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

Mariel C. Siravegna[†]

Department of Economics, Georgetown University
Job Market Paper

November 23, 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 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.
[†]Email: 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, & 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 then, 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 (2007); De la Rica, Dolado, and Llorens (2008); and Christofides, Polycarpou, and Vrachimis (2013). Arulampalam et al. (2007) 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. Women do not select themselves randomly to participate in the labor force and as a result, any analysis based solely on working women will yield biased estimates of the

gender wage gap. The importance of accounting for selection in measures of gender gap has been addressed at least since the seminal work of Heckman (1979) and recently highlighted by Maasoumi and Wang (2019).

In the case of Chile, Perticará 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.[1] In another paper, Perticará 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 gap is different across the wage distribution. 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 Protección Social* in Spanish), I study the gender wage gap in Chile by accounting for heterogeneity across the wage distribution and sample selection. My paper is an application of the approach presented in Maasoumi and Wang (2019). Their approach has two main components. First, they correct for selection across the distribution, i.e., not just at the mean. Specifically, I estimate quantile regressions using a new copula-based methodology, as proposed by Arellano and Bonhomme (2017) and applied by Maasoumi and Wang (2019) to the US, to model the joint distribution of the errors 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.[2] To validate the exclusion restriction's assumption in the selection equation, 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

[1] This methodology is similar to Heckman's (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.

[2] See Arellano and Bonhomme (2017)

equation. Second, following Maasoumi and Wang (2019), I apply the Machado and Mata (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: 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 to a maximum log wage difference of about 50%, 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, conditional to 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 as Maasoumi and Wang (2019) have shown in the context of the US. Working women are a selected sample and if selection is ignored the gap may be an underestimate of existing difference.

The rest of the paper is structured as follows: Section II describes the Arellano and Bonhomme (2017) quantile sample selection model and Machado and Mata (2005) counterfactual decomposition. Section III provides details about the data and descriptive statistics. Section IV presents the empirical results of the quantiles regressions for the unselected and selected sample, 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 (2017) 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 (2017) 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:

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

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 β 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:

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

where $1\{.\}$ 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 (2017),[3] 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 τ in (0,1) interval is:

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

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

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

where the numerator is an unconditional copula of *(U, V)* and the denominator is the propensity score. The copula parameter ρ 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 ρ indicates negative selection into employment, whereas negative values suggest positive selection.[4]

As stated by Arellano and Bonhomme (2017), G_x plays an important role in capturing the selection and mapping the rank τ 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 τ-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 τ-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.

[4]Positive selection implies that the wages of non-working women are lower than those who are working.

III Estimation

The Arellano and Bonhomme (2017) 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 desired quantile through rotated quantile regression which preserves the linear programming structure of the standard linear quantile regression (see Koenker & Bassett, 1978).

The first step consists of estimating the propensity score γ by a probit regression:

$$\hat{\gamma} = argmax_a \sum_{i=1}^{N} D_i ln\Phi(Z_i'a) + (1 - D_i) ln\Phi(-Z_i'a) \tag{4}$$

The second step is to estimate the copula parameter ρ 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 τ in the unit interval the following population moment restriction holds:

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

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

$$\hat{\rho} = argmin_c \Big\| \sum_{i=1}^{N} \sum_{l=1}^{L} D_{i\varphi}(\tau_l, Z_i)[\mathbf{1}\{Y_i \leq X_i'\hat{\beta}_{\tau_l}(c)\} - G(\tau_l, \Phi(Z_i'; \hat{\gamma}), c)] \Big\|$$

where $\|.\|$ is the Euclidean norm, $\tau_1 < \tau_2 < \cdots < \tau_L$ is a finite grid on $(0, 1)$, and $\varphi(\tau, Z_i)$ is

a vector of instruments where the dim $\varphi \leq$ dim ρ and:

$$\hat{\beta}_\tau(c) = argmin_{b(\tau)} \sum_{i=1}^{N} D_i[G(\tau, \Phi(Z_i'\hat{\gamma}); c)(Y_i - X_i'b(\tau))^+ + (1 - G(\tau, \Phi(Z_i'\hat{\gamma}); c))(Y_i - X_i'b(\tau))^-]$$

where $a^+ = max\{a, 0\}$, $a^- = max\{-a, 0\}$, and the grid of τ 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:

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

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

IV The Machado and Mata decomposition

The Machado and Mata (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 Albrecht et al., 2003; Blinder, 1973; Oaxaca, 1973) but instead of doing the decomposition at the mean, Machado and Mata (2005) decompose the gap between men and women at the quantiles.

