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

Alejandro Badel† Ximena Peña‡

WORK IN PROGRESS - August, 2007

Abstract

In Colombia, women earn lower wages than men. This contrasts with the increased participation, hours worked and scope of women in the labor force observed in the last two decades. We use quantile regression techniques to analyze the gender wage gap and account for self selection of women into work. We find (i) that male and female wages display a U-shape and have been extremely unequal for the past 20 years, (ii) that most of the gender gap represents differences in the returns to labor market characteristics rather than differences in characteristics and (iii) positive and increasing selection along the distribution, which implies that able women are increasingly pulled into the workforce by the high returns.

Keywords: Gender gap, semiparametric, quantile regression, selection.

JEL classification numbers: C21, J22, J31.

1 Introduction

In Colombia, women earn lower wages than men. This contrasts with the increased participation, hours worked and scope of women in the labor force observed in the last two decades. According to International Labor Organization calculations, the female participation rate in Colombia passed from 19% in 1950 to 55.8% in 2000 and is, in fact, among the highest in Latin America (Duryea, Jaramillo and Pagés, 2001). In addition, female labor market attachment is stronger, as evidenced by the increase in hours worked. Whereas in 1986 females worked an average of 185 hours per month and 193 in 1996, in 2006 they work 197. The existing differences in (potential) work experience between genders have receded since the increase in the employment of mothers with young children

*The authors are grateful to James Albrecht, Susan Vroman and participants at the NIP-Colombia Conference for their comments. The usual disclaimer applies.

†Georgetown University, Washington, D.C. http://www12.georgetown.edu/students/ab377.

‡Georgetown University, Washington, D.C. http://www12.georgetown.edu/students/xp. Please send comments to xp@georgetown.edu.

has outpaced the growth of any other large demographic group, as in the U.S. (Anderson and Levine, 1999). Furthermore, women surpassed men in educational attainment, not only in terms of average years of schooling (Duryea, Galliani, Piras and Ñopo, 2006), but also in College attainment (Peña, 2006). In terms of inequality, if the distribution of characteristics that women bring into the labor market is now similar to that of men, we study to what extent have the earnings of women caught up with the earnings of men in Colombia.

Estimations of conventional Mincerian equations indicate that men in urban Colombia conditional on labor market characteristics such as education and age are paid 12-20% more than women on average. We go beyond means estimates and use a Quantile Regression framework (QR in what follows) to capture differences in the location, shape and spread of the wage distributions. A wage gap is calculated by taking the differences in log wages of two distributions at different percentiles. Gaps are decomposed, using the Machado Mata (MM) decomposition technique, into a component due to differences in human capital characteristics such as education and age -composition effect- and differences in the *rewards* to these characteristics -price effect.

First we analyze the raw wage gap for 1986, 1996 and 2006, that is, the difference between the observed men and women distributions. Over the last tow decades male and female wages are extremely unequal, men are always paid significantly more than women and the raw gap displays a U-shape: women's wages fall behind men's more at the extremes of the distribution whereas they are closer around the middle of the distribution. The levels of the gap are similar for 1986 and 1996, but there is a decrease for the year 2006, especially at the top 70% of the distribution. Focusing henceforth in the year 2006, a decomposition exercise shows that the price effect explains most of the raw gap.

Men participate more in the labor market than women in Colombia. Since women select into the labor force in a non-random way, we following Buchinsky (1998) to correct for sample selection bias in a QR. Once we account for selection we build the *potential* distribution of female wages -the log wage distribution that would prevail if all women worked. The gender gap is thus the difference between the male log wage distribution and the potential women distribution. The *level* is significantly higher than what the raw gap suggested, especially at the upper-end 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%. However, it displays a U-shape similar to that of the raw gap. The price effect explains roughly two thirds of the gender gap.

The direct effect of selection, namely, the difference between the observed and the potential distribution of women's wages, is positive: able women are increasingly pulled into the workforce by the high returns. This implies that the raw gap underestimates the gender gap since women who actually work are those who would get the greatest return. The selection effect is decomposed into a portion due to observables, which accounts for roughly one third of the selection effect, while

the remainder is attributed to unobservables.

