universität wien

MAGISTERARBEIT | MASTER'S THESIS

Titel | Title

Quantile Regression with Censored Selection to examine the Gender Pay Gap in Austria

verfasst von | submitted by

Anna Ahaieva BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Magistra der Sozial- und Wirtschaftswissenschaften (Mag.rer.soc.oec.)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 951
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Magisterstudium Statistik
Betreut von | Supervisor:	Mag. Dr. Leopold Sögner

Abstract

Diese Arbeit untersucht den Gender Wage Gap in Österreich aus einer verteilungsbasierten Perspektive und berücksichtigt die Rolle der Selektion in Beschäftigung. Da Löhne nur für erwerbstätige Personen beobachtet werden, können übliche Vergleiche ein verzerrtes Bild liefern, wenn Erwerbstätigkeit mit potenziellen Löhnen zusammenhängt.

Die Analyse basiert auf dem Quantil-Selektionsansatz von Arellano und Bonhomme. Empirisch verbindet die Arbeit eine an österreichische Arbeitsmarktmerkmale angelehnte Simulationsstudie mit deskriptiver Evidenz aus veröffentlichten Lohnquantilen von Statistik Austria für das Jahr 2022.

Die Ergebnisse zeigen, dass Selektion die beobachteten geschlechtsspezifischen Lohnunterschiede über die Lohnverteilung hinweg verzerren kann und dass sich diese Unterschiede deutlich nach Quantil, Branche und Beschäftigungsform unterscheiden. Insgesamt legt die Arbeit nahe, dass sich der Gender Wage Gap nicht angemessen durch ein einziges Durchschnittsmaß zusammenfassen lässt.

Abstract

This thesis studies the gender wage gap in Austria from a distributional perspective and considers the role of selection into employment. Since wages are observed only for employed individuals, standard comparisons may give a distorted picture if employment is related to potential wages.

The analysis is based on the quantile-selection framework of Arellano and Bonhomme. Empirically, the thesis combines a simulation study calibrated to Austrian labour market features with descriptive evidence from published Statistik Austria wage quantiles for 2022.

The results show that selection can distort observed gender wage gaps across the wage distribution and that these gaps differ substantially by quantile, industry, and employment type. Overall, the thesis suggests that the gender wage gap cannot be adequately summarized by a single average measure.

Keywords: gender pay gap; quantile regression; sample selection; inverse probability weighting; copula; Austria

Contents

1 Introduction 4

2 Statistical modeling of the gender pay gap in Austria 5

3 Simulation study - assessment of model performance using simulated data (Monte Carlo simulation) 9

4 Benchmark application of the quantile-selection method 13

5 Analysis of the real data published by Statistik Austria 14

6 Results 18

7 Conclusion 28

8 Appendix 30

1 Introduction

The gender pay gap remains an important economic and social issue in Europe and in Austria. Even in 2022, women in Austria earned 18.4% less than men on average in terms of gross hourly wages. Compared with the EU average of about 12.7%, Austria belongs to the group of countries with a comparatively high gender pay gap [Ös24]. According to Statistik Austria, women perform almost two-thirds (65.1%) of unpaid work such as housework, childcare, volunteer work, or care for relatives, and about every second woman in Austria works part-time. These differences in labour-market attachment are likely to affect both current wages and future pension outcomes.

The gender pay gap is most often reported as an *unadjusted* indicator, namely the difference between men's and women's average gross hourly earnings expressed as a percentage of men's average gross hourly earnings. Such a measure is useful as a summary indicator, but it does not isolate a pure "discrimination" effect. Instead, it reflects both differences in observable characteristics, such as occupation, sector, working time, or career interruptions, and differences in the returns to these characteristics [Eur25].

More generally, non-random sample selection is a major issue in empirical work [Hec79, AB17]. In the labour-market context, wages and employment provide a classic example: wages are observed only for employed individuals, so conventional measures of wage inequality or wage gaps may be biased if employment participation is related to potential wages [AB17]. This problem is especially relevant when comparing men and women, since labour-market participation, part-time work, and family-related employment interruptions differ systematically by sex.

An additional limitation of standard gender pay gap measures is that they focus on averages. Much of the literature on sample selection has historically concentrated on conditional mean models [Hec79], while in many applications the whole distribution of outcomes is of interest [AB17]. Quantile regression, introduced by Koenker and Bassett [KB78], provides a natural way to study such distributional heterogeneity. In the context of gender wage inequality, this is important because the gap may differ substantially between the lower tail, the centre, and the upper tail of the wage distribution. Mean-based comparisons may therefore hide relevant differences across worker groups and across parts of the distribution. Quantile-based methods have become an important tool for studying distributional changes in wages and inequality [Buc94, Buc98].

However, quantile analysis in the presence of sample selection is methodologically challenging. Earlier approaches to selection correction in quantile models relied on strong assumptions, such as conditional independence and additivity, that may be restrictive in practice [HM15]. Arellano and Bonhomme [AB17] propose a more general framework in which selection is modeled through the dependence between latent wage ranks and participation ranks using a copula representation. Their approach allows selection corrections that are not simply additive, but may instead nonlinearly rotate observed distributional ranks [AB17].

This thesis studies the gender pay gap in Austria from a distributional perspective and explicitly accounts for selection into employment. In particular, the aim is to analyze how the wage gap varies across quantiles of the wage distribution and to assess how far observed wage differences may be distorted by non-random employment selection. Methodologically, the thesis builds on the quantile-selection framework of Arellano and Bonhomme [AB17] and adapts it to the Austrian setting.

The empirical analysis proceeds in three parts. First, a Monte Carlo study calibrated to stylized Austrian labour-market features is used to examine how strongly selection can distort observed quantile gaps and how well different estimators recover the true population gap under controlled scenarios. Second, a benchmark application based on a known reference dataset is used as an implementation check for the estimator. Third, descriptive evidence based on published Statistik Austria [Ös24] wage quantiles is used to document how gender wage gaps vary across industries, quantiles, and employment types.

The main contribution of the thesis is therefore twofold. On the one hand, it applies a modern quantile-selection framework to the question of gender wage inequality in Austria. On the other hand, it emphasizes that the gender pay gap should not be viewed as a single average number, but as a distributional object that may vary substantially across quantiles, industries, and employment groups.

Structure of the thesis:

Section 1 Introduction

Section 2 Statistical modelling of the gender pay gap in Austria - formal presentation of the quantile model with selection adjustment (Arellano & Bonhomme [AB17]) adapted to our case.

Section 3 Simulation study - assessment of model performance using simulated data (Monte Carlo simulation).

Section 4 Benchmark application of the quantile-selection method.

Section 5 Analysis of the real data published by *Statistik Austria*.

Section 6 Results - presentation and interpretation of empirical results on the gender pay gap across quantiles, industries, and employment types.

Section 7 Conclusion - summary of the results.

2 Statistical modeling of the gender pay gap in Austria

To study the gender pay gap in Austria, we use the quantile regression method developed in [AB17] by Arellano M. and Bonhomme S., which allows us to analyze salary differences across the entire distribution, while taking into account the fact that salaries are only counted for employed people.

Here we use the model as in Chapter 2.1 in [AB17].

$$Y^* = q(U, X) \tag{1}$$
$$D = 1\{V \leq p(Z)\} \tag{2}$$
$$Y = Y^* \quad \text{if } D = 1 \tag{3}$$

In our case:

- Y^* is *potential* log hourly wage of Austrian men and women, across industries and employment types. In other words, it is what a person can earn if working.

- Y observed wage - the actual log-hourly wage.
- q(U,X) is the structural quantile function: it tells us what a person's potential hourly wage Y^* would be if their rank in the wage distribution were U (e.g., the 25th, 50th, or 90th percentile) and their observed characteristics were X (gender, part-time, industry, ...). Evaluating q at a quantile index τ yields the τ-th conditional quantile of potential wages,

  $$Q_{Y^*|X}(\tau \mid X) \;=\; q(\tau, X).$$

  In our simulation, we specify $q(\tau, X)$ flexibly so that the effect of gender can vary by quantile. Because we only observe wages when $D = 1$ (employed), the naïve estimators of $q(\tau, X)$ based on the employed sample are generally biased. The IPW (inverse probability weighting) and AB (Arellano and Bonhomme) procedures correct for this selection by reweighting or rank relabeling among the employed, so that we can consistently recover $q(\tau, X)$ and, in particular, the gender wage gap

  $$q(\tau, X \mid \text{female} = 1) - q(\tau, X \mid \text{female} = 0) \tag{4}$$

  across the wage distribution.
- $D \in \{0, 1\}$: $D = 1$ means the individual is observed in our data set of employees either full-time or part-time, while $D = 0$ corresponds to the population not in dependent employment (self-employed, unemployed, inactive).
- $p(Z)$ measures the probability of employment in each subgroup. For example, a woman with young children may have a lower value of $p(Z)$ that shows a lower probability of employment.

  $$p(Z) = Pr(D = 1|Z) \qquad \text{(propensity score)}$$

- X are observed factors that influence salaries. According to our data, these factors include:
  - Gender (male/female),
  - Employment status (full-time or part-time),
  - Sector of activity (ÖNACE 2008 categories).
- $Z = (B, X)$. In the participation equation (1) we collect all variables that can influence whether someone is employed into one vector $Z = (B, X)$. Here, X collects observed wage covariates used in all quantile models (e.g., *female* indicator, *part-time* status, and the full set of industry dummies at the NACE section level). By contrast, B contains *exclusion shifters*: variables that plausibly affect employment but, conditional on X, do not change a worker's rank U in the potential wage distribution. Typical examples include instruments derived from potential non-employment income or benefits, partner-income proxies, child-care access, or local labor-demand shifters measured at a fine geographic or sectoral level. The spirit of the paper is that B should be credible for selection, but not directly affect wages once we condition on X.

  We parameterize the propensity score as

  $$p(Z) \;=\; \Pr(D = 1 \mid Z) \;=\; \Lambda\Big(\alpha + X^\top \gamma_X + B^\top \gamma_B\Big), \tag{5}$$

where $\Lambda(t) = \frac{e^t}{1+e^t}$ is the logistic link (a probit link $\Phi(\cdot)$ would be equivalent for our purposes). Thus X and B may both shift employment, but we require B to be excluded from the wage ranks.

- U (unobserved wage rank): A person's position in the *potential* wage distribution given X. Low U means a lower potential wage, high U a higher potential wage. The structural quantile function $q(\tau, X)$ returns the τ-th potential wage, so evaluating at $U = \tau$ yields $Q_{Y^*|X}(\tau \mid X) = q(\tau, X)$.

- $V \in (0, 1)$ (unobserved participation rank) indexes a person's latent propensity to be employed given (B, X). Employment is observed when the draw V falls below the propensity score

$$D \;=\; \mathbf{1}\{\, V \le p(Z)\,\}, \qquad p(Z) = \Pr(D = 1 \mid Z),$$

  so for a fixed Z a lower value of V makes employment more likely, while a larger $p(Z)$ shifts the employment threshold up and increases the probability of $D = 1$. We model $p(Z)$ with a standard binary-response link (logit/probit).

