RATIONALITY
& COMPETITION
CRC TRR 190

Quantile Selection in the Gender Pay Gap

Egshiglen Batbayar (University of Bonn)
Christoph Breunig (University of Bonn)
Peter Haan (DIW Berlin, FU Berlin)
Boryana Ilieva (DIW Berlin, European Central Bank)

Discussion Paper No. 560

January 09, 2026

Quantile Selection in the Gender Pay Gap*

Egshiglen Batbayar[†] Christoph Breunig[‡]
Peter Haan[§] Boryana Ilieva[¶]

January 9, 2026

Abstract

We propose a new approach to estimate selection-corrected quantiles of the gender wage gap. Our method employs instrumental variables that explain variation in the latent variable but, conditional on the latent process, do not directly affect selection. We provide semiparametric identification of the quantile parameters without imposing parametric restrictions on the selection probability, derive the asymptotic distribution of the proposed estimator based on constrained selection probability weighting, and demonstrate how the approach applies to the Roy model of labor supply. Using German administrative data, we analyze the distribution of the gender gap in full-time earnings. We find pronounced positive selection among women at the lower end, especially those with less education, which widens the gender gap in this segment, and strong positive selection among highly educated men at the top, which narrows the gender wage gap at upper quantiles.

JEL codes: C14, C31, C36, J16, J21, J31
Keywords: Quantile regression, sample selection, Roy model, rank invariance, semiparametric inference, gender wage gap, wage inequality

[*] We thank Martin Biewen, Richard Blundell, Joachim Freyberger, Björn Höppner, Maximilian Schaller, Jan Scherer, Arne Uhlendorff and numerous seminar and conference participants for helpful comments and illuminating discussions. Christoph Breunig gratefully acknowledges the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2047/1 – 390685813. Peter Haan acknowledges the support by the DFG through CRC TRR 190 (project number 280092119), FOR 5675 (project number 518302089) and HA 5526/7-1.

[†] University of Bonn, e-mail: ebatbayar@uni-bonn.de

[‡] University of Bonn, e-mail: cbreunig@uni-bonn.de

[§] DIW Berlin and Freie Universität Berlin, e-mail: phaan@diw.de

[¶] DIW Berlin and European Central Bank, e-mail: boryana.ilieva@ecb.europa.eu

1 Introduction

In all countries, men continue to earn higher wages than women. Crucially, the gender wage gap spans the entire wage distribution and affects all educational groups (e.g., Goldin [2014], Blau and Kahn [2017], Olivetti et al. [2024]). To meaningfully quantify and analyze this gap, it is essential to account not only for gender differences in observable characteristics such as education and labor market experience, but also differences in selection into employment. The impact of selection on the gender wage gap is ambiguous and depends on the sign and magnitude of gender-specific selection patterns (see, e.g., Mulligan and Rubinstein [2008], Arellano and Bonhomme [2017], or Blau et al. [2024]). Understanding why men consistently out-earn women and designing policies to address this disparity requires tackling the challenging task of identifying and quantifying the role of employment selection in shaping the wage distribution, and of deriving a selection-corrected measure of the gender wage gap distribution.

In this paper, we propose a new strategy to derive the selection-corrected wage distribution and to quantify the full distribution of the gender wage gap. Identification relies on exogenous variation that affects latent wages but, conditional on the latent wages and other observed variables, provides no additional information on the selection mechanism. In the context of a Roy model [Roy, 1951], as applied to labor supply decisions by Gronau [1974] and Heckman [1974], where individuals choose to work if their potential wage exceeds their reservation wage, we show that our exogeneity requirements have the interpretation of a *rank invariance* condition: Conditional on potential wages, the instrument does not provide additional information about an individual's position in the reservation wage distribution.

An important contribution of our paper is that we do not restrict the functional form of the conditional selection probability. We provide semiparametric identification of selection-corrected quantile regression when the outcome is selectively observed and selection depends on unobservables. Identification is achieved by inverse probability weighting, where selection probabilities may depend on latent potential outcomes but are otherwise unaffected by the excluded variable. We establish identification of linear quantile regression even when nonparametric selection probabilities are not identified via conditional moment restrictions implied by the exclusion restriction. The constructive identification arguments motivate a flexible B-spline estimator, combined with cone projection, to estimate the initial selection probabilities. The resulting inverse probability quantile regression exhibits excellent finite-sample properties, and we establish its asymptotic behavior.

We apply the proposed estimation strategy to derive selection-corrected wage distributions for men and women into full-time employment and to quantify the distribution of the selection-corrected gender wage gap among full-time workers. The empirical analysis is based on administrative social security data from Germany (SIAB), which provides detailed information on the complete employment and earnings histories of individuals in the labor force. We use the initial wage as the instrument, which captures persistent individual heterogeneity, such as ability, motivation or social skills, that affect potential earnings over the life cycle. This is in line with previous studies showing that initial wages have persistent effects on future outcomes (see, e.g., Devereux [2002], Oreopoulos et al. [2012], Schwandt and von Wachter [2019]). In the Roy model, the exclusion restriction implies that the instrument conditional on the potential wage provides no additional information about an individual's position in the reservation wage distribution. Intuitively, once potential wages and observed characteristics are held fixed, individuals with the same rank in the distribution of reservation wages remain comparable regardless of their value of the initial wage.

Our results highlight the critical role of selection in shaping the gender wage gap. Women are positively selected into full-time employment across the wage distribution, with the strongest effects for lower and median wages. For men, selection effects are smaller but still positive and statistically significant. Conditional analyses yield a more nuanced picture. After controlling for education, experience, age, and workplace characteristics, selection correction consistently lowers women's estimated wages, particularly at lower quantiles and among the less educated. This indicates that full-time working women in these groups tend to possess unobserved attributes (e.g., motivation, ability) associated with higher wages, leading to a positively selected sample. The selection effect diminishes at higher quantiles and is weakest among highly educated women, implying that top female earners are less affected by unobserved selection and that full-time workers in this group more closely represent the broader female population. For men, positive selection is also present across education groups but follows a different pattern. Among the highly educated, median wages decline by about 10.7 euros after correction, compared with 5.6 euros for women, resulting in a 21% selection-corrected gender wage gap, which is 2 percentage points lower than the gender wage gap without selection correction. In contrast, selection effects are minimal among less educated men. These different selection patterns have important implications for the distribution of the gender wage gap. While the gender wage gap among the low educated increases from 3.6% to 10.7% at the median after correcting for selection, the selection-corrected gender wage gap for the high educated is slightly lower than the uncorrected wage gap. Overall, these results underscore the importance of accounting for selection heterogeneity across quantiles and education levels. Ignoring selection biases can mask substantial differences in the gender wage gap, particularly in the lower part of the wage distribution. In robustness checks, we impose stricter requirements on the lag structure of the instrument and find that the results remain virtually unchanged; see Appendix E.

Related Literature. Our approach extends the existing literature on endogenous selection and nonignorable nonresponse, where instrumental variables are used to restore identification when selection depends on the outcome variable itself. IV-based strategies similar in spirit to ours have been developed in both the statistics (see, e.g., Tang et al. [2003], Zhao and Shao [2015]) and econometrics (see, e.g., D'Haultfoeuille [2010], Ramalho and Smith [2013], Breunig et al. [2018]) literatures, as well as by Breunig and Haan [2020] for the case of selectively missing covariates. More recent contributions by Zhang and Wang [2020] and Yu et al. [2023] extend these ideas to quantile regression, proposing likelihood-based estimators under nonignorable selection that rely on parametric assumptions about the selection mechanism. In contrast, our method avoids such parametric restrictions, offering greater flexibility and robustness to model misspecification. While most applications of these statistical methods have focused on medical or clinical settings, we demonstrate their potential for economic contexts, particularly in analyzing labor supply and wage inequality.

By using an exclusion restriction for the underlying latent process, our strategy extends the existing literature analyzing the gender wage gap. Most previous studies focus on the average wage and address selection using the classical Heckman correction [Heckman, 1979], which requires an exclusion restriction for the selection into employment. Typically, researchers have relied on variables such as the number and age of children (e.g., Mulligan and Rubinstein [2008]) or on variation in the tax and transfer system, as suggested by Blundell et al. [2023].[1] However, these exclusion restrictions are often difficult to justify in practice, particularly when analyzing selection among men. Several studies have extended the classical Heckman model. Das et al. [2003], for example, develop a nonparametric approach and apply it to estimate returns to schooling among young Australian females. Of particular importance for our paper is the extension by Arellano and Bonhomme [2017], who propose a framework that allows for selection correction across the entire wage distribution, recognizing that selection patterns and their effects may vary across wage levels. Using UK data, they find positive selection for both men and women – more pronounced for men – and therefore report that the selection-corrected gender wage gap is smaller than the uncorrected one. Their identification strategy relies on the assumption that out-of-work benefit income is independent of unobserved determinants of wages.

Maasoumi and Wang [2019] apply a similar approach for the US and show that selection-corrected gender wage gaps have increased over time, both at the mean and across various quantiles of the wage distribution. Blundell et al. [2023] focus on gender differences in the distribution of lifetime earnings. They find that gender disparities widened during the first half of individuals' working lives but that substantial convergence occurred after age 40 for more recent cohorts, particularly compared to those born before the 1950s. Several alternative approaches have been proposed to study the distributional aspects of the gender wage gap (for an overview, see Blau et al. [2024]).

Blundell et al. [2007] develop and estimate bounds on the wage distribution to address non-random selection into employment. Although this method does not yield point identification, the resulting bounds are relatively tight. Their findings suggest a narrowing of the gender wage gap among the less educated. Relatedly, Honoré and Hu [2020] examine identification in classical semiparametric sample selection models without exclusion restrictions. They derive sharp bounds for the parameters of interest and show that, even in the absence of valid instruments, the identified set can be narrow in practice.

Blau et al. [2024] further propose imputation-based methods to address missing wage information, following techniques similar to those in Olivetti and Petrongolo [2008]. This approach has the advantage of correcting for selection across the full wage distribution. However, imputation methods rely solely on observed characteristics and are therefore based on the Missing at Random (MAR) assumption. Using such techniques, Blau et al. [2024] show that the reduction in the gender wage gap in the US over time becomes more pronounced when accounting for selection effects, although substantial disparities persist.

The paper is structured as follows. Section 2 outlines the modeling framework and establishes semiparametric identification, introducing the main assumptions. We also illustrate the identification strategy using the Roy model and show under which conditions it is nested within our sample selection framework. Section 3 details the estimation procedure and derives the asymptotic properties of the estimator, as well

[1] For a survey of studies using this approach, see Blau and Kahn [2017] and Blau et al. [2024].

as its finite-sample performance based on Monte Carlo simulations. The empirical application is discussed in Section 4, including a description of the data, the empirical specification, and the instrument. Section 5 reports the empirical findings, and Section 6 concludes. All proofs are provided in Appendices A1–A3. Appendix B describes the data preparation process and reports summary statistics for the wage variable at different stages of data cleaning. Appendix C illustrates the application of our estimator to the estimation of the distribution function. Appendix D contains additional simulation results and a detailed description of the simulation design. Finally, Appendix E reports robustness checks.

2 Semiparametric Identification

This section begins by outlining the assumptions required for identification. We then establish the identification of the quantile function using inverse selection probability weighting. We further extend our identification results more specifically to the Roy model of labor supply.

2.1 Setup and Identification

Given a latent outcome Y^* (for example, potential earnings from employment) and the participation indicator D (e.g., employment), we consider the following model for $\tau \in (0,1)$:

$$Y^* = Z^\top \theta_\tau + U_\tau, \quad \text{where} \quad \mathbb{P}(U_\tau \leq 0 \mid Z) = \tau \tag{2.1}$$

$$D = \mathbb{1}\{V \leq p(Y^*, X)\} \tag{2.2}$$

$$Y = Y^* \quad \text{if} \quad D = 1,$$

for some function $p(\cdot)$ that maps to the unit interval and where $\mathbb{1}\{\cdot\}$ denotes the indicator function. Here, U_τ and V are unobservable variables, and $Z = (1, X^\top, W^\top)^\top$, where X is a vector of covariates and the random vector W is excluded from the selection equation, which will be formally imposed below. We refer to W as instruments. Throughout the paper, we assume that a random sample of $(Y, D, X^\top, W^\top)$ is available. In the following, we provide sufficient conditions to identify the quantile regression coefficient θ_τ, which is the d_z–dimensional parameter of interest.

Identification is based on instrumental variables W that explain variations in the latent variable Y^* but are not directly related to the selection mechanism. This contrasts with parametric and semiparametric versions of the sample selection model of Heckman [1979], which use variables that explain the selection equation but not the outcome (wage) equation, for example Ahn and Powell [1993], Donald [1995], Chen and Khan [2003], Das et al. [2003] and Arellano and Bonhomme [2017], the latter extending Heckman's framework to quantile selection models.

Assumption 1. *(i) $V \perp\!\!\!\perp W \mid (Y^*, X)$ and $V \mid (Y^*, X) \sim \mathcal{U}(0,1)$; (ii) the selection probability p is uniformly bounded away from zero on its support; (iii) p is identified through the conditional moment restriction $\mathbb{E}[D/p(Y,X) \mid W, X] = 1$.*

Assumption 1(i) imposes conditional independence of the instrument W and the unobservables V in the selection equation. This assumption further imposes a uniform distribution of the conditional distribution of V, which is an innocuous normalization condition. Under Assumption 1(i) we conclude that

$$\mathbb{P}(D = 1 \mid Y^*, X, W) = \mathbb{P}(V \leq p(Y^*, X) \mid Y^*, X) = p(Y^*, X).$$

In particular, $p(Y^*, X) = \mathbb{P}(D = 1 \mid Y^*, X)$ by the law of iterated expectations, and thus we refer to p as the selection probability in the remainder of the paper.

Also note that Assumption 1(i) excludes any relation between W and the selection variable D that is not channeled through (Y^*, X). Assumption 1(ii) extends to the usual assumption of overlap required for the identification of population parameters under data missing at random, since herein the probability of selection can depend on the latent outcome.

Throughout the paper, we define the inverse selection probability function as $g(y, x) := 1/\mathbb{P}(D = 1 \mid Y^* = y, X = x)$. Under Assumption 1(i)(ii), this definition directly implies the following conditional moment restriction:

$$\mathbb{E}[Dg(Y, X) \mid W, X] = 1, \tag{2.3}$$

which, importantly, only involves observable random variables. Assumption 1(iii) ensures identifiability of the selection probability through this moment condition. In parametric settings, this requirement corresponds to the familiar rank condition, whereas in the nonparametric case it relies on a bounded completeness assumption, as shown by D'Haultfoeuille [2010], which can be more challenging to justify. While our first result establishes identification of the quantile regression coefficient θ_τ under Assumption 1(iii), we show below that this assumption can also be avoided by imposing a mild variational condition on the quantile of interest.

Under the standard quantile regression assumption that, for a given $\tau \in (0, 1)$, the conditional quantile function satisfies $Q_{Y^*}(\tau \mid Z = z) = z^\top \theta_\tau$ for some $\theta_\tau \in \mathbb{R}^{d_z}$, the parameter θ_τ can be characterized, following the check function argument of Koenker and Bassett [1978], as the solution to

$$\theta_\tau = \underset{\theta \in \mathbb{R}^{d_z}}{\arg\min}\, \mathbb{E}[\rho_\tau(Y^* - Z^\top \theta)], \tag{2.4}$$

where $\rho_\tau(u) = u(\tau - \mathbb{1}\{u < 0\})$ denotes the check function. Under Assumption 1, this parameter is identified and uniquely determined, as stated in the following proposition.

Proposition 2.1. *Suppose Assumption 1 holds. Then for any $\tau \in (0, 1)$, the quantile regression coefficient θ_τ is uniquely determined as the solution to*

$$\theta_\tau = \underset{\theta \in \mathbb{R}^{d_z}}{\arg\min}\, \mathbb{E}\left[D\, g(Y, X)\, \rho_\tau\big(Y - Z^\top \theta\big)\right]. \tag{2.5}$$

The next theorem establishes identification of our quantile selection model even if the inverse selection probability function g is not identified through the conditional moment condition (2.3), as imposed in Assumption 1(iii). For this result, we adopt an insight from Severini and Tripathi [2012] regarding point identification of functionals of partially identified structural functions.

Theorem 2.2. *Suppose Assumption 1(i)(ii) holds. Then, for any $\tau \in (0, 1)$ for which a square integrable function μ_τ exists that satisfies*

$$\mathbb{E}\left[\mu_\tau(W, X) \mid Y^*, X\right] = \mathbb{E}\left[\rho_\tau\big(Y^* - Z^\top \theta_\tau\big) \,\big|\, Y^*, X\right], \tag{2.6}$$

the quantile regression coefficient θ_τ is identified by (2.5).

Equation (2.6) requires that the instrument W carry sufficient information so that the conditional expectation $\mathbb{E}\left[\rho_\tau\left(Y^* - Z^\top\theta_\tau\right) \mid Y^*, X\right]$ can be expressed as an expectation of some measurable function of (W, X) given (Y^*, X). In other words, the assumption postulates the existence of a bridge function linking the residual loss function to the observed variables. This corresponds to a proximal identification requirement, where, conditional on X, the instrument W spans the same information space for the relevant conditional moment as the latent outcome Y^* does.

