(0,02)	(0,02)	(0,01)	(0,01)	(0,01)
Age squared	0	0	0	0	0	0	0	0
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Married	0,21***	0,09***	0,11***	0,17***	0,11***	0,04***	0,05***	0,12***
	(0,07)	(0,05)	(0,03)	(0,05)	(0,03)	(0,02)	(0,02)	(0,02)
Head	0,25***	0,11***	0,12***	0,18***	0,19***	0,09***	0,1***	0,19***
	(0,08)	(0,05)	(0,03)	(0,06)	(0,04)	(0,02)	(0,02)	(0,02)
Education								
<Primary	-0,58***	-0,19*	-0,23***	-0,19***	-0,22***	-0,32***	-0,32***	-0,27***
	(0,23)	(0,12)	(0,07)	(0,06)	(0,06)	(0,04)	(0,03)	(0,04)
Secondary	0,34***	0,28***	0,27***	0,50***	0,55***	0,48***	0,36***	0,45***
	(0,08)	(0,06)	(0,03)	(0,06)	(0,04)	(0,02)	(0,02)	(0,02)
College	1,18***	1,14***	1,35***	1,59***	1,38***	1,27***	1,29***	1,48***
	(0,08)	(0,10)	(0,05)	(0,07)	(0,04)	(0,03)	(0,03)	(0,03)

*** Significant to 99%, ** Significant to 95%, *Significant to 90%

[16] For a description of standard QR estimation techniques, see Greene (2002).

Table 4: Mincer Equation, Men								
	Bogota				Elsewhere			
	20%	40%	60%	80%	20%	40%	60%	80%
Constant	7,09***	7,7***	7,36***	6,99***	6,74***	6,91***	6,83***	6,76***
	(0,60)	(0,40)	(0,35)	(0,49)	(0,17)	(0,14)	(0,13)	(0,21)
Age	0	-0,02	0	0,03	0,01	0,02	0,03***	0,04***
	(0,03)	(0,02)	(0,02)	(0,03)	(0,01)	(0,00)	(0,01)	(0,01)
Age squared	0	0	0	0	0	0	0	0
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Married	0,09	0,06	0,05	-0,06	0,1***	0,01	0	0,02
	(0,07)	(0,04)	(0,04)	(0,06)	(0,02)	(0,02)	(0,02)	(0,03)
Head	0,07	0,09	0,10***	0,20***	0,13***	0,12***	0,16***	0,18***
	(0,08)	(0,05)	(0,04)	(0,06)	(0,03)	(0,01)	(0,02)	(0,03)
Education								
<Primary	-0,28**	-0,34***	-0,25***	-0,22**	-0,28***	-0,25***	-0,18***	-0,23***
	(0,16)	(0,09)	(0,08)	(0,13)	(0,04)	(0,02)	(0,02)	(0,03)
Secondary	0,32***	0,24***	0,33***	0,47***	0,32***	0,27***	0,29***	0,40***
	(0,06)	(0,03)	(0,04)	(0,06)	(0,02)	(0,02)	(0,01)	(0,02)
College	1,01***	1,25***	1,52***	1,69***	1,00***	1,12***	1,27***	1,47***
	(0,11)	(0,09)	(0,05)	(0,08)	(0,03)	(0,02)	(0,03)	(0,04)

*** Significant to 99%, ** Significant to 95%, *Significant to 90%

We now study the gender gap from raw data, before conditioning on covariates (such as age and education) and before accounting for selection of women into the labor market. The raw gap is the difference between the log wage of a male at a specific quantile of their distribution and the log wage of a female at the same quantile of the female distribution. A gap of, say, 0.4 at the *i*-th percentile is interpreted as one group having a log-wage 40% higher than the other at that percentile.

To characterize the evolution of the gender wage gap along the distribution of wages, Figure (1) displays the raw gender gap for 2006. Several features are worth mentioning. First, male and female wages are extremely unequal, and men are always paid significantly more than women. Second, the gender gap displays a U-shape, that is, women's wages fall behind men's more at the extremes of the distribution whereas they are closer near the median.