In this empirical exercise, I consider two groups, men and women. Using Albrecht et al. (2003) notation, each group has stochastic vectors associated with the characteristics of each group together with its realizations,x_i for *i=male,female*. Define $G_i(x)$ as the distribution function of the characteristic and $F_i(w)$ as the distribution function for the log wage. The

τ_{th} quantile of the unconditional distribution of the log wage, w_τ is defined by

$$\tau \equiv F(w_\tau)$$

so the difference between the τ_{th} quantiles of the men's and women 's distribution is $w_\tau^m - w_\tau^f$. According to Arellano and Bonhomme (2017), after estimating the parameter $\hat{\rho}$ and the propensity score $\hat{\gamma}$ in the last step of the estimation algorithm, the β_τ 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. 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.

$$F_m(w) = \int F_m(w|x) dG_m(x)$$

and the same for female workers,

$$F_f(w) = \int F_f(w|x) dG_f(x)$$

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

$$F_c(w) = \int F_f(w|x) dG_m(x)$$

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

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

The differences in the distributions $F_m(w) - F_c(w)$ provide insights into the structural ef-

fect and it is the part of the gap due difference in rewards to workers' characteristics. As Maasoumi and Wang (2019) mention in their paper, the structural effects can be target of discrimination 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 due to differences in the distribution of workers' 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 and I deleted those observations who reported working more 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
Education Level			
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
Number of Observations	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
Education Levels			
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[5] of male workers who are out of the labor market, this would generate a lack of precision in wage gap estimates.[6]

IV Empirical results

In this section, I compute the quantile regressions without correcting for sample selection and then correcting for sample selection. Then, I apply the Machado and Mata (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 of the gender wage gap is $w_\tau^m - w_\tau^f$ where w_τ^m and w_τ^f denote the log of wages of male and female workers, respectively, at the corresponding quantile (τ). In the model of sample selection, the independent variables are marital status, effective experience and education

[5] 16% among all the men in the sample are not participating versus 40% of women.
[6] See Badel and Peña (2010).

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. In keeping with 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 female workers but it may influence the probability of participating in the labor market. Maasoumi and Wang (2019) argue that this variable is a valid instrument in the case of the US and it also has been used in other influential papers such as Heckman (1974) and Heckman and MaCurdy (1980). As discussed in Arellano and Bonhomme (2017), 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.

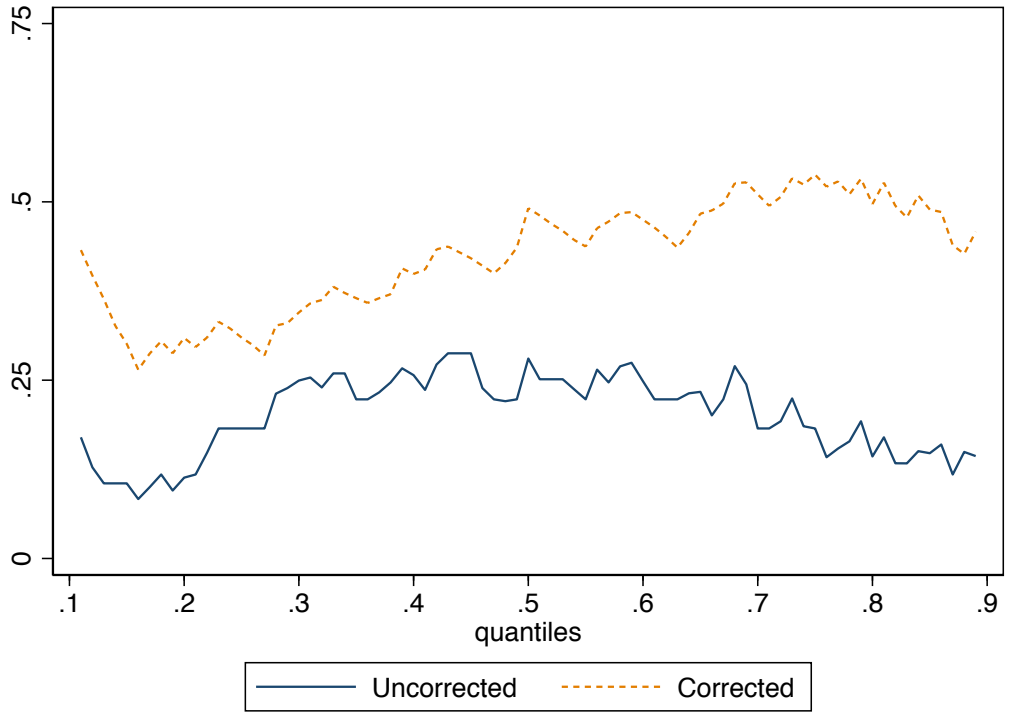
Figure 1 shows the raw gender wage gap without and controlling for selection. Male and female wages are unequal at every point of the wage distribution regardless of the sample selection bias. The gender wage gap without correction oscillated between 10 and 25 log points, which can be interpreted as 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 to a maximum log wage difference of about 50%.