2.2 Application to the Roy Model

To illustrate identification, we use the Roy model [Roy, 1951] applied to the decision on labor supply by Gronau [1974] and Heckman [1974].[2] In this model, Y^* is the potential wage of employment and D is the indicator of labor market participation. According to the model, individuals choose employment if their potential wage Y^* exceeds their reservation wage R:

$$D = \mathbb{1}\{Y^* \geq R\}. \tag{2.7}$$

In the following, we explicitly derive under which conditions the Roy model is nested within our sample selection framework.

We assume that the reservation wage R is continuously distributed conditional on the potential wage Y^* and the observed characteristics X. Formally, the cumulative distribution function $F_{R|Y^*,X}(\cdot)$ is continuous and strictly increasing.

In the Roy model, labor market participation is determined by the comparison of potential and reservation wages:

$$\mathbb{P}(D = 1 \mid Y^*, X) = \mathbb{P}(Y^* \geq R \mid Y^*, X) = F_{R|Y^*,X}(Y^*).$$

Since $F_{R|Y^*,X}(\cdot)$ is strictly increasing, we can interpret $F_{R|Y^*,X}(R)$ as the *conditional rank* (or quantile position) of the reservation wage within the distribution of reservation wages among individuals with the same (Y^*, X). Under the continuity assumption, this rank is uniformly distributed on $(0, 1)$ conditional on (Y^*, X).

Our key independence condition is that this reservation-wage rank is independent of the instrument W once we condition on the potential wage and observed characteristics:

$$F_{R|Y^*,X}(R) \perp\!\!\!\perp W \mid (Y^*, X). \tag{2.8}$$

This assumption states that, conditional on productive ability and observable characteristics, the instrument does not provide information about unobserved determinants of labor supply decisions. It thus rules out selection on unobservables with respect to the instrument while allowing for rich and continuous heterogeneity in reservation wages. This condition can be interpreted as a *rank invariance* assumption (on the reservation wage), analogous to that used in quantile IV models (see Chernozhukov and Hansen [2005]). It relaxes the conditional independence assumption on the reservation wage itself, imposed in Example 2 in D'Haultfoeuille [2010]. The rank invariance condition (2.8) implies that, given (Y^*, X), the instrument W does not

[2] Mulligan and Rubinstein [2008] define this model as the Gronau-Heckman-Roy (GHR) labor supply model.

provide additional information about an individual's position in the reservation wage distribution. Intuitively, once the potential wages and observed characteristics are fixed, individuals with the same rank in the distribution of reservation wages remain comparable regardless of their value of W. Consequently, we obtain the following identification result.

Corollary 2.3. *For the Roy model* (2.7)*, suppose that Assumption 1(ii) holds and that R given (Y^*, X) is continuously distributed. In addition, assume that the rank invariance condition* (2.8) *is satisfied. Then, for any $\tau \in (0, 1)$ that satisfies* (2.6) *the quantile regression coefficient θ_τ is identified by* (2.5)*.*

In our empirical application (see Section 4), we use an individual's initial wage as the instrument W. The initial wage captures unobserved individual heterogeneity, such as ability, diligence, and both hard (education, technical expertise) and soft (teamwork, communication, negotiation) skills, that shape potential wages over the life cycle. This argument is consistent with explanations in the literature showing that initial wages have persistent effects on future outcomes (see, e.g., Devereux [2002], Oreopoulos et al. [2012], Schwandt and von Wachter [2019]). For example, Oreopoulos et al. [2012] show that career development models or models of human capital accumulation can explain why high initial wages lead to persistent higher wages.

The exclusion restriction in Corollary 2.3 implies that, once we fix (Y^*, X), the initial wage W does not carry further information on an individual's position in the conditional distribution of reservation wages. In other words, among individuals with the same potential wage and observable characteristics, the initial wage does not predict whether someone is at a high or low quantile of the reservation wage distribution. Note that this exclusion restriction is far less restrictive than assuming that the initial wage has no effect on the reservation wage. In fact, the exclusion restriction is not violated if the initial wage would shift the level of the conditional reservation wage while not changing the rank distribution.

In the empirical application, we discuss in detail how we construct the initial wage and argue that it is important to use wage information of the distant past to relax strong assumptions of a static labor supply model. For details, see Section 4.2.

3 Estimation and Inference

In this section, we sketch an estimation procedure that builds on the identification result. We then analyze the asymptotic properties of the estimator and evaluate its finite-sample performance through simulation evidence.

3.1 Estimation

Following the identification result, the estimation of the conditional quantiles of the latent outcome variable Y^* proceeds in three steps, summarized in Algorithm 1.

The first step estimates the inverse selection probability function $g(Y, X)$ using the moment condition:

$$\mathbb{E}\left[Dg(Y, X) \mid W, X\right] = 1, \quad \text{where } Y = DY^*.$$

Let $\phi^J(\cdot)$ and $b^K(\cdot)$ be vectors of basis functions of dimensions J and $K \geq J$, respectively. Denote by Φ_J the linear span of $\{\phi_1, \ldots, \phi_J\}$. We assume that the function $g(Y, X)$ lies in this linear span, i.e., $g(Y, X) = \phi^J(Y, X)^\top \beta$ for some coefficient vector $\beta \in \mathbb{R}^J$, where $\phi^J(Y, X)$ denotes a transformation of (Y, X). Multiplying both sides of the moment condition by the spline basis functions of (W, X) yields:

$$\mathbb{E}\left[Db^K(W, X)\phi^J(Y, X)^\top\right]\beta = \mathbb{E}\left[b^K(W, X)\right]. \tag{3.1}$$

In the second step, we enforce a lower bound on the estimated inverse selection probabilities by projecting the unconstrained estimator $\widehat{g}_u$ onto the cone of functions bounded below by 1. This ensures $\widehat{g}(y, x) \geq 1$ for all (y, x). The projection is implemented using the `coneproj` package in R. Cone projection of series TSLS estimators for constrained hypothesis testing was also considered by Breunig and Chen [2024].

In the last step, we estimate the conditional quantile function of the latent outcome variable Y^* via weighted quantile regression, using weights $D\,\widehat{g}(Y, X)$. This extends the quantile regression framework of Koenker and Bassett [1978] to account for selection, with inverse probability weights that depend jointly on the selection indicator and the outcome. For the algorithm, let $\mathcal{Y} \times \mathcal{X}$ denote the support of (Y, X), and let $\mathbb{E}_n[\cdot]$ denote the empirical mean. For (Y, X), we define $L^2(Y, X) = \{h : \|h\|_{L^2(Y,X)} < \infty\}$, where $\|h\|_{L^2(Y,X)} := \sqrt{\mathbb{E}[h^2(Y, X)]}$. The conditional probability density function of a random variable Y given Z is denoted by $f_{Y|Z}(\cdot)$. Throughout the paper, we assume that a sample $\{S_i\}_{i=1}^n$ is observed, where $S_i = (D_i, Y_i, W_i^\top, X_i^\top)$.

Algorithm 1 Estimation Procedure with Inverse Probability Weighting

Input: Sample data $\{S_i\}_{i=1}^n$ and quantile level $\tau \in (0, 1)$.

1: **Estimate the inverse selection probability function.** Obtain the unconstrained estimator $\widehat{g}_u(\cdot) = \phi^J(\cdot)^\top\widehat{\beta}$, where $\widehat{\beta}$ is the 2SLS coefficient regressing the constant 1 on $D\phi^J(Y, X)$ with $b^K(W, X)$ as instruments.

2: **Impose the shape restriction.** Project $\widehat{g}_u$ onto the cone $\mathcal{C} := \{g \in L^2(Y, X) : g(y, x) \geq 1 \text{ for all } (y, x) \in \mathcal{Y} \times \mathcal{X}\}$ by solving

$$\widehat{g} = \underset{h \in \mathcal{C} \cap \Phi_J}{\arg\min}\ \mathbb{E}_n\Big[\big(\widehat{g}_u(Y, X) - h(Y, X)\big)^2\Big].$$

3: **Estimate the conditional quantile function of Y^*.** Estimate the quantile parameter θ_τ by

$$\widehat{\theta}_\tau = \underset{\theta \in \mathbb{R}^{d_Z}}{\arg\min}\, \mathbb{E}_n\Big[D\,\widehat{g}(Y, X)\,\rho_\tau(Y - Z^\top\theta)\Big]. \tag{3.2}$$

Output: Estimates $\widehat{\theta}_\tau$ and conditional quantile function $\widehat{Q}_{Y^*}(\tau \mid Z = z) = z^\top\widehat{\theta}_\tau$.

3.2 Asymptotic Properties

In this subsection, we present the asymptotic distribution for the quantile selection estimator $\widehat{\theta}_\tau$. To establish the result, we introduce some notation. Let $\Omega_g := Dg(Y, X)$

be the selection weights and let $\psi_\tau(u) := \tau - \mathbb{1}\{u < 0\}$ be the quantile score function. Define $T := \mathbb{E}\left[\phi^J(Y,X)DZ^\top\psi_\tau(U_\tau)\right]$, where $U_\tau = Y^* - Z^\top\theta_\tau$.[3] The matrices G and H capture key population moments involving the basis functions and the first-stage estimation errors: $G := \mathbb{E}[b^K(W,X)b^K(W,X)^\top]$, $H := \mathbb{E}[D\phi^J(Y,X)b^K(W,X)^\top]$. Let $U_J := 1 - D\phi^J(Y,X)^\top \beta$ be the first-stage residual. To obtain our asymptotic distribution result, we assume that the coefficient β is identified through the moment condition (3.1). Under additional assumptions as specified in Appendix A2, the influence function, as derived in the next theorem, can be written as:

$$\chi_{\tau g}(S_i) := M_{1\tau g}^{-1}\left(Z_i\Omega_{gi}\psi_\tau(U_{\tau i}) + T^\top(HG^{-1}H^\top)^{-1}HG^{-1}b^K(W_i,X_i)U_{Ji}\right),$$

where $M_{1\tau g} := \mathbb{E}\big[\Omega_g f_{Y|\Omega_g,Z}\left(Z^\top\theta_\tau\right)ZZ^\top\big]$. We now present the asymptotic distribution of our quantile selection estimator $\widehat{\theta}_\tau$.

Theorem 3.1. *Suppose Assumptions 1(i)(ii), A2, and A3 hold. Then, for any* $\tau \in (0,1)$ *that satisfies* (2.6) *we have*

$$\sqrt{n}\left(\widehat{\theta}_\tau - \theta_\tau\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\chi_{\tau g^*}(S_i) + o_p(1)$$

and in particular

$$\sqrt{n}\left(\widehat{\theta}_\tau - \theta_\tau\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\big[\chi_{\tau g^*}(S)\chi_{\tau g^*}^\top(S)\big]\right),$$

where g^* *satisfies* (2.3) *and minimizes* $\mathbb{E}\big[\chi_{\tau g}(S)\chi_{\tau g}^\top(S)\big]$.

This result provides the asymptotic distribution of the quantile regression estimator. The detailed regularity conditions required to establish this asymptotic result, together with its proof, are given in Appendix A2. The asymptotic covariance estimator used in our implementation is presented in Appendix A3.[4]

3.3 Monte Carlo Studies

We evaluate the finite-sample properties of the estimator compared to existing approaches, using 1,000 replications with $n = 1{,}000$ observations per replication.

We consider the outcome model

$$Y_i^* = \beta_0 + \beta_1 W_i + \beta_2 X_i + \epsilon_i(\tau), \quad i = 1, \ldots, n,$$

where $(W_i, X_i)^\top \sim \mathcal{N}\left(\begin{pmatrix}2\\1\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$ and $(\beta_0, \beta_1, \beta_2) = (1,1,2)$. The error term $\epsilon_i(\tau)$ ensures a zero conditional τ-quantile. In this section, we present results for $\epsilon_i(\tau)$ drawn from a t-distribution with 3 degrees of freedom, scaled by 0.7, while sensitivity to alternative error distributions is presented in Appendix D.

[3] Note that we can replace Y^* by Y whenever U_τ is multiplied by D, as $Y = DY^*$

[4] In our application, we estimate the conditional densities using the `npcdens` package in R, which implements the method of Racine and Li [2004], with bandwidths selected via the cross-validation procedure of Hall et al. [2004].

Selection into the observed sample is modeled through a binary indicator D_i, which equals one if individual i is observed and zero otherwise. Conditional on the latent outcome Y_i^* and covariates X_i, we assume a logistic selection model:

$$D_i \sim \text{Bernoulli}(p(Y_i^*, X_i)), \quad \text{where } p(Y_i^*, X_i) = \frac{1}{1 + \exp(-s(Y_i^*, X_i))}.$$

The function $s(\cdot)$ determines the selection mechanism. We consider three specifications that differ in how selection depends on observed covariates and the latent outcome. In the first mechanism (M1, MAR), selection depends only on observed covariates, with $s(Y_i^*, X_i) = \alpha_1 + \gamma_1 X_i$. The second (M2, MNAR-linear) introduces dependence on the latent outcome in a linear form, $s(Y_i^*, X_i) = \alpha_2 + \gamma_2 X_i + \xi_2 Y_i^*$. The third mechanism (M3, MNAR-nonlinear) allows for nonlinear dependence on the covariate, $s(Y_i^*, X_i) = \alpha_3 + \gamma_3 \sin^2(X_i) + \xi_3 Y_i^*$. The parameters $(\alpha_j, \gamma_j, \xi_j)$ are chosen such that the average missing-data rate is approximately 35%. The design mirrors our empirical application: W serves as an instrumental variable that affects the latent outcome but is excluded from the selection equation, paralleling the role of initial wages as instruments in our application. The exact parameter values and further implementation details are provided in Appendix D.

We compare four estimation methods at the median quantile, $\tau = 0.5$. First, the uncorrected complete-case estimator relies solely on the observed outcomes, ignoring the missing-data mechanism and therefore serving as a naive benchmark. Second, the MAR correction assumes that selection depends only on observable characteristics and is implemented through a two-step inverse probability weighting procedure. Third, we include the joint estimating equations (JEE) method of Yu et al. [2023], which, similar in spirit to ours, employs an IV but specifies a parametric form for the selection probability and jointly estimates the quantile and selection equations via an augmented likelihood framework. Finally, our semiparametric IV method corrects for selection using an instrument without imposing any parametric restrictions on the selection probability, providing greater flexibility and robustness to model misspecification.

Our estimation procedure follows Algorithm 1. In the first step, we estimate $\widehat{g}_u(Y, X)$ using quadratic B-splines – without interior knots for Y and with two interior knots for X – yielding three and five basis functions, respectively. Including one control variable, this yields dimensions $J = 4$ and $K = 6$. In the second step, we impose the lower-bound shape constraint using the `coneproj` package in R to obtain the constrained estimate $\widehat{g}(Y, X)$. Finally, quantile coefficients are estimated using weighted quantile regression with weights $D\widehat{g}(Y, X)$. The covariance matrix and confidence intervals are computed using the plug-in estimators detailed in Appendix A3.

While the full set of simulation results and the detailed description of the setup – covering all error distributions and selection mechanisms – are provided in Appendix D, Table 1 summarizes results for a representative setting with heavy-tailed errors: t-distributed errors with 3 degrees of freedom, scaled by 0.7. The proposed semiparametric IV estimator exhibits negligible bias and low RMSE across all selection mechanisms. Under M1, where selection is independent of the latent outcome given covariates, all estimators perform comparably. The proposed method tends to produce slightly wider confidence intervals, reflecting its robustness against more complex forms of selection. In contrast, when selection depends on the latent outcome (M2 and M3), both the MAR-assumed and complete-case estimates suffer from

Table 1: (Setting C) Simulation results by method with $(\tau, n) = (0.5, 1000)$ and 35% missing in the outcome

Method	M1			M2			M3		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Mean biases									
Uncorrected	0.005	-0.002	0.001	0.249	-0.025	-0.076	0.255	-0.025	-0.077
MAR-assumed	0.004	-0.001	0.001	0.317	-0.033	-0.108	0.296	-0.034	-0.089
JEE	0.001	-0.004	0.005	0.010	-0.012	0.000	0.005	0.001	-0.008
Semiparametric IV	-0.033	0.009	0.012	0.042	0.005	-0.018	0.056	0.003	-0.029
RMSE									
Uncorrected	0.094	0.044	0.043	0.271	0.048	0.090	0.275	0.050	0.090
MAR-assumed	0.100	0.047	0.047	0.377	0.072	0.156	0.362	0.077	0.135
JEE	0.120	0.102	0.125	0.841	0.362	0.330	0.714	0.265	0.179
Semiparametric IV	0.110	0.047	0.048	0.131	0.050	0.060	0.133	0.049	0.064
CI lengths									
Uncorrected	0.355	0.170	0.178	0.408	0.170	0.186	0.402	0.171	0.184
MAR-assumed	0.371	0.175	0.191	0.668	0.230	0.382	0.663	0.250	0.357
JEE	0.716	0.306	0.332	1.063	0.406	0.495	0.994	0.360	0.447
Semiparametric IV	0.701	0.344	0.318	0.636	0.316	0.256	0.674	0.327	0.263
Coverage probabilities									
Uncorrected	0.943	0.944	0.963	0.343	0.921	0.639	0.302	0.903	0.626
MAR-assumed	0.935	0.932	0.961	0.409	0.891	0.657	0.480	0.885	0.709
JEE	0.941	0.935	0.955	0.879	0.927	0.882	0.854	0.924	0.882
Semiparametric IV	0.989	1.000	0.990	0.965	0.994	0.956	0.977	0.993	0.948

substantial bias, high RMSE, and poor coverage. The joint EE method also fails under M2 and M3, with sharp increases in RMSE and confidence interval length. Overall, the semiparametric IV method demonstrates the most robust performance, consistently achieving the lowest RMSE and the highest coverage even in the presence of selection dependent on the latent outcome. Moreover, the joint EE method is computationally more intensive, as it relies on bootstrapping to estimate standard deviations. For large samples, as in our empirical application, this makes it difficult to implement the estimator, whereas our method is considerably faster.