De la Rica, Dolado and Llorens (2006) report a similar non-monotonicity in Spain, due to a composition effect: the gap for high education workers increases along the distribution while that of low education ones decreases. This is not the case in Colombian data, for any of the studied covariates. Rather, the minimum wage may be behind the lowers levels of the wage gap in the middle of the distribution. Since the people at the middle of the distribution earn around the minimum, the minimum wage may compress the gender gap intermediate wage quantiles.

In our sample, the median wage is 1,1 times the minimum wage, and approximately 40% of workers earn wages less than, or equal to the minimum wage. The minimum wage effect varies along the income distribution. It does not affect the wages of people earning less than the minimum, usually informal, unprotected workers. It is very binding at and around the level of the minimum wage, and it loses its grip as we move along the income distribution towards the high earners (Cunningham, 2007). Hence, even though the

minimum indexes the wage distribution, its evolution has little effect on very highly paid workers.

The raw gender gap is around 35% of the male log-wage at low levels of the distribution, near the median it is close to zero, and it increases towards the upper tail of the distribution gap to a maximum log wage difference of about 30%. Recall that a log-wage gap of 35% is equivalent to a 42% gap in the wage level. Finally, even though the gap increases in the second half of the distribution, the main increase is observed at the richest decile: at the 90th percentile the gap is around 10% and it increases to about 30% at the 99th percentile. However, given that there are higher standard deviations of the distribution of wages at either extreme, the results are measured less precisely.

The QR framework allows us to observe the variation across the distribution hidden behind means analysis. In Colombia, the gender gap is higher for women at the top and bottom of the distribution of log-wages. Since the gap widens at the top of the distribution, there is a *glass ceiling* effect suggesting a barrier to further advancement of women once they have attained a certain level. Albrecht et al.(2003) finds that the raw gap in Sweden increases to 40% around the top of the distribution. We find lower gaps at the top of the distribution for Colombia. Since the wage gap also widens at the bottom of the distribution, we calim there is a "quicksand floor effect" suggesting a barrier for the advancement of women with meager labor market characteristics.

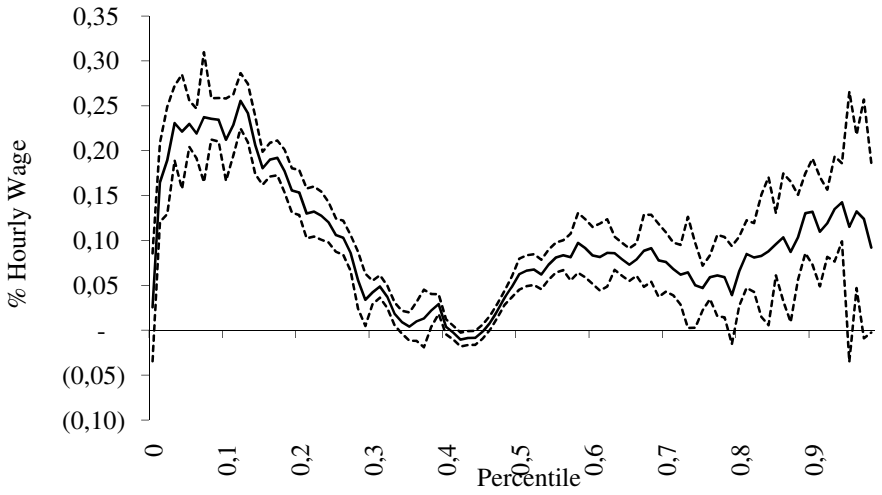
Figure 1. Raw Gender Gap 2006

Note: The solid line is the raw gap while the dashed lines are the 95% confidence intervals

Do observables explain the Colombian gender wage gap? Using the MM technique we can decompose the gap in Figure (1) into a component generated by differences in labor market characteristics, composition effect, and a component due to differences in the returns to these characteristics, price effect. We build the counterfactual distribution of men wages given their characteristics but paid the level of female returns. The difference

in the characteristics between men and women is accounted for by taking the difference between the observed male distribution and the proposed counterfactual distribution of `men paid as women'.