- Link between U and V: We allow wages and participation to be statistically dependent by specifying a copula for the pair of ranks:

$$(U, V) \sim C_\rho, \qquad (U, V) \sim \mathrm{Unif}(0, 1),$$

  where ρ summarizes the strength and direction of rank dependence (Gaussian copula in our simulation). When U (potential wage rank) and V (participation rank) are statistically dependent, the sample of employed people corresponds in composition to the general population. In words, $\rho > 0$ (our "positive selection" case) means that individuals with higher potential wage rank U tend to draw V's that more often satisfy $V \le p(Z)$, hence are observed among the employed; $\rho < 0$ ("negative selection") means the opposite. Because we only see Y when $D = 1$, this dependence generates selection bias in naive quantile comparisons within the employed sample.

  Under a copula C_ρ, observed (selected) wage quantiles can be related to latent quantiles through a rank-translation map $G(\tau, p; \rho)$:

$$Q_{Y|X,D=1}(\tau \,|\, X, Z) \;=\; q(G(\tau,\, p(Z); \rho),\, X,)\,. \tag{6}$$

  IPW (the inverse probability weighting) tackles selection by reweighting with $1/p(Z)$, while AB (Arellano and Bonhomme) procedure inverts the translation G ("rank relabeling") to recover the population quantiles $q(\tau, X)$ from the employed sample.

Following [AB17], we use four key assumptions for the model. We describe these assumptions and see if they also make sense in our case in the context of the gender pay gap in Austria using data from *Statistik Austria*.

A1. Exclusion restriction: Conditionally on X, the pair (U, V) is independent of Z, namely

$$(U, V) \perp\!\!\!\perp \mathbb{E}\Big[Z|X\Big].$$

Empirically, B can be taken as potential out-of-work benefits.

A2. Unobservables: (U,V) have uniform marginals and a (possibly X-dependent) copula. Dependence between U and V is exactly the selection problem.
In our case this is a normalization (ranks are uniform) and allows flexible dependence between wage rank U and participation rank V through a copula. Dependence ($\rho \neq 0$) *is* the selection channel we want to correct.
Model (1)–(3) depends on two sources of unobserved heterogeneity: the latent outcome rank U and the percentile rank V. In Assumption A2, we normalize their marginal distributions to be uniform on the unit interval, independent of Z.

A3. Continuous outcomes - wages are treated as continuous, although Statistik Austria presents them as grouped statistics (means and quartiles). This assumption ensures clarity and growth of the quantile functions. In practice, this assumption is valid, since the distribution of wages in large administrative data sets behaves as a continuous variable.

A4. *The propensity score satisfies* $p(Z) = \Pr(D = 1 \mid Z) > 0$ *with probability 1.*
In our case: every Z-cell has some employed workers (no structural zero participation), which is the usual overlap condition for selection models and is needed for both identification and for IPW (inverse probability weighting) weights to behave. A Z-cell is a subset of the sample with identical values of $Z = (X, B)$ (e.g., female, part-time, industry = CJ (ÖNASE 2008)).
Every Z-cell (i.e. a set of observations sharing the same value of $Z = (X, B)$, such as sex × part-time status × industry) contains employed workers, so the propensity score $p(Z) = \Pr(D = 1 \mid Z)$ is bounded away from zero. This is the usual *overlap* requirement in selection models and is needed both for identification and for inverse-probability weights $1/p(Z)$ to behave numerically.

Under A1-A4, quantiles in the selected sample ($D = 1$) can be related to latent quantiles via a *rank-translation map* $G_x(\tau, \rho) = \Pr\big(U \leq \tau \mid V \leq p, X\big)$, where $p = p(Z)$ and ρ indexes the copula of(U, V). With excluded variation in B, this link allows identification (or bounds) of $q(\tau, x)$. Operationally, the estimator in [AB17]:

- estimates $p(Z)$ (e.g., logit/probit) to model employment;
- fits the dependence parameter(s) ρ of the copula for (U, V);
- runs quantile regression on employment ($D = 1$) but *relabels* each observation's target quantile from τ to $G\{\tau, \hat{p}(Z_i; \hat{\rho}\}$), yielding selection-corrected $\hat{q}(\tau, x)$.

Two complementary strategies follow:

1. IPW (inverse probability weighting): under selection of Z, reweight observed wages by $1/p(Z)$ to target population quantiles. This does not use dependence between U and V beyond Z.
2. AB (rank relabeling - Arellano and Bonhomme): estimate $p(Z)$ and the copula parameter(s) ρ, invert G to relabel employed-sample ranks, and evaluate $q(\tau, X)$ at the corrected quantile indices. This explicitly corrects rank selection.

Results rely on credible excluded variation in B (A1) and overlap (A4); finite-sample performance can be fragile in the tails and under weak excluded variation, that is, when B induces only limited variation in $p(Z)$. We therefore accompany point estimates with diagnostics and bounds.

3 Simulation study - assessment of model performance using simulated data (Monte Carlo simulation)

In the Austrian microdata, the wage we want to study, Y, is defined only for employed individuals. If the probability of employment depends on factors also related to potential wages, a simplistic comparison of women's and men's wages in a sample of employed individuals could be misleading. We evaluate the methodology using simulated data calibrated to stylized Austrian labour-market features (sector structure, underemployment, gender composition). The goal is to understand the extent of selection bias and the ability of the estimators (IPW and Arellano-Bonhomme, AB), which account for it to reconstruct the true gender wage gap across the entire wage distribution.

Let

$$Y^* \text{ be latent log hourly wages}, \quad D \in \{0,1\} \text{ employment}, \quad Y = Y^* \text{ if } D = 1.$$

Let X collect observed controls (female, part–time, industry), and let $Z = (X, B)$ denote the covariates that determine selection, where B is an exclusion shifter. Define the propensity score

$$p(Z) \equiv \Pr(D = 1 \mid Z).$$

For each Monte Carlo simulation, a large synthetic population is drawn with:

- **Demography:** the simulated population is calibrated to stylized Austrian labour-market features rather than exact sample moments. In particular, we set the female share to about 48%, allow part-time work to be substantially more frequent among women, and choose broad industry shares of roughly 30% manufacturing, 45% services, and 25% other. These values are not estimated directly from Austrian microdata; instead, they are simple calibration choices intended to reflect the broad composition of the Austrian labour market while keeping the Monte Carlo design parsimonious.

- **Wages:** the structural quantile function $Y^* = q(U, X)$ depends on a latent rank $U \sim \mathsf{Unif}(0,1)$ via a flexible polynomial in U, allowing the gender gap to vary between quantiles. We specify a flexible, monotone structural quantile function

  $$q(u, X) \;=\; b(u)^\top \theta(X) \qquad u \in (0,1)$$

  where $b(u)$ is a basis (Bernstein/B-spline or polynomial) in u and $\theta(X)$ is linear in X (female, part-time, industry dummies and interactions). For a cubic polynomial example,

  $$q(u, X) = \beta_0(X) + \beta_1(X)u + \beta_2(X)u^2 + \beta_3(X)u^3,$$

  with $\beta_k(X)$ linear in X. We impose that $u \mapsto q(u, X)$ is nondecreasing for every X, so that $q(\cdot, X)$ is a valid conditional quantile function. In practice, this monotonicity can be enforced either by using a Bernstein basis with coefficients constrained to be ordered, or by applying a monotone rearrangement to the fitted curve.

Evaluating the structural quantile function at $u = \tau$ yields the conditional potential quantile

$$Q_{Y^*|X}(\tau \mid X) = q(\tau, X).$$

Let $F \in \{0, 1\}$ denote the female indicator and write $X = (F, X_f)$, where X_f collects the remaining covariates (e.g. part-time status, industry dummies, and interactions). Hence, the true gender gap at quantile τ is

$$\text{Gap}^{true}(\tau; X_f) = q(\tau, F = 1, X_f) - q(\tau, F = 0, X_f). \tag{7}$$

Because $q(\tau, \cdot)$ depends on τ and on covariates, the gap can vary across quantiles and across groups (e.g. part-time vs. full-time, industries).

- **Selection:** employment is generated from a logit in $Z = (X, B)$. The shifter B (e.g., proxies for childcare access/benefit eligibility) affects D but, by assumption, not wage ranks given X.
- **Rank-selection dependence:** the pair (U, V) follows a Gaussian copula with correlation ρ (where V indexes the selection equation). We consider three scenarios:
  1. **S0_Base**: no rank-selection ($\rho = 0$);
  2. **S1_PosSel**: positive rank-selection ($\rho > 0$);
  3. **S2_NegSel**: negative rank-selection ($\rho < 0$).

In the selection equation (5) we parameterize the propensity score as

$$p(Z) = \Lambda\Big(\alpha + X^\top \gamma_X \; + \; s \cdot B^\top \gamma_B\Big), \qquad \Lambda(t) = \frac{e^t}{1 + e^t},$$

where B are exclusion shifters. The scalar s controls the strength of the instrument:

- base: $s = 1$ (benchmark strength). B has a sizeable effect on D, creating clearer variation in employment at given X and generally better overlap for identification.
- weak: $s = \kappa$ with $0 < \kappa < 1$. Specifically, we set $s = 0.5$ in the weak case, so the weak design simply attenuates the effect of the excluded shifter B in the participation equation. B shifts D less, so the employed and non-employed look more similar in terms of $p(Z)$; overlap is poorer and methods relying on B (IPW/AB) face more sampling noise and potential bias.

Thus, in the base case the excluded factor noticeably affects employment probabilities, whereas in the weak case it shifts them only slightly.

Each rank-selection scenario is evaluated under two instrument-strength levels (*weak* vs. *base*) by changing the effect of B on D. For each scenario and τ we report

$$\text{Bias}(\tau) = \mathbb{E}\Big[\widehat{\text{Gap}}(\tau) - \text{Gap}^{\text{true}}(\tau)\Big], \tag{8}$$

$$\text{RMSE}(\tau) = \sqrt{\mathbb{E}\Big[\Big(\widehat{\text{Gap}}(\tau) - \text{Gap}^{\text{true}}(\tau)\Big)^2\Big]}, \tag{9}$$

where $\text{Gap}^{\text{true}}(\tau)$ is the difference between the female and male *structural* τ-quantiles of potential wages in the full population (not just the employed). In real data this object is unknown; in the MC design it is known from the data-generating process (DGP).