4 Empirical Application

In this section, we present the empirical application and outline the implementation procedure. We begin by describing the data and providing summary statistics of the key variables. Next, we discuss the construction and choice of the instrumental variable. We then detail the implementation of our proposed selection correction method, explaining how it is applied to the data alongside the benchmark estimations – both the uncorrected model and the selection correction under the MAR assumption.

4.1 Data Description and Summary Statistics

The empirical analysis is based on administrative social security data from Germany.[5] The data include individuals who have held employment subject to social security contributions or marginal part-time employment at least once during the observation period, as well as benefit recipients and job seekers.[6] It provides comprehensive information on the employment history, including employment spells, daily wages, and various demographic characteristics such as gender, age, education, work experience, and occupation.

Our analysis is based on the 2017 wave of the dataset. We focus on individuals between the ages of 25 and 50 who have held German nationality throughout their entire recorded labor history. Our sample includes i) individuals in part-time and full-time employment, ii) recipients of unemployment benefits and means-tested transfers, and iii) job seekers not receiving unemployment benefits. We exclude individuals who are still in education or vocational training, and those who are partially retired. Additionally, our analysis excludes self-employed individuals and civil servants, as the administrative records do not cover these employment groups. Finally, we do not observe individuals who are part of the working-age population but are not in the labor force because they are not actively seeking employment. Thus, we analyze the effects of non-random selection only of the individuals who are in the labor force.

The administrative records offer three major advantages for analyzing gender-specific wage distributions, and the distribution of the gender wage gap for full-time workers. First, the data cover a broad and representative segment of the population with a large sample size, enabling the use of nonparametric estimation techniques, which typically require larger datasets than parametric methods.[7] Second, the data are extensive and precise. Unlike self-reported survey data, these records avoid issues of misreporting and non-random missing observations in wage data. Finally, the data include information about the full earnings and working history with occupational details, which is required to construct the instrument.

Our key variable of interest is the gross daily wage in full-time employment, which is recorded for every spell of full-time employment.[8] To construct a panel by year and individual, we closely follow Dauth and Eppelsheimer [2020]: we record individuals with spells spanning June 30th of each year in the panel and use the wage from the longest employment spell within the respective year as the gross daily earnings for that year. The gross daily wage is top-coded due to social security insurance limits, meaning wages above this threshold are not fully observed.[9] This censoring restricts the analysis of the full wage distribution. Therefore, we trim the top 5% of the wage distribution by gender, year, and region, removing individuals whose earnings ever

[5] We use the factually anonymous version of the Regional File of the Sample of Integrated Labor market Biographies 1975-2017 (SIAB-R 7517), a 2% random sample drawn from the Integrated Employment Biographies (IEB), provided by the Institute for Employment Research (IAB) [Antoni et al., 2019].

[6] Marginal part-time employment is recorded from 1999 onward. Following the German unification, for East Germany recorded information is assumed to be complete from 1993 onward.

[7] While the large sample size is beneficial, our estimation strategy remains applicable even in smaller samples.

[8] The data do include information about working hours. Therefore, we focus on the gender wage gap in full-time employment.

[9] The limit's assessment ceiling varies by year and between East and West Germany. In 2017, it was 76,200 euros in West Germany and 68,400 euros in East Germany.

fall within the top 5%. Additionally, to avoid outliers, we trim the bottom 5% of the wage distribution. The data preparation process and the sample restriction are discussed in detail in Appendix B. Overall, we use 119,930 observations for females and 124,588 observations for males in our empirical analysis.

Table 2 provides information about gross daily wages in full-time employment and an overview of the sample's descriptive statistics.[10] Consistent with previous evidence, we observe a clear raw gender wage gap in full-time earnings. On average, men earn about 30 euros more per day than women, corresponding to a raw gender wage gap of approximately 25%. The median wage gap is slightly smaller, around 20%.[11]

Education is classified into three groups based on the completion of a university degree or vocational training.[12] Educational attainment is generally similar across genders: the majority have completed vocational training, while a smaller proportion hold a university degree.

The employment status is derived from mandatory social security notifications submitted by employers. The distinction between full- and part-time work is based on a comparison of contracted hours with the standard working hours at the respective establishment. Non-employment spells are identified through the receipt of unemployment benefits or means-tested transfers. A notable disparity in employment status is observed: more than half of women (52%) work part-time, compared to only 10% of men. In contrast, the majority of men (83%) are employed full-time, while only 42% of women are. The share of individuals in the labor force but currently out of employment is similar for men and women. To account for regional differences in labor market trajectories, we classify individuals by workplace region, distinguishing between East and West Germany, which evolved differently after reunification. In 2017, around 76% of both men and women in our sample were employed in West Germany.

The data enable us to reconstruct detailed employment and wage histories for all individuals. We aggregate past employment into part-time and full-time experience and account for periods of non-employment.[13] Table 2 shows that women have significantly more part-time work experience – five years on average compared to one year for men – while men have more full-time experience, averaging 11.6 years versus 7.3 years for women. The median part-time experience for men is zero, whereas women have accumulated a median of three years.[14] To account for initial career starting points, occupational differences, and potential variation in wage trajectories across occupations, we classify individuals based on the difficulty level of their earliest job. This job also determines their initial wage, which we use as an instrument in our

[10] Wages are adjusted for inflation, using 2015 as the base year.

[11] The gender gap in full-time earnings in our data is slightly higher than that reported in Ilieva and Wrohlich [2022], who focus on hourly wages based on SOEP data.

[12] For more details, see Dauth and Eppelsheimer [2020]. Note that the proportion of university graduates in the general population is higher than in our sample, as the top 5% of the wage distribution has been trimmed, potentially excluding individuals with higher educational qualifications.

[13] Some unrecorded periods or data gaps may reflect time spent abroad or other forms of employment. Since we cannot verify these activities, we do not classify them as full-time, part-time, or non-employment.

[14] In the empirical analysis, we control for experience but do not interpret the coefficients. Experience is highly correlated with age, and part-time employment is often linked to the presence and age of children, which are not observed in the data, making a meaningful interpretation challenging.

Table 2: Descriptive Statistics

	Women		Men	
	Mean/Share	Median	Mean/Share	Median
Gross Daily Full-Time Wage (in Euro)	91.26	91.18	112.86	107.84
Employment Status (%)				
Non-employed	6.23		6.67	
Part-time	52.18	✓	10.09	
Full-time	41.59		83.24	✓
Age	38.74	39.00	37.95	38.00
Education Level (%)				
No vocational training	5.18		5.55	
Vocational training	77.88	✓	77.79	✓
University	16.94		16.66	
Workplace (%)				
Non-employed	6.23		6.67	
Employed in East Germany	17.70		17.35	
Employed in West Germany	76.07	✓	75.97	✓
Experience (in years)				
Non-employed	0.77	0.00	0.87	0.00
Part-time	4.86	3.00	0.96	0.00
Full-time	7.30	6.00	11.59	10.00
Earliest job difficulty (%)				
Unskilled/semiskilled task	6.38		6.75	
Skilled task	80.03	✓	82.11	✓
Complex task	4.22		3.65	
Highly complex task	9.37		7.49	
Individuals	119,930		124,588	

Notes: SIAB-R data, 2017 cross-section. The sample includes individuals aged 25-50, all German nationals. Education reflects the highest degree attained. Employment status is based on employers' mandatory social security notifications. Wages represent gross daily full-time earnings from each individual's longest employment spell in 2017.

analysis.

4.2 Instrument

As mentioned above, we use an instrument W based on the individual's wage history constructed to satisfy the exclusion restriction. The instrument explains variation in the potential wage but, conditional on this latent outcome and observable characteristics, does not directly affect the employment decision. We argue that early wage history satisfies this restriction because it captures persistent individual heterogeneity – such as ability, diligence, and both hard (education, technical knowledge) and soft (teamwork, communication, negotiation) skills – that shape potential wages over the life cycle. Under this assumption, the initial wage influences labor market participation only indirectly, through its effect on the potential wage. This interpretation aligns with theories of career development, which suggest that higher initial wages provide greater opportunities for human capital accumulation and are therefore positively correlated with future skill acquisition and long-term earning potential [Oreopoulos et al., 2012].

Arguably, the assumptions of a static labor supply model may be violated due

Table 3: Instrument: Initial wage

	Women			Men		
	Mean	Min	Max	Mean	Min	Max
Initial wage (in euro)	50.26	4.29	150.00	60.25	4.46	192.81
Age at initial wage	22.22	17	48	21.97	17	48
Years to 2017	16.52	2	33	15.97	2	33.00
Share full-time (%)	93.26			98.38		
Individuals	119,930			124,588		

Notes: SIAB-R data, 1975-2017. The initial wage is the daily wage at the earliest available point in the individual's wage history within the SIAB dataset.

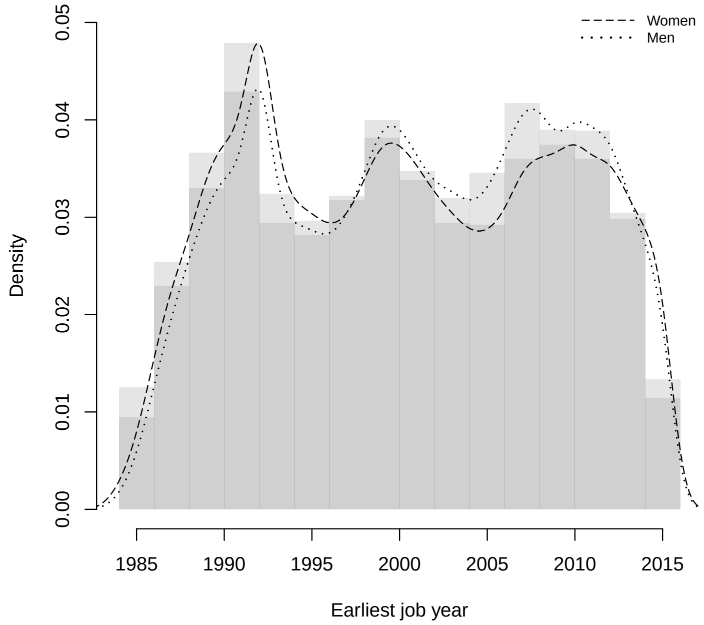
to factors such as state dependence in preferences [Heckman, 1991], labor market dynamics, wage persistence [Meghir and Pistaferri, 2011], or intertemporal financial incentives. Specifically, wages from recent periods may influence current employment decisions. For example, unemployment insurance payments or other transfers are often determined by recent earnings and employment histories, which can directly affect reservation wages and thereby the selection process in the current period. Similarly, persistent earnings shocks or mean reversion may directly influence the current period's offered wage, again linking recent wages to labor market participation.

To address these dynamic channels within our identification strategy, we exploit detailed information on each individual's employment and wage history. Specifically, we use the earliest observed wage as our main instrument, referring to it as the *initial wage*. This choice mitigates concerns that contemporaneous or short-lagged wages could directly affect current employment decisions.[15] When constructing the instrument, we prioritize full-time earnings from the earliest available date; if an individual has never worked full-time, we instead use their earliest part-time earnings.[16] Finally, we require that the initial wage be observed in 2015 or earlier, i.e., at least two periods before the current observation $t-2$, to ensure that the conditional independence assumption underlying our identification is plausibly satisfied.

[15] In Appendix E, we show that our results are robust when considering only wages from 2011 or earlier (i.e., at least five years back) as valid instruments.

[16] Ancillary analyses in Appendix E confirm that our results are robust when restricting the instrument to only past full-time wages.

Figure 1: Years of initial wages



Note: The figure shows the histogram and density of the earliest job years, with dashed lines representing women and dotted lines representing men.

Table 3 summarizes key characteristics of the instrumental variable, which captures each individual's earliest observed wage in the dataset. On average, men's initial wages are higher than women's, at approximately 60 and 50 euros per day, respectively. The instruments originate from an average age of 22 and reach back around 16 years for both genders. Moreover, 98% of men's and 93% of women's instruments are based on full-time wages, reflecting the higher incidence of part-time employment among women. Figure 1 displays the distribution of the instrument years, indicating that women's wage histories extend slightly further back in time than men's.

In Table 4, we assess the relevance of the instrument by regressing log full-time wages observed in 2017 on the initial wage variable. We additionally control for age, educational attainment, workplace region, and total experience, all measured as of 2017, as well as the difficulty level of the earliest job. Total experience is calculated as the sum of full-time experience and half the part-time experience (i.e., one year of part-time work counts as 0.5 years). The coefficient on the initial wage is strongly positive and statistically significant for both men and women, underscoring the predictive power of early labor market conditions for later-life earnings.

4.3 Implementation Details

We implement quantile regression using three different approaches. First, we apply our proposed IV-based selection correction method. Second, we consider a commonly used selection correction approach that relies on the MAR assumption. Third, we estimate quantile regressions based solely on the observed full-time wages, without correcting for selection. These three methods were compared in detail in the simulation study presented in Section 3.3.

Table 4: OLS Regression Results

	Women:		*Men:*	
Initial Wage (log)	0.216 (0.004)	0.183 (0.004)	0.272 (0.004)	0.183 (0.002)
Age		-0.012 (0.000)		-0.011 (0.000)
Education Level		0.197 (0.004)		0.250 (0.002)
Workplace		0.078 (0.004)		0.205 (0.003)
Total Experience		0.018 (0.000)		0.019 (0.000)
Earliest Job Difficulty		0.029 (0.003)		0.066 (0.002)
Constant	3.580 (0.016)	3.327 (0.021)	3.537 (0.010)	3.017 (0.012)
Observations	49,876		103,703	
R^2	0.103	0.130	0.058	0.285

Note: The dependent variable is the log of full-time wages in 2017. The regression sample includes only individuals who were in full-time employment in 2017.

IV-based selection correction We apply our three-stage procedure described in Section 3.1 to estimate selection-corrected wage quantiles separately for men and women. As outlined in Algorithm 1, the first two stages involve estimating the constrained inverse selection probabilities using the initial wage (log) as an instrument, while controlling for the same covariates X as in Table 4. In the third stage, we use the estimated weights to recover the selection-corrected conditional quantile function of the latent outcome, based on the observed full-time wages. This estimation is carried out conditional on both the covariates X and the instrument, in line with Equation (2.1). For unconditional quantile estimation, no covariates are included in the third stage. All estimations are conducted separately by gender. Since our analysis focuses on selection into full-time employment, we define $D = 1$ when full-time wage information is observed. Wages of part-time workers and the absence of labor earnings among non-employed individuals are treated as unobserved, corresponding to $D = 0$.

Comparison to MAR correction and uncorrected estimates A common approach assumes that, conditional on covariates, data are missing at random (MAR), as described in Section 3.3. Under this assumption, selection depends only on observable characteristics and not on unobserved outcomes. We implement the MAR-based selection correction using a two-step procedure: first, we estimate inverse probability weights via a probit model; second, we use these weights to estimate the quantile functions. Following prior studies that apply imputation-based methods to address selection, such as Olivetti and Petrongolo [2008] and Blau et al. [2024], we use the same set of variables as in the IV correction but exclude the instrument (initial wage) and its related variable (earliest job difficulty) from both stages of the MAR correction method. For comparison, we also report uncorrected estimates based solely on individuals with observed wages. These are obtained from standard quantile regressions using the same covariates as in the MAR specification. Although straightforward to implement, such estimates may be biased if sample selection is non-random.

5 Results

We begin by presenting the overall patterns in the unconditional wage quantiles of men and women, comparing our IV-based selection-corrected estimates to both uncorrected estimates and those derived under the MAR assumption. We then turn to the conditional wage quantiles, which account for gender differences in education, experience, and other relevant covariates. These analyses allow us to assess how selection operates across the wage distribution within and between groups. Finally, we quantify the magnitudes of the selection effects and their implications for the gender wage gap across three selected quantiles.

