

# *The Gender Gap across the Wage Distribution in Chile: An Application of Copula-Based Methods\**

Mariel C. Siravegna<sup>†</sup>

Department of Economics, Georgetown University  
*Job Market Paper*

November 12, 2020

[\(Click here for the most recent version\)](#)

## **Abstract**

In this paper, I analyze the gender pay gap in Chile by considering two main issues, heterogeneity across wage distribution and selection into the labor force. I apply a quantile regression technique and correct for sample selection using a copula-based methodology. My results highlight the importance of heterogeneous effects and selective participation in gender pay gaps. If men's and women's rates of employment were equal, the gap would be approximately 30 percentage points larger in all quantiles. My analysis reveals that the gender wage gap oscillates between 25 and 35 log points at the bottom half of the hourly wage distribution but increases to approximately 50 log points in the upper quantiles, evidencing a "glass ceiling" effect. Finally, I decompose the gender pay gap into "structural" and "composition" effects, concluding that the gap is explained mainly by differences in rewards for observable labor market characteristics and not by differences in the distribution of those characteristics.

**JEL-Codes:** C21, D6, I3, J1, J3, J7

**Keywords:** *Gender wage gap, sample selection, glass ceiling, heterogeneous effects.*

---

\*I am deeply thankful to my advisor James Albrecht for his continuous support and encouragement. I am also grateful with Anna Maria Mayda and Susan Vroman for their invaluable support in this project. I also want to thank Georgetown University EGSO, APPAM, Stata Conference and MVEA 2020 seminar participants. All errors are my own.

<sup>†</sup>Email: [mcs92@georgetown.edu](mailto:mcs92@georgetown.edu)

# I Introduction

Men earn on average more than women in Chile. According to OECD statistics, the gender wage gap, defined as the difference between the median wage of men and the median wage of women, was 12.5% as of 2017. That gap has remained stable over the last few decades, and it has been attributed mainly to inadequate job training, limited childcare options, and traditional attitudes toward gender roles (Fort, John-Abraham, Orlando, and Piras 2007).

The gender gap can vary significantly across the wage distribution. For example, the seminal work of Albrecht, Björklund, and Vroman (2003) demonstrates the glass ceiling effect in Sweden (i.e., the gender gap increases in the upper tail of wage distribution). Since 2003, a growing body of literature has shown evidence that the gap at the mean (median) is very different from the gap at the various percentiles. Related studies of other European countries include, among others, Arulampalam, Booth, and Bryan (2005); De la Rica, Dolado, and Llorens (2008); and Christofides, Polycarpou, and Vrachimis (2013). Arulampalam et al. (2005) study 11 European countries and conclude that differences in upper quantiles are more prevalent than at the bottom of the distribution for sample countries. De la Rica et al. (2008) find that in Spain, the gender gap is increasing across the distribution for highly educated women, whereas the gap decreases for less-educated women. Lastly, Christofides et al. (2013) research 23 European countries, and they find both a glass ceiling effect and a “sticky floors” effect (i.e., greater differences at the lowest quantiles) in some of those countries. Similarly, but for an emerging economy, Badel and Peña (2010) find that the gender wage gap in an emerging economy like Colombia is wider at the bottom and at the top of the wage distribution. The variety of results across these studies show the importance of extending the focus of analysis beyond the mean and median gender pay gap and analyzing the entire distribution of wages instead.

In addition, it is well known that men are more likely than women to participate in the labor market. Although the participation gap has been shrinking over time, women do not select themselves randomly to participate in the labor force. As a result, any analysis

based solely on working women will yield biased estimates of the gender wage gap. Hence, correcting for sample selection is important for analyzing the mean (median) gender gap and the gap across the wage distribution.

In the case of Chile, Perticar  and Bueno (2009) study the gender gap across the distribution of wages in Chile for the period from 2002 to 2006. Using the Lee (1998) methodology, they find an hourly gender wage gap between 11% and 18%, after controlling for occupational selection.<sup>1</sup> In another paper, Perticara and Astudillo (2010) use the Melly (2006) technique to decompose the portion of the gap explained by the difference in characteristics between male and female workers and perceived discrimination. This last component is negative across all the wage distributions, and it gets larger for the higher quantiles, although there is no glass ceiling effect in their results. In general, papers that analyze the gender wage gap in Chile find that the differences at the top quantiles are nothing like the differences at the lowest quantiles (Montenegro 2001). However, none of them address the potential bias from self-selection for participation in the labor force.

Using the longitudinal Social Protection Survey (*Encuesta de Protecci n Social* in Spanish), I study the gender wage gap in Chile by accounting for heterogeneity across the wage distribution and sample selection. I estimate quantile regressions using a new copula-based methodology, as proposed by Arellano and Bonhomme (2017b), to model the joint distribution of error in the outcome and participation equations. This method is suited for a model such as quantile regression, where the quantile curves are non-additive in the propensity score and covariates. To validate the exclusion restriction’s assumption of my instrument, I apply the Huber and Mellace (2014) test, which indicates that there is no significant relationship between the excluded regressor and the error term of the outcome equation. Then I apply the Mata and Machado (2005) technique to simulate the entire distribution of the wages that female workers would have earned if all women worked and had men’s distribution of characteristics. This last step allows me to decompose the gender wage gap into two parts:

---

<sup>1</sup>This methodology is similar to Heckmans (1979), but it uses a multinomial logit to predict the propensity score; however, like Buchinsky (1998), this approach is only effective for additive models, as explained later.

the composition effect and the structural effect. The composition effect is the portion of the gap attributable to gender differences in labor market characteristics, and the structural effect is the portion of the gap attributable to gender differences in the rewards for those characteristics.

I find that after controlling for selection into employment, the gender wage gap in Chile increases significantly across the whole wage distribution. If men's and women's rates of employment were equal, the gender wage gap would be around 25% to 35% at the low levels of the distribution, but it increases toward the upper tail of the distribution gap to a maximum log wage difference of about 50%. The difference between men's and women's wages is larger at the top of the distribution, evidencing a glass ceiling effect. Moreover, these differences across wage distributions are not primarily explained by the observable labor market characteristics of workers. Instead, they are explained by the differences in rewards offered for those labor market characteristics.

This paper contributes to the existing literature by measuring the gender wage gap in Chile using a quantile-copula methodology to control for non-random selection. Previous papers have measured the gender wage gap using quantile regression and corrected for sample selection using the Buchinsky (1998) technique, but this methodology has been shown to be ill-suited in this context. Buchinsky (1998) proposes a control function and assumes that the errors are independent of the regressors, given the selection probability. According to Huber and Melly (2015) this assumption implies that the quantile slope coefficients and the mean slope coefficient are identical, limiting the usefulness of heterogeneity across the distribution. Moreover, my paper adds to the literature that emphasizes the importance of self-selection in measuring gender gaps. Working women are a selected sample of women and if selection is ignored the gap would be considerably lower at all quantiles.

The rest of the paper is structured as follows: Section II describes the Arellano and Bonhomme (2017b) quantile sample selection model and Mata and Machado (2005) counterfactual decomposition. Section III provides details about the data and descriptive statistics.