The MM technique has been adopted in several papers to decompose de wage gaps across the distribution in several developed economies. Examples of this are Albrecht, Van Vuuren and Vroman (2007, AVV in what follows) for Denmark, Albrecht, Bjorklund and Vroman (2003) for Sweden and de la Rica, Dolado and Llorens (2007) for Spain[1]. The analysis has not been applied to any developing country. It is thus interesting not only to see what the results are for the Colombian case, but also to compare how results from a developing economy contrast with those of developed ones. However, since sample selection is an issue for the Colombia case, we follow the extension of the MM technique to account for selection proposed by AVV, who prove that the procedure yields consistent and asymptotically normal estimates of the quantiles of the counterfactual distributions. To our knowledge, this is the only application other than AVV of the MM technique adjusting for selection.

The rest of the paper is organized as follows. The next section presents the sample selection as well as some descriptive statistics while the methodology is described in Section 3. Section 4 presents the results and Section 5 concludes.

2 Descriptive Statistics and Data

We use the Colombian Household Survey (Encuesta Continua de Hogares-ECH) for the second quarter of 2006. The ECH is a repeated cross-section carried out by the Statistics Department that collects information on demographic and socioeconomic characteristics of the whole population, such as gender, age, marital status and educational attainment, together with labor market variables for the population aged 12 or more including occupation, job type, income and sector of employment. We also use the June 1986 and June 1996 shifts to perform the historical comparison of the raw gap.

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[2]. 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 Bogota and the other cities, we find that even though men participate the same, women participation is

[1] Only AVV account for selection in a QR framework, in a similar fashion to ours.

[2] Bogotá accounts for 45% of the population in the 7 main cities but given the design of the ECH, the sample size corresponds to only 15%. Sample weights are used to get representative results. Instead of using sample weights we perform calculations for Bogotá and Elsewhere separately, and then we build the weighted distribution as follows:

1. Let q_i be the percetiles of the log wage distributions for $i = \{Bogotá, Elsewhere\}$
2. Calculate at the j distribution the percentile levels at which q_i lies and call these P_i. E.g. $P_{bog} = F_{bog}(q_{else})$.
3. 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.
4. Obtain the country percentiles by linear interpolation.

higher in Bogota: 75% vs 65%.

We use only observations with a complete set of covariates and restrict our sample to individuals between 25 and 55 years of age who report working between 16 and 84 hours per week[3] and earn more than one dollar per day. This leaves 15,423 observations, equivalent to 3,978,580 using weights, 47% of which are female (See Table 1 in the Appendix for the details). In addition to the documented differences in participation rates, men and women also display important differences in hours worked per month. In our sample both have median hours of 208, on average men work 220 hours while women work only 197 hours per month.

The dependent variable is log hourly wage. The explanatory variables included in the estimations are: age and its square[4], 4 education groups[5], a dummy for marital status and another for head of household. Men earn higher mean hourly wages than women, as shown in Table (2) the average log wage for men is 7.86 and 7.72 for women. Working men and women have similar average age, whereas non-working women are nearly 2 years older than working ones. Schooling attainment differs between the groups and 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. Finally, the dummy variable for Bogotá captures an important difference. Even though 45% of the population in the seven main cities live in Bogotá, the city holds a disproportionately large percentage of working women, 47%, while only 38% of non-working women live in Bogotá.

The additional variables included in the selection equation, that are believed to determine the decision to work but not the wage, are 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). There are significant differences between working and non-working women in 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.

NEI is defined as income not related to labor market activities: accrued interest rates, rentals,

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

[4] 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. Therefore, we use age and its square to proxy for expericence instead of a transformation of age and schooling.

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

pensions, remittances and other concepts. A low percentage of agents report positive levels of this variable, which highlights the relevance of labor market earnings in an individual's total income. For example, 19% of women who do not work report strictly positive levels, while 14% of working women do. Of those who report strictly positive NEI, women not working report higher levels in average than those working. Finally, OFI is defined as the total household income minus the individual's total income. Again, a higher percentage of non-working women report strictly positive levels vis-à-vis working ones, 87% vs. 80%. However, the average OFI report for working women is higher than for non-working women.

We estimate Quantile Regression (QR) equations where the log hourly wage is regressed on the specified set of covariates and results are reported in Tables (3) and (4) for women and men, respectively.