Figure 2. Price Effect



Note: The solid line is the raw gap while the dashed lines are the 95% confidence intervals

Figure 2 displays the price effect, that is, the gender gap that remains after we purge the effect of differences in labor market characteristics. The result is quantitatively similar to the raw gap in Figure (1), except at the extremes. In the bottom 10% of the distribution, the remaining gap after controlling for observables is nearly 25%, 10 percentage points lower than the raw gap in Figure (1). At the top 5% of the distribution, the difference in characteristics explains the steep increase of the raw gap between roughly 15% and 30%. Therefore observable characteristics only account for some of the observed gender wage gap at the extremes of the income distribution. The price effect accounts for most of the gender gap between the 10^{th} and 95^{th} percentiles. This result is in line with results from other studies (see for example Albrecht, Van Vuuren and Vroman, 2009).

To confirm the previous result, we also calculated the composition effect: the difference between the observed male distribution and the (counterfactual) distribution of women's wages that would have prevailed if women retained their labor market characteristics but were paid for them as men: the `women paid as men' distribution[17]. The previous conclusion is confirmed by the additional exercise: the price effect accounts for most of the raw gender gap.

[17] Results available from authors upon request.

Wage Gap Decompositions, Controlling for Selection[18]

While male participation rates are very high -approximately universal, the proportion of working women is smaller. In addition, working and non-working women differ in labor market characteristics such as age and schooling. This suggests that selection bias is an issue in this estimation since women select into the labor force in a non-random way.

We begin by estimating QR wage equations where the log hourly wage is regressed on the specified set of covariates, after controlling for selection. Because we only control for selection for women, we present the adjusted coefficients only for this group, in Table (5) As before, log hourly wage is regressed on the specified set of covariates. The included variables have the expected sign, and the results are comparable to coefficients before accounting for selection, reported in Table (3). Education retains its explanatory power and monotonic effect on wages, even though primary seems insignificant for most of the distribution of cities other than Bogotá. After adjusting for selection, being married or the head of household are significant only for some percentiles.

Table 5: Mincer Equation after Controlling for Selection, Women

	Bogota				Elsewhere			
	20%	**40%**	**60%**	**80%**	**20%**	**40%**	**60%**	**80%**
Constant	7,86**	6,96***	7,87***	8,82***	5,77***	5,53***	5,66***	5,44***
	(3,45)	(2,20)	(1,48)	(2,72)	(1,33)	(1,03)	(0,96)	(1,04)
Age	0,05	0,02	0,01	0,04	0,05	0,03	0,02	0,04
	(0,06)	(0,04)	(0,03)	(0,05)	(0,10)	(0,09)	(0,08)	(0,07)
Age squared	7E-04	2E-04	6E-06	3E-04	-6E-04	4E-04	2E-04	3E-04
	(7E-04)	(5E-04)	(4E-04)	(6E-04)	(1E-03)	(1E-03)	(1E-03)	(1E-03)
Education								
<Primary	-0,53***	-0,21*	-0,23***	-0,22***	-0,20	-0,30	-0,32	-0,26***
	(0,12)	(0,12)	(0,07)	(0,06)	(0,19)	(0,18)	(0,21)	(0,08)
Secondary	0,38***	0,26***	0,26***	0,51***	0,58***	0,58***	0,37***	0,49***
	(0,11)	(0,05)	(0,03)	(0,06)	(0,12)	(0,06)	(0,06)	(0,07)
College	1,27***	1,14***	1,35***	1,65***	1,47***	1,39***	1,39***	1,64***
	(0,11)	(0,08)	(0,06)	(0,09)	(0,20)	(0,16)	(0,12)	(0,18)
Married	0,10	0,07	0,09	0,11	0,08	0,01	0,00	0,05
	(0,08)	(0,05)	(0,06)	(0,08)	(0,09)	(0,05)	(0,08)	(0,10)
Head	0,26***	0,11	0,12**	0,21**	0,23**	0,14*	0,15**	0,26**
	(0,14)	(0,08)	(0,07)	(0,11)	(0,09)	(0,07)	(0,07)	(0,13)

*** Significant to 99%, ** Significant to 95%, *Significant to 90%

Stardard errors calculated using bootstrap, 100 repetitions.