5.1 Selection Effects in the Unconditional Wage Quantiles

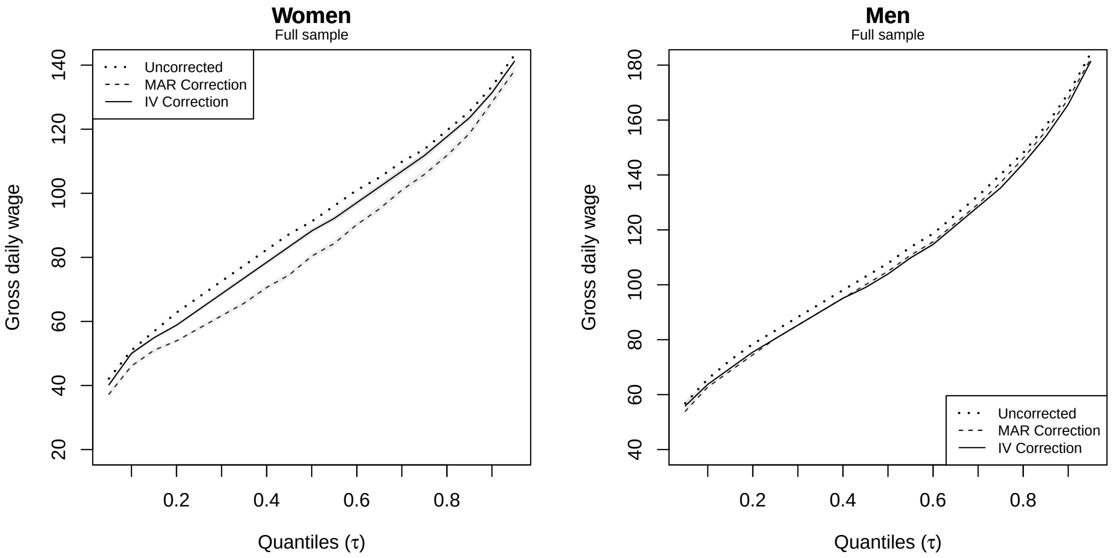
The results document sizable and statistically significant positive selection effects for both men and women throughout the wage distribution. Figure 2 displays estimated wage quantiles for men and women, comparing our IV-based selection-corrected estimates with both the uncorrected and MAR-corrected counterparts. The corresponding uncorrected and IV-corrected wage estimates at the 25th, 50th, and 75th percentiles, along with the implied gender wage gaps, are presented in the first panel of Table 5.

The selection-corrected unconditional wage quantiles for both women and men lie consistently below the empirical quantiles of wages of observed full-time workers. This indicates positive selection into full-time employment for both genders: individuals who work full-time tend to have higher potential wages than those who do not. As a result, the observed distribution overstates overall wage levels by disproportionately reflecting high-earners.

Among women, the selection effect is strongest at the lower quantiles and around the median, gradually tapering off toward the upper tail. For example, the empirical median daily wage for women is approximately 91 euros, while the IV-based selection-corrected estimate is about 88 euros, reflecting a 3.2% reduction. At the 25th percentile, the correction lowers wages by 5.8%, whereas at the 75th percentile the effect is more modest at 1.7%. For men, the magnitude of selection is more uniform across the distribution, averaging around 3.5%. This pattern of positive selection echoes classic insights from Heckman [1979] and empirical findings such as those by Wang et al. [2014]. It is consistent with the Roy model of labor supply, according to which individuals with higher potential earnings are more likely to select into full-time work, while those with lower potential wages tend to work part-time or remain non-employed. Since both men and women are positively selected, but to different degrees across the distribution, the gender wage gap tends to increase below the median and decrease above it. The findings generally align with and refine the distributional evidence in Arellano and Bonhomme [2017], who also find positive selection for both genders but with a larger bias for men in the UK. Our German data reveal that selection among women is particularly influential in the lower half of the distribution, which helps explain why the median gap barely changes while the 25th-percentile gap widens.

Comparing the MAR and IV correction results, we find that both correction methods indicate positive selection for women, with stronger effects in the MAR correction. For men, both approaches imply positive selection effects of similar magnitude. Discrepancies between the methods' result can arise because the MAR approach assumes that selection depends only on observed characteristics, ignoring unobserved factors that jointly affect wages and labor supply decisions. The problem is amplified for women, who exhibit substantially higher rates of non-employment and part-time work. As shown in the Monte Carlo simulations in Section 3.3, the MAR approach performs well only when selection is driven solely by observables, but fails when unobserved heterogeneity plays a role. Consequently, the MAR correction tends to over-adjust in the presence of unobserved selection, producing biased estimates, as illustrated in Figure 2.

Figure 2: Quantile Selection Effects (Unconditional)

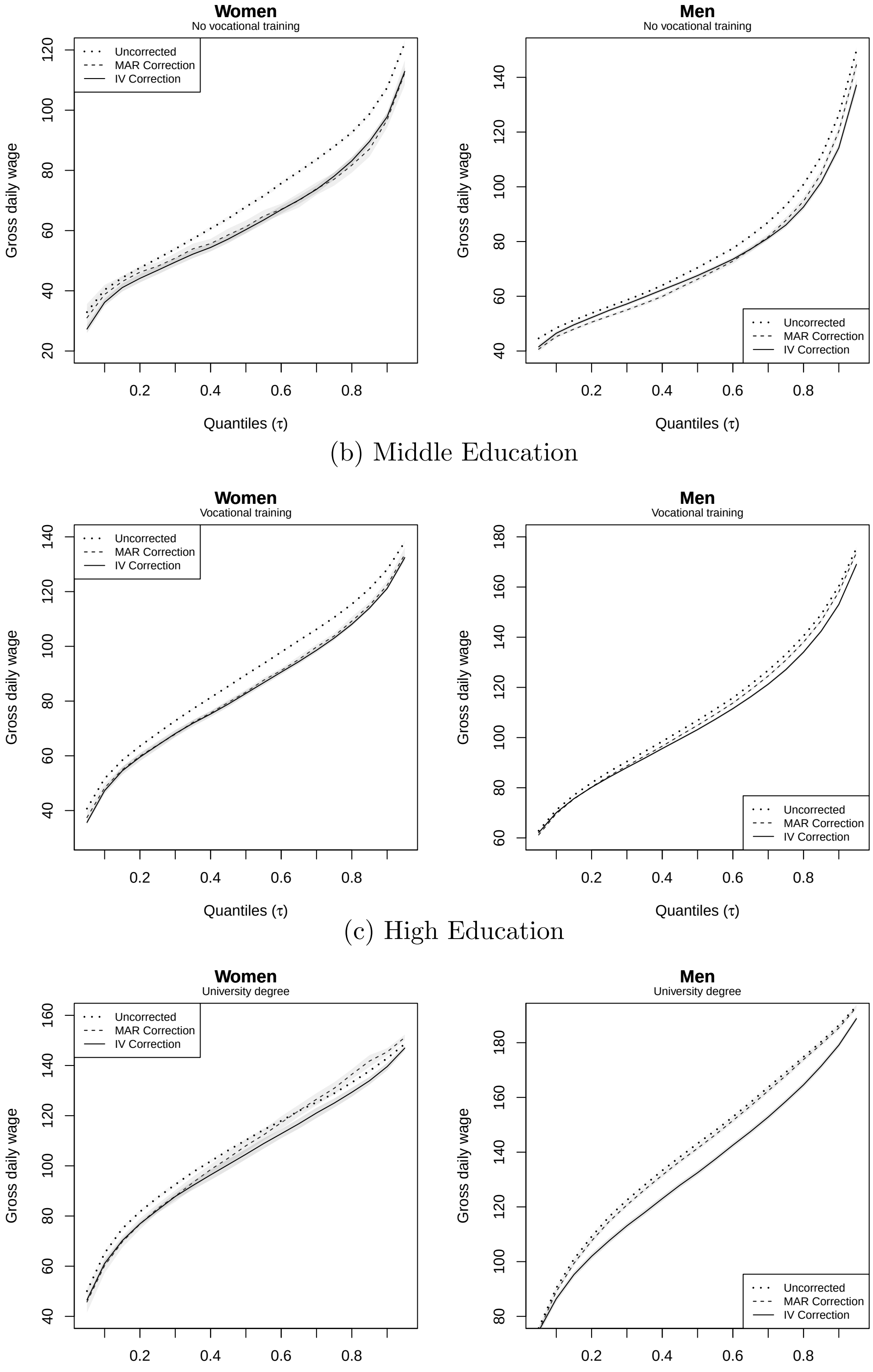


Note: SIAB-R data, 2017. Curves display wage point estimates from unconditional quantile regressions across quantiles, connected by lines for visualization: uncorrected (dotted), MAR correction (dashed), and IV correction (solid). The IV correction uses initial log wage as an instrument and controls for education, age, workplace region, total experience, and earliest job difficulty in the first two stages. The MAR correction uses the same controls except earliest job difficulty. The final-stage regressions include no covariates for all methods. Shaded areas represent 95% pointwise confidence intervals.

5.2 Selection Effects in the Conditional Wage Quantiles

Unconditional wage analyses offer only limited insights, as they do not control for differences in observable characteristics like education, experience, or workplace factors. This can confound true wage disparities with compositional differences across groups. To address this, we turn to conditional quantile regression, which allows us to examine how wages vary across the distribution while accounting for important covariates. Thus, this method captures heterogeneous effects and selection patterns that are often hidden in mean-based or unconditional approaches, providing a more detailed understanding of wage inequality.

Figure 3: Conditional wage quantiles by education groups



(a) Low Education

(b) Middle Education

(c) High Education

Note: SIAB-R data, 2017. Curves show wage point estimates from unconditional quantile regressions across quantiles, connected by lines for visualization: uncorrected (dotted), MAR correction (dashed), and IV correction (solid). The IV correction uses initial log wage as an instrument and controls for education, age, workplace region, total experience, and earliest job difficulty. The MAR correction uses the same controls, excluding the instrument and earliest job difficulty. The uncorrected estimates include the same controls as in MAR. Shaded areas represent 95% pointwise confidence intervals.

Figure 3 displays the conditional wage quantiles by gender and education level, comparing uncorrected estimates based on observed full-time wages with selection-corrected estimates obtained under the MAR assumption and our IV-based method. For women, both correction methods yield broadly similar results across most of the distribution, except at the upper quantiles among the highly educated, where the MAR correction indicates a small negative selection effect beyond the 60th percentile, while the IV-based approach still points to positive selection. For men, selection patterns differ by education level. Among those with low or medium education, differences between the MAR and IV-based methods are relatively minor and occur primarily in the upper parts of the wage distribution. The MAR correction indicates little to no selection effect, while our IV-based method reveals a substantial and statistically significant positive selection.

Overall, wage estimates for women based on the IV method are consistently lower than the uncorrected values across all education groups, with the largest corrections observed among low-educated women and smaller adjustments among the highly educated. This pattern suggests that unobserved characteristics strongly influence selection into full-time employment, especially among lower-educated women. Women in these groups who are observed working full-time appear to have unobserved traits associated with higher wages, leading to a positively selected sample. The IV correction adjusts for this by providing a more representative picture of the full population, including women not observed in full-time work who likely earn lower wages due to their unobserved characteristics.

Men exhibit a distinctly different pattern. For low-educated men, selection effects are modest and become even smaller among those with middle education. By contrast, highly educated men experience substantial positive selection, suggesting that those with higher earning potential are disproportionately represented among full-time workers. This aligns with a setting where full-time participation among highly educated men reflects stronger labor market attachment and career-oriented incentives rather than financial necessity. For this group, employment decisions are likely influenced by unobserved traits such as ambition, productivity, or access to high-return jobs that amplify the benefits of continuous full-time work. In contrast, for less educated men, participation decisions may be driven more by short-term income needs or local labor demand conditions, making selection less sensitive to unobserved earning potential.

While the IV-based and MAR corrections yield broadly similar results for low- and medium-educated women and men, the two approaches diverge most clearly at the top of the wage distribution among the highly educated. This pattern suggests that unobserved factors such as ambition or career orientation play a stronger role in shaping high-end wages, particularly for highly educated men and women. In this group, the MAR correction misestimates selection effects because it fails to account for selection driven by such unobservables, whereas the IV-based approach corrects for them more effectively.

5.3 Gender wage gaps by education and wage quantile

Table 5 reports predicted daily wages for men and women and the corresponding wage gaps at selected quantiles (25th, 50th, and 75th), setting education to one of three levels (low, middle, high) while averaging over all other covariates. The first

Table 5: The distribution of the gender wage gap by education

	Uncorrected			IV selection correction		
	25th	50th	75th	25th	50th	75th
Unconditional						
Men	83.33 [83.06,83.60]	107.84 [107.55,108.13]	140.20 [139.74,140.65]	80.39 [80.08,80.71]	103.92 [103.47,104.37]	135.29 [134.54,136.05]
Women	67.65 [67.24,68.06]	91.18 [90.77,91.58]	113.73 [113.42,114.03]	63.73 [62.96,64.50]	88.24 [87.50,88.97]	111.76 [110.99,112.54]
% Difference	18.82	15.45	18.88	20.73	15.09	17.39
Low Education						
Men	56.26 [55.83,56.68]	70.44 [69.90,70.99]	93.15 [92.37,93.93]	54.88 [54.27,55.49]	67.57 [66.91,68.23]	86.02 [85.12,86.94]
Women	50.71 [50.04,51.40]	67.90 [67.08,68.74]	87.94 [87.16,88.73]	46.80 [45.44,48.21]	60.32 [59.04,61.62]	78.10 [76.84,79.37]
% Difference	9.86	3.60	5.59	14.72	10.73	9.22
Middle Education						
Men	86.44 [86.19,86.69]	106.85 [106.55,107.16]	133.13 [132.72,133.55]	84.25 [83.87,84.63]	103.17 [102.70,103.65]	127.07 [126.44,127.70]
Women	68.23 [67.83,68.63]	89.61 [89.18,90.06]	110.59 [110.18,111.00]	63.85 [62.46,65.27]	82.81 [81.83,83.79]	103.01 [102.23,103.80]
% Difference	21.07	16.13	16.93	24.21	19.74	18.93
High Education						
Men	116.47 [115.73,117.21]	143.16 [142.43,143.89]	169.18 [168.38,169.99]	107.70 [106.70,108.70]	132.48 [131.50,133.46]	158.57 [157.53,159.63]
Women	87.38 [86.47,88.31]	110.24 [109.39,111.09]	128.90 [128.19,129.61]	82.42 [80.78,84.08]	104.61 [103.14,106.10]	125.00 [123.32,126.70]
% Difference	24.97	23.00	23.81	23.47	21.04	21.17

Note: SIAB-R data, 2017. The table presents quantile regression estimates for the 0.25, 0.50, and 0.75 quantiles. The IV correction uses initial log wage as an instrument and controls for education, age, workplace region, total experience, and earliest job difficulty in the first two stages. In the first panel (unconditional) quantile regressions include no covariates. In the remaining panels (conditional), the IV correction carries the same variables to the third stage. The uncorrected estimates use the same controls, excluding the instrument and earliest job difficulty. Conditional estimates fix controls at their median within each education group. 95% confidence intervals are shown in square brackets.

panel of the table additionally presents the unconditional results, providing a baseline comparison before adjusting for observable characteristics. We compare estimates that do not correct for selection and selection corrected estimates using the proposed IV method. A central and straightforward observation is that men consistently earn higher wages than women across all quantiles and education groups, indicating a persistent gender wage gap that holds regardless of educational attainment, quantiles, and estimation method. However, there exist interesting differences between the groups.

The unconditional estimates in the first panel, corresponding to Figure 2, capture the differences in wage distributions without adjusting for covariates. At the median, men earn about 108 euros per day, compared to 91 euros for women, implying a raw wage gap of 15.5%. After IV-based selection correction, median wages fall to 104 euros for men and 88 euros for women, leaving the gap virtually unchanged.

At the lower quantile, however, the correction increases the wage gap by roughly 2 percentage points to 20.7%, while at the upper quantile it reduces the gap by about 1.5 percentage points to 17.4%.

In the second panel, where education is set to no vocational training (low education), we find strong positive selection for women and significant but less pronounced positive selection for men. The uncorrected daily wages for men are 56 euros, 70 euros, and 93 euros at the 25th, 50th, and 75th percentiles, respectively. For women, the corresponding figures are 51, 68, and 88 euros. After applying the IV-based selection correction, median wages fall to 68 euros for men and 60 euros for women, increasing the gender wage gap from 3.6% to 10.7%. At the 25th percentile, strong positive selection for both genders widens the gap by 4.8 percentage points, whereas at the 75th percentile the effect is smaller, raising the gap by only 3.6 percentage points.

Among workers with vocational training (middle education), the gender wage gap widens across all quantiles, although the changes are less dramatic than for the low-education group. At the median, men's wages fall from 107 to 103 euros and women's from 90 to 83 euros, widening the gender wage gap by 3.6 percentage points to 19.7%. At the lower end of the distribution, the gap increases by 3.1 percentage points to 24.2%, while at the 75th percentile, the correction adds only 2 percentage points.

For university graduates, selection patterns differ sharply between men and women. Women with higher education are only marginally affected by selection, whereas highly educated men experience the strongest positive selection in the sample. At the upper quantiles, men's daily wages are about 11 euros lower after correction, compared to a smaller decline of 4–5 euros for women. These opposing effects even reduce the gender wage gap at the top of the distribution, by roughly 1–2 percentage points.