3 Methodology

3.1 Machado-Mata Technique

The Machado-Mata (MM hereafter) technique is a decomposition in the spirit of Oaxaca-Blinder extended to the analysis of full distributions. It uses Quantile Regressions to partition the observed differences between the male and female log wage distributions into a 'price' component (due differences in the wage coefficients) and a 'quantity' component (attributed to differences in labor market characteristics). We use a partial equilibrium assumption and thus assume away the effect of changes in aggregate quantities of skills on skill prices. To simulate the impact of changing the distribution of characteristics on the gender gap we build a counterfactual distribution of the female wage density that would arise if we could endow women with men's labor market characteristics, but were paid like women. Similarly, to simulate the effect of changing prices, we build the density that would prevail if women retained their own labor market characteristics but were paid like men.

Let us illustrate the first counterfactual distribution with an example. Suppose that for each individual i in either population of females, F, or males, M, we observe a the log wage w_i and a vector of covariates x_i. Further, assume that for each population $j = M, F$, the conditional $\theta-$quantile of w_i^j, conditional on the set of covariates x_i^j, is given by $Q_\theta\left(w_i^j\right) = x_i^j \beta_\theta^j$. Then we can define the error term as $e_{\theta i}^j = y_i^j - x_i^j \beta_\theta^j$ where $e_{\theta i}^j$ is a random disturbance that satisfies $Q_\theta\left(e_{\theta i}^j\right) = 0$ by construction. The QR model for population F is $w_i^F = x_i^F \beta_\theta^F$ and similarly for population M.

The conditional distribution of wages given x is fully characterized by the conditional quantile process, that is, $Q_\theta\left(w|x\right)$ as a function of θ. Hence, realizations of w_i given x_i can be taken as independent draws from $Q_\theta\left(w_i|x_i\right)$ where θ is a uniform random variable in $(0, 1)$.

Let w^{FM} be the counterfactual distribution of female log wages that would prevail if we maintained the returns to observable characteristics of women, but endowed women with the male

distribution of labor market characteristics. Using estimates of β_θ^F and the empirical distribution of x_i^M we can simulate the conditional distribution of w given x by applying the probability integral transformation[6]. In particular,

$$w_i^{FM} = \beta_\theta^F x_i^M + e_i^{FM} \text{ with } Q_\theta\left(e_{\theta i}^{FM}\right) = 0$$

To generate the (counterfactual) conditional distribution of w given x, we generate random draws as follows. Pick at random man i with covariates x_i^M from the empirical distribution and quantile θ_i from the uniform $(0,1)$ distribution. With the consistent estimator form $\widetilde{w}_i^{FM} \equiv \widehat{\beta}_{\theta_i}^F x_i^M$. This way we can generate an arbitrarily large sample of draws from the conditional distribution of male labor market characteristics but paid as women, w^{FM}.

Notice that under the true $\beta_{\theta_i}^F$, we have $w_i^{FM} - \widetilde{w}_i^{FM} = e_{i\theta_i}^{FM}$. Therefore $Q_{\theta_i}\left(w^{FM}\right) = Q_{\theta_i}\left(\widetilde{w}^{FM}\right)$. Thus, for a consistent estimator of $\widehat{\beta}_\theta^F$, the empirical distribution of $\widetilde{w}^{FM}$ is a consistent estimator of the empirical distribution of w^{FM} since, as shown in Albrecht, Van Vuuren, and Vroman (2006), the quantiles of $\widetilde{w}^{FM}$ converge in probability to the quantiles of w^{FM}.

The wage gap, $Q_\theta\left(w^M\right) - Q_\theta\left(w^F\right)$, is decomposed as $\left[Q_\theta\left(w^M\right) - Q_\theta\left(w^{FM}\right)\right] + \left[Q_\theta\left(w^{FM}\right) - Q_\theta\left(w^F\right)\right]$ where the first term in squared brackets is the price effect while the second is the composition effect.

3.2 Estimating a quantile regression of Female Wages

Since women self-select into work, the usual problem of sample selection bias applies to the estimation of $\widehat{\beta}_\theta^F$. If for higher and higher quantiles of the potential wage distribution, the fraction of women actually participating increases, observed data under-samples the low potential earners and oversamples the high potential ones.