What would the distribution of female wages be if all women worked? The MM procedure described above is used to build this counterfactual distribution. We generate a random sample of female wages using the female sample selection adjusted coefficients combined with the labor market characteristics of all women -not just those who work. Hence, in what follows `accounting for selection' refers to the use of this potential distribution of female wages.

[18] We are grateful to Albrecht, Van Vuuren and Vroman for making their code available. The standard errors reported in this section were calculated using their codes.

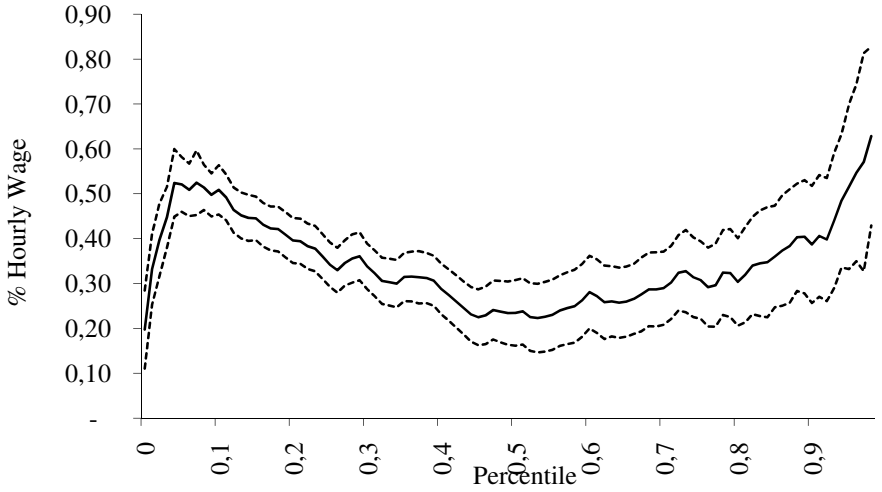
Table 6. Selection Equation				
	Bogota		Rest	
	Probit	Klein&Spady	Probit	Klein&Spady
Age squared	-1,172***	-1,22***	-1,103***	-1,087***
	(0,06)	(0,04)	(0,02)	(0,01)
< Primary	-0.01	0.00	-0,02*	-0.02
	(0,03)	(0,01)	(0,01)	(0,03)
Secondary+	0.00	0.03	0,12***	0,07***
	(0,03)	(0,03)	(0,02)	(0,02)
University	0,13***	0,17***	0,25***	0,18***
	(0,05)	(0,02)	(0,03)	(0,02)
Married	-0,133***	-0,287***	-0,173***	-0,143***
	(0,05)	(0,05)	(0,02)	(0,02)
Head	0,22***	0,23***	0,20***	0,12***
	(0,07)	(0,05)	(0,02)	(0,04)
# Children <1	-0.04	-0,09***	-0,04***	-0,023***
	(0,03)	(0,03)	(0,01)	(0,01)
# Children <6	-0,07***	-0,04***	-0,02***	-0,01***
	(0,04)	(0,02)	(0,01)	(0,02)
Home Ownership	-0,08***	-0,13***	-0,04***	-0,03***
	(0,04)	(0,03)	(0,01)	(0,01)
Non-Earned Income	-0,10***	-0,20***	-0,18***	-0,13***
	(0,04)	(0,04)	(0,02)	(0,02)
Other Family Income	0,06*	0,09***	-0,02**	0.02
	(0,03)	(0,02)	(0,01)	(0,03)
Hausman	Test	95% Critical Value	Test	95% Critical Value
	36.163	19.675	35.018	19.675
Note: All the coefficients are calculated relative to the absolute value of the coefficient of age. Standard errors in parentheses.				
*** Significant to 99%, ** Significant to 95%, *Significant to 90%				