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 increases across the distribution 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

Figure 1: Raw Gender Wage Gap



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 who are in the top of the wage distribution. When returns to 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.

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 is still 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 at all the different quantiles. Being married has a positive coefficient at the tails of the wage distribution for women but it is not statistically significant for the 10th and 90th percentiles. However, for male workers, being married positively affects their wages across the entire distribution. The big difference between the uncorrected and corrected quantile regression for woman seems to be the effect of being married.

Table 3: Quantile regressions for men

VARIABLES	(1) log_wage	(2) log_wage	(3) log_wage	(4) log_wage	(5) log_wage	(6) log_wage	(7) log_wage	(8) log_wage	(9) log_wage
Experience	0.00163*	0.000844*	0.00114**	0.000384	-6.65e-05	0.000142	0.000266	0.000428	-0.000458
	(0.000834)	(0.000482)	(0.000527)	(0.000452)	(0.000453)	(0.000411)	(0.000752)	(0.00102)	(0.00108)
Exp. Squared	-2.96e-06**	-1.24e-06	-1.77e-06*	-5.43e-07	5.37e-07	2.23e-07	1.82e-07	1.05e-07	1.86e-06
	(1.47e-06)	(8.20e-07)	(9.69e-07)	(8.09e-07)	(8.21e-07)	(7.78e-07)	(1.41e-06)	(1.79e-06)	(2.09e-06)
High School	0.145***	0.174***	0.258***	0.287***	0.294***	0.326***	0.309***	0.316***	0.302***
	(0.0561)	(0.0214)	(0.0319)	(0.0252)	(0.0289)	(0.0313)	(0.0519)	(0.0528)	(0.0787)
Technical Education	0.423***	0.460***	0.529***	0.557***	0.533***	0.608***	0.602***	0.655***	0.554***
	(0.0804)	(0.0374)	(0.0516)	(0.0340)	(0.0503)	(0.0491)	(0.0912)	(0.0755)	(0.105)
Bachelor Degree	0.577***	0.785***	0.952***	0.965***	1.029***	1.070***	1.130***	1.316***	1.302***
	(0.119)	(0.0999)	(0.0617)	(0.0415)	(0.0544)	(0.0582)	(0.0993)	(0.139)	(0.131)
Graduate Degree	1.285***	1.323***	1.341***	1.509***	1.432***	1.352***	1.235***	1.312	1.638***
	(0.102)	(0.170)	(0.160)	(0.259)	(0.0413)	(0.0338)	(0.134)	(1.410)	(0.124)
Married	0.0882**	0.110***	0.104***	0.116***	0.104***	0.0965***	0.0842*	0.0772	0.0489
	(0.0419)	(0.0251)	(0.0325)	(0.0253)	(0.0311)	(0.0232)	(0.0447)	(0.0541)	(0.0642)
Constant	6.837***	6.991***	7.026***	7.190***	7.307***	7.366***	7.467***	7.606***	7.991***
	(0.123)	(0.0666)	(0.0678)	(0.0614)	(0.0624)	(0.0507)	(0.102)	(0.139)	(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) log_wage	(2) log_wage	(3) log_wage	(4) log_wage	(5) log_wage	(6) log_wage	(7) log_wage	(8) log_wage	(9) log_wage
Experience	0.00142	0.000703	6.62e-06	0.000554	0.000334	0.00101***	0.00121**	0.00138**	0.00150**
	(0.00109)	(0.000457)	(0.000378)	(0.000347)	(0.000447)	(0.000334)	(0.000514)	(0.000647)	(0.000758)
Exp. Squared	-2.03e-06	-8.60e-07	5.09e-07	-4.55e-07	4.18e-08	-1.14e-06*	-1.79e-06	-1.71e-06	-1.79e-06
	(2.41e-06)	(9.08e-07)	(8.54e-07)	(7.82e-07)	(1.04e-06)	(6.58e-07)	(1.11e-06)	(1.46e-06)	(1.98e-06)
High School	0.283***	0.195***	0.116**	0.117***	0.152***	0.183***	0.164***	0.144*	-0.205*
	(0.0927)	(0.0428)	(0.0453)	(0.0257)	(0.0303)	(0.0356)	(0.0423)	(0.0873)	(0.120)
Technical Degree	0.495***	0.326***	0.279***	0.361***	0.442***	0.497***	0.491***	0.513***	0.113
	(0.109)	(0.0568)	(0.0565)	(0.0619)	(0.0608)	(0.0545)	(0.0781)	(0.104)	(0.115)
Bachelor Degree	0.842***	0.872***	0.849***	0.965***	1.031***	1.062***	1.067***	1.060***	0.591***
	(0.0960)	(0.0850)	(0.0674)	(0.0523)	(0.0445)	(0.0393)	(0.0592)	(0.0902)	(0.115)
Graduate Degree	1.488***	1.273***	1.238***	1.267***	1.326***	1.399***	1.492***	1.421***	1.028***
	(0.129)	(0.0628)	(0.102)	(0.0684)	(0.0976)	(0.238)	(0.177)	(0.112)	(0.384)
Married	0.0901	0.0599**	0.0284	0.0529**	0.0325	0.0592*	0.0782**	0.0710	0.129**
	(0.0648)	(0.0258)	(0.0240)	(0.0268)	(0.0284)	(0.0334)	(0.0344)	(0.0482)	(0.0537)
Constant	6.512***	6.874***	7.090***	7.097***	7.168***	7.168***	7.272***	7.402***	8.010***
	(0.136)	(0.0650)	(0.0568)	(0.0339)	(0.0475)	(0.0400)	(0.0629)	(0.0998)	(0.127)
Observations	2,233	2,233	2,233	2,233	2,233	2,233	2,233	2,233	2,233