Section IV presents a measurement of the gender wage gap that controls for sample selection, and I discuss implications for Chile. Section V analyzes the validity of my instrument, and Section VI is the conclusion.

## **II Empirical methods**

### **I Copula-based approach for quantile sample selection models**

A common approach in the literature to deal with non-random selection into the labor force is to impute the wages of non-working women using information of working women. Heckman (1979) proposed a two step estimator which assumes that errors in the selection and outcome equations are jointly normally distributed. However this implicit control function approach can not be applied in a quantile regression context since the latter is a non-additive model. As Huber and Melly (2015) have shown, the assumption of additivity does not hold in general in quantile models. Regarding the effect of selection across all the wage distribution, Arellano and Bonhomme (2017a) proposed a methodology that is compatible with sample selection in a non-additive model, which precludes the use of the control function approach. In their proposed estimator, sample selection is modeled using a copula, which is a function that couples a multivariate distribution to its marginal distribution functions. As a result, using information from the marginal distribution of the error from the participation decision and error of the outcome equation, it is possible to recover the gap across the distribution of wage offers for the entire female population (those who work and those who do not).

### **II The model**

In their paper, Arellano and Bonhomme (2017a) suggest to modeling sample selection using a bivariate cumulative distribution function or copula of the percentile error in the latent outcome equation and the error in the sample selection equation.

Consider a general outcome equation specification where the quantile functions are linear:

$$Y^* = Q(\tau, X) = x'\beta(\tau) \quad (1)$$

where the function  $Q$  is the  $\tau$ -th conditional quantile of  $Y^*$  given  $X$ . In this context,  $Y^*$  is the latent outcome variable (in this case, wage offers) and  $X$  are the covariates (e.g. education, experience, etc.). I specify  $\beta$  as a function increasing in  $U$ , where  $U$  is the error term of the outcome equation that is distributed uniformly and independent of the covariates.

The participation equation is defined as:

$$D = 1\{V \leq p(Z)\} \quad (2)$$

where  $1\{\cdot\}$  denotes a selection indicator  $D$  which takes values equal to one when the latent variable is observable (e.g. employment),  $Z$  contains  $X$  and at least one covariate  $B$  that does not appear in the outcome equation (e.g. a determinant of employment that does not affect wages directly),  $p(Z)$  is a propensity score, and  $V$  is an error term of the selection equation which is uniformly distributed on the unit interval and independent of  $Z$ .

Under the set of assumptions detailed in Arellano and Bonhomme (2017b),<sup>2</sup> the conditional cumulative distribution function of  $Y^*$  given  $Z = z$  for those observations that participate in the labor market is, evaluated at  $x'\beta(\tau)$  for some  $\tau$  in  $(0,1)$  interval is:

$$Pr(Y^* \leq x'\beta(\tau)|D = 1, Z = z) = Pr(U \leq \tau|V \leq p(z), Z = z) = G_x(\tau, p(z))$$

where  $G_x \equiv C(\tau, p)/p$  is defined as the conditional copula function of  $U$  given  $V$ , which measures the dependence between  $U$  and  $V$  as:

$$G_x(\tau, p) \equiv G(\tau, p; \rho) = \frac{C(\tau, p; \rho)}{p} \quad (3)$$

---

<sup>2</sup>Assumptions: 1)  $Z$  is independent of  $(U, V)|X$  (exclusion restriction), 2) absolutely continuous bivariate distribution of  $(U, V)$ , 3) continuous outcome, and 4) propensity score,  $p(z) > 0$ .

where the numerator is an unconditional copula of  $(U, V)$  and the denominator is the propensity score. The copula parameter  $\rho$  governs the correlation between the error in the outcome equation and the error in the participation decision and captures the degree of selection. In the context of this paper, a positive value for  $\rho$  indicates negative selection into employment, whereas negative values suggest positive selection<sup>3</sup>.

$G_x$  plays an important role in capturing the selection and mapping the rank  $\tau$  in the distribution of latent outcomes (given  $X=x$ ) to ranks  $G_x(\tau, p(z))$  in the distribution of observed outcomes conditional on participation (given  $Z=z$ ). Namely, the conditional  $G_x(\tau, p(z))$ , quantile of observed outcome (that is when  $D = 1$ ) coincides with the conditional  $\tau$ -th of latent outcome, and this is true for each  $\tau \in (0, 1)$ . The key implication from this is if it is possible to estimate the mapping  $G_x(\tau, p)$  from latent to observed ranks, it is possible to estimate the quantiles of observed outcomes corrected for selection. It follows from equation 3 that the  $\tau$ -th conditional quantile of  $Y^*$  given  $D=1$  and  $Z$  is

$$Q^s(\tau, Z) = X'\beta(\tau^*(Z))$$

where  $\tau^*(Z)$  is the inverse of the conditional copula with respect to its first argument and  $Q^s$  refers to conditional on selection. Hence, this strategy can be used in quantile selection models that are non-additive in the propensity score  $p(Z)$  and the covariates  $X$ , as is the case of the application in this paper.

### III Estimation

The Arellano and Bonhomme (2017b) estimation algorithm can be summarized in 3 steps: estimation of the propensity score, estimation of the degree of selection via the cumulative distribution function of the percentile error in the outcome equation and the error in the participation decision, and then, using the estimated parameters, the computation of any

---

<sup>3</sup>Positive selection implies that the wages of non-working women are lower than those who are working.

desired quantile through rotated quantile regression which preserves the linear programming structure of the standard linear quantile regression (see Koenker and Bassett Jr (1978)).

The first step consists of estimating the propensity score  $\gamma$  by a probit regression:

$$\hat{\gamma} = \underset{a}{\operatorname{argmax}} \sum_{i=1}^N D_i \ln \Phi(Z'_i a) + (1 - D_i) \ln \Phi(-Z'_i a) \quad (4)$$

The second step is to estimate the copula parameter  $\rho$  by generalized method of moments, which allows us to obtain an observation-specific measure of dependence between the rank error in the equation of interest and the rank error in the selection equation. This step consists of working with a parametric copula and deriving moment restrictions on the copula parameter. For every  $\tau$  in the unit interval the following population moment restriction holds:

$$\mathbb{E} \left[ \mathbf{1}\{Y \leq X'\beta(\tau)\} - G(\tau, \Phi(Z'\gamma); \rho) | D = 1, Z \right] = 0$$

This is then used to create a sample counterpart where  $\rho$  minimizes the following objective function:

$$\hat{\rho} = \underset{c}{\operatorname{argmin}} \left\| \sum_{i=1}^N \sum_{l=1}^L D_{i\varphi}(\tau_l, Z_i) [\mathbf{1}\{Y_i \leq X'_i \hat{\beta}_{\tau_l}(c)\} - G(\tau_l, \Phi(Z'_i \hat{\gamma}), c)] \right\|$$

where  $\|\cdot\|$  is the Euclidean norm,  $\tau_1 < \tau_2 < \dots < \tau_L$  is a finite grid on  $(0, 1)$ , and  $\varphi(\tau, Z_i)$  is a vector of instruments where the  $\dim \varphi \leq \dim \rho$  and:

$$\begin{aligned} \hat{\beta}_{\tau}(c) = \underset{b(\tau)}{\operatorname{argmin}} \sum_{i=1}^N D_i [ & G(\tau, \Phi(Z'_i \hat{\gamma}); c) (Y_i - X'_i b(\tau))^+ + \\ & (1 - G(\tau, \Phi(Z'_i \hat{\gamma}); c)) (Y_i - X'_i b(\tau))^- ] \end{aligned}$$

where  $a^+ = \max\{a, 0\}$ ,  $a^- = \max\{-a, 0\}$ , and the grid of  $\tau$  values on the unit interval as well as the instrument function are chosen by the researcher.