Given an estimator $\mathcal{E}$ (Observed, Naive QR, IPW, AB - see definitions later in this chapter), let $\widehat{\text{Gap}}_{\mathcal{E}}(\tau)$ denote the *estimated* gender gap at quantile τ produced by method $\mathcal{E}$ from the simulated (or empirical) sample. In the simulations, we summarize accuracy by

$$\text{Bias}_{\mathcal{E}}(\tau) = \mathbb{E}\Big[\widehat{\text{Gap}}_{\mathcal{E}}(\tau) - \text{Gap}^{\text{true}}(\tau)\Big], \qquad \text{RMSE}_{\mathcal{E}}(\tau) = \sqrt{\mathbb{E}\Big[\Big(\widehat{\text{Gap}}_{\mathcal{E}}(\tau) - \text{Gap}^{\text{true}}(\tau)\Big)^2\Big]}.$$

We also record the overlap of $\hat{p}(Z)$ between $D=1$ and $D=0$, percentiles of IPW weights and effective sample size, the share of QR fallbacks at extreme τ, the average width and coverage of the cellwise Horowitz-Manski quantile bounds $\Big[\text{LB}(\tau), \text{UB}(\tau)\Big]$ (10).

To clarify the notation, let $p = \Pr(D=1 \mid Z=z)$ and $F_1(y) = \Pr(Y \le y \mid D=1, Z=z)$ be the CDF of observed wages in a given Z-cell. With no assumptions about wages when $D=0$, the population CDF at $Z=z$ is bounded by

$$F^L(y) = p, F_1(y) \quad \le \quad F(y \mid Z=z) \quad \le \quad F^U(y) = p, F_1(y) + (1-p),$$

where the function F^L is the *worst-case* (lower) bound for the population CDF: it assumes that all wages for non-employed individuals ($D=0$) lie *above* y, so they contribute 0 to $\Pr(Y \le y)$; and F^U is the *best-case* (upper) bound: it assumes that all wages for non-employed individuals ($D=0$) lie *below* y, so they contribute 1 to $\Pr(Y \le y)$.

The corresponding *Horowitz-Manski* quantile bounds are

$$\text{LB}(\tau) = \inf\{y : F^U(y) \ge \tau\}, \qquad \text{UB}(\tau) = \inf\{y : F^L(y) \ge \tau\}. \tag{10}$$

Intuitively, $\text{LB}(\tau)$ assumes all non-employed would have wages *below* y (worst case for the quantile), while $\text{UB}(\tau)$ assumes they would all be *above* y (best case).

Next, for each $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ we compute five objects:

1. **Observed (selected).** Difference of empirical τ–quantiles among $D=1$:

$$\widehat{\text{Gap}}^{\text{obs}}(\tau) = Q_\tau(Y \mid \texttt{female}{=}1, D{=}1) - Q_\tau(Y \mid \texttt{female}{=}0, D{=}1).$$

   This treats the employed as representative and *ignores selection*.

2. **Naive QR (with X).** Koenker-Bassett QR on the employed sample

$$Q_\tau(Y \mid X, D{=}1) = \beta_0(\tau) + \beta_f(\tau)\,\texttt{female} + \beta_X(\tau)^\top X,$$

   and we take $\widehat{\text{Gap}}^{\text{nq}}(\tau) = \hat{\beta}_f(\tau)$. This controls for X (female, part-time, industry) but still *uses only* $D{=}1$ and thus inherits selection bias when present.

3. **IPW (inverse probability weighting).** Estimate $\hat{p}(Z) = \Pr(D{=}1 \mid Z)$ by logit in (X, B); for $D{=}1$ units use weights $w_i = 1/\max\{\hat{p}(Z_i), \varepsilon\}$ with a small floor ε for stability, and compute *weighted* female and male τ-quantiles. The IPW gap $\widehat{\text{Gap}}^{\text{ipw}}(\tau)$ targets the *population* quantile gap under the exclusion $(U, V) \perp Z \mid X$.

4. **AB (oracle).** Arellano-Bonhomme rank-relabeling using the *true* copula parameter ρ:
   (a) invert $q(\cdot \mid X)$ to obtain a feasible $\widehat{U}_i \in [0,1]$ with minimal reconstruction error,
   (b) compute the selected rank r_i within (part-time × industry) cells and p–bins,
   (c) map $r_i \mapsto \widehat{\tau}_i$ using an interpolated grid of $r(\tau, p; \rho)$ implied by the copula,
   (d) for each target τ, run a kernel-weighted QR of Y on (female, X) with weights $K_h(\widehat{\tau}_i - \tau)$; the coefficient on female is $\widehat{\text{Gap}}^{\text{ab}}(\tau)$.
5. **Bounds.** Within each (part-time × industry) cell compute Horowitz-Manski-type lower/upper quantile bounds $\text{LB}(\tau), \text{UB}(\tau)$ and average across cells; these require weaker assumptions (no copula) and deliver a plausible range for the gap.

Each scenario-instrument pair is repeated $K = 200$ times with $N = 120000$ per replication. For every estimator and quantile we compute

$$\text{Bias}(\tau) = \mathbb{E}\Big[\widehat{\text{Gap}}(\tau) - \text{Gap}^{\text{true}}(\tau)\Big], \qquad \text{RMSE}(\tau) = \sqrt{\mathbb{E}\Big[(\widehat{\text{Gap}}(\tau) - \text{Gap}^{\text{true}}(\tau))^2\Big]}.$$

We also record: (i) the overlap of $\hat{p}(Z)$ between D=1 and D=0 groups (propensity deciles and densities), (ii) percentiles of IPW weights and the effective sample size $\text{ESS} = (\sum w_i)^2 / \sum w_i^2$, (iii) the share of replications that needed a Naive-QR fallback (e.g. when a weighted QR failed at extreme τ), and (iv) the average bound width $\mathbb{E}[\text{UB}(\tau) - \text{LB}(\tau)]$ and the coverage indicator $\mathbf{1}\{\text{Gap}^{\text{true}}(\tau) \in [\text{LB}(\tau), \text{UB}(\tau)]\}$.

Component	Value
Population size per replication (N)	120000
Replications per scenario (K)	200
Quantile grid (τ)	0.10, 0.25, 0.50, 0.75, 0.90
Scenarios (rank-selection ρ)	$S0_Base(\rho = 0)$, $S1_PosSel(\rho > 0)$, $S2_NegSel(\rho < 0)$
Instrument strengths	weak, base
Estimators	Observed (selected), Naive QR (X), IPW, AB, Bounds

Table 1: Monte Carlo design. Source: own simulation design in R

Table 1 summarizes the baseline Monte Carlo setup used throughout this Section. Each replication draws a synthetic population of size $N = 120{,}000$, and for each scenario we repeat the simulation $K = 200$ times. We evaluate distributional effects at the quantile grid $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$. Selection is varied across three rank-selection scenarios indexed by the copula correlation parameter ρ (see Table 2), and across two levels of instrument strength (*weak* vs. *base*) that change the impact of the exclusion shifter B on employment D (hence the amount of overlap

Scenario	ρ	Description
$S0_Base(\rho = 0)$	0.00	No rank-selection (benchmark)
$S1_PosSel(\rho > 0)$	0.25	Positive rank-selection: high-U more likely employed
$S2_NegSel(\rho < 0)$	-0.25	Negative rank-selection: low/median-U more likely employed

Table 2: Scenarios used in the Monte Carlo study. Source: own simulation design in R

in the propensity score $p(Z)$). For each design cell, we compare five estimators: the *Observed (selected)* gap, *Naive QR* on the employed sample using X, IPW, the AB estimator, and *bounds*.

Table 2 defines the three data-generating scenarios for rank dependence between the latent wage rank U and the participation rank V. The case *S0_Base* sets $\rho = 0$, implying no rank-selection (employment is independent of the wage rank conditional on covariates). In *S1_PosSel* we set $\rho = 0.25$, which induces positive rank-selection: individuals with higher potential wage ranks U are more likely to be employed and therefore observed. In *S2_NegSel* we set $\rho = -0.25$, corresponding to negative rank-selection: relatively lower (or median) wage ranks are more likely to be employed/observed. Together, these scenarios allow us to assess how estimator performance changes as selection moves from absent, to positive, to negative while keeping the rest of the design fixed.

4 Benchmark application of the quantile-selection method

As an additional implementation check, we apply the quantile-selection procedure described in Chapter 2 to a benchmark dataset associated with Koenker's *Quantile Selection Models: An R Vignette* [Koe17] and the related reference application of Arellano and Bonhomme [AB17]. The purpose of this exercise is methodological rather than substantive. It is not intended to study the Austrian gender pay gap directly, but to verify that the estimator can be implemented successfully on a known reference dataset.

The benchmark application follows the same general three-step logic as the estimator introduced by Arellano and Bonhomme [AB17]. First, the propensity score of participation is estimated. Second, the copula parameter governing the dependence between latent wage ranks and participation ranks is estimated. Third, conditional on these estimates, quantile parameters are obtained by minimizing a rotated check function with observation-specific corrected ranks. In this sense, the benchmark serves as a practical application of the estimation strategy presented in Chapter 2.

To implement this benchmark in `R`, we started from the reference data files used in the Koenker-Bonhomme environment and transformed the raw benchmark files into two compact working datasets, `M.Rda` and `F.Rda` [Koe17]. In this step, blocks of binary indicators were collapsed into factor variables in order to simplify the empirical specification and make the benchmark data more suitable for estimation. The resulting datasets were then split by sex and marital status into four subgroups:

- married men,
- single men,

- married women,
- single women.

This subgroup structure follows the benchmark application in Koenker's vignette [Koe17].

For each subgroup, we implemented the estimator in three explicit stages. First, we estimated the employment propensity score by a binary response model. Second, we estimated the copula dependence parameter ρ, which captures the dependence between the latent wage rank and the latent participation rank. In the benchmark implementation, we considered both a method-of-moments criterion and a likelihood-based criterion for this step, in line with the two estimation approaches discussed in the vignette [Koe17]. Third, conditional on the estimated propensity score and copula parameter, we ran a rotated quantile regression in which the usual fixed quantile index τ is replaced by the individual-specific corrected rank

$$G(\tau, p_i; \rho) = \frac{C(\tau, p_i; \rho)}{p_i}.$$

This corresponds to the key insight of the Arellano-Bonhomme method [AB17] correcting for selection amounts to rotating the check function through observation-specific rank adjustments.

The benchmark is therefore not used to claim an original empirical contribution on this external dataset. Rather, its role is to show that the estimator described in Chapter 2 can be implemented in practice on a known reference application, using our own structured implementation and producing benchmark outputs that can be compared with the reference implementation.

The resulting benchmark estimates and their interpretation are presented in Chapter 6.

5 Analysis of the real data published by Statistik Austria

We use Statistik Austria published gross quantiles (hourly earnings in euros) by industry (NACE), sex, and type of employment (full/part time) for $\tau \in \{0.25, 0.50, 0.75\}$. We harmonize industry (`ONVE10`) and employment-type (`VZTZ`) codes to English labels using the official dictionaries. The data report cell quantiles rather than micro records, so all plots below are descriptive.