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

Heckman's two-step sample selection procedure first estimates a probit binary model of participation and then uses the results to construct a correction term to be included in the wage equation. In Buchinsky (1998), the first step estimates are obtained from a single-index semiparametric estimator and the selection correction term is a polynomial in the inverse mills ratio implied by the normal distribution and the estimates from the single-index procedure.

The model is summarized as follows. Let y_i be a participation dummy, z_i the vector of variables

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

that influence the participation decision and G an unknown function of the single index $z_i'\gamma$. The probability of participating is given by

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

We then construct the selection correction term, $P(\cdot)$, as a polynomial of the index,

$$P(z_i'\gamma) = \lambda_0 + \lambda_1 r(a + b(z_i'\gamma)) + \lambda_2 r(a + b(z_i'\gamma))^2 + ... + +\lambda_q r(a + b(z_i'\gamma))^q,$$

where a and b are location and scale parameters and $r(\cdot)$ 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. To see this, note that for any pair (a, b) and a function $G(a + b(z_i'\gamma))$ there is a function $\widehat{G}(z_i'\gamma)$ such that $\widehat{G}(z_i'\gamma) = G(a + b(z_i'\gamma))$ for all z_i. Following Buchinsky, we estimate a and b by running a probit regression of y_i on the semiparametrically estimated index $z_i'\widehat{\gamma}$.

We can now estimate the Mincer equation for working women correcting for selection

$$w_i = \mu_\theta + x_i^{F\prime}\beta_\theta^F + P(z_i'\gamma) + e_{\theta i} \text{ for } i \in \{j : y_j = 1\}, \tag{1}$$

where each quantile is given by

$$Q_\theta(w|\cdot) = \mu_\theta + x^{F\prime}\beta_\theta^F + P(z'\gamma) + e_\theta.$$

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 are evident from the data section.

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[7]. 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. While Buchinsky (1998) uses 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 (1986) and Cosslet (1987).

Last, note that μ and λ_0 are not separately identified in the quantile regression model above. We follow Buchinsky and estimate by the method developed in Andrews and Schafgans (1998). The

[7] 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.

method works as follows. First estimate equation (1) using quantile regression without separating μ and λ_0, let the combined constant be β_0. Then, construct the counterfactual residual $\widetilde{e}_i = w_i - x_i^{F\prime}\widehat{\beta}_\theta^F$. Note that this residual includes the constant and the selection correction term. Then, choose observations for which the probability of working is very high. In this subsample, the selection correction term is presumably zero. Then, μ is estimated as the $\theta-$Quantile of $\widetilde{e}_i$ within this subsample. As the total sample size increases, the fraction of observations used in this calculation should tend to zero. At infinity, the women chosen for the estimation of μ have probability 1 of working and μ equals the constant term in the original quantile regression model.

Once $\beta_{\theta(k)}^F$ has been consistently estimated, the MM procedure is conducted as above. However, the covariance matrix has to account for the variability coming from the estimation of γ, a, b and μ. Albrecht, Van Vuuren and Vroman (2006) prove asymptotic normality of the MM quantiles in this context, and extend the covariance matrix estimator in Buchinsky for quantile regression with selection correction to the MM quantiles.

4 Results

4.1 Raw Gap Decompositions, without Selection Correction

The first step is to study the gender gap from raw data, before conditioning on covariates (such as age and education) and before accounting for selection. The 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 earning 40% more than the other at that percentile. Figure (1) displays three observations of the raw gender gap, one decade apart. First note that male and female wages are extremely unequal, and men are always paid 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. Finally, even though 1986 and 1996 are very similar, there is a noticeable difference with respect to 2006: the gap shifted downwards in the top 70% of the distribution.

[Insert Figure 1 here]

The QR framework allows us to observe the variation across the distribution hidden behind means analysis. There is a *glass ceiling*, that is a gap that widens at the top of the distribution, suggesting a barrier to further advancement of women once they have attained a certain level. Albrecht, Bjorklund and Vroman (2003) find that the raw gap in Sweden increases to 40% in 1992 and 1998. We find similar levels at the top of the distribution for Colombia in 1986 and 1996. We also observe a widening gap at the bottom of the distribution.

In what follows, we will focus on the year 2006. As mentioned above, the raw gap for 2006 (Figure 2) displays a U-shape[8]. Whereas at low levels of the distribution the gap is around 35%, near the median it is close to zero, and it increased towards the upper tail of the distribution gap to a maximum log wage difference of about 35%. 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.