Lastly, using  $\hat{\gamma}$  and  $\hat{\rho}$  obtained before, the third step consists in using  $\hat{G}_{\tau i} = G(\tau, \Phi(Z'_i \hat{\gamma}); \hat{\rho})$



to estimate  $\beta(\tau)$  for any desired  $\tau \in (0,1)$  by minimizing a rotated check function of the form:

$$\hat{\beta}(\tau) = \underset{b(\tau)}{\operatorname{argmin}} \sum_{i=1}^N D_i [\hat{G}_{\tau i} (Y_i - X_i' b(\tau))^+ + (1 - \hat{G}_{\tau i}) (Y_i - X_i' b(\tau))^-] \quad (5)$$

where  $\hat{\beta}(\tau)$  will be a consistent estimator of the  $\tau$ -th quantile regression coefficient. Note that the third step is unnecessary if the quantiles of interest are included in the set  $\tau_1 < \tau_2 < \dots < \tau_L$  used in the second step.

## IV Mata-Machado Decomposition

The Mata and Machado (2005) technique decomposes the gender wage gap into two components: the size of the gap that can be explained by observable worker characteristics and the size of the gap that can be explained by the compensation of those characteristics. This technique has the same flavor as the Oaxaca-Blinder decomposition (see Oaxaca (1973) and Blinder (1973)) but instead of doing the decomposition at the mean, Mata and Machado (2005) decompose the gap between men and women at the quantiles. In order to explain Mata and Machado (2005), consider the conditional cumulative distribution function of the wages given  $x=x_i$  as  $F_{w|x_i}$  and the corresponding  $\tau^{th}$  quantile of the conditional distribution as  $Q_\tau(w|x_i)$ . The relationship between both is the following:

$$F_{w|x_i} = \int_0^1 1[F_{w|x_i}^{-1}(\tau) \leq w] d\tau = \int_0^1 1[Q_\tau(w|x_i) \leq w] d\tau \quad (6)$$

That is, the conditional CDF is related to its inverse, the conditional quantile function. According to Arellano and Bonhomme (2017a), after estimating the parameter  $\hat{\rho}$  and the propensity score  $\hat{\gamma}$ , in the last step of the estimation algorithm, the  $\beta_\tau$ , for  $\tau \in (0,1)$ , are computed by minimizing a rotated check function that has been rescaled for  $\hat{G}_{\tau i} = G(\tau, \Phi(Z_i' \hat{\gamma}); \hat{\rho})$ . Therefore, any procedure developed for standard linear quantile regression could be used in the presence of sample selection. In particular, the unconditional cumulative

distribution function of the outcome may be estimated by

$$\hat{F}_w(w) = \frac{1}{N} \sum_{i=1}^N \int_0^1 1[Q_\tau(w|x_i) \leq w] d\tau$$

and the unconditional quantile can be computed by inverting the marginal cumulative function  $Q_\tau(w) = \inf\{y : \hat{F}_w \leq \tau\}$ .

In this empirical exercise, I consider two groups, men and women. Each group has stochastic vectors associated with the characteristics of each group together with its realizations,  $x_m$  and  $x_f$ , respectively. Each group has a distribution function that is indicated by  $F_{x_i}$  where  $i=male, female$ . From equation 6, the distribution of wages for males can be written as the conditional distribution of wages for males given males' characteristics integrated against the distribution of the characteristics for males.

$$F_m(w) = \int \left\{ \int_0^1 1[Q_\tau(w_m|x_m) \leq w] d\tau \right\} dF_{x_m}(x)$$

and the same for female workers,

$$F_f(w) = \int \left\{ \int_0^1 1[Q_\tau(w_f|x_f) \leq w] d\tau \right\} dF_{x_f}(x)$$

The methodology generates a counterfactual distribution where women have men's characteristics but those characteristics are paid the way women are rewarded.

$$F_c(w) = \int \left\{ \int_0^1 1[Q_\tau(w_f|x_f) \leq w] d\tau \right\} dF_{x_m}(x)$$

In other words, the counterfactual distribution is constructed using the female wage structure but holding the distribution of men's human capital characteristics unchanged. The Mata

and Machado (2005) decomposition consists of the following subtraction:

$$F_m(w) - F_f(w) = \underbrace{F_m(w) - F_c(w)}_{\text{differences in return}} + \underbrace{F_c(w) - F_f(w)}_{\text{differences in characteristics}}$$

The differences in the distributions  $F_m(w) - F_c(w)$  provide insights into the structural effect, which is target of policies that aim to equalize pay structure between female and male workers with the same set of skills. The differences between  $F_c(w) - F_f(w)$  will explain the composition effect, which is the part of the gap that is due to differences in the distribution of human capital characteristics.

### III Data and descriptive statistics

I use data from the Social Protection Survey (*Encuesta de Proteccion Social* in Spanish) of Chile. This longitudinal survey was initiated in 2002, and it has been conducted roughly every two years. The Social Protection Survey uses a representative nationwide sample of approximately 17,000 individuals, and it contains comprehensive socio-economic information such as labor history, family history, wages, assets, and health. An important feature of this self-reported survey is that the data related to the labor market is very detailed. The working history of the individuals encompass the time when they were employed, unemployed and inactive, so it is feasible to calculate the effective experience of the workers. This is very important because women have intermittent participation in the labor market during their life cycle. In a hypothetical case, if experience were measured as age subtracting education, it would not be reflecting the gender differences in the timing of experience acquisition that exists between men and women.

For my analysis, I use data from 2016's wave, which contains variables that can be used to explain women's participation in the labor market. The sample size is 16,906 individuals, where 47% of them are men and 52% are women. I restrict the sample to individuals between

25 and 55 years old who work less than 60 hours per week and receive a positive salary, ending up with 6,771 observations.

The descriptive statistics for all women, those who do not participate in the labor market and those who participate, can be found in Table 1. Women who are working are less likely to be married and to live with children under 3 years old. Additionally, they are more educated and have more accumulated experience.