Figure 1 displays the published gender gap (measured in Euro per hour - rescaled to avoid decimal, i.e. multiplied by 1000 (e.g. 12500 corresponds to 12.50 per hour).

$$\mathrm{Gap}^{\mathrm{pub}}(\tau) = Q_\tau^{\mathrm{female}} - Q_\tau^{\mathrm{male}}$$

by industry (rows) and quantile $\tau \in \{0.25, 0.50, 0.75\}$ (columns). The panels split full-time (left) and part-time (right) employment. Warmer colours indicate $\mathrm{Gap}^{\mathrm{pub}}(\tau) > 0$ (female quantile above male), cooler colours $\mathrm{Gap}^{\mathrm{pub}}(\tau) < 0$ (male quantile above female).

Full-time. Most industries are blue across all three quantiles, indicating predominantly *negative* published gaps for full–time workers. The blue tone often deepens between $\tau = 0.25$ and $\tau = 0.75$, suggesting that the male-advantage widens towards the upper part of the full–time wage distribution (a pattern of

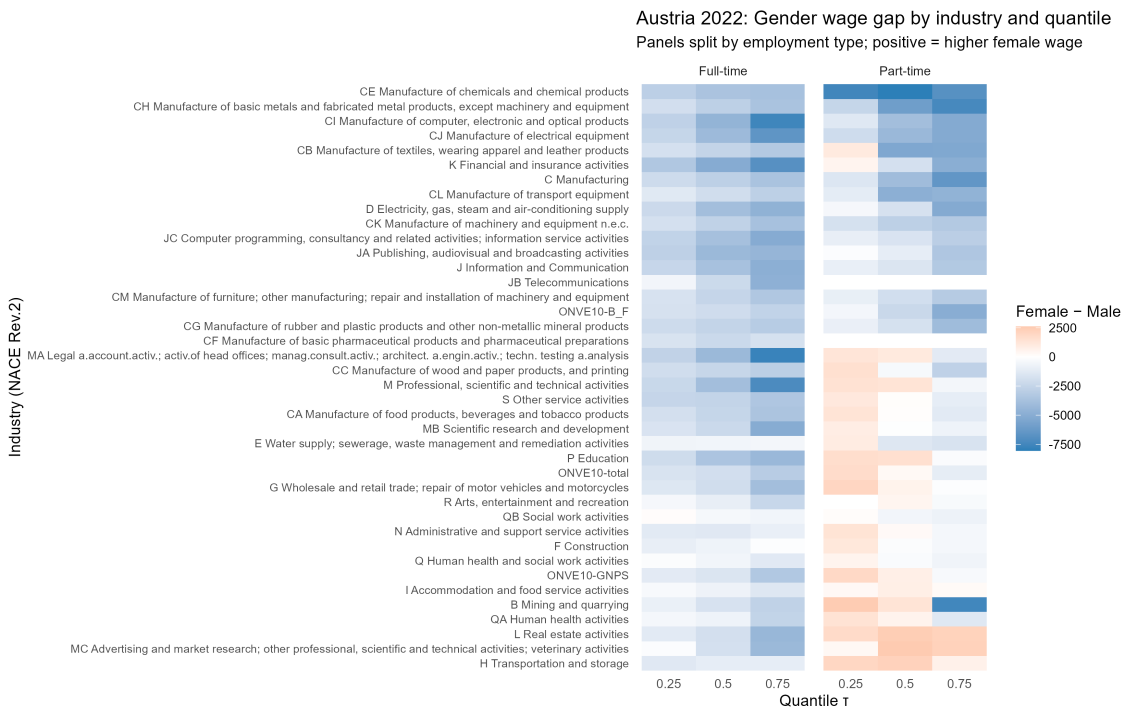


Figure 1: Austria 2022: Gender wage gap (in Euro) by industry and quantile. Source: own R code using published Statistik Austria data

"glass-ceiling" type). Tradable and technical sections (e.g. chemicals, basic metals, computer/electronic and electrical equipment, energy, finance) are among the darker rows, while several public or non-market services show lighter (closer to zero) shades.

Part-time. The right panel is more mixed. Many service industries show warmer or near-zero tones at $\tau = 0.25$ and $\tau = 0.50$, with some rows turning cooler at $\tau = 0.75$. This pattern is consistent with small or even *positive* published gaps in the lower middle of the part–time distribution, but a reversion towards male advantage at the upper tail. In other words, the dispersion by sex within part-time jobs appears weaker at the lower quantiles and stronger at the top.

Comparing panels, full-time cells are systematically more blue than part-time cells, indicating larger negative gaps among full-time workers. Aggregated rows (e.g. economy-wide or broad sections) mirror the same contrast: small/variable in part-time, clearly negative in full-time.

The colour gradients vary substantially across rows and across τ, confirming that any single, pooled gap masks important distributional and sectoral differences.

Figure 2 is a dumbbell plot showing, for $\tau = 0.5$ (the median), the 15 industries with the largest absolute within-industry gender gaps. Wages are gross hourly earnings in Euro (VESTE-1). For readability, published values are shown as in the CSV, i.e. multiplied by 1000 (e.g. 12500 corresponds to Euro 12.50 per hour).

Each row represents one industry. The orange dot marks the female median wage, and the blue dot marks the male median wage. The grey segment between the two dots shows the within-industry median gap. Industries are ordered by the absolute size of this gap, so the largest disparities appear at the bottom of the figure. The ordering makes the plot a quick "league table" of median disparities: the bottom rows are those where the female-male difference is largest in magnitude.

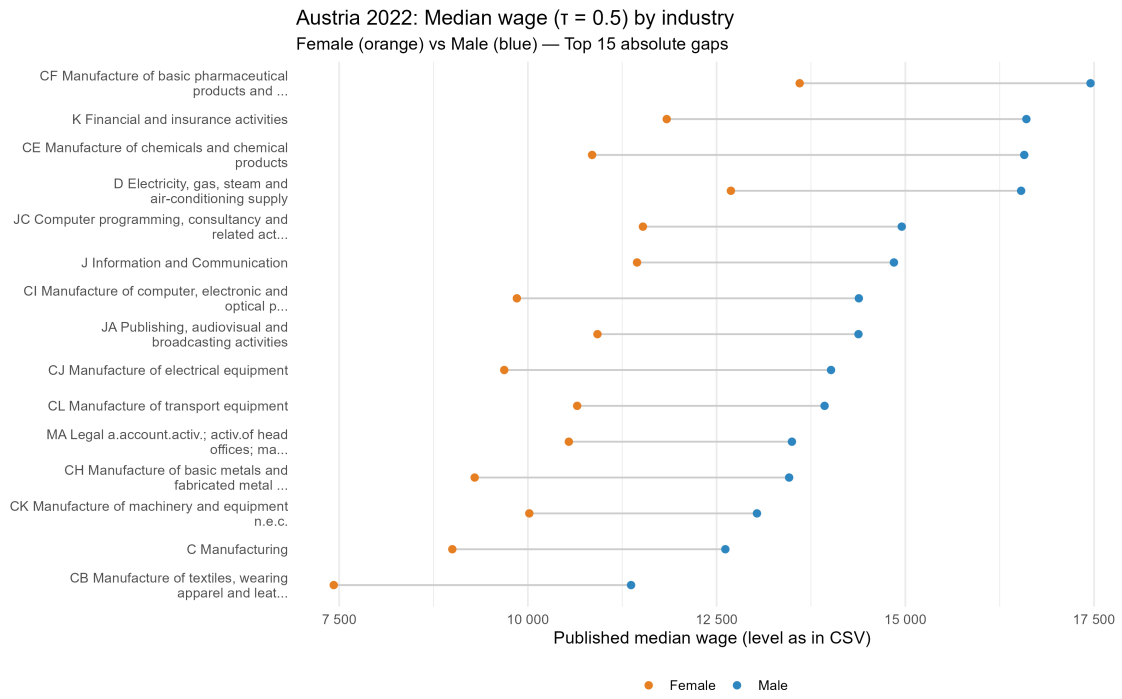


Figure 2: Austria 2022: median wage by industry (in Euro), Top 15 industries by absolute within-industry gender gap. Source: own R code using published Statistik Austria data

The figure shows that, in almost all of the displayed industries, the male median wage exceeds the female median wage. The largest median gaps appear in several high-pay sectors, but substantial differences are also visible in some manufacturing industries with more moderate overall wage levels. Hence, the figure suggests that median gender wage differences are not confined to a single part of the industry distribution.

In nearly all of the top-15 industries, the blue dot lies to the right of the orange dot, indicating male median wages exceed female median wages within the same industry. The longest segments appear in high-pay sectors (e.g., chemicals/pharmaceuticals, finance/insurance, electricity and gas, ICT-related activities), implying that high levels often coexist with large gender gaps at the median. Several manufacturing branches (e.g., machinery, basic metals) also show substantial gaps despite moderate overall levels, highlighting that large within-industry disparities are not confined to the very top of the wage distribution. Right-shifted pairs of dots indicate industries with high pay for both genders; the length of the segment isolates the gender difference in that industry irrespective of the overall level.

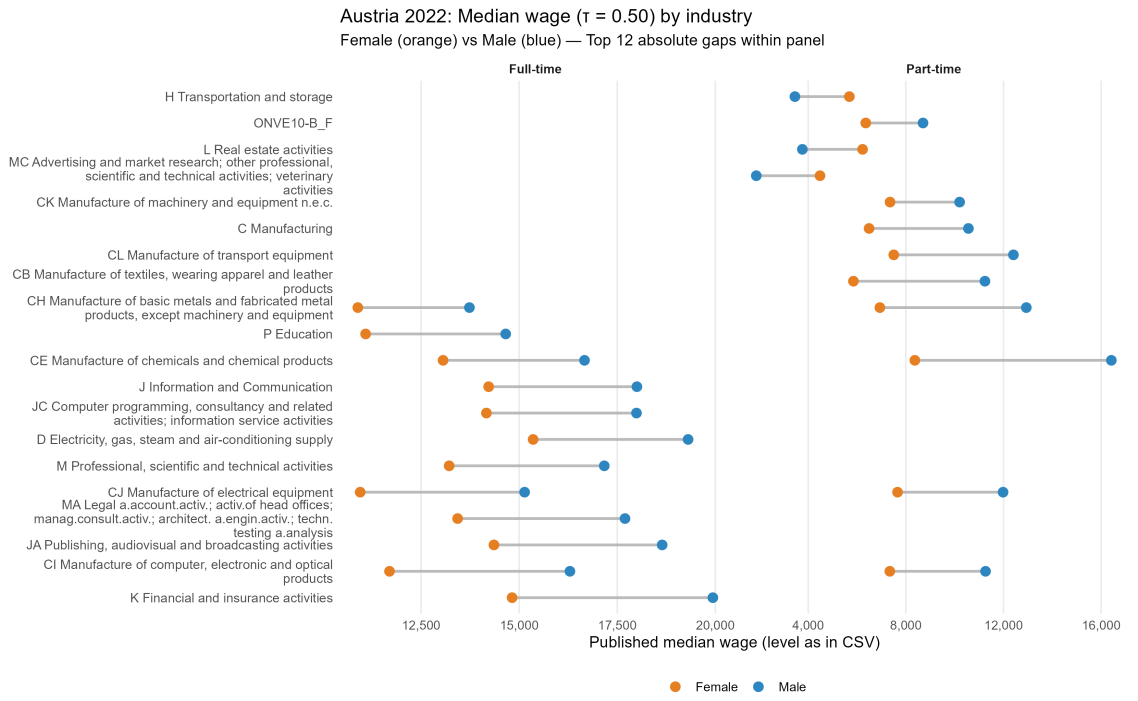


Figure 3: Median wage by industry and employment type. Top 12. Source: own R code using published Statistik Austria data

Figure 3 shows median ($\tau = 0.5$) gaps split by employment type. The figure shows two dumbbell panels - *Full-time* on the left and *Part-time* on the right. Within each panel, rows are industries; the orange (blue) dot marks the female (male) published median. The grey segment joining the dots is the *within-industry median gap*. Each panel lists the 12 industries with the largest absolute gap inside that panel, so rankings are independent across full-time and part-time.