[Insert Figure 2 here]

Using the MM technique we can decompose the gap in Figure (2) into a component generated by differences in labor market characteristics and a component due to differences in the returns to said characteristics. Figure (3) presents 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 as men: the 'women paid as men' distribution. The price effect accounts for most of the raw gender gap since purging the differences payoffs the wage gap would be less than one third at the bottom half of the distribution and *zero* in the top half. This in line with results from other studies.

[Insert Figure 3 here]

4.2 Wage Gap Decompositions, Controlling for Selection

While male participation rates are very high -approximately universal, the proportion of women working is smaller. This suggests that selection bias is an issue in this estimation since women select into the labor force in a non-random way. Therefore, the raw gap is not a good measure of the differences in pay between genders since we're comparing the universe of men with a selected sample of women. We need to account for the selection bias in women's distribution, to make the male and female distributions comparable, and *then* calculate the gender gap.

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 coefficients adjusted à la Buchinsky 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.

After accounting for selection, we calculate the wage gap as the difference between the male wage distribution and the potential women distribution. Figure (4) shows that the gender gap displays a U-shape, as did the raw gap. However, the *level* is significantly higher, especially at the

[8] 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.

upper-end 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%. Again, the gender gap increases substantially at the top tenth of the distribution, this time passing from 40% to 60%. AVV find that after acounting for selection the gender gap is increasing and it reaches 40% at the top of the distribution. Therefore, the glass ceiling in Colombia is steeper.

[Insert Figure 4 here]

Note that the gender gap is equivalent to 'adding up' the raw gap (Figure 2) and the selection effect (Figure 6).

With equality in mind, if productivity was neutral to gender, two identical workers who differ only by their gender should earn the same. Therefore, we build the following counterfactual: what would the distribution of men wages be if they retained their characteristics but were disguised as women, and hence were paid the selection-adjusted returns of women. 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 in disguise' (see Figure 5). Note that over two-thirds of the wage gap, attributable to the price effect, remains after accounting for differences in characteristics.

[Insert Figure 5 here]

4.3 Decomposing the Selection Term

Let us now turn to the direct effect of selection. The difference between the observed and the potential distribution of women's wages is the selection term. Selection is positive and rather high in our application, around 20%. This is twice what AVV find the direct effect of selection to be in the Netherlands. Additionally, the QR analysis allows us to discover that the selection term is increasing along the distribution, which simply cannot be observed from traditional analysis. This implies that able women are increasingly pulled into the workforce by the high returns. Hence, not only does the raw gap underestimate the gender gap, since women who actually work are those who would get the greatest return, but it does so especially at the top of the distribution.

[Insert Figure 6 here]

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 selection effect is decomposed, using MM techniques 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. 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. Even though the effect of observables is not homogeneous along the distribution, it accounts to roughly one quarter of the selection effect until the 70th percentile, and in the top 30% it explains about half.

[Insert Figure 7 here]

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. What we do here is hold the distribution of observable characteristics constant -that of working women- and change the returns to characteristics from the ones observed in the market to the selection adjusted ones.

[Insert Figure 8 here]

Clearly, adding up the portions due to observables and unobservables yields the selection effect.

5 Concluding Remarks

The raw gap has remained relatively stable over the past 20 years. Men are always paid more than women and the gap displays a U-shape. Correcting for selection in a QR framework is key since we find that the selection effect is not only positive and significant but also increasing: able women are increasingly pulled into the workforce. Since last two effects can only be captured in a QR framework, it is not only necessary but also *interesting* to go beyond means analysis.

The results for Colombia, a developing country, are similar to those of other countries since the price effect explains a big portion of the gender gap. However, the levels are in general higher for Colombia.

6 References

Albrecht, J. A. Van Vuuren and S. Vroman (2006) "Counterfactual Distributions with Sample Selection Adjustments: Econometric Theory and an Application to the Netherlands", Mimeo, Georgetown University.

Andrews, D. W. and M. Schafgans (1998) "Semiparametric Estimation of the Intercept of a Sample Selection Model", *The Review of Economic Studies*, Vol. 65, No. 3 (Jul., 1998), pp. 497-517

	No. Observations	Weighted	% Men
13 main cities, 12+ years	81,339	14,200,850	0.44
7 main cities, 12+ years	46,439	11,585,058	0.44
Agents between 25 and 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

Table 1: Sample Selection, April-June 2006

Autor, D., L. Katz and M. Kearney (2005) "Rising Wage Inequality: the Role of Composition and Prices" NBER Working Paper 11628, September.