Table 1: Descriptive Statistics for Women

	All	Non-Participate	Participate
Age	40	41	39
Married	.498	.644	.399
Experience	11.09	7.34	15.10
Children Under 3 Living at Home	.117	.149	.095
<i>Education Level</i>			
Non-Edu or Elementary	.020	.031	.012
Middle School	.198	.286	.138
High School	.508	.534	.490
Technical Degree	.134	.093	.163
Bachelor Degree	.124	.053	.172
Graduate Degree	.013	.001	.022
<i>Number of Observations</i>	3,742	1,509	2,233

Notes: The reported numbers correspond to weighted sample averages.

In the case of men, the descriptive statistics of the sample are shown in Table 2. In contrast to women, men who work are more likely to be married and live with children under the age of 3, and they seem to be younger than women and with more experience. However, for those who are working, their level of education on average is lower than working women.

Table 2: Descriptive Statistics for Men

	All	Non-Participate	Participate
Age	39	37	39
Married	.545	.368	.579
Experience	19.5	14.09	20.60
Children Under 3 Living at Home	.115	.088	.120
<i>Education Levels</i>			
Non-Educ	.020	.032	.018
Middle School	.214	.300	.198
Technical Education	.491	.440	.501
High School	.127	.1005	.132
Bachelor Degree	.134	.123	.136
Graduate Degree	.010	.002	.012
Number of observations	3,029	486	2,543

Notes: The reported numbers correspond to weighted sample averages.

In this paper, I model selection only for women and the potential sample selection for men is ignored. If sample selection for men were also controlled, since it is a small fraction<sup>4</sup> of male workers who are out of the labor market, this would generate a lack of precision in wage gap estimates.<sup>5</sup>

## IV Empirical results

In the following subsections, I consider the impact of female selection into employment. First, I compute the quantile regressions without correcting for sample selection and then correcting for sample selection. After that, I apply the Mata and Machado (2005) technique to decompose the gender wage gap corrected for sample selection.

In the empirical analysis, the dependent variable is the log hourly wage and the definition

<sup>4</sup>16% among all the men in the sample are not participating versus 40% of women.

<sup>5</sup>See Badel and Peña (2010).

of the gender wage gap is the following:

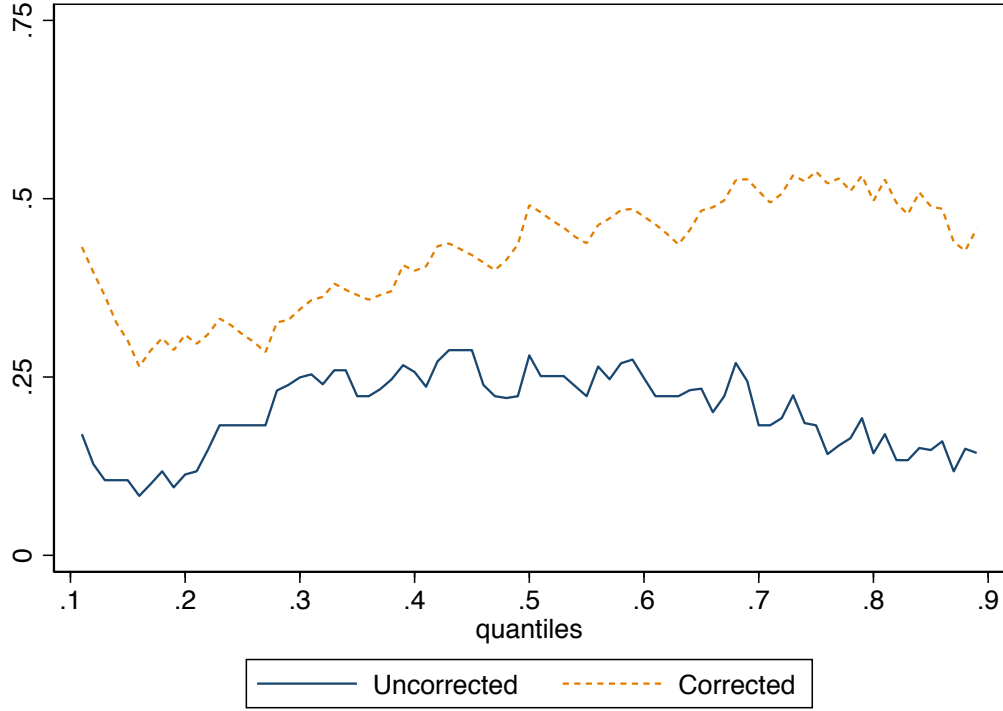
$$GenderWageGap(\tau) = w_{\tau}^m - w_{\tau}^f$$

where  $w_{\tau}^m$  and  $w_{\tau}^f$  denote the log of wages of male and female workers, respectively, at the corresponding quantile ( $\tau$ ). Heterogeneity across the wage distribution requires measuring the gap at the same set of skills, i.e., the  $\tau$  quantile of men's and women's wage sets are invariant to skills (Koenker & Bassett, 1978; Maasoumi & Wang, 2019). In the model of sample selection, the independent variables are marital status, effective experience and education attainment (4 categories: high school, technical degree, bachelor degree and graduate degree) and the exclusion restriction variable is the number of children under the age of 3 years old. Following the traditional econometric strategy, the number of children will have the role of an instrument because it is assumed that it would not affect the wage of the female workers but it may influence the probability of participate in the labor market. This variable has been used in the seminal works on female selection as Heckman (1974), Heckman and MaCurdy (1980) and recently by Maasoumi and Wang (2019), and has not been refuted as a valid instrument in this type of setting. As discussed in Maasoumi and Wang (2019) and Arellano and Bonhomme (2017a), the identification is given by the copula and not from the exclusion restriction. Despite that, I validate my IV using the test by Huber and Mellace (2014), which is explained in section V.

Figure 1 shows the raw gender wage gap without and controlling for selection. Male and female wages are unequal at every point of the wage distribution regardless of the sample selection bias. The gender wage gap without correction oscillated between 10 and 25 log points, which can be interpreted as men having a wage between approximately 10% and 25% higher than women at that percentile respectively. With sample selection correction the gender wage gap is around 25% and 35% at the low levels of the distribution, but it increases towards the upper tail of the distribution gap to a maximum log wage difference

of about 50%.

Figure 1: Raw Gender Wage Gap



This exercise highlights how important it is for gender gap analyses to consider sample selection into employment and heterogeneity in wages. As Figure 1 shows, the gender wage gap shifts upward across the whole wage distributions after controlling for selection bias. There is a bigger difference at the highest quantiles, which means that women who are above the median face a glass ceiling effect.

## I Quantile regression without sample selection adjustment

The quantile regression results for men and women are displayed in Table 3 and Table 4, respectively. For both groups, human capital covariates have the expected sign. It appears that experience is statistically significant for men only at the bottom quantiles, and for women only at the upper quantiles. Education has a positive coefficient for men and for women except for the coefficient associated with high school for the 90th percentile for

female workers. It seems that it makes virtually no difference to have no formal education or just a high school education at least for those women which are in the top of the wage distribution. When returns on education are compared between the two groups across the wage distribution, it appears that education is more valued for men than for women, except for the 10th percentile, where the coefficients for the education covariates are higher for women than for men. This situation changes slightly when the results are corrected for sample selection as described in the next section.