The results of the estimation of the selection equation are presented in Table 6. Hausman test results suggest that for the Colombian data the single index (as opposed to the probit) estimator should be used in the first step of the correction model. After accounting for selection using the single index estimator proposed by Klein and Spady, we calculate the true gender gap (gender gap in what follows) as the difference between the male wage distribution and the potential women distribution. Figure (3) shows that the gender gap displays a U-shape, as did the raw gap. However, the level is significantly higher, especially at the upper-end of the distribution. The lowest wage gap is around 25% of log wages and it is observed in the middle of the distribution. Whereas the maximum levels recorded by the raw gap were around 35% in the lower end of the distribution and 30% in the upper end, the respective maxima for the gender gap are 50% and 60%. Recall that a log wage gap of 60% is equivalent to a wage gap in the level of wages of over 80%.

The gender gap increases substantially at the top tenth of the distribution, this time passing from 40% to 60%. AVV find that after accounting for selection the gender gap is

increasing and it reaches 40% at the top of the distribution in the Netherlands. According to our calculations, the glass ceiling in Colombia after adjusting for selection is steeper.

Figure 3. Gender Gap



Note: The solid line is the raw gap while the dashed lines are the 95% confidence intervals

Clearly, the selection correction is important and sizable. Given that working women are a selected sample of women, the raw gap underestimates the existing gender gap in the country. Note that the gender gap is equivalent to `adding up' the raw gap (Figure 1) and the selection effect (which will be discussed later and is portrayed in Figure 5).

Again, we perform a decomposition using the MM technique and accounting for selection. As in the previous section, we build the distribution of male wages that we would observe if they retained their characteristics but were disguised as women, and hence were paid the selection-adjusted returns of women: the `men paid as women' distribution. We then calculate the price effect after accounting for selection, that is, we substract this counterfactual distribution from the men wage distribution to purge the effect of differences in the distribution of observable characteristics between men and women (Figure 4). Note that roughly two-thirds of the wage gap, attributable to the price effect, remains after accounting for differences in characteristics.

Figure 4. Price effect, accounting for selection

Note: The solid line is the raw gap while the dashed lines are the 95% confidence intervals

Decomposing the Selection Term

Let us now turn to the direct effect of selection by characterizing the non-randomness of the participation decision of women. Are less able women forced to work because of need or, on the contrary, is there a positive selection and able women participate more? We already saw that in Colombia working women are younger and more educated than those who don't work. They are also less likely to own the house they live in than non-working women.

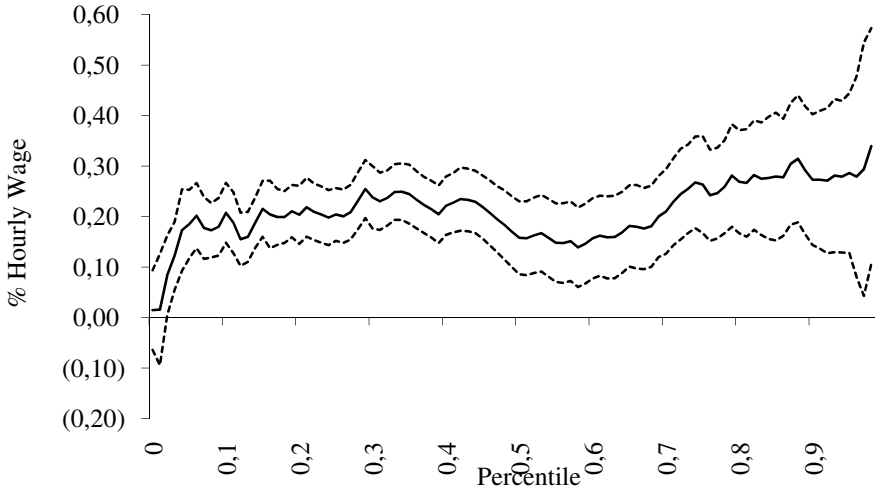
The selection effect is calculated as the difference between the distributions of observed and the potential women's wages. Specifically, the selection effect is given by

$$Q_\theta\left(w^F\right) - Q_\theta\left(w^{AF}\right)$$
$$w^{AF} = \beta_\theta^F x^{AF} + e^{AF} \quad \text{with} \quad Q_\theta\left(e^{AF}\right) = 0$$