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

De la Rica, S., J. Dolado and V. Llorens (2005) "Glass Ceiling or Floors?: Gender Wage Gaps by Education in Spain" IZA Discussion Paper No. 1483, January.

Duryea, Suzanne, Sebastian Galiani, Claudia Piras and Hugo Ñopo (2007) "The Reversal of the Schooling Gender Gap in LAC", IADB, Mimeo.

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

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

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

Peña, X. (2006) "Assortative Matching and the Education Gap", Mimeo Georgetown University.

7 Appendix

	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	0.7	0.7	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
Bogota	0.43	0.47	0.38
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	0.04
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

Table 2: Descriptive Statistics, Wage Equation

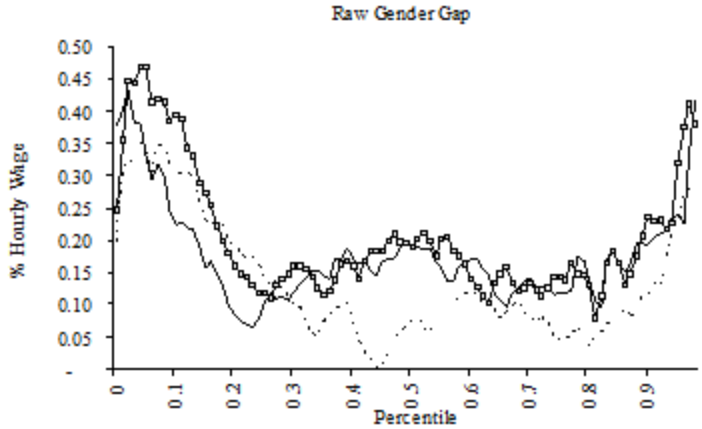

Figure 1: Raw Gap for 1986, 1996 and 2006. The solid line with markers is the 1986 gap, while the solid line and dashed lines represent the gaps for 1996 and 2006, respectively.

	Bogota				Elsewhere			
	20%	40%	60%	80%	20%	40%	60%	80%
Constant	6.62	7.18	7.33	7.20	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	0.00	0.00	0.01	0.04	0.02	0.00	0.01
	(0.05)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Age^2	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00
	(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	0.10	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)

Table 3: Mincer Equation, Women

	Bogota				Elsewhere			
	20%	40%	60%	80%	20%	40%	60%	80%
Constant	7.09	7.70	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	0.00	-0.02	0.00	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^2	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00
	(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	0.10	0.01	0.00	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)

Table 4: Mincer Equation, Men

	Bogota		Rest	
	Probit	**Klein & Spady**	**Probit**	**Klein & Spady**
Age^2	-1.172	-1.22	-1.103	-1.087
	(0.058)	(0.041)	(0.017)	(0.013)
Edu1	-0.008	0.001	-0.019	-0.017
	(0.029)	(0.011)	(0.012)	(0.034)
Edu3	-0.001	0.028	0.120	0.074
	(0.032)	(0.026)	(0.017)	(0.017)
Edu4	0.128	0.170	0.248	0.177
	(0.046)	(0.029)	(0.029)	(0.018)
Married	-0.133	-0.287	-0.173	-0.143
	(0.046)	(0.048)	(0.021)	(0.015)
Head	0.216	0.231	0.196	0.117
	(0.071)	(0.050)	(0.025)	(0.038)
# Children <1	-0.038	-0.089	-0.036	-0.023
	(0.031)	(0.026)	(0.012)	(0.008)
# Children <6	-0.069	-0.042	-0.023	-0.010
	(0.035)	(0.015)	(0.012)	(0.020)
Home Ownership	-0.078	-0.125	-0.035	-0.033
	(0.035)	(0.029)	(0.011)	(0.009)
Non-Earned Income	-0.102	-0.198	-0.184	-0.125
	(0.040)	(0.038)	(0.023)	(0.024)
Other Family Income	0.055	0.091	-0.023	0.021
	(0.034)	(0.023)	(0.013)	(0.031)
Hausman	Test	95% Critical Value	Test	95% Critical value
	36.163	19.675	35.018	19.675

Table 5: Selection Equation: Probit and Single Index Estimation
Note: all the coefficients are calculated relative to the absolute value of the coefficient of age.