Relative to the results of Blau et al. [e.g., 2024], our IV-based estimates attribute more of the observed female advantage among full-time workers to selection on unobservables, especially at the bottom of the distribution. In this sense, our results caution against interpreting MAR-corrected improvements as purely structural convergence: once non-ignorable selection is accommodated, distributional gaps at the lower tail are larger, while top-end gaps can be smaller due to stronger male selection.

Taken together, our estimates both reconcile and extend existing literature. In addition to documenting positive selection throughout the wage distribution, we uncover two distinctive patterns: (i) pronounced selection at the lower end for women, especially those with less education, which widens the gender gap in this segment; and (ii) strong positive selection among highly educated men at the top of the distribution, which narrows the gender wage gap at upper quantiles. These findings rationalize why selection-corrected gender wage gaps can widen, as shown by Maasoumi and Wang [2019], while in other cases they may narrow at the upper end of the distribution, consistent with Arellano and Bonhomme [2017].

6 Conclusion

In this paper, we propose a novel estimation strategy to recover the selection-corrected wage distribution and to quantify the full distribution of the gender wage gap. Identification is obtained from exogenous variation that shifts latent wages but does not, conditional on these latent and other observed variables, provide additional information on the selection mechanism. The approach does not impose parametric restrictions on the selection probability and, within the Roy model, can be interpreted as a rank invariance condition on conditional reservation wages.

We formally derive an estimation strategy for selection-corrected quantile regression when the outcome is selectively observed and selection depends on unobservables. We use inverse probability weighting in combination with an instrumental variable approach, thereby relaxing the MAR assumption that selection depends only on observables. We use shape constraints via cone projection and B-spline approximations in the estimation of inverse probabilities, which stabilize the weights and improve inference.

Using administrative social security data from Germany, we apply the proposed estimation strategy to derive selection-corrected wage distributions for men and women in full-time employment. The results of our analysis highlight the critical role of selection in shaping the gender wage gap. We document substantial selection at the lower tail for women, amplifying gaps specifically among the less educated, and strong positive selection among highly educated men, compressing gaps at the top. Despite these differences in the selection pattern, men consistently earn higher full-time wages than women across all quantiles and education groups. Thus, the gender wage gap persists regardless of educational attainment, quantiles, and selection patterns.

References

H. Ahn and J. L. Powell. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1):3–29, 1993.

M. Antoni, A. Schmucker, S. Seth, P. Vom Berge, et al. Sample of integrated labour market biographies (SIAB) 1975-2017. Technical report, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg, 2019.

M. Arellano and S. Bonhomme. Quantile selection models with an application to understanding changes in wage inequality. *Econometrica*, 85(1):1–28, 2017.

F. D. Blau and L. M. Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of economic literature*, 55(3):789–865, 2017.

F. D. Blau, L. M. Kahn, N. Boboshko, and M. Comey. The impact of selection into the labor force on the gender wage gap. *Journal of Labor Economics*, 42(4): 1093–1133, 2024.

R. Blundell, A. Gosling, H. Ichimura, and C. Meghir. Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica*, 75(2):323–363, 2007.

R. Blundell, J. P. Ziliak, and H. Lopez. Labour market inequality and the changing life cycle profile of male and female wages. IFS Working Papers W23/16, Institute for Fiscal Studies, Apr. 2023.

C. Breunig and X. Chen. Adaptive, rate-optimal hypothesis testing in nonparametric IV models. *Econometrica*, 92(6):2027–2067, 2024.

C. Breunig and P. Haan. Nonparametric regression with selectively missing covariates. *Journal of Econometrics*, 2020.

C. Breunig, E. Mammen, and A. Simoni. Nonparametric estimation in case of endogenous selection. *Journal of Econometrics*, 202(2):268–285, 2018.

S. Chen and S. Khan. Semiparametric estimation of a heteroskedastic sample selection model. *Econometric Theory*, 19(6):1040–1064, 2003.

X. Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 2007.

X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9 (1):39–84, 2018.

V. Chernozhukov and C. Hansen. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

M. Das, W. K. Newey, and F. Vella. Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70(1):33–58, 01 2003.

W. Dauth and J. Eppelsheimer. Preparing the sample of integrated labour market biographies (SIAB) for scientific analysis: A guide. *Journal for Labour Market Research*, 54(1):1–14, 2020.

P. J. Devereux. The importance of obtaining a high-paying job. MPRA Paper 49326, University Library of Munich, Germany, Jan 2002.

X. D'Haultfoeuille. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154(1):1–15, 2010.

S. G. Donald. Two-step estimation of heteroskedastic sample selection models. *Journal of Econometrics*, 65(2):347–380, 1995.

C. Goldin. A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119, April 2014.

R. Gronau. Wage comparisons – A selectivity bias. *Journal of Political Economy*, 82 (6):1119–1143, 1974. doi: 10.1086/260246.

P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468): 1015–1026, 2004.

J. Heckman. Shadow prices, market wages, and labor supply. *Econometrica: Journal of the Econometric Society*, pages 679–694, 1974.

J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

J. J. Heckman. Identifying the hand of past: Distinguishing state dependence from heterogeneity. *The American Economic Review*, 81(2):75–79, 1991.

N. L. Hjørt and D. Pollard. Asymptotics for minimizers of convex processes. Statistical Research Report, 1993.

B. E. Honoré and L. Hu. Selection without exclusion. *Econometrica*, 88(3):1007–1029, 2020.

B. Ilieva and K. Wrohlich. Gender gaps in employment, working hours and wages in Germany: Trends and developments over the last 35 years. *CESifo Forum*, 23(02): 17–19, March 2022.

K. Knight. Limiting distributions for L1 regression estimators under general conditions. *The Annals of Statistics*, 26(2):755–770, 1998.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

E. Maasoumi and L. Wang. The gender gap between earnings distributions. *Journal of Political Economy*, 127(5):2438–2504, 2019.

C. Meghir and L. Pistaferri. Earnings, consumption and life cycle choices. In *Handbook of Labor Economics*, volume 4 of *Handbook of Labor Economics*, chapter 9, pages 773–854. Elsevier, 2011.

C. B. Mulligan and Y. Rubinstein. Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics*, 123(3):1061–1110, 2008.

C. Olivetti and B. Petrongolo. Unequal pay or unequal employment? A cross-county analysis of gender gaps. *Journal of Labor Economics*, 26(4):621–654, 2008.

C. Olivetti, J. Pan, and B. Petrongolo. The evolution of gender in the labor market. NBER Working Papers 33153, National Bureau of Economic Research, Inc, Nov. 2024.

P. Oreopoulos, T. von Wachter, and A. Heisz. The short- and long-term career effects of graduating in a recession. *American Economic Journal: Applied Economics*, 4 (1):1–29, January 2012.

D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

J. Racine and Q. Li. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130, 2004.

E. A. Ramalho and R. J. Smith. Discrete choice non-response. *The Review of Economic Studies*, 80(1):343–364, 2013.

A. D. Roy. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146, 1951.

H. Schwandt and T. von Wachter. Unlucky cohorts: Estimating the long-term effects of entering the labor market in a recession in large cross-sectional data sets. *Journal of Labor Economics*, 37(S1):161–198, None 2019.

T. A. Severini and G. Tripathi. Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, 170(2):491–498, 2012.

G. Tang, R. J. Little, and T. E. Raghunathan. Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90(4):747–764, 2003.

S. Wang, J. Shao, and J. K. Kim. An instrumental variable approach for identification and estimation with nonignorable response. *Statistica Sinica*, 24(3):1097–1116, 2014.

A. Yu, Y. Zhong, X. Feng, and Y. Wei. Quantile regression for nonignorable missing data with its application of analyzing electronic medical records. *Biometrics*, 79 (3):2036–2049, 2023.

T. Zhang and L. Wang. Smoothed empirical likelihood inference and variable selection for quantile regression with nonignorable missing response. *Computational Statistics and Data Analysis*, 144:106888, 2020.

J. Zhao and J. Shao. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110(512):1577–1590, 2015.

A1 Proofs of the Identification Results in Section 2

Proof of Proposition 2.1. For any $\tau \in (0,1)$, it is well known (see Koenker and Bassett [1978]) that the quantile regression coefficient θ_τ minimizes the check function as given in Equation (2.4), i.e.,

$$\theta_\tau = \underset{\theta \in \mathbb{R}^{d_Z}}{\arg\min}\, \mathbb{E}\big[\big(Y^* - Z^\top \theta\big)\big(\tau - \mathbb{1}\{Y^* < Z^\top \theta\}\big)\big].$$

Multiplying and dividing by $\mathbb{P}(D = 1 \mid Y^*, X)$ gives

$$\theta_\tau = \underset{\theta \in \mathbb{R}^{d_Z}}{\arg\min}\, \mathbb{E}\left[\big(Y^* - Z^\top \theta\big)\big(\tau - \mathbb{1}\{Y^* < Z^\top \theta\}\big)\frac{\mathbb{P}(D = 1 \mid Y^*, X)}{\mathbb{P}(D = 1 \mid Y^*, X)}\right].$$

By the exclusion restriction on W, we have $\mathbb{P}(D = 1 \mid Y^*, X, W) = p(Y^*, X)$, so that

$$\theta_\tau = \underset{\theta \in \mathbb{R}^{d_Z}}{\arg\min}\, \mathbb{E}\left[\big(Y^* - Z^\top \theta\big)\big(\tau - \mathbb{1}\{Y^* < Z^\top \theta\}\big)\frac{\mathbb{P}(D = 1 \mid Y^*, X, W)}{\mathbb{P}(D = 1 \mid Y^*, X)}\right].$$

Applying the law of iterated expectations and noting that $Y = Y^*$ whenever $D = 1$, with $p(Y^*, X) = p(Y, X)$ on this support, we obtain

$$\begin{aligned}\theta_\tau &= \underset{\theta \in \mathbb{R}^{d_Z}}{\arg\min}\, \mathbb{E}\left[\big(Y - Z^\top \theta\big)\big(\tau - \mathbb{1}\{Y < Z^\top \theta\}\big)\frac{D}{p(Y, X)}\right]\\ &= \underset{\theta \in \mathbb{R}^{d_Z}}{\arg\min}\, \mathbb{E}\left[\frac{D}{p(Y, X)}\rho_\tau\big(Y - Z^\top \theta\big)\right],\end{aligned}$$

which proves the claim. □

Proof of Theorem 2.2. We first introduce the linear operator T by $T : f \mapsto \mathbb{E}[Df(Y^*, X) \mid W, X]$ for any square integrable function f. Then, the conditional mean equation in (2.3), that is, $\mathbb{E}[Dg(Y^*, X) \mid W, X] = 1$, can be written as the operator equation $Tg = 1$. Furthermore, the quantile regression coefficient θ_τ minimizes a population criterion as in Equation (2.5) which can be rewritten as

$$\mathbb{E}\left[g(Y^*, X) m_\tau(Y^*, X)\right], \quad \text{with} \quad m_\tau(Y^*, X) := \mathbb{E}\left[D\rho_\tau\big(Y^* - Z^\top \theta_\tau\big) \,\big|\, Y^*, X\right].$$

This population criterion thus immediately implies uniqueness of the quantile regression coefficient θ_τ.

We can now apply Lemma 3.1 in Severini and Tripathi [2012]: Identification of the population criterion $\mathbb{E}\left[g(Y^*, X) m_\tau(Y^*, X)\right]$ requires only that m_τ belongs to the orthogonal complement (in the L^2 sense) of the operator T. Equivalently, there exists a function μ_τ such that $T^* \mu_\tau = m_\tau$, where the adjoint operator is defined as $T^* : \phi \mapsto \mathbb{E}[D\phi(W, X) \mid Y^*, X]$. This condition can be written as

$$\mathbb{E}\left[D\mu_\tau(W, X) \mid Y^*, X\right] = \mathbb{E}\left[D\rho_\tau\big(Y^* - Z^\top \theta_\tau\big) \mid Y^*, X\right].$$

Using the selection equation $D = \mathbb{1}\{V \leq p(Y^*, X)\}$ and the conditional independence $V \perp\!\!\!\perp W \mid (Y^*, X)$, we can factor out the probability of selection $p(Y^*, X) = \mathbb{E}[D \mid Y^*, X]$, yielding

$$p(Y^*, X)\mathbb{E}\left[\mu_\tau(W, X) \mid Y^*, X\right] = p(Y^*, X)\mathbb{E}\left[\rho_\tau\big(Y^* - Z^\top \theta_\tau\big) \mid Y^*, X\right].$$

By the overlap condition $p(Y^*, X) > 0$, we divide both sides by $p(Y^*, X)$ to obtain the equivalent expression

$$\mathbb{E}\left[\mu_\tau(W, X) \mid Y^*, X\right] = \mathbb{E}\left[\rho_\tau\left(Y^* - Z^\top \theta_\tau\right) \mid Y^*, X\right],$$

which completes the proof. □

Proof of Corollary 2.3. Since R conditional on (Y^*, X) is continuously distributed, the conditional cdf. $F_{R|Y^*,X}(\cdot)$ is strictly increasing. Consequently, we may write the Roy selection rule as

$$\begin{aligned} D &= \mathbb{1}\left\{Y^* \geq R\right\} \\ &= \mathbb{1}\left\{F_{R|Y^*,X}(Y^*) \geq F_{R|Y^*,X}(R)\right\} \\ &= \mathbb{1}\left\{p(Y^*, X) \geq F_{R|Y^*,X}(R)\right\}. \end{aligned}$$

This corresponds exactly to the general selection model (2.2) with $V = F_{R|Y^*,X}(R)$. Hence, we can apply the proof of Theorem 2.2, replacing V by $F_{R|Y^*,X}(R)$ to obtain identification of the quantile regression coefficients θ_τ. □

A2 Technical Assumptions and Proof of Theorem 3.1

The following conditions are required to establish Theorem 3.1. We observe a random sample $\{S_i\}_{i=1}^n$, where $S_i = (D_i, Y_i, W_i^\top, X_i^\top)$. Define the identified set of inverse selection probability functions as $\mathcal{I} := \{g \in L^2(Y, X) : \mathbb{E}[Dg(Y, X) \mid W, X] = 1\}$. Let $H = \mathbb{E}[D\phi^J(Y, X) b^K(W, X)^\top]$ and define the quantile score function $\psi_\tau(u) = \tau - \mathbb{1}\{u < 0\}$. Recall that $M_{1\tau g} = \mathbb{E}\left[\Omega_g f_{Y|\Omega_g,Z}\left(Z^\top\theta_\tau\right) ZZ^\top\right]$.

Assumption A2. *(i) H is of full rank for all $J, K \in \mathbb{N}$; (ii) The conditional distribution function $F_{Y|\Omega_g,Z}$ is absolutely continuous, with continuous densities $f_{Y|\Omega_g,Z}(\xi)$ uniformly bounded away from 0 and ∞ at the points $Z_i^\top\theta_\tau, i = 1, \ldots, n$, uniformly for $g \in \mathcal{I}$; (iii) $\mathbb{E}[ZZ^\top]$ is positive definite.*

Assumption A2(i) ensures that the matrix H is invertible and that the projection of the inverse selection functions onto the sets of basis functions considered is unique. Assumption A2(ii) requires the conditional distribution of the outcome to be smooth and well-behaved at the linear quantile model points. Together, Assumptions A2(ii) and A2(iii) imply that the matrix $M_{1\tau g}$ is invertible.

For each $g \in \mathcal{I}$ recall the definition $\Omega_g = Dg(Y, X)$ of the selection weights. Note that by Assumption A2(i) is uniquely determined when the function g belongs to the sieve space under consideration. The corresponding estimator is denoted by $\widehat{\Omega} := D\widehat{g}(Y, X)$. From Equation (2.1), $U_\tau = Y^* - Z^\top\theta_\tau$, which reduces to $U_\tau = Y - Z^\top\theta_\tau$ whenever multiplied by D, since $Y = Y^*$ holds for $D = 1$. We write $\|\cdot\|_\infty$ for the supremum norm and, for any $\delta \in \mathbb{R}^{d_z}$, define

$$R_g(\delta) := \left(\beta^\top\phi^J(Y, X) - g(Y, X)\right) D\, Z^\top\delta\, \psi_\tau(U_\tau),$$

which captures the weighted approximation remainder from replacing $g(Y, X)$ by its series approximation $\beta^\top\phi^J(Y, X)$.

Assumption A3. *(i)* $\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}} \|\phi^J(y,x)\| = O(\sqrt{J})$ *such that* $J\log(J)/n = o(1)$, *and* $K = cJ$ *for some constant* $c \geq 1$*; (ii)* $\sqrt{n}\,\mathbb{E}[R_g(\delta)] = o\big(\sqrt{\mathbb{V}ar(R_g(\delta))}\big)$ *uniformly for* $g \in \mathcal{I}$*; (iii)* $\mathbb{E}\left[\left(1 - D\phi^J(Y,X)^\top\beta\right)^2 \mid W, X\right)\right]$ *is bounded away from zero and infinity, and* $\mathbb{E}\left[\left(1 - D\phi^J(Y,X)^\top\beta\right)^4\right] < \infty$*; (iv)* $\max_{1\leq i\leq n} \|\Omega_{gi} Z_i\| = o(\sqrt{n})$ *almost surely, uniformly for* $g \in \mathcal{I}$*; (v)* $\sup_{g\in\mathcal{I}} \|\widehat{g} - g\|_\infty = o_p(1)$.