Where w^F is the observed distribution of female wages and w^{AF} is a counterfactual distribution that assumes that working women have the labor market characteristics of all women and the estimated returns of women. Note that in this exercise β_θ^F needs to be estimated controlling for sample selection.

We find that the selection effect is positive and rather high in our application, around 20%, as shown in Figure 5. This is roughly twice the size of the effect AVV for the Netherlands. Able women, those with better observable and unobservable labor market characteristics, are pulled into the workforce by the high returns. Hence, the raw gap underestimates the true gender gap by roughly 20% -the selection effect- since women who actually work are those who would get the greatest return.

Figure 5. Selection Effect



Note: The solid line is the raw gap while the dashed lines are the 95% confidence intervals

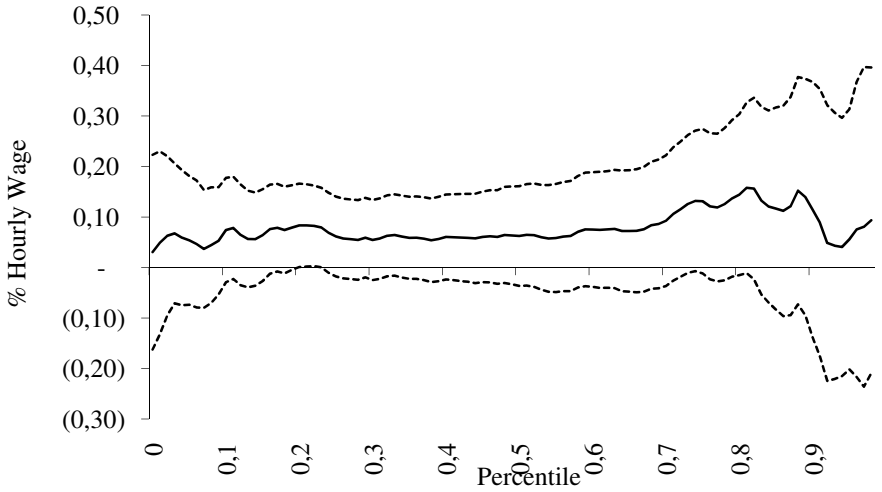
Selection is due both to differences in the labor market characteristics between women who work and those who don't and to unobserved characteristics. The MM methodology allows us to decompose selection effect into a portion due to observables labor market characteristics, and the remainder due to unobservables. In doing so we build another counterfactual distribution: the distribution of women's wages that would have prevailed if prices accounted for selection, but women had the distribution of labor market characteristics of working women -not of all women. Specifically, selection gap is decomposed as follows:

$$Q_\theta\left(w^F\right) - Q_\theta\left(w^{AF}\right) = \left[Q_\theta\left(w^F\right) - Q_\theta\left(w^{SF}\right)\right] - \left[Q_\theta\left(w^{AF}\right) - Q_\theta\left(w^{SF}\right)\right]$$

$$w^{SF} = \beta_\theta^F x^F + e^{SF} \quad \text{with} \quad Q_\theta\left(e^{SF}\right) = 0$$

Notice that, here too, β_θ^F is estimated controlling for sample selection bias. The first term in brackets is interpreted as the effect of unobservables, while the second term is interpreted as the effect of observables.

Figure 6. Selection due to observables



Note: The solid line is the raw gap while the dashed lines are the 95% confidence intervals

The difference between this `working women adjusting for selection' counterfactual and the potential distribution tells us how much of the selection effect can be explained by differences in the distribution of characteristics between women who work and those who don't (Figure 6). Observable characteristics not explain much of the selection effect.