Table 3: Quantile regressions for men

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	log_wage	log_wage	log_wage	log_wage	log_wage	log_wage	log_wage	log_wage	log_wage
Experience	0.00163* (0.000834)	0.000844* (0.000482)	0.00114** (0.000527)	0.000384 (0.000452)	-6.65e-05 (0.000453)	0.000142 (0.000411)	0.000266 (0.000752)	0.000428 (0.00102)	-0.000458 (0.00108)
Exp. Squared	-2.96e-06** (1.47e-06)	-1.24e-06 (8.20e-07)	-1.77e-06* (9.69e-07)	-5.43e-07 (8.09e-07)	5.37e-07 (8.21e-07)	2.23e-07 (7.78e-07)	1.82e-07 (1.41e-06)	1.05e-07 (1.79e-06)	1.86e-06 (2.09e-06)
High School	0.145*** (0.0561)	0.174*** (0.0214)	0.258*** (0.0319)	0.287*** (0.0252)	0.294*** (0.0289)	0.326*** (0.0313)	0.309*** (0.0519)	0.316*** (0.0528)	0.302*** (0.0787)
Technical Education	0.423*** (0.0804)	0.460*** (0.0374)	0.529*** (0.0516)	0.557*** (0.0340)	0.533*** (0.0503)	0.608*** (0.0491)	0.602*** (0.0912)	0.655*** (0.0755)	0.554*** (0.105)
Bachelor Degree	0.577*** (0.119)	0.785*** (0.0999)	0.952*** (0.0617)	0.965*** (0.0415)	1.029*** (0.0544)	1.070*** (0.0582)	1.130*** (0.0993)	1.316*** (0.139)	1.302*** (0.131)
Graduate Degree	1.285*** (0.102)	1.323*** (0.170)	1.341*** (0.160)	1.509*** (0.259)	1.432*** (0.0413)	1.352*** (0.0338)	1.235*** (0.134)	1.312 (1.410)	1.638*** (0.124)
Married	0.0882** (0.0419)	0.110*** (0.0251)	0.104*** (0.0325)	0.116*** (0.0253)	0.104*** (0.0311)	0.0965*** (0.0232)	0.0842* (0.0447)	0.0772 (0.0541)	0.0489 (0.0642)
Constant	6.837*** (0.123)	6.991*** (0.0666)	7.026*** (0.0678)	7.190*** (0.0614)	7.307*** (0.0624)	7.366*** (0.0507)	7.467*** (0.102)	7.606*** (0.139)	7.991*** (0.150)
Observations	2,543	2,543	2,543	2,543	2,543	2,543	2,543	2,543	2,543

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 4: Quantile regressions for women without corrections for selectivity

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	log_wage	log_wage	log_wage	log_wage	log_wage	log_wage	log_wage	log_wage	log_wage
Experience	0.00142 (0.00109)	0.000703 (0.000457)	6.62e-06 (0.000378)	0.000554 (0.000347)	0.000334 (0.000447)	0.00101*** (0.000334)	0.00121** (0.000514)	0.00138** (0.000647)	0.00150** (0.000758)
Exp. Squared	-2.03e-06 (2.41e-06)	-8.60e-07 (9.08e-07)	5.09e-07 (8.54e-07)	-4.55e-07 (7.82e-07)	4.18e-08 (1.04e-06)	-1.14e-06* (6.58e-07)	-1.79e-06 (1.11e-06)	-1.71e-06 (1.46e-06)	-1.79e-06 (1.98e-06)
High School	0.283*** (0.0927)	0.195*** (0.0428)	0.116** (0.0453)	0.117*** (0.0257)	0.152*** (0.0303)	0.183*** (0.0356)	0.164*** (0.0423)	0.144* (0.0873)	-0.205* (0.120)
Technical Degree	0.495*** (0.109)	0.326*** (0.0568)	0.279*** (0.0565)	0.361*** (0.0619)	0.442*** (0.0608)	0.497*** (0.0545)	0.491*** (0.0781)	0.513*** (0.104)	0.113 (0.115)
Bachelor Degree	0.842*** (0.0960)	0.872*** (0.0850)	0.849*** (0.0674)	0.965*** (0.0523)	1.031*** (0.0445)	1.062*** (0.0393)	1.067*** (0.0592)	1.060*** (0.0902)	0.591*** (0.115)
Graduate Degree	1.488*** (0.129)	1.273*** (0.0628)	1.238*** (0.102)	1.267*** (0.0684)	1.326*** (0.0976)	1.399*** (0.238)	1.492*** (0.177)	1.421*** (0.112)	1.028*** (0.384)
Married	0.0901 (0.0648)	0.0599** (0.0258)	0.0284 (0.0240)	0.0529** (0.0268)	0.0325 (0.0284)	0.0592* (0.0334)	0.0782** (0.0344)	0.0710 (0.0482)	0.129** (0.0537)
Constant	6.512*** (0.136)	6.874*** (0.0650)	7.090*** (0.0568)	7.097*** (0.0339)	7.168*** (0.0475)	7.168*** (0.0400)	7.272*** (0.0629)	7.402*** (0.0998)	8.010*** (0.127)
Observations	2,233	2,233	2,233	2,233	2,233	2,233	2,233	2,233	2,233

Robust standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

## II Quantile regression with sample selection adjustment

Table 5 shows the results of estimating quantile regression for women controlling for selection. This table is not shown for men since I am controlling for selection bias and adjusting the coefficients only for women. The specified set of covariates is the same as the prior section, and they also have the expected sign. After adjusting for selection, in general the variables that represent education have a monotonic effect on wages for bachelor and graduate degrees. Coefficients for education are all positive showing evidence that for female and male workers, the more educated earn more, but education still being relatively more valued for men at the highest paid jobs. In addition, the fact that men's return on schooling is increasing with the quantile suggests that a high level of education has a positive impact on wage dispersion. Experience appears not to be statistically significant for female workers on all the different quantiles. Being married impacts positively at the tails of the wage distribution for women but it is not statistically significant for the 10th and 90th percentile. However, for male workers, being married positively affects their wages in all levels of wages.