Axes show markedly lower wage levels for part-time rows (right panel) than for full-time rows (left), consistent with part-time work being concentrated in lower-pay schedules and occupations. Thus the same industry can appear at very different levels across panels even before considering gender gaps. In both panels, most segments are directed to the right (male > female at the median). The longest segments in full-time arise in higher-pay, skill-intensive sectors (e.g., chemicals/pharmaceuticals, finance, utilities/energy, ICT-related activities), while manufacturing branches also contribute significant gaps. In the part-time panel, large gaps are visible even in service sectors where the level is low, indicating that median disparities persist for part-time workers and are not only a high-pay phenomenon.

Several industries appear among the top-gap group in one panel but not the other. This indicates that the size of the median gender gap is *panel-specific*: the same industry may be relatively more unequal among full-timers than part-timers (or vice versa), suggesting different task mixes, hours structures, or sorting patterns by employment type.

The horizontal distance between dots is the gender difference at $\tau = 0.5$ holding the industry and the panel fixed. The absolute position of the pair reflects the industry's pay level for that panel; comparing two rows with similar segment lengths but different locations separates "large gap at low level" from "large gap at high level".

Median gender gaps are pervasive in both full-time and part-time work, with sectoral composition and employment type jointly shaping their magnitude.

6 Results

This chapter presents and interprets the empirical results of the thesis. The goal is to quantify the gender wage gap not only at the mean or at the median, but *across the wage distribution*. In the framework introduced in Chapter 2, wages are observed only for employed individuals ($D = 1$), while the object of interest is the population (potential) wage distribution and the associated *population gender quantile gap*. This creates a selection problem: if employment is systematically related to wage potential, then a comparison based only on the employed sample can be misleading. The methods considered in this thesis (IPW and Arellano-Bonhomme rank-relabeling, AB) are designed to correct for this selection and to recover population quantile effects under explicit assumptions.

We report results for three complementary empirical components introduced earlier in the thesis (Chapter 3 - 5). The first component is a Monte Carlo simulation study, where the data-generating process is controlled by the researcher and the true population gender quantile gap is known by construction. This setting allows us to evaluate estimator performance in a transparent way: we can directly compare an estimated gap $\widehat{\text{Gap}}(\tau)$ to the truth $\text{Gap}^{true}(\tau)$ for each quantile τ and under different selection environments. The simulation design follows the model structure in Chapter 2: potential wages are generated through a structural quantile function $Y^* = q(U, X)$ with latent wage rank $U \sim \text{Unif}(0, 1)$, employment is generated via a propensity score $p(Z) = \Pr(D = 1 \mid Z)$ with $Z = (X, B)$, and the dependence between wage ranks and participation ranks is governed by a copula parameter ρ. We consider three scenarios: no rank-selection ($\rho = 0$), positive rank-selection ($\rho > 0$), and negative rank-selection ($\rho < 0$), and we study two levels of instrument strength ("base" vs. "weak") which affect overlap in the propensity score. In the simulation design, X contains the observed wage covariates, namely the female indicator, part-time status, and industry indicators. The variable B is an excluded shifter that affects employment but does not enter the structural wage equation. In each replication, B is generated as an exogenous random variable and enters the participation equation through the propensity score $p(Z)$, where $Z = (X, B)$. Instrument strength is varied by changing the coefficient on B: in the base case, B has a stronger effect on employment, whereas in the weak case its effect is attenuated, leading to poorer overlap and more difficult identification. The Monte Carlo results therefore answer a precise question: when selection is present, how large is the distortion in observed quantile gaps, and how well do the correction methods recover the population gap?

The second component is the benchmark application introduced in Section 4. Its purpose is methodological rather than substantive. The benchmark is not used to study the Austrian gender pay gap directly, but to verify that the quantile-selection procedure can be implemented successfully on a known reference dataset and to compare our implementation with the reference R implementation documented in Koenker's vignette [Koe17]. Since the benchmark dataset does not reveal the true latent population gap, it does not permit direct bias or RMSE calculations. Instead, it allows us to examine how strongly the correction changes the naive fitted quantile curves and how sensitive the corrected benchmark results are to the choice of estimation criterion.

The third component uses published Austrian wage quantiles from Statistik Austria [Ös24] by industry, gender, and employment type (full-time vs. part-time). These data are available only at the cell level (published quantiles), not as micro-

records. Consequently, this part of the analysis is descriptive: it documents distributional and sectoral heterogeneity of the gender wage gap in Austria and motivates why a quantile-based view is informative. In particular, the published quantiles already show that the magnitude and even the sign of the gap can vary across industries, employment types, and quantiles. This heterogeneity motivates the methodological focus of the thesis: a single "average" gap can mask substantial distributional differences.

Throughout this chapter we use the notation established earlier. For each quantile τ, the population gender quantile gap is the difference between female and male potential wage quantiles evaluated at the same covariates (and in particular at different gender indicators). In the simulation study this object is observed as $\text{Gap}^{true}(\tau)$, while in practice it is estimated by $\widehat{\text{Gap}}(\tau)$. We evaluate estimators using the two standard performance measures reported in Chapter 3:

$$\text{Bias}(\tau) = \mathbb{E}\Big[\widehat{\text{Gap}}(\tau) - \text{Gap}^{true}(\tau)\Big], \qquad \text{RMSE}(\tau) = \sqrt{\mathbb{E}\Big[\Big(\widehat{\text{Gap}}(\tau) - \text{Gap}^{true}(\tau)\Big)^2\Big]}.$$

Bias captures systematic over- or underestimation of the true gap, while RMSE combines bias and variability and therefore reflects overall accuracy.

The results are presented in a way that mirrors the methodological discussion in Chapter 2. We first report Monte Carlo performance for the observed (selected) gap, naive quantile regression on the employed sample, IPW-corrected quantile estimates, and the AB correction. We then turn to the benchmark application and compare both copula estimates and corrected quantile fits with the reference implementation. Finally, we connect these findings to the descriptive patterns in the published Austrian quantiles by industry and employment type.

In sum, Chapter 6 brings together the model-based selection correction logic from Chapter 2 and the empirical setups from Chapter 3 - 5. The simulation study provides a controlled environment for assessing the size and direction of selection bias and the performance of correction methods, the benchmark application serves as an implementation check on a known reference dataset, and the published Austrian data illustrate the economically relevant heterogeneity of gender wage gaps across industries, quantiles, and employment types.

Figure 4 summarizes Monte Carlo *bias* of each estimator of the gender quantile gap. For each quantile level $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$, the vertical axis reports

$$\text{Bias}(\tau) = \mathbb{E}\Big[\widehat{Gap}(\tau) - Gap^{true}(\tau)\Big],$$

so values close to zero indicate that an estimator recovers the true population gap on average, while positive (negative) values indicate systematic overestimation (underestimation).

The figure is organized as a 3×2 panel grid. Rows correspond to the rank-selection dependence scenarios: (i) $S0$ (no rank-selection, $\rho = 0$), (ii) $S1$ (positive rank-selection, $\rho > 0$), and (iii) $S2$ (negative rank-selection, $\rho < 0$).

Columns compare instrument strength: base (left) versus weak (right). In the simulation design, instrument strength is controlled by the scalar s in the participation equation

$$p(Z) = \Lambda\Big(\alpha + X^\top \gamma_X + s\, B^\top \gamma_B\Big).$$

In the base case, $s = 1$, so the excluded shifter B has a stronger effect on employment; in the weak case, $0 < s < 1$ (s=0.5 in our simulation), so the effect of B is

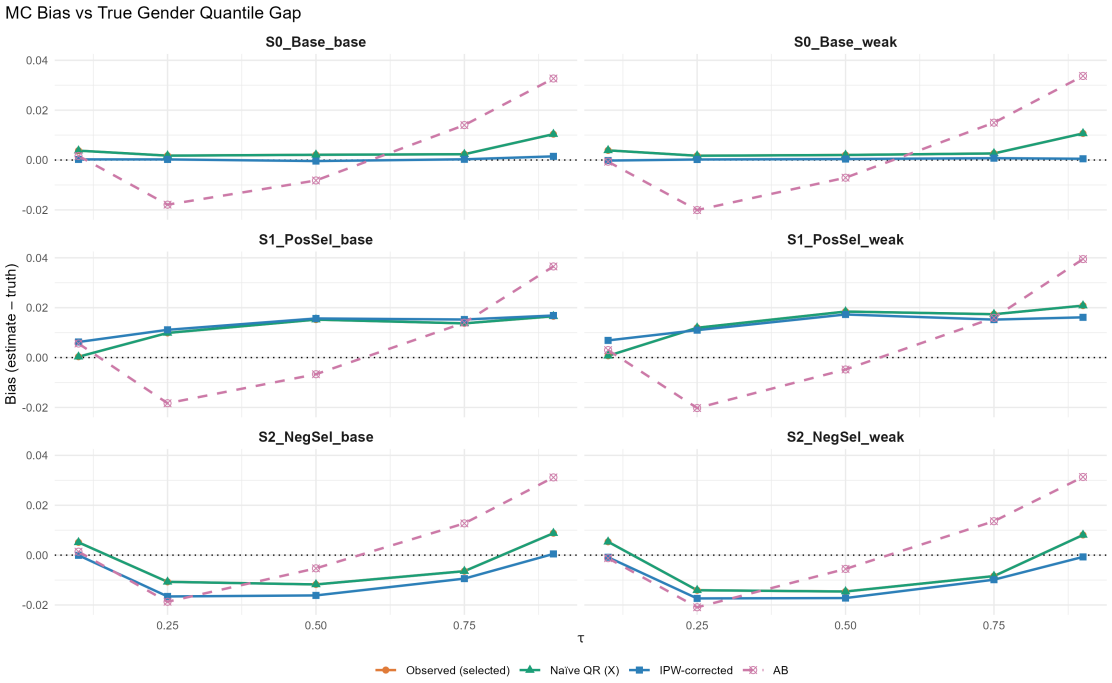


Figure 4: MC Bias vs True Gender Quantile Gap. Source: own Monte Carlo design in R

attenuated. This reduces variation in $p(Z)$, worsens overlap, and makes selection correction more difficult.