The remainder of the selection effect is the attributed to unobservables. It is calculated as the difference between the actual distribution of female wages and the `working women adjusting for selection' counterfactual distribution. Unobservables account for roughly three quarters of the selection effect until the 70th percentile, and half in the top 30% of the distribution, and the effect is statistically significant..

Because the selection effect due to observables is not statistically strong, we conclude that selection in Colombia is mostly due to unobservables. This is in contrast with Albrecht, Van Vuuren and Vroman (2009) who find that most of the selection effect in the Netherlands is due to observables. This may be due to the fact that they have additional available variables, such as whether the woman agrees that parents should decrease their work hours, and whether the person is religious.

V. Concluding Remarks

The gender wage gap in Colombia is still substantial, despite the strong convergence in men and women labor market characteristics and the existence of legal provisions aimed at gender equality. Men are always paid more than women and underpaid women are prevalent at the extremes of the distribution. Since this can only be captured in a QR framework, it is not only necessary but also interesting to go beyond means analysis for the Colombian case.

Our first finding is that the gender gap, controlling for differences in observable characteristics, is largest at the extremes of the wage distribution, both before and after adjusting for sample selection. Our result complements the international literature on gender gap distributions. The closest papers are De la Rica, Dolado and Llorens (2006), which finds a similar U shaped gap for Spain, and AVV, which finds that in the Netherlands the gap is larger at the top of the distribution. A substantial gender gap at the top quantiles of the distribution is commonly referred to as a "glass ceiling effect". We refer to the sizable gap observed at the bottom of the distribution in Spain and Colombia as a "quicksand floor effect".

We pose a hypothesis about the source of this phenomenon. Minimum wage policy may be an important factor compressing the wage gap in the middle of the distribution. On one hand, low productivity workers are prevalent at the bottom quintiles. Most of these workers are employed in the so called "informal sector" so that minimum wage policy is unlikely to affect their wages.[19] On the other hand, high productivity workers, such as professionals, are prevalent at the top percentiles. This suggests that two usual explanations for a "glass ceiling" effect on women's wages are at play. First, high-skilled women sacrifice some mobility along the corporate ladder and forego top paying jobs in order to balance their family and work lives – especially if they have young children. Second, firms may be reluctant to promote women to the top earning positions because of gender bias.

The U-shape of the gender gap and the differences between agents suggests that different measures may be considered to target the gender gap at either extreme. At the bottom of the wage distribution, our hypothesis suggests that any policy promoting formal labor contracts will have the side effect of enforcing minimum wage policy, and therefore tightening the gender gap at low skilled occupations.

At the top of the distribution, policies that reduce the cost of taking highly demanding jobs for skilled women. Extended maternity leave policies are an example. Carneiro, Loken and Salvanes (2008) find that, in Norway, extended maternity leave policies improve the likelihood that women stay in the labor force in the long run, without having a negative impact in their earnings profile. In the Colombian formal sector, the maternity leave is paid by the health insurance provider, not by the firm directly. Therefore, it is unlikely that the employers would pass-on the cost of the maternity leave to women via lower wages. We conjecture that policies focusing on childcare provision or on providing incentives for market provision of childcare may have positive effects. Finally, Blau and Kahn (1992) argue that anti-discrimination laws may have played a role in the reduction of the US gender gap. This type of policy could play a significant role, especially in reducing the "glass ceiling effect" in Colombia.

The size of the selection effect even in a country with a relatively high female participation, implies that non-random selection is an issue in the calculation of gender gaps, and should be taken seriously. Correcting for selection in a QR framework is

[19] Hoyos, Ñopo and Peña (2010) document this fact using the same data and roughly similar sample selection criteria as this paper.

important since we find that the selection effect is positive and significant: able women are pulled into the workforce.