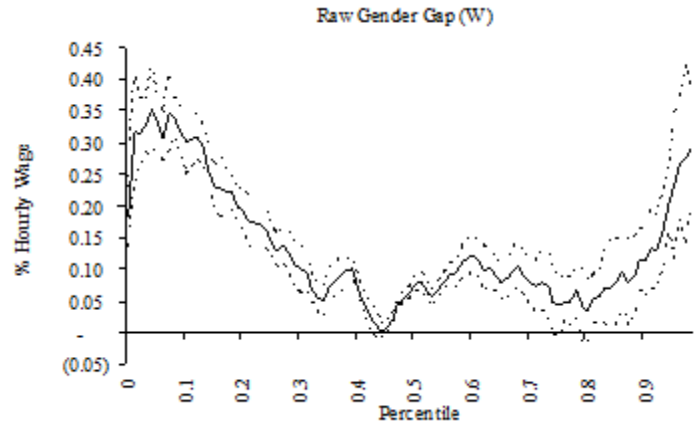

Figure 2: Raw Gap for 2006. The solid line is the raw gap while the dashed lines are the 95% confidence intervals.

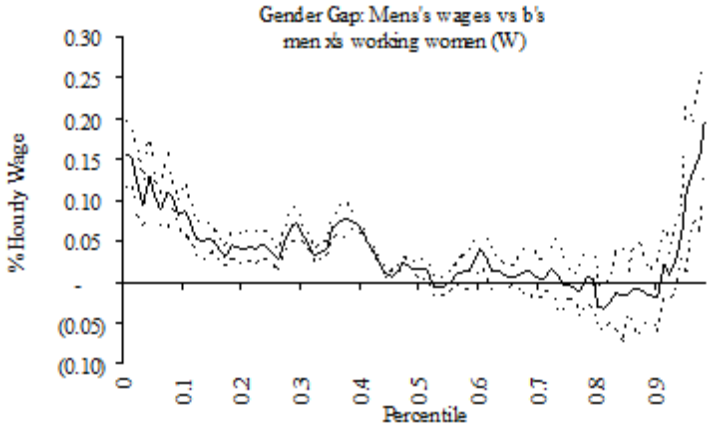

Figure 3: Difference between the male distribution and the distribution of 'women paid like men'. This is the portion of the raw gap that remains after controlling for differences in the rewards to labor market characteristics.

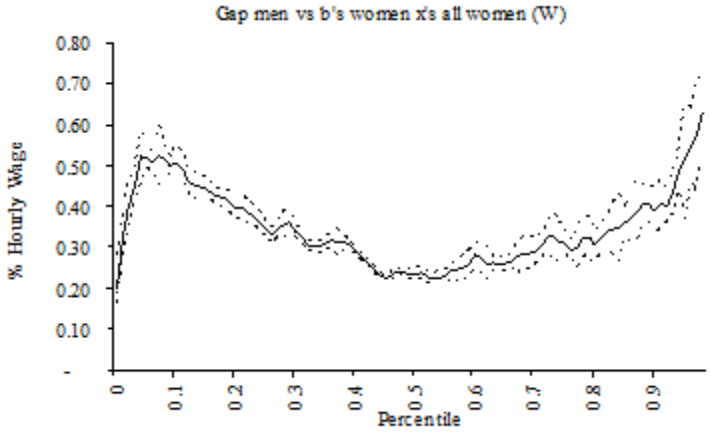

Figure 4: Wage Gap after accounting for selection: the difference between male distribution and the potential distribution of women -female distribution of wages if all women worked.