Across all panels we compare four estimators: (i) *Observed (selected)* quantile differences within the employed sample, (ii) *Naive QR (X)* on $D = 1$ only, (iii) *IPW-corrected* quantile gaps using estimated $p(Z)$, and (iv) AB rank-relabeling (oracle) based on the copula parameter ρ used in the DGP.

A horizontal reference at zero bias is the benchmark: if an estimator is centered around zero across τ, it is approximately unbiased in finite samples under the given scenario. Deviations from zero show the direction and magnitude of systematic error, and changes across τ reveal whether the error is concentrated in the lower tail, center, or upper tail of the wage distribution.

In $S0$ (top row), rank-selection is absent by construction ($\rho = 0$), so selection does not distort wage ranks. Consistent with this, biases are relatively small across quantiles and estimators, and the lines remain close to zero. This scenario serves as a benchmark, validating that when selection is uncorrelated with U, the estimators do not exhibit large systematic distortions in the estimated quantile gap.

In $S1$ (middle row), $\rho > 0$ induces *positive rank* selection: people with a higher latent wage rank U are more likely to be observed with $D = 1$. In this setting, estimators that rely only on the employed sample, the observed (selected) gap, and the Naive QR exhibit non-negligible bias that varies with τ. This pattern reflects the fact that the employed subsample is not representative of the population wage distribution: selection changes the composition of U among observed wages, and the resulting distortion differs between quantiles. The IPW-corrected estimator typically moves the bias closer to zero relative to the Naive estimators, consistent with reweighting the employed observations to approximate the population distribution.

In $S2$ (bottom row), $\rho < 0$ implies *negative rank* selection, so selection favors lower or median values of U. Accordingly, the direction of bias shifts compared to

$S1$ in several quantiles: the employed subsample now reflects a different distortion of wage ranks, and the induced bias in the estimated gender gap can change sign or change curvature across τ. As in $S1$, IPW generally attenuates the magnitude of bias relative to the observed and Naive QR estimators, indicating that reweighting helps mitigate the effect of selection-driven composition. The IPW estimator is generally closer to zero bias than the uncorrected alternatives, especially in the negative-selection scenarios. By contrast, the AB estimator exhibits larger deviations in several panels, especially toward the upper quantiles. The weak-instrument cases also tend to show slightly less favourable performance than the base cases, which is consistent with weaker identifying variation in the participation equation.

Overall, Figure 4 demonstrates that (i) selection can substantially distort quantile gender gaps, (ii) the distortion is quantile-dependent, and (iii) correction methods improve accuracy when rank-selection is present, but their performance depends on identification strength and distributional location.

When comparing the strength of the instrument (left vs right column), the *weak* setting of the instrument tends to produce larger and less stable biases, particularly for estimators that depend on the estimated propensity score $p(Z)$. This is consistent with weaker overlap and noisier reweighting: when the shifter B moves selection only weakly, the inverse-probability weights become more variable and finite-sample distortions may increase, especially in the tails. Overall, the contrast between *base* and *weak* columns highlights that identification strength and overlap are key drivers of finite-sample performance in selection-correction procedures.

Across scenarios, biases are often more pronounced at extreme quantiles than around the center. This reflects two common forces in finite samples: (i) tail quantiles are estimated with fewer effective observations (and are thus more sensitive to sampling variation), and (ii) selection and weighting distort the distribution most strongly in regions where overlap is weakest, which typically occurs toward the tails of $p(Z)$ and, consequently, toward the tails of the implied wage ranks.

Finally, the AB rank-relabeling estimator exhibits a systematic pattern across τ in several panels. A likely explanation is that, even though AB uses the copula structure of the data-generating process, it remains a multi-step procedure that relies on rank reconstruction, interpolation on (τ, p) grids, and kernel smoothing. In finite samples, these features may introduce additional approximation error, especially in the tails and under weaker overlap. For this reason, the figure should be interpreted primarily as evidence on finite-sample performance under the chosen implementation and tuning parameters, rather than as a purely asymptotic statement about the estimator.

Figure 5 reports the corresponding RMSE values:

$$\mathrm{RMSE}(\tau) = \sqrt{\mathbb{E}\Big[(\widehat{Gap}(\tau) - Gap^{true}(\tau))^2\Big]}.$$

Since RMSE combines bias and variability, it provides a broader measure of estimator performance than bias alone.

In S0 (top row), RMSE is generally small across quantiles, reflecting the absence of rank-selection. Once rank-selection is introduced, estimator performance becomes more heterogeneous. In S1 and S2, the observed and naive procedures no longer perform uniformly well, and errors tend to increase in the tails. The IPW estimator often improves on the uncorrected alternatives, although it also becomes more variable when overlap deteriorates. The AB estimator remains informative, but its RMSE is often larger, especially at higher quantiles and under weaker instruments.

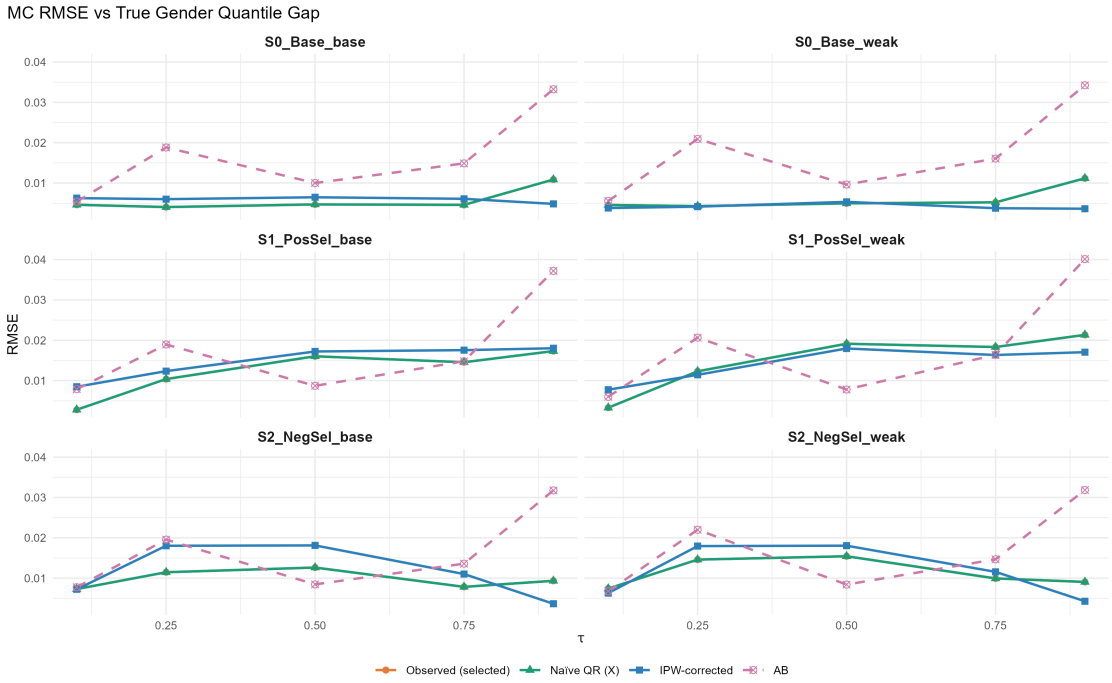


Figure 5: MC RMSE vs True Gender Quantile Gap. Source: own Monte Carlo design in R

Thus, Figure 5 reinforces the main message of Figure 4 correction methods can improve performance, but they do not do so uniformly across quantiles and scenarios. Their success depends critically on overlap, instrument strength, and the part of the distribution under consideration.

The benchmark application provides a complementary implementation check. Table 3 compares the copula estimates from the original benchmark/reference implementation with those from our own implementation.

Subgroup	Original benchmark (MoM)	(Lik.)	Our implementation (MoM)	(Lik.)
Married men	-1.397	-0.280	-1.47	-1.49
Single men	-8.486	-0.198	-8.38	-2.32
Married women	-0.960	-0.205	-0.99	-1.33
Single women	-0.469	-0.099	-0.40	-2.06

Table 3: Benchmark application: copula parameter estimates by subgroup. Comparison of original benchmark/reference results and own implementation

Several conclusions follow from Table 3. The first two columns report the earlier reference/package-based benchmark values, while the last two columns report the estimates from our own implementation. First, all copula estimates are negative across all four subgroups. In this framework, a negative copula parameter indicates positive selection into employment: individuals with higher latent wage ranks are more likely to be observed among workers. Second, the subgroup ranking is qualitatively stable across implementations. In both the original benchmark values and our own implementation, the strongest negative dependence is found for single men, while the estimates for married men and married women are more moderate.

At the same time, the exact magnitudes differ across estimation criteria and implementations. For married men, the method-of-moments and likelihood-based estimates are fairly close in our implementation, whereas for single men and single women the differences are larger. Thus, the benchmark is qualitatively robust, but the exact estimated strength of dependence may remain sensitive to the copula estimation criterion.