We find that the bulk of the gender gap (both before and after controlling for selection) is explained by differences in the returns to characteristics. This is similar to what has been found in other countries (see for example, AVV). However, in contrast with AVV, we find that the selection effect in Colombia is mostly due to unobservables. This difference may stem from the fact that AVV have some additional variables that capture individual attitudes that may well affect the decision to participate, and that are unavailable in our dataset. Future research should try to include variables in the same spirit of the ones used by AVV. If, however, we suppose that the variables included are the ones that determine productivity, then there may be variables over which employers discriminate such as race or height. If this is the case, anti-discrimination policies would help level the ground for minorities, and increase their labor participation.

Given the size of the selection effect, we conjecture that the observed persistence of raw gender gaps in Colombia is partly due to the increase in female labor force participation itself. As female participation increases, the marginal female entrant into the labor force becomes, on average, less productive. This force tends to generate increasing raw gender gaps. This suggests that the evolution of gender gaps with sample selection correction should enable us to quantify the true dynamics of gender differences in the Colombian labor market. We see this as a fruitful avenue for future work.

References

Albrecht, James, Anders Bjorklund and Susan Vroman, 2003 "Is there a glass ceiling in Sweden?", Journal of Labor Economics, 21, 145--177.

Albrecht, James, Aico Van Vuuren and Susan Vroman, 2009 "Counterfactual Distributions with Sample Selection Adjustments: Econometric Theory and an Application to the Netherlands", Labour Economics.

Andrews, Donald and Marcia Schafgans, 1998 "Semiparametric Estimation of the Intercept of a Sample Selection Model", The Review of Economic Studies, Vol. 65, No. 3, Jul., pp. 497-517.

Autor, David, Lawrence Katz and Melissa Kearney, 2005 "Rising Wage Inequality: the Role of Composition and Prices" NBER Working Paper 11628, September.

Blau, Moshe and Melllisa Kahn, 1998 "Race and Gender Pay Differentials", NBER Working Paper # 4120.

Buchinsky, Moshe, 1998 "The Dynamics of Changes in the Female Wage Distribution in the USA: a Quantile Regression Approach", Journal of Applied Econometrics, 13, 1-30.

Carneiro, Pedro, Katrine Loken and Kjell Salvanes, 1998 "A flying start? Maternity leave and long-term outcomes for mother and child ", University College of London, unpublished manuscript.

Cunningham, Wendy, 2007 *Minimum Wages and Social Policy: Lessons from Developing Countries*. World Bank Publications.

De la Rica, Sara, Juan Dolado and Vanesa Llorens, 2005 "Glass Ceiling or Floors?: Gender Wage Gaps by Education in Spain" IZA Discussion Paper No. 1483, January.

Duryea, Suzanne, Olga Jaramillo and Carmen Pagés, 2001. *Latin American Labor Markets in the 1990s: Deciphering the Decade*. Inter-American Development Bank.

Fernández, Pilar, 2006 "Determinantes del diferencial salarial por género en Colombia, 1997-2003" Desarrollo y Sociedad, #58, septiembre.

Ganguli, Ina and Katherine Terrell, 2005 "Wage Ceilings and Floors: The Gender Gap in Ukraine's Transition" IZA Discussion Paper #1776.

Klein , Roger and Richard Spady, 1993 "An Efficient Semi-Parametric Estimator for Binary Response Models". Econometrica, Vol. 61, No. 2, March, 387-421.

Hoyos, Alejandro, Hugo Ñopo and Ximena Peña, 2010 "The Persistent Gender Earnings Gap in Colombia, 1994-2010" RES Working Papers 4673, Inter-American Development Bank Research Department.

Machado, José and José Mata, 2005 "Counterfactual Decomposition of Changes in Wage Distribution Using Quantile Regression" Journal of Applied Econometrics, 20, 445-65.

Ñopo, Hugo, 2006 "The Gender Wage Gap in Chile 1992-2003 from a Matching Comparisons Perspective" Research Department Working paper series # 562, Inter-American Development Bank.

Peña, Ximena, 2006 "Assortative Matching and the Education Gap", Georgetown University Working Paper.