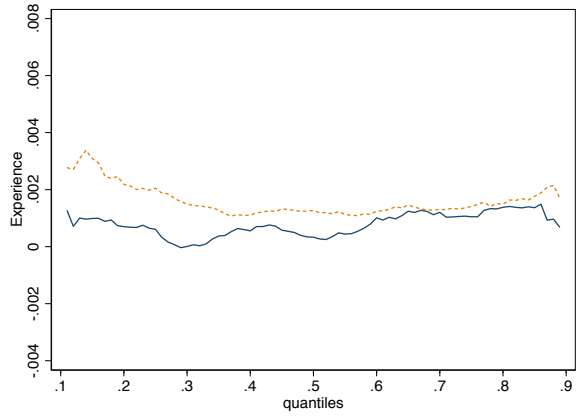
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5: Quantile regressions for women with corrections for selectivity

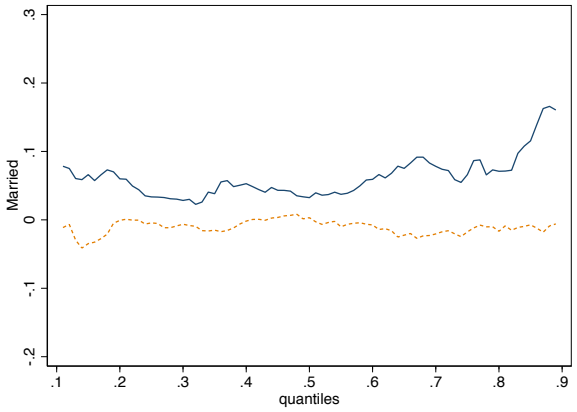
VARIABLES	(1) log_corrected	(2) log_corrected	(3) log_corrected	(4) log_corrected	(5) log_corrected	(6) log_corrected	(7) log_corrected	(8) log_corrected	(9) log_corrected
Experience	0.0033	0.0022	0.0015	0.0011	0.0013	0.0012	0.0013	0.0015	0.0018
	(0.0023)	(0.0019)	(0.0016)	(0.0012)	(0.0011)	(0.0010)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