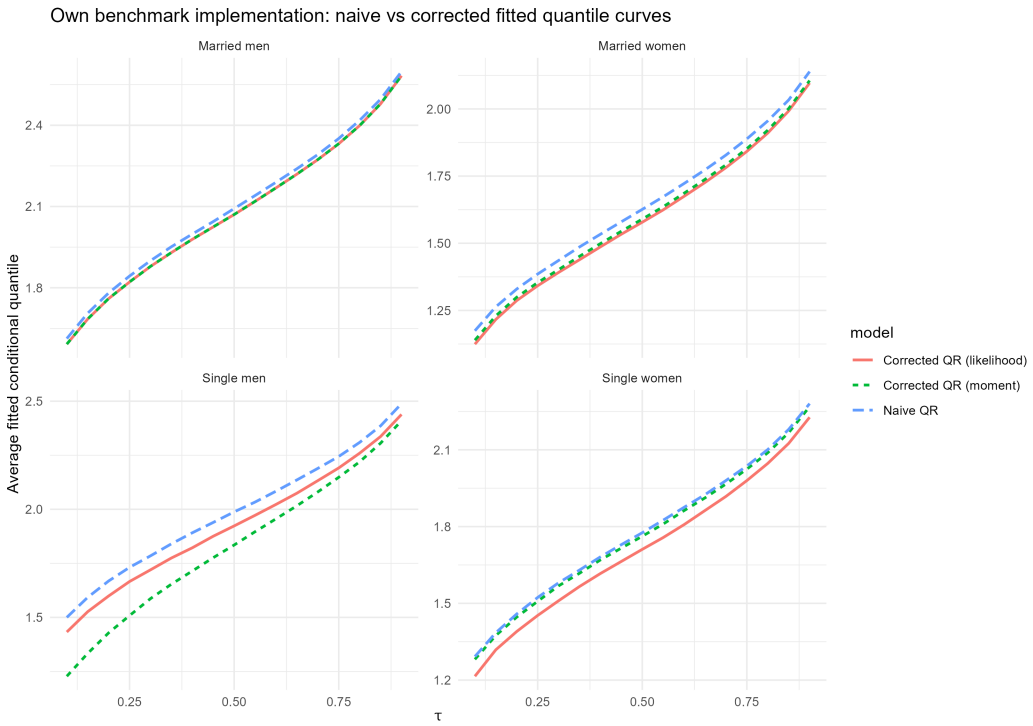


Figure 6: Benchmark application: naive and selection-corrected fitted quantile curves by subgroup. Source: own implementation of the benchmark procedure using the reference benchmark data

Figure 6 provides the most direct visual summary of the benchmark correction. The figure compares the naive quantile regression fit with the corrected fits obtained under the method-of-moments and likelihood-based copula estimates. The resulting patterns are broadly consistent with the benchmark behaviour documented in Koenker's reference implementation. Across all four subgroups, the selection-corrected curves differ from the naive quantile regression fit, although the magnitude of the adjustment varies substantially. For married men, the corrected curves remain very close to the naive fit, indicating that the benchmark correction is small and robust to the estimation criterion. For married women and single women, the corrected curves lie systematically below the naive fit, with the likelihood-based correction generally stronger than the moment-based one. The largest deviations appear for single men, where the benchmark correction is clearly non-trivial and the distance between the two corrected fits is also largest. Hence, Figure 6 confirms that the benchmark adjustment is subgroup-specific: it is weakest for married men, moderate for women, and strongest for single men.

Unlike the Monte Carlo study, the benchmark dataset does not reveal the true latent population gap. Direct bias and RMSE measures therefore cannot be computed. As a benchmark analogue of Figure 4, Figure 7 reports the size of the

correction relative to naive quantile regression:

$$q^{corr}(\tau) - q^{naive}(\tau).$$

Values close to zero indicate little correction, while larger absolute values indicate a stronger benchmark adjustment.

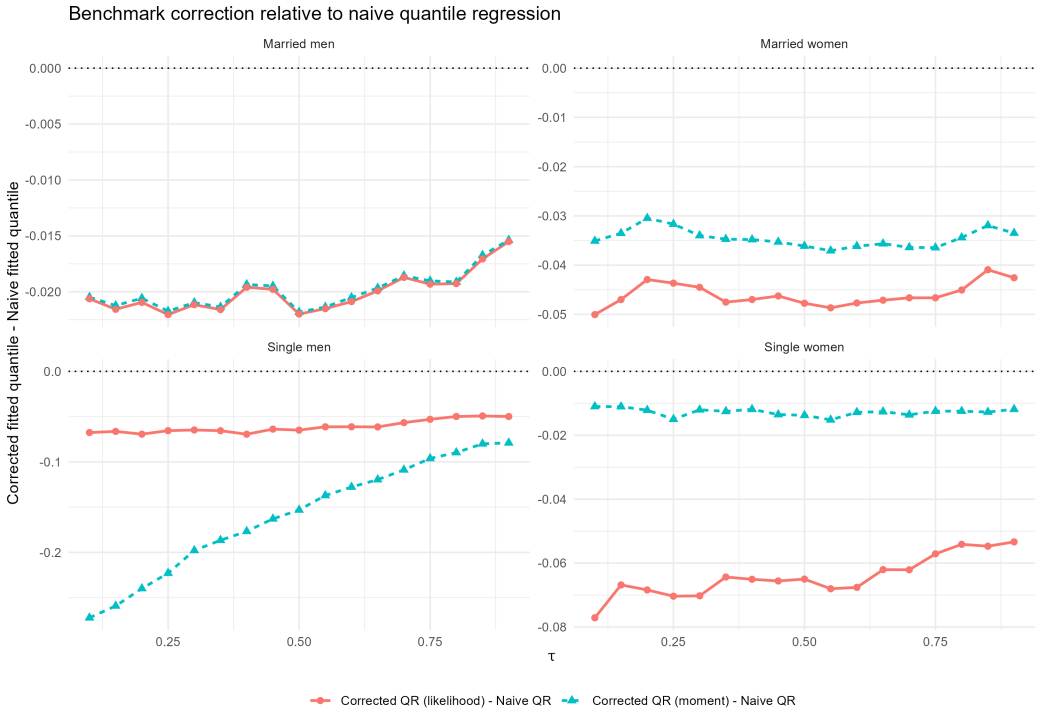


Figure 7: Benchmark application: correction relative to naive quantile regression by subgroup. Source: own implementation of the benchmark procedure using the reference benchmark data.

Figure 7 plots the difference between the selection-corrected fitted quantile curves and the naive quantile regression fit, separately for the method-of-moments and likelihood-based copula estimates. Values close to zero indicate little correction, while larger absolute values indicate stronger benchmark adjustments. The figure shows that the benchmark correction is not equally important across groups. For married men, both corrected fits lie only slightly below the naive fit, and the method-of-moments and likelihood corrections are almost indistinguishable. Hence, in this subgroup the benchmark adjustment is small.

For married women, the correction is more pronounced. Both corrected curves lie below the naive curve, which means that the selection correction shifts the fitted quantile profile downward relative to naive quantile regression. The likelihood-based correction is stronger than the moment-based correction, so the benchmark adjustment is not only non-trivial, but also somewhat method-sensitive.

For single women, the downward correction is stronger still. Again, both corrected fits lie below the naive fit, and the likelihood-based curve implies a larger correction than the moment-based one. Thus, single women display a more substantial benchmark adjustment than the married groups.

The strongest benchmark adjustment appears for single men. In this subgroup, the corrected fit differs substantially from naive quantile regression, especially for the method-of-moments estimate. This confirms graphically what Table 3 already suggests: selection matters most strongly in the single-men subgroup.

As a benchmark analogue of Figure 5, Figure 8 reports the difference between the two corrected benchmark fits:

$$q^{corr,MoM}(\tau) - q^{corr,Lik}(\tau).$$

This figure measures how sensitive the benchmark correction is to the choice of estimation criterion.

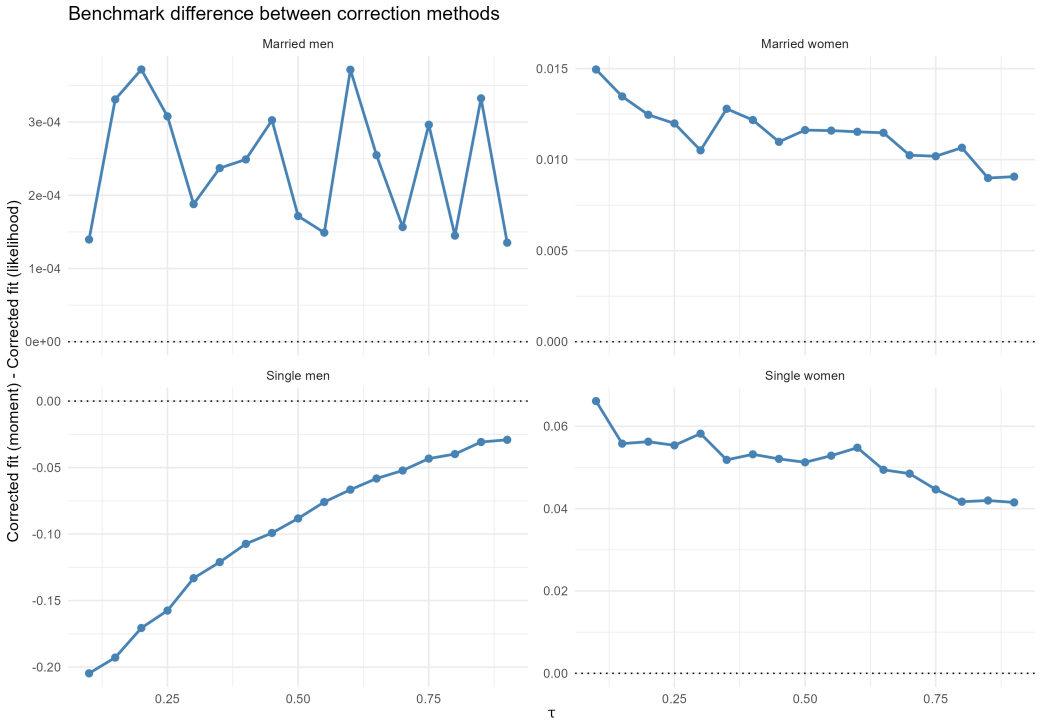


Figure 8: Benchmark application: difference between corrected fits obtained under the method-of-moments and likelihood criteria. Source: own implementation of the benchmark procedure using the reference benchmark data

The figure plots the subgroup-specific difference between the two corrected fitted quantile curves across quantiles. Values close to zero indicate that the two correction methods produce similar benchmark fits, whereas larger deviations indicate greater sensitivity to the copula estimation criterion. The figure shows that this sensitivity is negligible for married men, moderate for married and single women, and strongest for single men.

For married men, the difference between the two corrected fits is extremely small across all quantiles. Hence, in this subgroup the benchmark result is practically insensitive to whether the copula is estimated by moments or by likelihood. For married women, the difference is positive but moderate. The two correction methods do not coincide exactly, but they remain fairly close throughout the quantile range. For single women, the difference becomes noticeably larger. This indicates that the corrected benchmark fit is more sensitive to the estimation criterion in this subgroup than in the married groups.

The largest divergence between the two correction methods is observed for single men. This means that the subgroup with the strongest benchmark correction is also the subgroup in which the benchmark result is most method-sensitive.

Taken together, Table 3, Figure 7, and Figure 8 convey a coherent message. The benchmark correction is real and non-trivial, but its quantitative importance

differs substantially across subgroups. It is strongest for single men, moderate for single and married women, and weakest for married men. At the same time, the benchmark remains qualitatively stable across estimation criteria, since all subgroup estimates point to positive selection into employment and the subgroup ranking is broadly preserved.

The Monte Carlo results (Chapter 3) provide a useful conceptual reference for interpreting the Austrian descriptive evidence. In the simulations, selection into employment can distort quantile gender gaps in a quantile-dependent way, especially when participation is related to latent wage ranks. This shows that observed quantile gaps do not necessarily coincide with the corresponding gaps in the population, even if the descriptive patterns appear systematic.