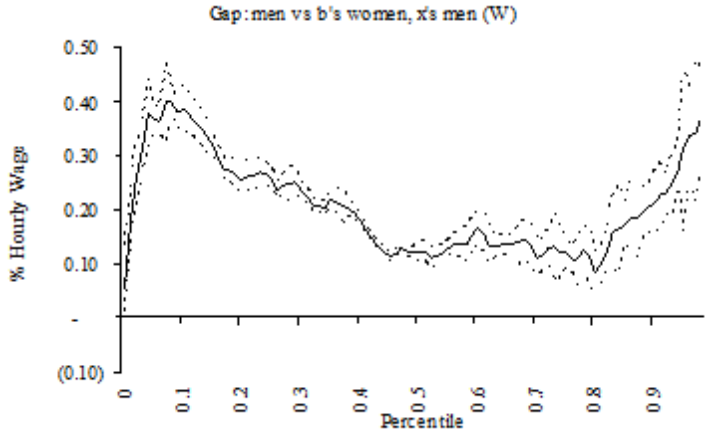

Figure 5: Difference between the observed wage distribution for men and distribution of 'men in disguise': the female distribution of wages if they had the labor market characteristics of men. The remainder is the portion of the wage gap due to differences in the returns.

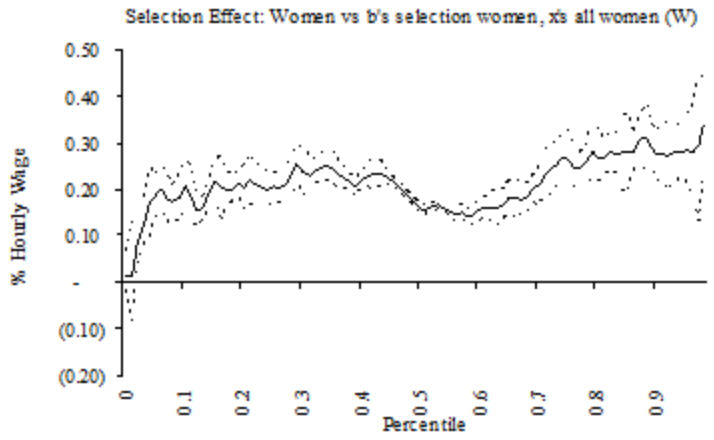

Figure 6: Difference between the observed female wage distribution and the potential women distribution.

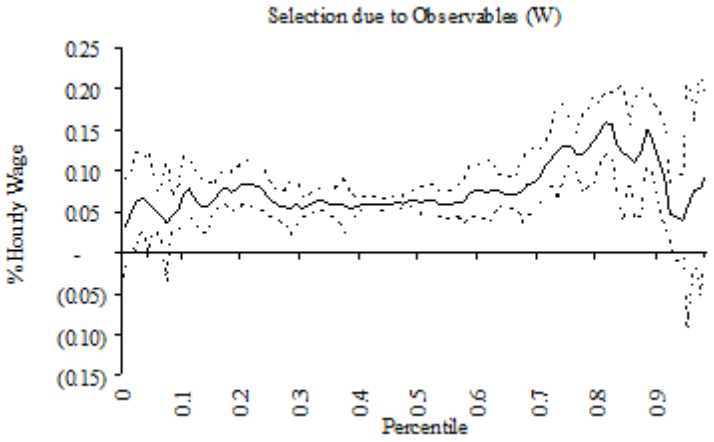

Figure 7: Portion of Selection Term due to observable characteristics, that is, the difference between 'working women adjusting for selection' and the potential distribution.

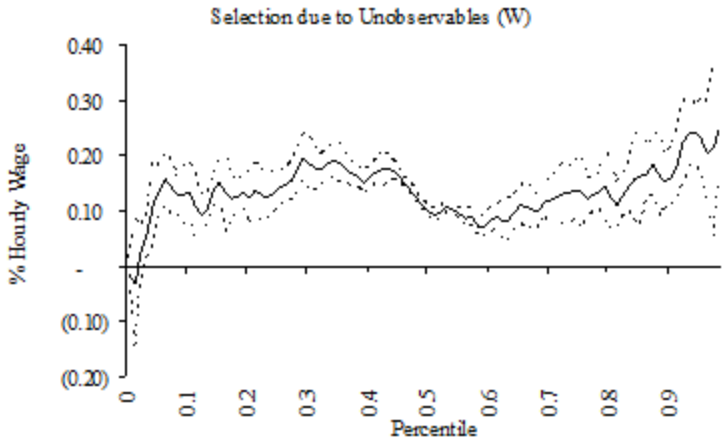

Figure 8: Portion of Selection Term due to unobservable characteristics, namely, the difference between the actual distribution of female wages and the distribution of 'working women adjusting for selection'.