Assumption A3 captures standard regularity conditions for establishing inference for semi-/nonparametric series estimators (see, e.g., Chen [2007], Koenker [2005], Knight [1998], Hjørt and Pollard [1993]). Assumption A3(i) imposes an upper bound on the growth of the series dimension J. This ensures that the series approximation remains well-behaved as the sample size grows. Assumption A3(ii) controls the relative size of approximation errors when $g(Y_i, X_i)$ is approximated by $\beta^\top\phi^J(Y_i, X_i)$, ensuring they vanish as the sample grows. Assumption A3(iii) guarantees the existence of key moments and that certain conditional variances are positive, which underpins the asymptotic normality of $\widehat{\beta}$. Assumption A3(iv) ensures that no single observation dominates the weighted sums, effectively satisfying a Lindeberg condition. Finally, Assumption A3(v) controls the estimation error of the inverse probabilities and the resulting weights, ensuring that the estimated weights converge uniformly to the true weights (see Chen and Christensen [2018]).

Proof of Theorem 3.1. Consider an objective function

$$S_n(\delta) = \sum_{i=1}^n \widehat{\Omega}_i \Big[\rho_\tau\left(U_{\tau i} - Z_i^\top\delta/\sqrt{n}\right) - \rho_\tau\left(U_{\tau i}\right)\Big]$$

where $U_{\tau i} = Y_i - Z_i^\top\theta_\tau$. The function $S_n(\delta)$ is convex and is minimized at

$$\widehat{\delta} = \sqrt{n}\left(\widehat{\theta}_\tau - \theta_\tau\right)$$

as we know from Equation (3.2) that

$$\min_{\theta\in\mathbb{R}^{d_z}} \sum_{i=1}^n \widehat{\Omega}_i\,\rho_\tau\big(Y_i - Z_i^\top\theta\big) = \sum_{i=1}^n \widehat{\Omega}_i\,\rho_\tau\big(Y_i - Z_i^\top\widehat{\theta}_\tau\big),$$

which is equivalent to $\sum_{i=1}^n \widehat{\Omega}_i\,\rho_\tau\big(U_{\tau i} - Z_i^\top\widehat{\delta}/\sqrt{n}\big)$.

Following Knight [1998], it can be shown that the asymptotic distribution of $\widehat{\delta}$ is determined by the limiting behavior of the function $S_n(\delta)$. Using Knight's identity,

$$\rho_\tau(u-v) - \rho_\tau(u) = -v\left(\tau - \mathbb{1}\{u<0\}\right) + \int_0^v \left(\mathbb{1}\{u\leq s\} - \mathbb{1}\{u\leq 0\}\right)ds$$

we may write

$$S_n(\delta) = S_{1n}(\delta) + S_{2n}(\delta),$$

where

$$S_{1n}(\delta) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\Omega}_i Z_i^\top\delta\left(\tau - \mathbb{1}\{U_{\tau i}<0\}\right)$$

$$S_{2n}(\delta) = \sum_{i=1}^n \widehat{\Omega}_i \int_0^{Z_i^\top\delta/\sqrt{n}} \left(\mathbb{1}\{U_{\tau i}\leq s\} - \mathbb{1}\{U_{\tau i}\leq 0\}\right)ds =: \sum_{i=1}^n \widehat{\Omega}_i S_{2ni}(\delta)$$

and we control each term separately in the following.

The first term We can write $S_{1n}(\delta)$ as

$$\begin{aligned}
S_{1n}(\delta) &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Omega_{gi}Z_i^\top\delta\left(\tau-\mathbb{1}\{U_{\tau i}<0\}\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\widehat{\Omega}_i-\Omega_{gi}\right)Z_i^\top\delta\left(\tau-\mathbb{1}\{U_{\tau i}<0\}\right)\\
&= -\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Omega_{gi}Z_i^\top\delta\psi_\tau(U_{\tau i})+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\widehat{\Omega}_i-\Omega_{gi}\right)Z_i^\top\delta\psi_\tau(U_{\tau i})\right)\\
&=: -(A_1+A_2),
\end{aligned}$$

where we used the definition $\psi_\tau(u)=\tau-\mathbb{1}\{u<0\}$ in the second line.

Consider A_1. From the definition of θ_τ in Equation (2.4), Proposition 2.1 ensures that θ_τ minimizes the expected weighted check-loss function $\mathbb{E}\left[\Omega_g\rho_\tau\left(Y-Z^\top\theta\right)\right]$ for all $g\in\mathcal{I}$. This implies $\mathbb{E}\left[\Omega_g Z\psi_\tau(U_\tau)\right]=0$, and, in particular, for some vector δ and any $g\in\mathcal{I}$, that

$$\mathbb{E}\left[\Omega_g Z^\top\delta\psi_\tau(U_\tau)\right]=0.$$

Now consider A_2. We write A_2 as

$$A_2=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\widehat{\Omega}_i-\Omega_{gi}\right)Z_i^\top\delta\psi_\tau(U_{\tau i})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\widehat{\Omega}_i-\widetilde{\Omega}_i+\widetilde{\Omega}_i-\Omega_{gi}\right)Z_i^\top\delta\psi_\tau(U_{\tau i}),$$

where $\widetilde{\Omega}_i:=\beta^\top\phi^J(Y_i,X_i)D_i$. By writing $\widehat{\Omega}_i=\widehat{\beta}^\top\phi^J(Y_i,X_i)D_i$, we can further write

$$\begin{aligned}
A_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(\widehat{\beta}-\beta)^\top\phi^J(Y_i,X_i)+\left(\beta^\top\phi^J(Y_i,X_i)-g(Y_i,X_i)\right)D_iZ_i^\top\delta\psi_\tau(U_{\tau i})\right]\\
&= \sqrt{n}(\widehat{\beta}-\beta)^\top\left(\frac{1}{n}\sum_{i=1}^{n}D_i\phi^J(Y_i,X_i)Z_i^\top\delta\psi_\tau(U_{\tau i})\right)\\
&\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\beta^\top\phi^J(Y_i,X_i)-g(Y_i,X_i)\right)D_iZ_i^\top\delta\psi_\tau(U_{\tau i}).
\end{aligned}$$

By Assumption A3(ii), the last summand vanishes asymptotically. For the first summand we have

$$\frac{1}{n}\sum_{i=1}^{n}\phi^J(Y_i,X_i)D_iZ_i^\top\psi_\tau(U_{\tau i})=\mathbb{E}\left[\phi^J(Y,X)DZ^\top\psi_\tau(U_\tau)\right]+o_p(1)=T+o_p(1),$$

and thus A_2 becomes

$$A_2=\delta^\top T^\top(HG^{-1}H^\top)^{-1}HG^{-1}b^K(W_i,X_i)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_{Ji}+o_p(1).$$

Hence, $S_{1n}(\delta)$ becomes

$$S_{1n}(\delta)=-\frac{1}{\sqrt{n}}\delta^\top\sum_{i=1}^{n}M_{0\tau g}(S_i)+o_p(1),$$

where we define $M_{0\tau g}(S_i):=Z_i\Omega_{gi}\psi_\tau(U_{\tau i})+T^\top(HG^{-1}H^\top)^{-1}HG^{-1}b^K(W_i,X_i)U_{Ji}$.

In particular, for all $g\in\mathcal{I}$ we obtain that

$$S_{1n}(\delta)\xrightarrow{d}-\delta^\top\mathcal{Z}_{\tau g},$$

where $\mathcal{Z}_{\tau g}\sim\mathcal{N}\left(0,\mathbb{E}\left[M_{0\tau g}(S)M_{0\tau g}^\top(S)\right]\right)$.

The second term We write the second term for all $g \in \mathcal{I}$ as

$$\begin{aligned}
S_{2n}(\delta) &= \sum_{i=1}^{n} \Omega_{gi} S_{2ni}(\delta) + \sum_{i=1}^{n} \left(\widehat{\Omega}_i - \Omega_{gi}\right) S_{2ni}(\delta) \\
&= \sum_{i=1}^{n} \mathbb{E}[\Omega_{gi} S_{2ni}(\delta) \mid \Omega_{gi}, Z_i] + \sum_{i=1}^{n} (\Omega_{gi} S_{2ni}(\delta) - \mathbb{E}[\Omega_{gi} S_{2ni}(\delta) \mid \Omega_{gi}, Z_i]) \\
&+ \sum_{i=1}^{n} \left(\widehat{\Omega}_i - \Omega_{gi}\right) S_{2ni}(\delta) \\
&=: T_1 + T_2 + T_3.
\end{aligned}$$

Consider T_1. Under Assumption A2(ii), we have for $g \in \mathcal{I}$

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}[\Omega_{gi} S_{2ni}(\delta) \mid \Omega_{gi}, Z_i] \\
&\quad = \sum_{i=1}^{n} \Omega_{gi} \mathbb{E}\left[\int_0^{Z_i^\top \delta/\sqrt{n}} \left(\mathbb{1}\{Y_i \leq Z_i^\top \theta_\tau + s\} - \mathbb{1}\{Y_i \leq Z_i^\top \theta_\tau\}\right) ds \,\middle|\, \Omega_{gi}, Z_i\right] \\
&\quad = \sum_{i=1}^{n} \Omega_{gi} \int_0^{Z_i^\top \delta/\sqrt{n}} \left(F_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau + s\right) - F_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau\right)\right) ds \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Omega_{gi} \int_0^{Z_i^\top \delta} \left(F_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau + t/\sqrt{n}\right) - F_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau\right)\right) dt \\
&\quad = n^{-1} \sum_{i=1}^{n} \Omega_{gi} \int_0^{Z_i^\top \delta} \sqrt{n}\left(F_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau + t/\sqrt{n}\right) - F_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau\right)\right) dt,
\end{aligned}$$

where in the second line we plugged in $U_{\tau i} = Y_i - Z_i^\top \theta_\tau$, and in the third line we used the fact that $\mathbb{E}[\mathbb{1}\{Y_i \leq Z_i^\top \theta_\tau + s\} \mid \Omega_{gi}, Z_i] = F_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau + s\right)$. The fourth line follows from substitution. Next, by applying a Taylor expansion to the term $F_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau + t/\sqrt{n}\right)$, we get

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}[\Omega_{gi} S_{2ni}(\delta) \mid \Omega_{gi}, Z_i] &= n^{-1} \sum_{i=1}^{n} \Omega_{gi} \int_0^{Z_i^\top \delta} f_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau\right) t dt + o_p(1) \\
&= (2n)^{-1} \sum_{i=1}^{n} \Omega_{gi} f_{Y|\Omega_g,Z}\left(Z_i^\top \theta_\tau\right) \delta^\top Z_i Z_i^\top \delta + o_p(1) \\
&= o_p\left(\frac{1}{2} \delta^\top M_{1\tau g} \delta\right)
\end{aligned}$$

for all $g \in \mathcal{I}$, where we define $M_{1\tau g} := \mathbb{E}\left[\Omega_g f_{Y|\Omega_g,Z}\left(Z^\top \theta_\tau\right) Z Z^\top\right]$.

Now consider T_2. We obtain

$$\mathbb{E}|T_2|^2 = \sum_{i=1}^{n} \mathbb{E}[(\Omega_{gi} S_{2ni}(\delta) - \mathbb{E}[\Omega_{gi} S_{2ni}(\delta) \mid \Omega_{gi}, Z_i])^2] \leq \sum_{i=1}^{n} \mathbb{E}\left[(\Omega_{gi} S_{2ni}(\delta))^2\right].$$

Since $\Omega_{gi} S_{2ni}(\delta) \geq 0$ and $S_{2ni}(\delta) \leq Z_i \delta/\sqrt{n}$ for all $i = 1, \ldots, n$, we can further bound

T_2 by

$$\begin{aligned}
\mathbb{E}|T_2|^2 &\leq \max_i\{\Omega_{gi} S_{2ni}(\delta) : 1 \leq i \leq n\} \sum_{i=1}^{n} \mathbb{E}\left[\Omega_{gi} S_{2ni}(\delta)\right] \\
&\leq \frac{1}{\sqrt{n}} \max_i \left|\Omega_{gi} Z_i^\top \delta\right| \sum_{i=1}^{n} \mathbb{E}\left[\Omega_{gi} S_{2ni}(\delta)\right].
\end{aligned}$$

For the sum we have

$$\sum_{i=1}^{n} \mathbb{E}\left[\Omega_{gi} S_{2ni}(\delta)\right] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[\Omega_{gi} S_{2ni}(\delta) \mid \Omega_{gi}, Z_i\right]\right],$$

which can be controlled in the same way as in T_1. Hence, by Assumption A3(iv), T_2 converges in probability to zero for all $g \in \mathcal{I}$.

Now consider the final summand T_3. Since $S_{2ni}(\delta) \geq 0$ for all $i = 1, \ldots, n$, the bias due to the estimated weights vanishes as:

$$\sum_{i=1}^{n} \mathbb{E}\left[\left(\widehat{\Omega}_i - \Omega_{gi}\right) S_{2ni}(\delta)\right] \leq \sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}} \left|\widehat{\Omega}(y,x) - \Omega_g(y,x)\right| \sum_{i=1}^{n} \mathbb{E}\left[S_{2ni}(\delta)\right],$$

where the difference $\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}} |\widehat{\Omega}(y,x) - \Omega_g(y,x)| \xrightarrow{p} 0$ uniformly for $g \in \mathcal{I}$ by Assumption A3(v). The second factor converges to a finite limit as in T_1:

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}[S_{2ni}(\delta) \mid Z_i] &= (2n)^{-1} \sum_{i=1}^{n} f_{Y|Z}\left(Z_i^\top \theta_\tau\right) \delta^\top Z_i Z_i^\top \delta + o_p(1) \\
&\xrightarrow{p} \frac{1}{2} \delta^\top \mathbb{E}\left[f_{Y|Z}\left(Z_i^\top \theta_\tau\right) Z_i Z_i^\top\right] \delta.
\end{aligned}$$

Thus, $T_3 \xrightarrow{p} 0$.

Combining all three terms we obtain $S_{2n}(\delta) = T_1 + T_2 + T_3 \xrightarrow{p} \frac{1}{2}\delta^\top M_{1\tau g}\delta$, and consequently,

$$\begin{aligned}
S_n(\delta) &= -\frac{1}{\sqrt{n}} \delta^\top \sum_{i=1}^{n} M_{0\tau g}(S_i) + \frac{1}{2}\delta^\top M_{1\tau g}\delta + o_p(1) \\
&\xrightarrow{d} S_0(\delta) = -\delta^\top \mathcal{Z}_{\tau g} + \frac{1}{2}\delta^\top M_{1\tau g}\delta.
\end{aligned}$$

The convexity of the limiting objective function, $S_0(\delta)$, assures the uniqueness of the minimizer and, consequently, that

$$\sqrt{n}\left(\widehat{\theta}_\tau - \theta_\tau\right) = \widehat{\delta} = \operatorname{argmin} S_n(\delta) \xrightarrow{d} \operatorname{argmin} S_0(\delta).$$

(See, Koenker [2005]; Pollard [1991]; Knight [1998]). Finally, we conclude the proof by noting that the limiting objective function $S_n(\delta)$ is minimized at

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \chi_{\tau g}(S_i) + o_p(1),$$

recalling the definition of the influence function

$$\chi_{\tau,g}(S_i) = M_{1\tau g}^{-1} M_{0\tau g}(S_i).$$

Therefore, the asymptotic distribution results hold for all $g \in \mathcal{I}$, in particular for $g = g^*$, where g^* minimizes $\mathbb{E}\big[\chi_{\tau g}(S)\chi_{\tau g}^\top(S)\big]$, that is,

$$\sqrt{n}\Big(\widehat{\theta}_\tau - \theta_\tau\Big) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\big[\chi_{\tau g^*}(S)\chi_{\tau g^*}^\top(S)\big]\right).$$

This completes the derivation of the asymptotic distribution of the quantile selection estimator $\widehat{\theta}_\tau$ under the stated regularity conditions. □

A3 Estimation of Covariance Matrix

In finite samples, the asymptotic covariance matrix $\mathbb{E}\big[\chi_{\tau g^*}(S)\chi_{\tau g^*}^\top(S)\big]$ is estimated using the corresponding sample analogs. Note that the estimators are based on an empirical analog of sieve projections and are thus identified and do not vary with the inverse selection probability g. Let

$$\widehat{U}_{i\tau} := Y_i - Z_i^\top \widehat{\theta}_\tau,$$

and let $\widehat{f}_{Y|\Omega_g,Z}(\cdot)$ denote a kernel estimator of the conditional density of Y given (Ω_g, Z). The first component is estimated by

$$\widehat{M_{1\tau}} = \mathbb{E}_n\left[\widehat{\Omega}_i\, \widehat{f}_{Y|\Omega_g,Z}\Big(Z_i^\top \widehat{\theta}_\tau\Big)\, Z_i Z_i^\top\right].$$

The second component of the influence function is

$$\widehat{M_{0\tau}}(S_i) = Z_i \widehat{\Omega}_i\, \psi_\tau(\widehat{U}_{i\tau}) + \widehat{T}^\top \big(\widehat{H}\widehat{G}^{-1}\widehat{H}^\top\big)^{-1} \widehat{H}\widehat{G}^{-1} b^K(W_i, X_i)\, \widehat{U}_{J,i}.$$

The first-stage quantities are estimated as

$$\begin{aligned} \widehat{H} &= \mathbb{E}_n\big[D_i\, \phi^J(Y_i, X_i)\, b^K(W_i, X_i)^\top\big], \\ \widehat{G} &= \mathbb{E}_n\big[b^K(W_i, X_i)\, b^K(W_i, X_i)^\top\big], \\ \widehat{U}_{J,i} &= 1 - D_i\, \phi^J(Y_i, X_i)^\top \widehat{\beta}. \end{aligned}$$

The plug-in covariance estimator of $\mathbb{E}\big[\chi_{\tau g^*}(S)\chi_{\tau g^*}^\top(S)\big]$ is then

$$\widehat{\Sigma}_\tau = \widehat{M_{1\tau}^{-1}} \mathbb{E}_n\left[\widehat{M_{0\tau}}(S_i)\, \widehat{M_{0\tau}}(S_i)^\top\right] \widehat{M_{1\tau}^{-1}}.$$

Consistency of the plug-in estimators follows from the standard large-sample arguments for quantile regression in Koenker [2005], together with the rate condition on J in Assumption A3(i).