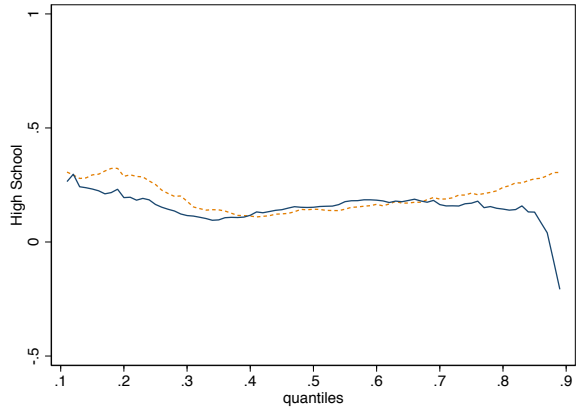
Figure 2: Quantile Regression Estimates



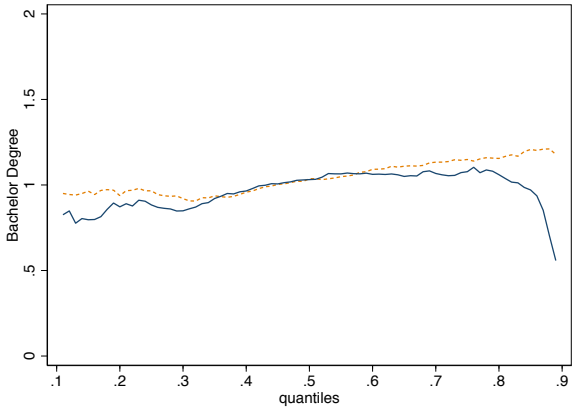
(a) Experience



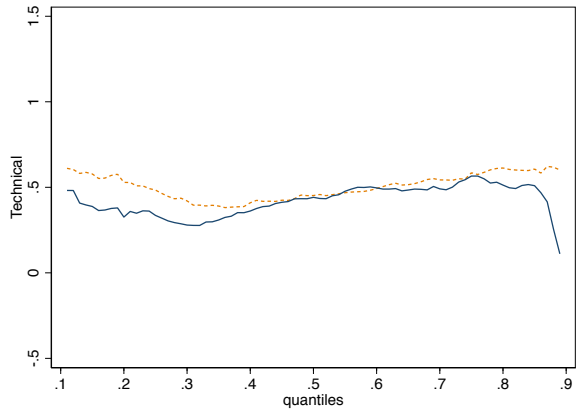
(b) Married



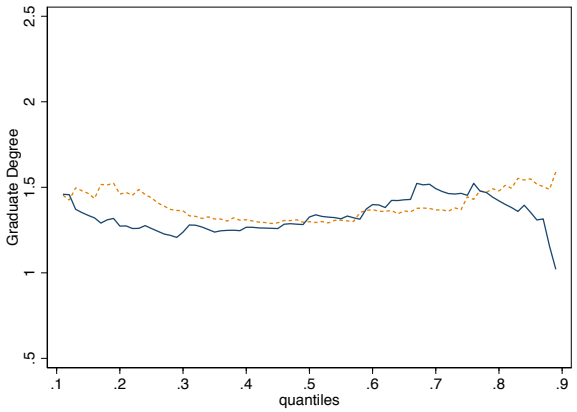
(c) High School



(d) Bachelor Degree



(e) Technical Degree



(f) Graduate Degree

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

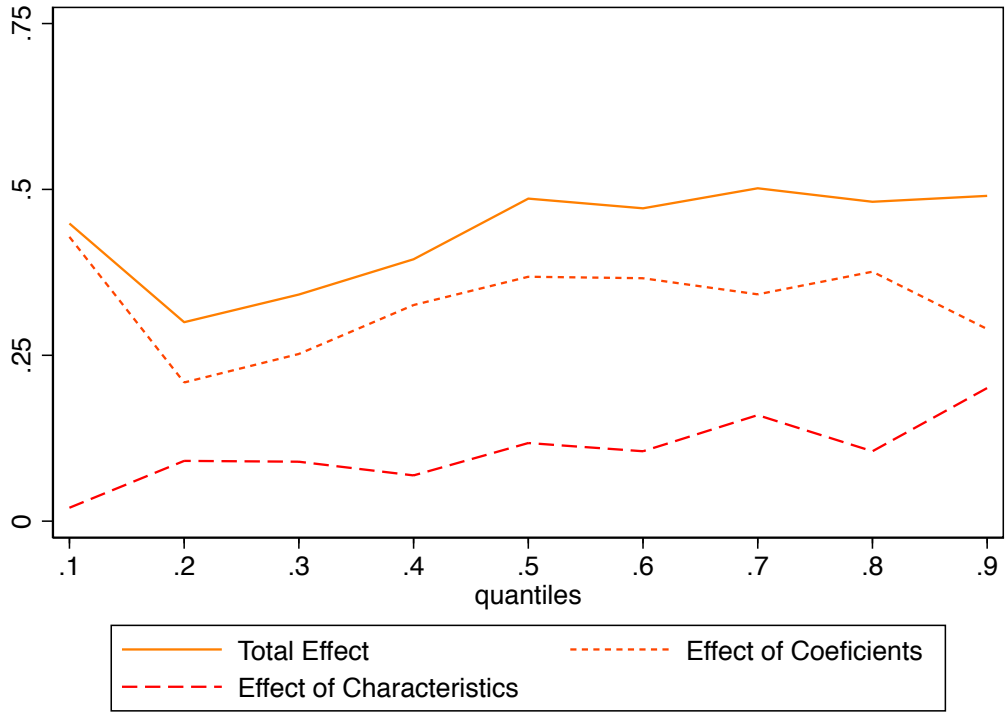
Figure 2 shows the coefficients of the quantile regression for women before and after controlling for selection into employment across the quantiles. The dashed line and the solid line stand for the coefficients corrected and uncorrected for selection respectively. The main difference in human capital 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 wages where there is a glass ceiling effect. This result shows that the selection into employment and heterogeneous effects are important and need to be consider any time that the gender wage gap is computed. Working woman are a selected sample and if the selection is ignored the gap may be an underestimate of the existing difference.

III The Machado and Mata decomposition

To better understand the difference between male and female log wage distribution, I applied the Machado and Mata (2005) technique to the corrected sample to decompose 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 structural effect, which is the return to the labor market characteristics. This result is in line with the structure of pay that is revealed in Tables 3 and 5. The return to high level of education is higher for men than for women at the upper quantiles. This situation is reversed for low paying work, where women's education is more valued than men. However, for higher quantiles, characteristic effects play an important role explaining the differences. This could suggest that, While women are more highly educated than men (as Table 1 shows), their degrees are not as well rewarded.

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



V Validity of Assumptions

To validate the exclusion restriction's assumption, I apply the Huber and Mellace (2014) test which reports that there is not a significant relationship between my excluded regressor and the error term of the outcome equation. This test is also applied by Maasoumi and Wang (2019) in similar context.

Huber and Mellace (2014) propose a test that validates two assumptions: i) the existence of at least one variable that affects selection but not the log wage, ii) additive separability of the errors in the selection process. To do that, they classify the population into four subgroup according to the reaction of selection to the instrument (Z). Using similar notation, the types are the following: "always selected', those individuals who participate in the labor force regardless of the instrument,"compliers" those observations that are selected under Z=1 but not under Z=0,"defiers" are selected under Z=0 but not under Z=1 and the never