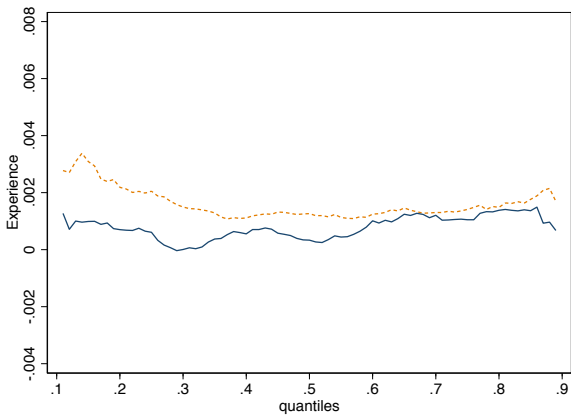
Table 5: Quantile regressions for women with corrections for selectivity

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	log_corrected	log_corrected	log_corrected	log_corrected	log_corrected	log_corrected	log_corrected	log_corrected	log_corrected
Experience	0.0033 (0.0023)	0.0022 (0.0019)	0.0015 (0.0016)	0.0011 (0.0012)	0.0013 (0.0011)	0.0012 (0.0010)	0.0013 (0.0011)	0.0015 (0.0011)	0.0018 (0.0014)
Exp. Squared	-5.01e-06 (0.0023)	-3.02-e06 (3.06e-06)	-2.08e-06 (2.54e-06)	-1.27e-06 (2.05e-06)	-1.49e-06 (2.09e-06)	-9.72e-07 (1.85e-06)	-9.64e-07 (1.88e-06)	-1.36e-06 (1.94e-06)	-1.69e-06 (2.51e-06)
High School	0.2465 (0.1594)	0.2879** (0.1251)	0.1785* (0.1040)	0.1139* (0.0621)	0.1438*** (0.0337)	0.1653*** (0.0308)	0.1880*** (0.0342)	0.2395*** (0.0636)	0.3080** (0.1420)
Technical Education	0.5677** (0.2400)	0.5289*** (0.1796)	0.4207*** (0.1413)	0.4097*** (0.0896)	0.4519*** (0.0550)	0.4932*** (0.0496)	0.5432*** (0.0538)	0.6130*** (0.1001)	0.6007*** (0.1951)
Bachelor Degree	0.9513*** (0.2924)	0.9375*** (0.2056)	0.9203*** (0.1435)	0.9583*** (0.0962)	1.0369*** (0.0707)	1.0917*** (0.0802)	1.1336*** (0.1019)	1.1551*** (0.1450)	1.2042*** (0.2569)
Graduate Degree	1.3221*** (0.3448)	1.4609*** (0.2500)	1.3629*** (0.1835)	1.3108*** (0.1317)	1.2991*** (0.1079)	1.3672*** (0.1172)	1.3674*** (0.1299)	1.4782*** (0.1856)	1.5485*** (0.2908)
Married	0.0199 (0.0603)	-0.0007 (0.0554)	-0.0064 (0.0485)	-0.0018 (0.0534)	0.0032 (0.0506)	-0.0075 (0.0546)	-0.0206 (0.0663)	-0.0167 (0.0863)	0.0065 (0.1041)
Constant	6.1848*** (0.5539)	6.5336*** (0.4553)	6.8108*** (0.3609)	6.9746*** (0.2524)	7.0069*** (0.1974)	7.0525*** (0.1904)	7.1245*** (0.2050)	7.2132*** (0.2406)	7.3439*** (0.3739)
Observations	3,743	3,743	3,743	3,743	3,743	3,743	3,743	3,743	3,743

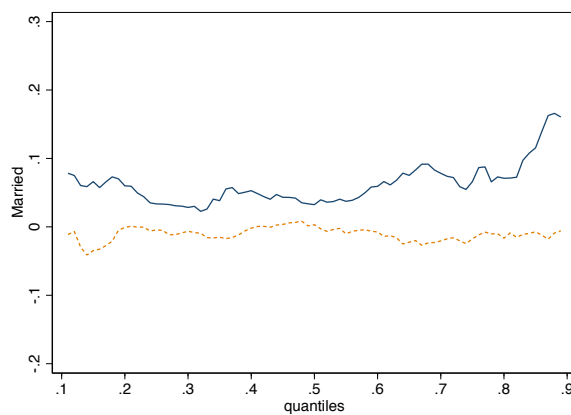
Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

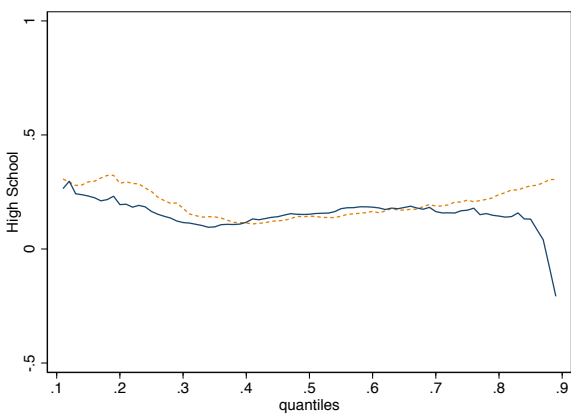
Figure 2: Quantile Regression Estimates



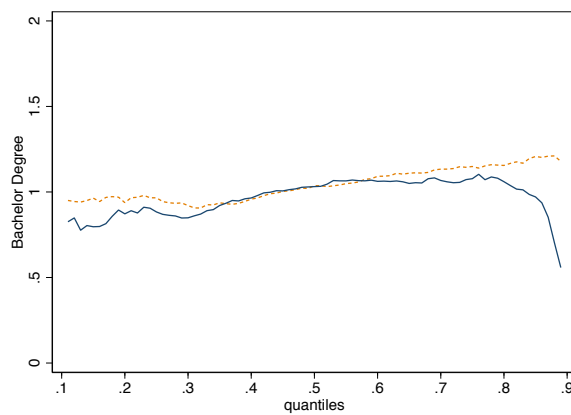
(a) Experience



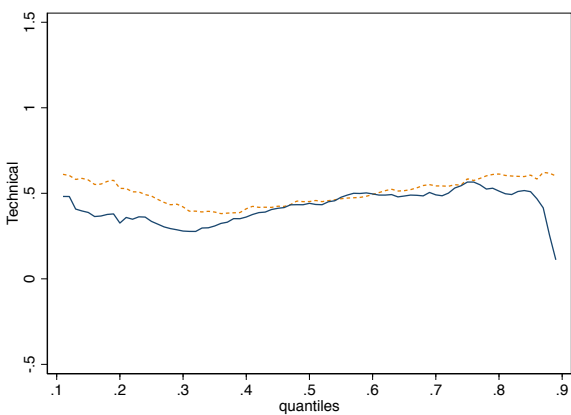
(b) Married



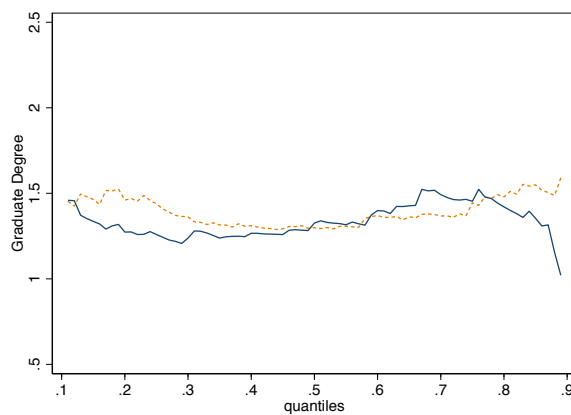
(c) High School



(d) Bachelor Degree



(e) Technical Degree



(f) Graduate Degree

Notes: Dashed line: corrected by selection. Solid line: uncorrected by selection

Figure 2 shows the coefficients of the quantile regression for women before and after