The benchmark application (Chapter 4) strengthens this interpretation at the methodological level. It shows that the quantile-selection procedure can be implemented successfully on a known reference dataset and that the resulting correction is non-trivial in practice. In particular, the benchmark confirms that selection adjustment may change fitted quantile profiles in a subgroup-specific way.

The published Austrian results (Chapter 5) reveal significant heterogeneity across industries, employment types, and quantiles. Specifically, the observed gaps differ between full-time and part-time employment, and also vary across the lower, middle, and upper segments of the wage distribution. These differences are informative as descriptive patterns among the observed workers.

At the same time, the Austrian data used in this thesis do not allow for full structural correction for selection at the micro-level. Therefore, the simulation and benchmark results should not be viewed as direct evidence that the Austrian patterns are caused by selection. Rather, they demonstrate that selection is a plausible mechanism that can influence the interpretation of the observed quantile gaps.

Hence, the Austrian descriptive analysis should be interpreted with appropriate caution. The published quantile gaps are empirically informative, but they cannot on their own distinguish between structural wage differences and differences related to employment selection.

This chapter delivers three complementary sets of findings: methodological insights from the controlled Monte Carlo study, methodological evidence from the benchmark implementation, and descriptive evidence from Austrian wage data. Together, they lead to several central conclusions.

First, selection into employment can distort gender wage gaps in a quantile-dependent manner. The simulation results show that when employment depends on latent wage ranks ($\rho \neq 0$), the observed gender gap does not simply shift uniformly across the distribution. Instead, the distortion varies across τ, and may even differ in sign between the lower and upper quantiles. This confirms that selection reshapes the wage distribution rather than affecting only its mean.

Second, ignoring selection leads to systematic bias. The observed gap and the Naive QR estimator tend to misrepresent the true quantile gap whenever rank-selection is present. The magnitude of this distortion increases in the tails of the distribution, where overlap and identification become weaker. Thus, descriptive quantile gaps should be interpreted with caution when participation decisions are non-random.

Third, selection-aware estimators improve accuracy under non-zero selection. Both IPW and the AB correction reduce bias relative to non-corrected methods. However, their performance is not uniform across quantiles and depends on instrument strength. In particular, weak overlap leads to higher variability, as reflected in

increased RMSE values. This highlights that correction methods critically rely on credible exclusion variation.

Fourth, the benchmark application shows that the estimation procedure can be implemented successfully on a known reference dataset. The estimated copula parameters are economically interpretable, and the corrected fitted quantile curves differ in a systematic way from the naive quantile regression fit. The benchmark correction is strongest for single men and weakest for married men, and the benchmark results are most sensitive to the estimation criterion in the single-men subgroup.

Fifth, the Austrian descriptive results reveal substantial heterogeneity of gender wage gaps across industries, employment types, and quantiles. Differences between full-time and part-time employment, as well as variation across sectors, indicate that a single aggregate measure of the gender pay gap masks important distributional and structural patterns. The magnitude of the gap differs not only across sectors but also across the wage distribution within sectors.

Overall, the empirical findings emphasize that the gender wage gap is a distributional phenomenon. It varies across quantiles, across employment types, and across industries, and may be affected by selection mechanisms. Consequently, analyzing only average wage differences provides an incomplete picture. A quantile-based and selection-aware framework offers a more nuanced understanding of gender inequality in wages.

While the results presented in this chapter provide valuable insights into the distributional structure of the gender wage gap in Austria, several limitations should be acknowledged.

First, the empirical Austrian analysis relies on published cell-level quantiles rather than individual microdata. As a consequence, the selection-correction procedures described in Chapter 2 cannot be fully implemented in the real-data application. The Monte Carlo study and the benchmark application illustrate how selection may distort quantile gaps and how correction methods operate, but the Austrian descriptive results remain conditional on the observed employed population. Therefore, the empirical findings should be interpreted as distributional patterns among observed workers rather than fully selection-adjusted population effects.

Second, industry-level aggregation may conceal substantial within-industry heterogeneity. Even though the analysis distinguishes between full-time and part-time employment and considers detailed NACE sectors, variation within occupations, firms, or career stages is not observable in the published data. Hence, the measured gaps combine multiple underlying mechanisms, including occupational sorting, hours differences, career interruptions, and potential differences in returns.

Third, the performance of selection-correction estimators in the simulation study depends on the strength and credibility of exclusion variation. In practice, identifying valid and sufficiently strong instruments for participation decisions is challenging. Weak overlap or misspecification may lead to instability, particularly in the tails of the distribution. The simulation evidence highlights this sensitivity and underlines that empirical implementation requires careful diagnostics.

Fourth, the benchmark application supports the practical feasibility of the estimator, but it is not itself an Austrian empirical result. Its role is methodological: it shows that the implementation works on a known reference dataset and produces interpretable outputs, including non-trivial subgroup-specific corrections.

Finally, the quantile framework captures distributional heterogeneity, but it does not by itself decompose the gap into explained and unexplained components. The estimated differences therefore reflect combined effects of characteristics, sorting,

and potential structural inequalities, without isolating specific causal channels.

Despite these limitations, the combined evidence from simulation, benchmark implementation, and Austrian descriptive data consistently points to one central insight: the gender wage gap is not a single number, but a distributional phenomenon. Its magnitude varies across quantiles, sectors, and employment types, and its interpretation may be affected by participation behavior.

7 Conclusion

This thesis set out to study the gender wage gap in Austria from a distributional perspective and to assess how far its observed shape may be influenced by selection into employment. Rather than treating the gap as a single average difference, the analysis approached it as a phenomenon that may vary across the wage distribution and across labour-market segments. This perspective is especially relevant in a context such as Austria, where employment patterns, part-time work, and labour-market attachment differ substantially between men and women.

A central message of the thesis is that distribution and selection cannot be separated conceptually when wage inequality is studied through observed earnings alone. Because wages are only observed for employed individuals, the empirical wage distribution need not coincide with the distribution of potential wages in the underlying population. The Monte Carlo results discussed in Chapter 6, based on the simulation design in Section 3, illustrate this point clearly. When participation depends on latent wage ranks, observed quantile gaps can deviate from the true population gaps, and this deviation is not constant across the distribution. Instead, it may differ from one quantile to another and become particularly pronounced in the tails. Selection therefore has the potential not merely to shift observed wage differences, but to reshape their entire distributional profile.

This also clarifies why mean-based approaches are too limited for the question considered here. A single average measure can summarize the overall level of inequality, but it cannot reveal whether the gap is relatively small at the bottom of the wage distribution, wider in the middle, or especially pronounced at the top. Quantile methods are valuable precisely because they allow these differences to appear. Once selection is added to the picture, such a framework becomes even more important: it enables the analysis to distinguish between variation across the distribution and distortions arising from non-random employment participation.

Within the controlled simulation design, this thesis shows that estimators based only on the observed employed sample do not generally recover the true quantile gap when rank-selection is present. By contrast, selection-aware procedures such as inverse probability weighting and the Arellano-Bonhomme correction often reduce these distortions and improve accuracy. At the same time, the simulations make clear that the practical success of such corrections depends on favourable empirical conditions, especially sufficient overlap and sufficiently strong exclusion variation. In weakly identified settings, and particularly at extreme quantiles, the correction itself may become unstable. Thus, the thesis does not suggest that selection correction is automatic or costless. Rather, it shows both its importance and its empirical demands.

In addition, the benchmark results show that the quantitative importance of the correction differs across subgroups: it is strongest for single men, weakest for married men, and the benchmark is most sensitive to the estimation criterion in the single-men subgroup. In this way, the benchmark links the theoretical framework to an

established empirical example and strengthens the credibility of the implementation used in the simulation and Austrian application.

The Austrian evidence, in turn, gives the thesis its substantive context. Even at a purely descriptive level, the published quantiles show substantial heterogeneity across industries, employment forms, and parts of the wage distribution. This alone is an important finding. It suggests that the gender wage gap in Austria is not sufficiently captured by a single aggregate statistic. Instead, it differs across labour-market segments and across quantiles in ways that matter for interpretation. Some of these differences may reflect structural features of wage-setting, occupational sorting, or working-time arrangements; others may reflect differences in who appears in observed employment in the first place. The available data do not allow these channels to be cleanly separated, but they strongly suggest that the gap is both distributional and context-specific.

This is also where the limits of the present analysis must be acknowledged. Because the Austrian application relies on published aggregated quantiles rather than individual microdata, it remains necessarily descriptive. It cannot fully implement the selection-correction strategy developed in the methodological chapters, nor can it isolate causal mechanisms behind the observed patterns. For this reason, the Austrian findings should not be interpreted as definitive estimates of selection-adjusted wage inequality. Their value lies instead in showing how much heterogeneity is present in the observed data and why that heterogeneity deserves a framework richer than standard average comparisons.

The broader contribution of the thesis is therefore methodological as well as interpretive. Methodologically, it shows how a quantile-selection framework can be brought to bear on the study of gender wage inequality. Interpretively, it argues that observed gender wage gaps should be read with greater caution whenever employment participation is itself selective. The thesis does not claim to resolve all empirical difficulties surrounding this problem. It does, however, provide a coherent framework for thinking about them and for analysing them in a way that respects both the distributional nature of wages and the role of selection into employment.

Taken together, the results of this thesis suggest that the gender wage gap in Austria is better understood as a distributional phenomenon than as a single number. A quantile-based and selection-aware perspective does not eliminate all ambiguity, but it yields a more nuanced and more informative picture of wage inequality. For that reason, it provides a stronger basis for both empirical interpretation and future research than mean-based comparisons alone.

8 Appendix

Disclosure

AI was used to support the linguistic quality of this document, including grammar checks and stylistic refinement. In addition, AI was occasionally used as a technical support tool for debugging R code and identifying programming errors.

References

[AB17] Manuel Arellano and Stephane Bonhomme. Quantile selection models with an application to understanding changes in wage inequality. *Econometrica*, 85(1):1–28, 2017.

[Buc94] Moshe Buchinsky. Changes in the u.s. wage structure 1963–1987: Application of quantile regression. *Econometrica*, 62(2):405–458, 1994.

[Buc98] Moshe Buchinsky. The dynamics of changes in the female wage distribution in the usa: A quantile regression approach. *Journal of Applied Econometrics*, 13(1):1–30, 1998.

[Eur25] Eurostat. Gender pay gap statistics, 2025.

[Hec79] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[HM15] Martin Huber and Blaise Melly. A test of the conditional independence assumption in sample selection models. *Journal of Applied Econometrics*, 30(7):1144–1168, 2015.

[KB78] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[Koe17] Roger Koenker. Quantile selection models: An r vignette, 2017. Version January 27, 2017, vignette for the R package `quantreg`.

[Ös24] Bundesanstalt Statistik Österreich. Frauen verdienten 2022 brutto weniger als maenner. *Statistik Austria, Pressemitteilung: 13 284-050/24*, (1):1–5, 2024.