B Outcome variable

The empirical analysis focuses on the full-time daily wages of German men and women who are active in the labor force, excluding individuals in education, the self-employed, and civil servants. The construction of the wage variable for individuals employed in 2017 follows a multi-stage procedure, largely based on Dauth and Eppelsheimer [2020], with additional modifications tailored to the analysis in this paper. Table B.1 reports summary statistics for the relevant daily wage variables at each stage of data preparation. All values are conditional on full-time employment. Only the most pertinent wage variables from each step are shown, without repeating intermediate versions.

- **Stage I** : Raw data – `tentgelt`, the original SIAB daily earnings variable (Tagesentgelt / täglicher Leistungssatz).
- **Stage II** : Inflation adjustment – `wage_defl`, the daily wage deflated to 2015 euros using the CPI.
- **Stage III** : Censoring and ceiling adjustment — `wage`, top-coded at four euros below the contribution assessment ceiling. Also, the censored wage `ln_wage_cens` equals the log wage if uncensored and 0 otherwise.
- **Stage IV** : Trimmed wage – `tln_wage`, obtained after excluding non-German individuals and trimming at the 5th and 95th percentiles.
- **Stage V** : Panel construction — transformation of the dataset into a yearly panel format.
- **Stage VI** : Final sample selection — `ln_wage_2017`, restricted to individuals aged 25–50 in 2017 with a valid instrument available.

Table B.1: Data preparation

	Men					Women				
	Mean	Median	Max	SD	Obs.	Mean	Median	Max	SD	Obs.
Stage I										
tentgelt_gr	75.30	67.00	208	40.19	16,394,987	58.33	51.00	208	36.39	8,613,552
Stage II										
tentgelt	75.30	67.00	208	40.19	16,394,987	58.33	51.00	208	36.39	8,613,552
wage_defl	97.91	93.02	203.92	42.36	16,394,987	74.77	70.50	208	38.83	8,613,552
Stage III										
wage	97.37	93.02	203.92	41.53	16,394,987	74.61	70.50	203.92	38.46	8,613,552
ln_wage_cens	4.42	4.48	5.32	0.48	14,497,003	4.18	4.25	5.32	0.56	8,143,881
Stage IV										
wage	100.73	96.37	203.92	42.54	13,488,223	75.91	71.93	203.92	38.79	7,376,056
ln_wage	4.53	4.57	5.32	0.49	13,324,958	4.22	4.29	5.32	0.57	7,180,494
ln_wage_cens	4.45	4.52	5.32	0.47	11,751,680	4.20	4.27	5.32	0.55	6,961,682
tln_wage	4.48	4.52	5.30	0.41	11,453,354	4.17	4.25	5.05	0.47	6,333,783
Stage V										
wage	108.78	103.48	200.75	38.48	9,129,764	81.48	77.93	200.75	37.19	4,830,944
ln_wage	4.63	4.64	5.30	0.39	9,088,967	4.30	4.36	5.30	0.50	4,769,570
ln_wage_cens	4.55	4.58	5.30	0.36	7,784,179	4.28	4.35	5.30	0.48	4,615,327
tln_wage	4.56	4.59	5.30	0.34	7,736,565	4.24	4.31	5.05	0.43	4,231,846
Stage VI										
ln_wage_2017	4.65	4.67	5.30	0.37	112,826	4.45	4.21	5.05	0.39	52,506

Notes: SIAB-R, 1975 - 2017. Summary statistics of the daily earnings variable at each stage of the data preparation. All values are conditional on being full-time employed. Only relevant wage variables from each step, without repeating everything again, are shown.

C Estimation of the Distribution Function

Using the estimator $\widehat{g}_n(y,x)$ from the first stage as discussed in Section 3.1 we can estimate

$$\widehat{F}_n(y) = \sum_{i=1,D_i=1}^{n} \mathbb{1}\{Y_i \leq y\}\widehat{\omega}_n(Y_i, X_i) \tag{C.1}$$

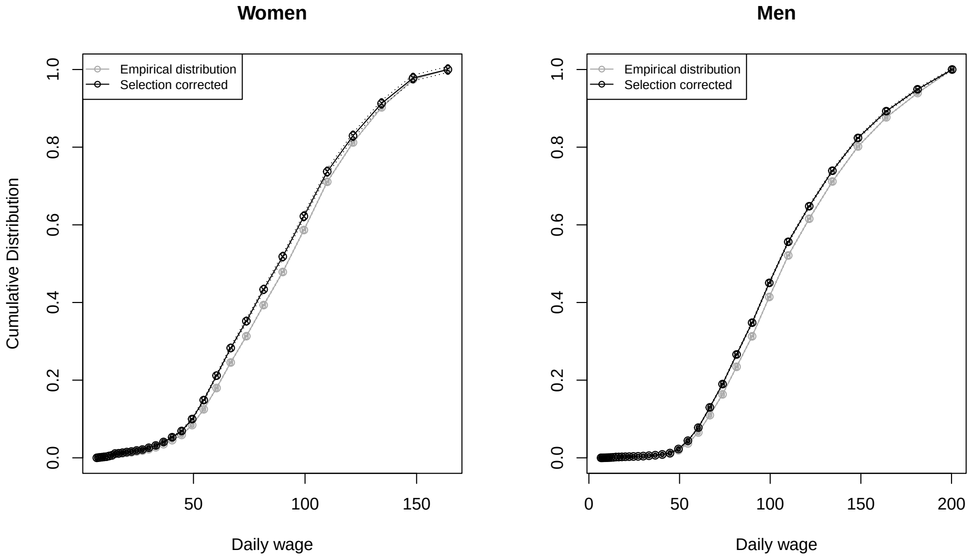
where the normalized inverse probability estimators are given by

$$\widehat{\omega}_n(y,x) = \frac{\widehat{g}_n(y,x)}{\sum_{i=1,D_i=1}^{n} \widehat{g}_n(Y_i, X_i)}, \tag{C.2}$$

which accounts for potential selection bias by weighing the contribution of each observation based on its probability of being observed.

In Figure C.1, we illustrate the empirical full-time wage distribution for women (left panel) and men (right panel), alongside the selection-corrected distributions shown in lighter colors, which are estimated using the inverse probability weighting method described in Equation C.1. The controls used in the first stage are detailed in Section 4.3. The comparison between the two distributions highlights important differences in the wage structure for men and women.

Figure C.1: Empirical and selection-corrected distributions by gender



Note: SIAB-R data, 2017 cross-section. The gray lines represent the empirical distributions of observed full-time wages, while the black lines depict the selection-corrected unconditional distributions, estimated using the IV-based selection correction method. All results are shown with 95% confidence bands.

D Simulation Design and Full Results

This section provides a detailed description of the Monte Carlo study introduced in Section 3.3. We compare its finite-sample performance to existing approaches in the literature. For each scenario, we conduct 1,000 Monte Carlo replications with a sample size of $n = 1,000$.

We consider the following model for the outcome:

$$Y_i^* = \beta_0 + \beta_1 W_i + \beta_2 X_i + \epsilon_i(\tau), \quad i = 1, 2, \ldots, n,$$

where $(W_i, X_i)^\top \sim \mathcal{N}\left(\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$. The true parameter vector is set to $(\beta_0, \beta_1, \beta_2) = (1, 1, 2)$. The error term $\epsilon(\tau)$ is defined as $\epsilon_i(\tau) = \epsilon_i - F_\epsilon^{-1}(\tau)$, where $F_\epsilon(\tau)$ is the τ-quantile of the error distribution. Following closely the simulation design of Zhang and Wang [2020], we consider five different error distributions: (**A**) ϵ_i is distributed as $\mathcal{N}(0, 1)$; (**B**) ϵ_i follows a mixture of normals $0.4\mathcal{N}(0, 1.5^2) + 0.6\mathcal{N}(0, 1)$; (**C**) ϵ_i follows a t-distribution with 3 degrees of freedom, scaled by 0.7; (**D**) ϵ_i follows a uniform distribution $\sim \mathcal{U}(-1.5, 1.5)$; (**E**) ϵ_i is distributed as $\mathcal{N}(0, (0.5 \cdot (1 + |X_i|))^2)$. Note that settings (**A**)-(**D**) are homoskedastic errors and (**E**) is heteroskedastic error.

We generate D_i from the Bernoulli distribution with probability $p(Y_i^*, X_i)$, considering three selection mechanisms:

M1 $p(Y_i^*, X_i) = 1/(1 + \exp(-(\alpha_1 + \gamma_1 X_i)))$ with $(\alpha_1, \gamma_1) = (-0.1, 0.8)$

M2 $p(Y_i^*, X_i) = 1/(1 + \exp(-(\alpha_2 + \gamma_2 X_i + \xi_2 Y_i^*)))$ with $(\alpha_2, \gamma_2, \xi_2) = (-2.4, 0.6, 0.6)$

M3 $p(Y_i^*, X_i) = 1/(1 + \exp(-(\alpha_3 + \gamma_3 \sin^2(X_i) + \xi_3 Y_i^*)))$ with $(\alpha_3, \gamma_3, \xi_3) = (-2.6, 1.2, 0.6)$.

The coefficients in the selection models are calibrated to achieve an average missing data rate of approximately 35%. In M1, the selection probability is independent of the latent outcome Y_i^*, reflecting a missing-at-random scenario. In contrast, M2 and M3 introduce dependence on the latent outcome, representing a missing-not-at-random setting. Notably, W_i serves as an instrumental variable, influencing the latent outcome but being excluded from the selection equation.

We consider the four estimators as described in Section 3.3 at the median quantile ($\tau = 0.5$): the uncorrected estimator, the MAR-assumed correction, the joint estimating equations (JEE) approach of Yu et al. [2023], and our proposed semiparametric IV estimator. For completeness, we briefly summarize them here. The uncorrected estimator uses only observed outcomes; the MAR-assumed estimator corrects for selection under the MAR assumption via inverse probability weighting; the JEE method employs an IV and specifies a parametric selection model, jointly estimating the quantile and selection equations; and the semiparametric IV estimator proposed in this paper avoids parametric restrictions on the selection mechanism.

We assess the finite-sample estimation performance in terms of the mean biases and the RMSE. Additionally, we report the coverage probabilities of the 95% confidence intervals along with their average lengths. The uncorrected and MAR-assumed correction methods are implemented using the `quantreg` package in R, while the JEE method follows the implementation described in Yu et al. [2023]. Tables D.1-D.4 present the simulation results.

Table D.1: (Setting A) Simulation results by method with $(\tau, n) = (0.5, 1000)$ and 35% missing in the outcome

Method	M1			M2			M3		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
	Mean biases								
Uncorrected	0.000	-0.002	0.003	0.342	-0.035	-0.106	0.351	-0.039	-0.101
MAR-assumed	0.002	-0.002	0.001	0.391	-0.040	-0.128	0.386	-0.045	-0.112
JEE	0.002	-0.001	0.000	0.051	-0.009	-0.012	0.055	-0.011	-0.008
Semiparametric IV	-0.047	0.010	0.022	0.048	0.007	-0.027	0.082	0.001	-0.045
	RMSE								
Uncorrected	0.122	0.057	0.057	0.365	0.064	0.122	0.376	0.068	0.118
MAR-assumed	0.126	0.059	0.061	0.439	0.085	0.175	0.444	0.092	0.169
JEE	0.134	0.06	0.063	0.273	0.083	0.134	0.318	0.102	0.137
Semiparametric IV	0.137	0.059	0.068	0.156	0.062	0.079	0.170	0.062	0.086
	CI lengths								
Uncorrected	0.464	0.223	0.233	0.523	0.219	0.241	0.519	0.220	0.237
MAR-assumed	0.485	0.230	0.250	0.612	0.226	0.326	0.519	0.198	0.183
JEE	0.527	0.240	0.263	0.934	0.314	0.418	0.972	0.341	0.444
Semiparametric IV	0.696	0.342	0.317	0.683	0.325	0.272	0.735	0.347	0.282
	Coverage probabilities								
Uncorrected	0.939	0.956	0.963	0.256	0.917	0.600	0.252	0.899	0.620
MAR-assumed	0.923	0.948	0.955	0.377	0.892	0.651	0.431	0.883	0.702
JEE	0.941	0.949	0.944	0.886	0.935	0.884	0.853	0.900	0.872
Semiparametric IV	0.979	0.995	0.968	0.948	0.959	0.955	0.958	0.989	0.877

Table D.2: (Setting B) Simulation results by method with $(\tau, n) = (0.5, 1000)$ and 35% missing in the outcome

Method	M1			M2			M3		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
	Mean biases								
Uncorrected	0.006	-0.003	0.000	0.246	-0.022	-0.079	0.257	-0.026	-0.078
MAR-assumed	0.005	-0.003	-0.001	0.287	-0.026	-0.099	0.294	-0.032	-0.092
JEE	0.004	-0.002	-0.002	0.020	-0.001	-0.008	0.044	-0.009	-0.010
Semiparametric IV	-0.029	0.005	0.016	0.044	0.002	-0.019	0.055	0.003	-0.034
	RMSE								
Uncorrected	0.101	0.050	0.051	0.272	0.054	0.095	0.280	0.055	0.093
MAR-assumed	0.106	0.051	0.056	0.338	0.073	0.136	0.352	0.077	0.138
JEE	0.111	0.052	0.057	0.233	0.076	0.103	0.269	0.084	0.118
Semiparametric IV	0.110	0.050	0.056	0.135	0.052	0.063	0.139	0.053	0.068
	CI lengths								
Uncorrected	0.393	0.189	0.197	0.446	0.187	0.206	0.446	0.189	0.203
MAR-assumed	0.408	0.194	0.211	0.672	0.268	0.414	0.710	0.275	0.377
JEE	0.437	0.204	0.222	0.758	0.264	0.344	0.803	0.291	0.376
Semiparametric IV	0.622	0.308	0.287	0.599	0.280	0.237	0.642	0.304	0.245
	Coverage probabilities								
Uncorrected	0.933	0.936	0.946	0.415	0.918	0.703	0.374	0.922	0.666
MAR-assumed	0.941	0.943	0.942	0.509	0.896	0.723	0.517	0.895	0.744
JEE	0.939	0.946	0.937	0.887	0.921	0.906	0.842	0.920	0.899
Semiparametric IV	0.985	0.996	0.981	0.961	0.992	0.949	0.968	0.992	0.917

Table D.3: (Setting D) Simulation results by method with $(\tau, n) = (0.5, 1000)$ and 35% missing in the outcome

Method	M1			M2			M3		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
	Mean biases								
Uncorrected	0.004	-0.002	-0.001	0.381	-0.035	-0.123	0.377	-0.037	-0.113
MAR-assumed	0.006	-0.002	-0.001	0.434	-0.039	-0.151	0.403	-0.043	-0.118
JEE	0.003	-0.002	0.000	0.059	-0.004	-0.025	0.054	-0.009	-0.013
Semiparametric IV	-0.045	0.012	0.023	0.062	0.007	-0.033	0.079	0.004	-0.049
	RMSE								
Uncorrected	0.135	0.065	0.071	0.409	0.073	0.140	0.405	0.074	0.134
MAR-assumed	0.144	0.069	0.077	0.487	0.093	0.192	0.467	0.101	0.178
JEE	0.152	0.069	0.080	0.318	0.101	0.143	0.345	0.112	0.155
Semiparametric IV	0.160	0.071	0.078	0.192	0.077	0.088	0.194	0.071	0.098
	CI lengths								
Uncorrected	0.543	0.262	0.274	0.604	0.255	0.281	0.603	0.257	0.277
MAR-assumed	0.568	0.271	0.294	0.855	0.326	0.480	0.850	0.351	0.472
JEE	0.621	0.285	0.311	1.005	0.355	0.448	1.021	0.379	0.465
Semiparametric IV	0.661	0.328	0.301	0.722	0.336	0.284	0.781	0.361	0.298
	Coverage probabilities								
Uncorrected	0.944	0.962	0.936	0.289	0.924	0.597	0.294	0.922	0.633
MAR-assumed	0.938	0.951	0.934	0.383	0.900	0.633	0.484	0.888	0.697
JEE	0.935	0.953	0.923	0.863	0.924	0.866	0.817	0.906	0.828
Semiparametric IV	0.952	0.965	0.940	0.922	0.960	0.874	0.935	0.979	0.847