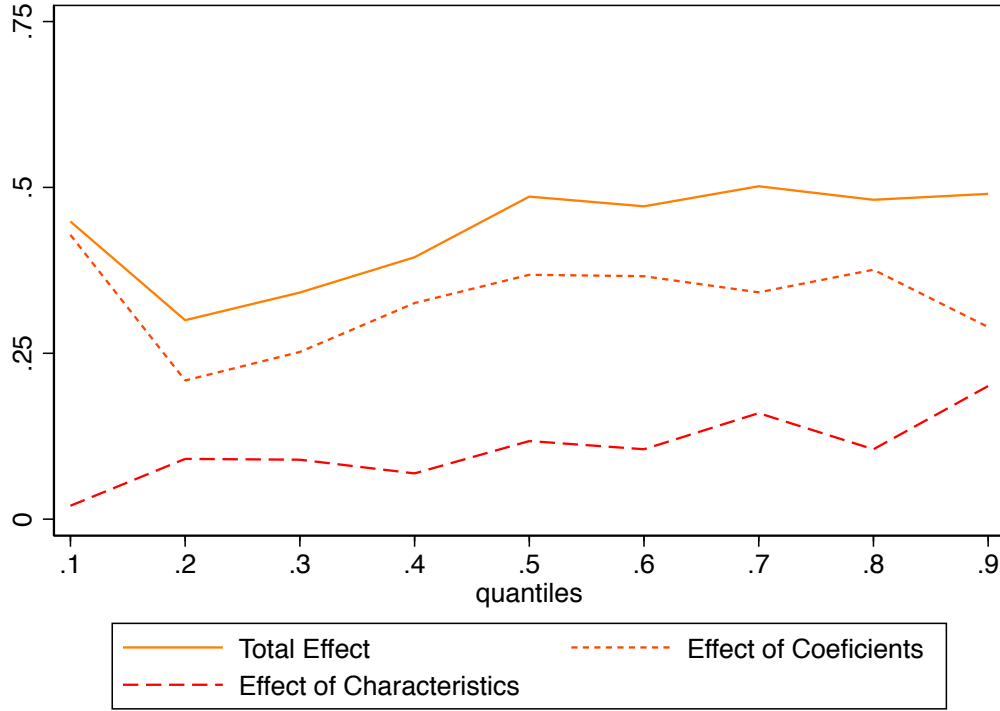
controlling for selection into employment across the quantiles. The dashed line and the solid line stand for the coefficients corrected and uncorrected for selection respectively. The main difference in human capital covariates appears to be at the bottom and upper quantiles. This could mean that the sample selection increases in the tails relative to the median in the distribution.

According to the results, the gender wage gap in Chile after controlling for sample selection is around 50% at the higher levels of the distribution of wage where there is a glass ceiling effect. This result shows that the selection into employment and heterogeneous effects are important and need to be considered any time that the gender wage gap is computed. Working women are a selected sample of women and if the selection is ignored the gap may be an underestimate of the existing difference.

### **III Mata-Machado decomposition**

Additionally, as part of understanding the difference between male and female log wage distribution, I applied the Mata and Machado (2005) technique to the corrected sample to analyze what explains the gender wage gap. Figure 4 shows the part of the gender wage gap that is explained by characteristics and coefficients respectively. Most of the differences in salary are explained by the price effect, that is the return of the labor market characteristics. These results are in line with the structure of pay that is revealed in (table 3 and 5). The return to school for a highly educated worker is relatively more valued for men than for women at the upper quantiles. This situation is reversed for low paying work, where women's achievements are relatively more valued than for men. However, for higher quantiles, characteristics effects play an important role explaining the differences. This could suggest that women may have more education (as Table 1 shows) than men, but those degrees are not well paid.

Figure 4: Mata-Machado decomposition of the gap corrected for sample selection



## V Robustness of the results

To validate the exclusion restriction's assumption of my instrument, I apply the Huber and Mellace (2014) test which reports that there is not a relationship between my excluded regressor and the error term of the outcome equation. Following the literature, I use the number of children under 3 years old as IV. This variable may affect women's decisions to participate in the labor force because it could influence their value of leisure since taking care of a child is time consuming and costly (Maasoumi and Wang (2019)). However, the exclusion restriction for this variable would fail if women with relatively high (low) wage potential had on average lower (higher) fertility (Huber and Mellace (2014)). The proposed test validate two assumptions: i) the existence of at least one variable that affects selection but not the outcome, ii) additive separability of the errors in the selection equation. In his paper, Huber and Mellace (2014) classified the observed values in subgroup: always

selected, compliers, defiers and never selected.<sup>6</sup> Those workers that have a salary can be described as a mixture of “always selected” (those individuals who participate in the labor force regardless of the instrument) and “compliers” (those who their selection state reacts to the instrument), because the “defiers” and the “never selected” are never observed. The two assumptions before mentioned imply two testable inequality constraints that come from point identifying and bounding the outcome distribution of the sub-population that is always selected/observed.

The intuition behind this test is that the outcome distribution of the sub-population of observations that are always takers can be point identified and bounded under the verified exclusion restriction and additive separability assumption. If the point identified does not lie within its bounds, the assumptions are necessarily violated.

Following the procedure suggested by the authors, under the assumption of validity/monotonicity, the following inequalities hold:

$$\begin{aligned}\mathbf{E}[\ln(w)|Z = 1, S = 1, \ln(w) \leq y_q] &\leq \mathbf{E}[Y|Z = 0, S = 1] \\ &\leq \mathbf{E}[\ln(w)|Z = 1, s = 1, \ln(w) \geq y_{1-q}]\end{aligned}$$

where  $y_q$  is the  $q$ th conditional quantile in the conditional outcome distribution when  $Z=1$  and  $S=1$  and  $q$  is the proportion of always selected in the mixed population of the individual with and without children under the age of 3 years old (70% in my sample). This is translated into the following null hypothesis:

$$H_0 : \begin{pmatrix} \mathbf{E}[\ln(w)|Z = 1, S = 1, \ln(w) \leq y_q] - \mathbf{E}[\ln(w)|Z = 0, S = 1] \\ \mathbf{E}[\ln(w)|Z = 1, S = 1, \ln(w)] - \mathbf{E}[\ln(w)|Z = 0, S = 1, Y \geq y_{1-q}] \end{pmatrix} \equiv \begin{pmatrix} \Omega \\ \Omega \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which means that the point identified probability measure of the outcome among always selected given  $Z=0$  and  $S=1$  must lie within the bounds of the conditional probability in the

---

<sup>6</sup>For more details of this procedure, see Huber and Mellace (2014).



mixed population with  $Z=1$  and  $S=1$ . If this is not the case, either the exclusion restriction or monotonicity are necessarily violated. Using their method, I run the test and I find that the standardized mean constraints are negative which means that the inequalities are never binding and a p-value of 0.9669. Therefore, there is not enough statistical evidence to reject the validity of the presence of children under the age of three as an excluded instrument.

## VI Conclusion

An extensive body of literature suggests that there is a difference between the wages of women and men. Nevertheless, the robustness of the results depends on the data set and econometric specification that a researcher uses. The majority of the gender applications that use quantile regression in the presence of non-random sample selection assume that the errors are independent of the regressors given the selection probability. However, as discussed in Huber and Melly (2015), this assumption implies parallel quantile curves across the distribution, limiting the usefulness of heterogeneity in the analysis.