selected who outcome is never observed.[7] Denote S(z) as an observable binary variable which indicates the potential selection state and it takes values of 1 if the log wage is observed, 0 otherwise. According to Huber and Mellace (2014), the workers who perceive a salary can be described as a mixture of always selected and compliers. The log wage for the other two types, defiers and never selected, is never observed (i.e., they are not employed).

The intuition behind this test is that, under the verified exclusion restriction and additive separability assumption, the point identified outcome distribution of the always takers in absence of the instrument lies within the bounds in the presence of the instrument, which implies two inequality constraints that can be tested.

Following the procedure suggested by the authors, under the assumption of validity/monotonicity, the following inequalites 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 qth 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 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

[7] For more details of this test, see Huber and Mellace (2014).

the standardized mean constraints are negative which means that the inequalities are never binding with 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 3 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. However, after correcting for selection, the gender wage gap is around 25% to 35% at the lower quantiles, and the gap is larger in the higher quantiles up 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 also decompose the selection-corrected 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 across wages distribution. 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. (2017). Quantile Selection Models With an Application to Understanding Changes in Wage Inequality. *Econometrica*, *85*(1), 1–28.

Arulampalam, W., Booth, A. L., & Bryan, M. L. (2007). Is there a glass ceiling over europe? exploring the gender pay gap across the wage distribution. *ILR Review*, *60*(2), 163–186.

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.

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.

Fort, L., John-Abraham, I., Orlando, M. B., & Piras, C. (2007). Chile-reconciling the gender paradox.

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.

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.

Machado, J. A., & Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics*, *20*(4), 445–465.

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.

Perticará, 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 and so on, 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:

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

where the parameter ρ 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 ellipses 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