Table D.4: (Setting E) Simulation results by method with $(\tau, n) = (0.5, 1000)$ and 35% missing in the outcome

Method	M1			M2			M3		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
					Mean biases				
Uncorrected	0.007	-0.004	0.002	0.231	-0.037	-0.033	0.247	-0.045	-0.021
MAR-assumed	0.006	-0.004	0.002	0.279	-0.038	-0.075	0.311	-0.044	-0.080
JEE	0.006	-0.003	0.001	0.061	-0.011	-0.016	0.068	-0.016	-0.010
Semiparametric IV	-0.032	0.007	0.037	-0.052	0.024	0.045	-0.016	0.018	0.001
					RMSE				
Uncorrected	0.108	0.059	0.069	0.265	0.070	0.079	0.284	0.078	0.076
MAR-assumed	0.109	0.058	0.070	0.349	0.086	0.155	0.381	0.096	0.161
JEE	0.113	0.058	0.070	0.260	0.086	0.125	0.287	0.094	0.146
Semiparametric IV	0.118	0.058	0.081	0.155	0.067	0.106	0.151	0.069	0.092
					CI lengths				
Uncorrected	0.370	0.214	0.211	0.479	0.232	0.246	0.464	0.231	0.222
MAR-assumed	0.370	0.207	0.201	0.768	0.288	0.468	0.687	0.296	0.356
JEE	0.452	0.230	0.286	0.764	0.296	0.398	0.866	0.325	0.453
Semiparametric IV	0.698	0.398	0.337	0.717	0.410	0.294	0.767	0.419	0.308
					Coverage probabilities				
Uncorrected	0.901	0.928	0.862	0.527	0.900	0.867	0.466	0.855	0.856
MAR-assumed	0.891	0.927	0.844	0.530	0.888	0.743	0.525	0.864	0.714
JEE	0.946	0.941	0.944	0.889	0.947	0.915	0.859	0.911	0.885
Semiparametric IV	0.995	0.998	0.956	0.975	0.989	0.852	0.977	0.991	0.900

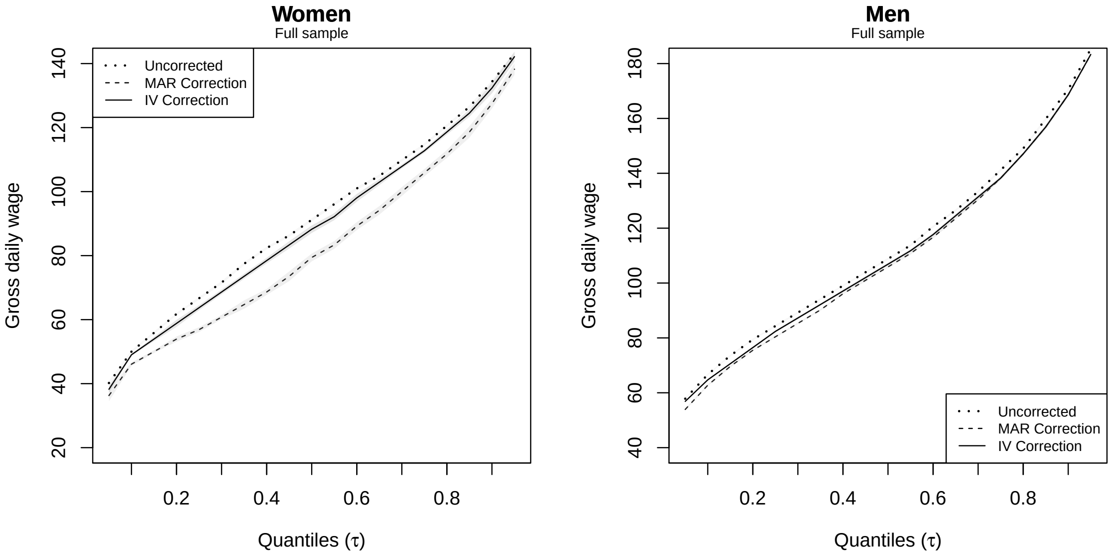
E Robustness

This section presents several robustness checks. In these analyses, we impose stricter requirements on the lag structure of the instrument. In the main specification, the initial wage instrument includes wages observed up to 2015, using full-time wages where available and part-time wages when no full-time record exists. The robustness checks restrict the instrument in two alternative ways: (i) by using only wages from 2011 or earlier (i.e., at least five years before the outcome period) as valid instruments, and (ii) by limiting the instrument to the earliest observed full-time wage only.

E.1 Instrument based on wage history from 2011 or prior

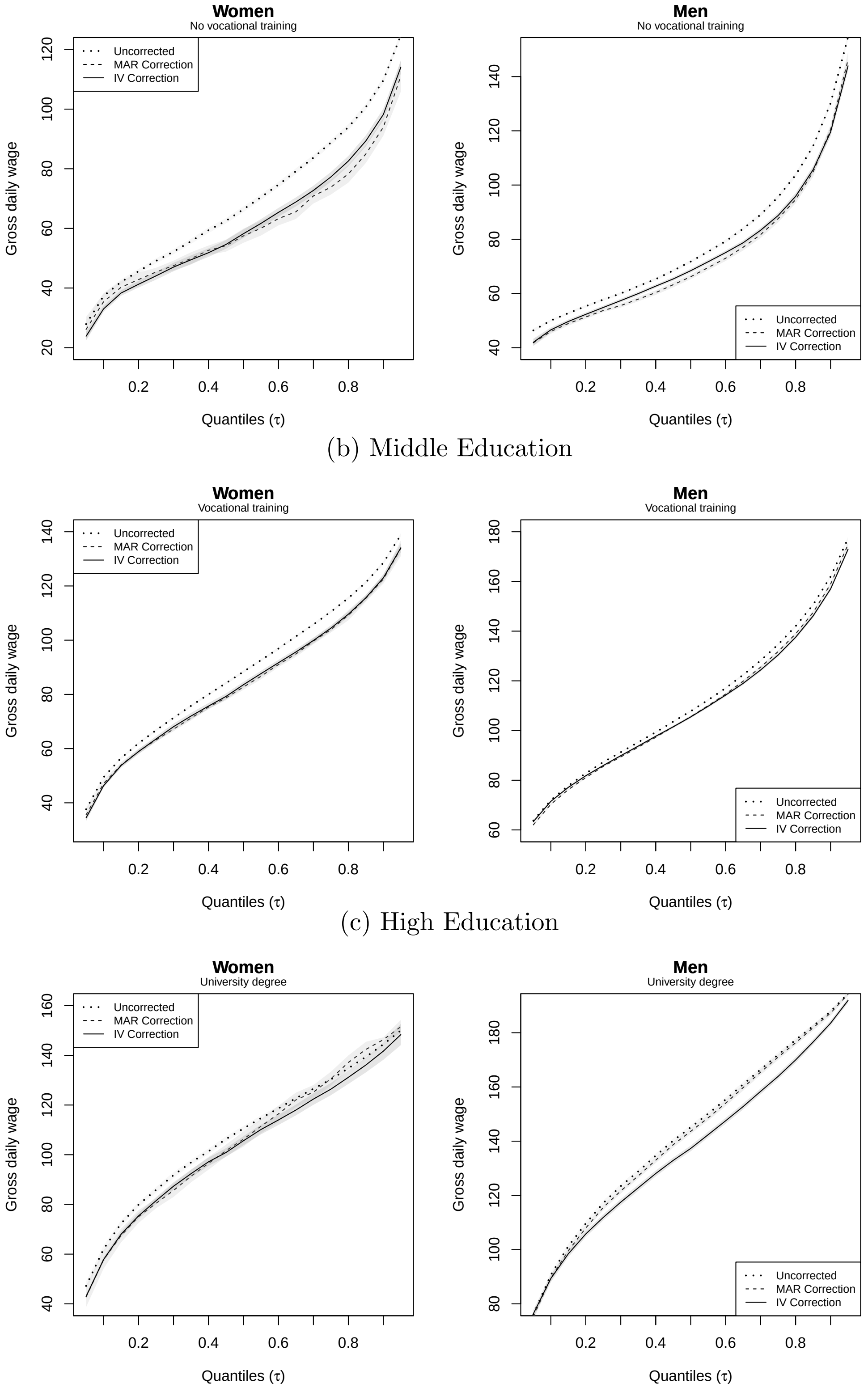
Figures E.1 and E.2 replicate Figures 2 and 3, respectively. The estimated quantile patterns remain virtually unchanged, with effect sizes and gender differences closely mirroring those in the main specification. This confirms that the results are robust to using earlier wage information as the instrument.

Figure E.1: Quantile Selection Effects (Unconditional) – only IV up to 2011



Note: SIAB-R data, 2017. Curves display wage point estimates from unconditional quantile regressions across quantiles, connected by lines for visualization: uncorrected (dotted), MAR correction (dashed), and IV correction (solid). The IV correction uses initial log wage as an instrument and controls for education, age, workplace region, total experience, and earliest job difficulty in the first two stages. The MAR correction uses the same controls except earliest job difficulty. The final-stage regressions include no covariates for all methods. Shaded areas represent 95% pointwise confidence intervals.

Figure E.2: Conditional wage quantiles by education groups – only IV up to 2011



(a) Low Education

(b) Middle Education

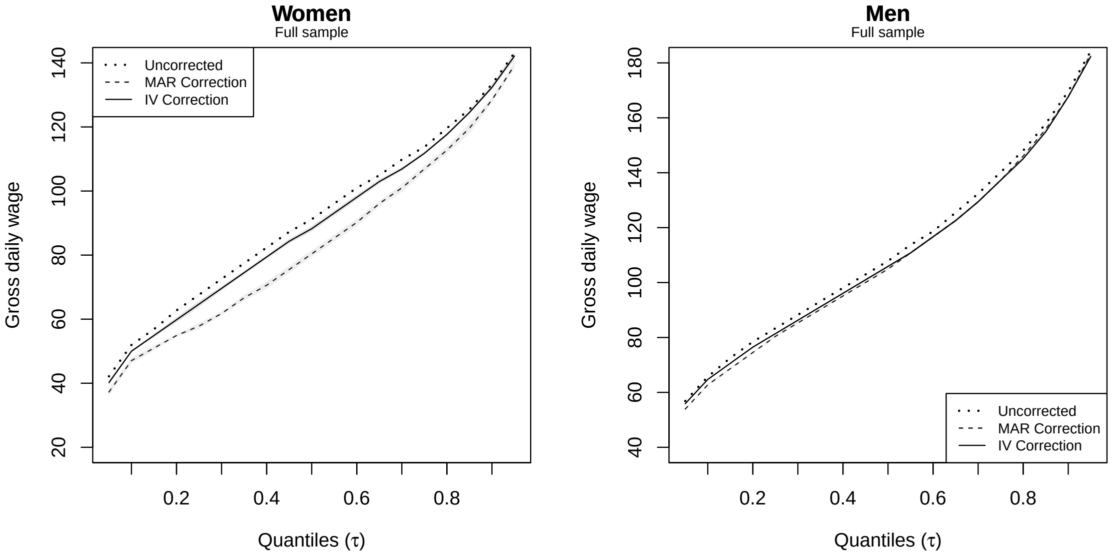
(c) High Education

Note: SIAB-R data, 2017. Curves show wage point estimates from unconditional quantile regressions across quantiles, connected by lines for visualization: uncorrected (dotted), MAR correction (dashed), and IV correction (solid). The IV correction uses initial log wage as an instrument and controls for education, age, workplace region, total experience, and earliest job difficulty. The MAR correction uses the same controls, excluding the instrument and earliest job difficulty. The uncorrected estimates include the same controls as in MAR. Shaded areas represent 95% pointwise confidence intervals.

E.2 Instrument based on only full-time wage history

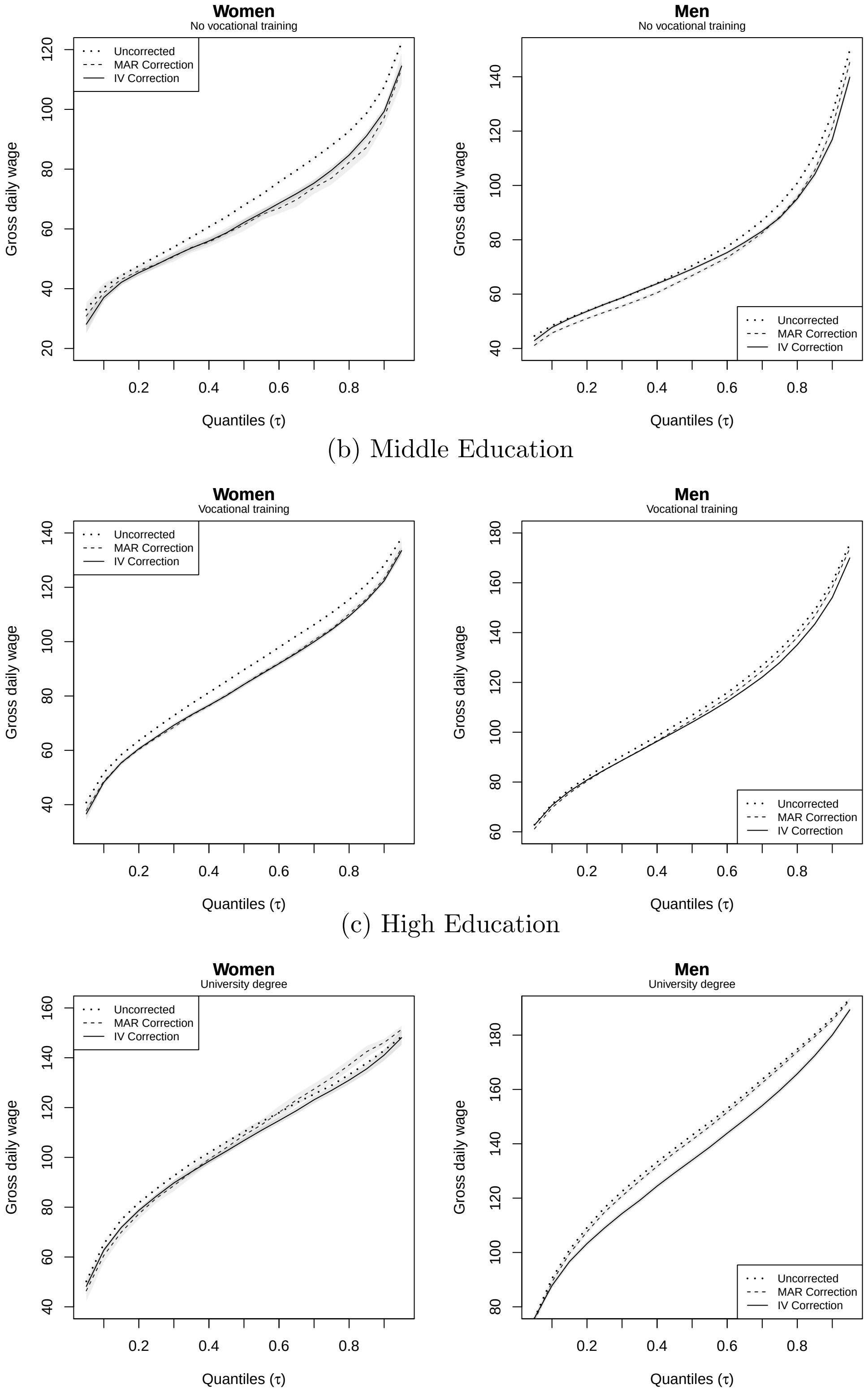
Figures E.3 and E.4 replicate Figures 2 and 3, respectively. The estimated quantile patterns remain virtually unchanged, with effect sizes and gender differences closely mirroring those in the main specification. This confirms that the results are robust to using only full-time wage information as the instrument.

Figure E.3: Quantile Selection Effects (Unconditional) – only full-time instruments



Note: SIAB-R data, 2017. Curves display wage point estimates from unconditional quantile regressions across quantiles, connected by lines for visualization: uncorrected (dotted), MAR correction (dashed), and IV correction (solid). The IV correction uses initial log wage as an instrument and controls for education, age, workplace region, total experience, and earliest job difficulty in the first two stages. The MAR correction uses the same controls except earliest job difficulty. The final-stage regressions include no covariates for all methods. Shaded areas represent 95% pointwise confidence intervals.

Figure E.4: Conditional wage quantiles by education – only full-time instruments



(a) Low Education

(b) Middle Education

(c) High Education

Note: SIAB-R data, 2017. Curves show wage point estimates from unconditional quantile regressions across quantiles, connected by lines for visualization: uncorrected (dotted), MAR correction (dashed), and IV correction (solid). The IV correction uses initial log wage as an instrument and controls for education, age, workplace region, total experience, and earliest job difficulty. The MAR correction uses the same controls, excluding the instrument and earliest job difficulty. The uncorrected estimates include the same controls as in MAR. Shaded areas represent 95% pointwise confidence intervals.