I address this problem by using a new quantile-copula methodology to account for female self-selection into employment and to analyze the gender wage gap between men and women across the distribution of wages in Chile. My results show that women receive significantly less pay than men across the whole wage distribution. Without correction, the gender wage gap oscillated between 10 and 25 log points, which can be interpreted as men having a wage approximately 10% to 25% higher than women. Furthermore, if men's and women's rates of employment were equal, the gender wage gap would be around 25% to 35% at the low levels of the distribution, but it increases toward the upper tail of the distribution gap to a maximum log wage difference of about 50%. The fact that the gender wage gap is larger in the upper quantiles of the hourly wage distribution is evidence of a glass ceiling effect in Chile.

I subsequently decompose the gender wage pay into structural and composition effects.

My results suggest that the differences in wages between female and male workers in Chile can be explained mainly by differences in the rewards for the workers' characteristics, such as education and age, and not by differences in the distribution of those characteristics. It is worth emphasizing that structural effects can be changed through policies that aim to equalize pay structures between female and male workers with the same set of skills.

In general, measuring the gender gap across the distribution has important consequences for gender wage equality. The main goal of my paper is to provide evidence about the inequality in Chile and to highlight how important it is for gender wage gap analyses to consider sample selection into employment and heterogeneity in wages. Accounting for rigorous indicators in gender issues is crucial for promoting gender equality and empowering women, especially in developing countries.

# References

- Albrecht, J., Björklund, A., & Vroman, S. (2003). Is there a glass ceiling in sweden? *Journal of Labor economics*, 21(1), 145–177.
- Arellano, M., & Bonhomme, S. (2017a). Quantile Selection Models With an Application to Understanding Changes in Wage Inequality. *Econometrica*, 85(1), 1–28.
- Arellano, M., & Bonhomme, S. (2017b). Sample Selection in Quantile Regression: A Survey. In R. Koenker, V. Chernozhukov, X. He, & L. Peng (Eds.), *Handbook of quantile regression* (1st ed., p. 463). Chapman and Hall/CRC.
- Arulampalam, W., Booth, A., & Bryan, M. (2005). Is there a glass ceiling over europe? an exploration of asymmetries in the gender pay gap across the wages distribution. *Institute for Social and Economic Research (ISER) Working Paper*, 25.
- Badel, A., & Peña, X. (2010). Decomposing the gender wage gap with sample selection adjustment: Evidence from colombia. *Revista de Análisis Económico*, 25(2), 169–191.
- Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, 436–455.
- Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the usa: a quantile regression approach. *Journal of Applied Econometrics*, 13(1), 1–30. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-1255%28199801%02%2913%3A1%3C1%3A%3AAID-JAE474%3E3.0.CO%3B2-A>
- Christofides, L. N., Polycarpou, A., & Vrachimis, K. (2013). Gender wage gaps, sticky floors and glass ceilings in europe. *Labour Economics*, 21, 86–102.
- De la Rica, S., Dolado, J. J., & Llorens, V. (2008). Ceilings or floors? gender wage gaps by education in spain. *Journal of Population Economics*, 21(3), 751–776.
- Heckman, J. J. (1974). Effects of child-care programs on women’s work effort. *Journal of Political Economy*, 82(2, Part 2), S136–S163.
- Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1), 153–161.
- Heckman, J. J., & MaCurdy, T. E. (1980). A life cycle model of female labour supply. *The Review of Economic Studies*, 47(1), 47–74.
- Huber, M., & Mellace, G. (2014). Testing exclusion restrictions and additive separability in sample selection models. *Empirical Economics*, 47(1), 75–92.
- Huber, M., & Melly, B. (2015). A Test of the Conditional Independence Assumption in Sample Selection Models. *Journal of Applied Econometrics*, 30(7), 1144–1168.

- Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1), 33–50.
- Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.
- Lee, W. K. M. (1998). Gender inequality and discrimination in singapore. *Journal of contemporary asia*, 28(4), 484–497.
- Maasoumi, E., & Wang, L. (2019). The gender gap between earnings distributions. *Journal of Political Economy*, 127(5), 2438–2504.
- Mata, J., & Machado, J. A. F. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20(4), 445–465. Retrieved from <https://ideas.repec.org/a/jae/japmet/v20y2005i4p445-465.html>
- Melly, B. (2006). Estimation of counterfactual distributions using quantile regression.
- Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International economic review*, 693–709.
- Perticara, M., & Astudillo, A. (2010). ¿existen brechas salariales por género en chile? descomposición de las diferencias salariales entre hombres y mujeres en el contexto de regresiones por cuantiles. *Latin American Research Review*, 191–215.
- Perticará, M., & Bueno, I. (2009). Brechas salariales por género en chile: un nuevo enfoque. *Revista de la CEPAL*, 2009(99), 133–149.

## **VII    Appendix**

### **VIII   Labor history**

Each survey wave collects the work history of the interviewee, where he or she gives details of the years and months in which they were under one of the following situations: employment, unemployment, searching for a job for the first time, and inactive. To create a variable that represents the effective labor experience, I took from each wave the number of months when the individual was actively working. For those individuals who have gaps, I assumed that they have zero experience during the period they did not answer (e.g. an individual who answered in the 2006 wave and 2015 wave but was not asked to provided information about labor status between 2006 and 2009). For the first interview which was in 2002, the labor history was covered from the year 1980 and then I extended it using the rest of the waves. For example, the experience variable for a respondent interviewed in 2015 will have a complete labor force history from 1980-2002, then it will be added to the labor history from 2002-2004 from 2004's wave for the case of individuals previously interviewed or asked since 1980 for new interviewees, and the completion continuous similarly with the rest of the years using later waves until the employment history is completed. The survey data has been merged using an identification number that is related to the interviewee.

## IX Copula approach to sample selection models

The copula approach adds more flexibility to model specifications and captures dependence more broadly than a standard multivariate normal framework. The normality assumption is often too strong and it can be the case that the log of wages may have thicker tails than normal distribution implies, for that reason, I modeled sample selection using a Frank copula. According to the Sklar theorem, any multivariate distribution function with continuous margins has a unique copula representation. In this exercise, the Frank copula will generate a joint distribution given marginal distribution from the errors of the sample and outcome equation. One-parameter Frank copula has the following function:

$$\rho^{-1} \log \left\{ 1 + \frac{(e^{-\rho U} - 1)(e^{-\rho V} - 1)}{(e^{-\rho} - 1)} \right\}$$

where the parameter  $\rho$  governs the degree of dependence and it can take values from  $-\infty \leq \rho \leq \infty$ .

Figure 5 shows the contour plot of the Frank copula in different regions of the (U,V) plane. The level curves suggest the magnitude of dependency which goes from low to high, where the smaller ellipse display stronger dependency. In my empirical implementation, the negative correlation indicates positive selection into employment which means that the wage of non-working women are lower than those who are working.

Figure 5: Contour plot of the frank copula

