Gini Index Regression *

Vedant Bhardwaj[1] and David Díaz-Villarejo[1,2]

[1]Universidad Carlos III de Madrid, *PhD Candidate*

[2]*Job Market Paper*

October 28, 2026

Click here for the latest version of the paper

Abstract

Many economic questions concern how and why inequality varies between populations with different characteristics. Comparisons of average outcomes do not reveal such variation. This paper develops Gini Index Regression (GIR), a framework in the spirit of mean or quantile regression that directly models the Gini of the conditional distribution of a nonnegative outcome, without choosing a parametric functional form for that distribution. In the linear case, the parameter of interest is shown to have a closed-form solution and, under model misspecification, the fitted function is a weighted best linear approximation to the conditional Gini. An omitted variable bias formula shows that changing controls affects both the projection and the ranking of outcomes. The estimator combines nonparametric series estimation of conditional ranks with weighted least squares, and reproduces empirical group Ginis in saturated group-indicator specifications. We establish $\sqrt{n}$-consistency, asymptotic normality, and validity of the bootstrap, accounting for both stages. A prediction characterization of the conditional Gini is also used to propose an R^2-type goodness-of-fit measure. An application to U.S. data finds that the wage inequality gap by union status remains significant even after controlling for relevant characteristics, and that college-educated workers indeed possess the lowest levels of inequality.

Keywords: Conditional inequality, parametric conditional Gini, regression, semiparametric inference.

*We are deeply grateful to Carlos Velasco for his invaluable guidance, patience, and generous time. We also thank Juan Carlos Escanciano, Jesús Gonzalo, Jan Stuhler, Felix Wellschmied, the members of the Econometrics Reading Group at the University of Mannheim, with particular thanks to Christoph Rothe. Generative AI was used for improving the readability of text, for helping with coding the statistical model, and for checking the final paper for technical accuracy (`Refine.ink`).

1 Introduction

Many economic questions concern why and how the inequality or relative dispersion in an outcome Y differs across subpopulations defined by observed characteristics or exposure to a treatment X. These comparisons go beyond what can be captured by a difference-in-means. For instance, does trade liberalization reduce firm markup dispersion? Does dispersion change differently across industries with different trade exposure? (Lu and Yu, 2015) How do changes in firm hierarchies following a trade shock change workers' wage inequality? (Friedrich, 2022). Are earnings more unequal among highly educated workers than among less educated? (Brunello et al., 2009).

The natural approach is to compare measures of inequality or spread by subpopulations, i.e., inequality of the conditional distribution of the outcome given treatment and population characteristics. For example, regarding the first two questions, one could compare markup dispersion for groups and periods with low versus high levels of trade exposure, and further condition on industry to unveil potential heterogeneity. Regression methods accomplish these comparisons succinctly and, through a parametric specification targeted to the distributional feature of interest, also allow simple comparisons between groups while holding covariates fixed. Mean (OLS) and quantile regression (QR) provide well-developed frameworks for modeling those specific features of the conditional distribution. Comparable regression methods remain under-explored for inequality or relative dispersion measures of a conditional distribution, with the Gini being the most popular. We develop its corresponding framework, placing the parametric specification directly on inequality and providing both coefficient interpretations and feasible inference.

Contribution. This paper turns the Gini of a conditional distribution $\mathbb{G}(Y|X)$ into a regression target with its own specification. Our formulation enables precise coefficient interpretation and inference without requiring the researcher to commit to a particular distribution function (or indexed family) generating the data, and without parameterizing its conditional mean or quantiles. We do so by deriving an equation involving a general parametric model $m(X;\theta)$ and an additive error term on which we impose a conditional moment restriction. Our characterization is based on a unified representation of the Gini of a conditional distribution of outcomes Y (either continuous, discrete, mixed, or singular) given covariates X, which is shown to be equivalent to other well-known representations of the Gini (Lorenz curve, pairwise differences, covariance, etc.). We then provide the econometric foundations for the linear specification $\mathbb{G}(Y|X) = X'\theta$. We find that, under the proposed identifying moment and a standard rank assumption, the parameter admits

a familiar closed-form solution in terms of moments of the joint distribution of Y and X.

Furthermore, we allow the regression equation to be misspecified, in which case it is shown that it can be interpreted as a weighted, best linear approximation to the conditional Gini function, analogous to the weighted approximation interpretation of QR (Angrist et al., 2006). We also show that omitting variables changes the estimand through two channels: the familiar projection channel, as in OLS and QR, and a change in the conditional population used to rank outcomes; so that for inequality, the choice of controls does more than determine the regression specification: it determines how and which outcomes are compared. We refer to our approach as Gini Index Regression (GIR)[1], aimed at describing how relative dispersion varies among observationally comparable workers, households, firms, or transactions, without imposing a particular welfare interpretation on that dispersion.

The paper then proposes a feasible two-step semiparametric GIR estimator for the linear-in-parameters model based on regression methods, where the first stage involves flexible, nonparametric estimation of the conditional *rank advantage* $R(y|x)$ of the observed outcomes, that is, the share of the subpopulation, within group $X = x$, with values below the observed outcome minus the share above, $R(y|x) := P(Y < y|X = x) - P(Y > y|X = x)$. It determines where each observation lies in the outcome distribution of units with similar characteristics. We do not impose parametric assumptions on the conditional distribution $F(y|x)$, either directly or indirectly via quantile or hazard functions. First-stage estimation relies on a least-squares series estimator of the conditional rank advantage. The second stage regresses the estimated rank advantage on X, with weights proportional to Y. The estimator is simple to implement and is shown to have desirable properties, namely the scale invariance in the outcome variable, directly inherited from the Gini; and the capability to reproduce empirical group Ginis with saturated group dummies in both stages, as in OLS and QR. Therefore, the proposed framework naturally extends the familiar Gini calculation to covariate profiles involving continuous characteristics and interactions.

Although GIR estimation reduces to regression operations, inference must account for the nonparametric estimation of conditional ranks, which induces a generated dependent variable whose estimation error contributes to the GIR estimator's first-order behavior and invalidates ordinary regression formulas for inference, which treat the ranks as observed. We provide asymptotic results for *i.i.d.* data: (i) we show that the GIR estimator is $\sqrt{n}$-consistent and asymptotically normal, and (ii) we prove the validity of a computationally convenient bootstrap for the series-based procedure. We do this by extending the uniform

[1] As opposed to Gini Regression (see Yitzhaki and Schechtman (2013), Chapters 7–10), which is unrelated to the conditional Gini.

asymptotic linearization results, including the weighted bootstrap, for least-squares series estimators in Belloni et al. (2015) to cover the series approximation to the family of indexed rank-advantage maps $\{x \mapsto R(y|x) : y \in \mathcal{Y}\}$ uniformly over (y, x), and combining the extensions with the classical semiparametric inference setup of Newey (1994a). Our asymptotic results cover both the continuous and discrete outcome cases.

Finally, connecting the regression framework to the target, we show that given the characteristics defining a comparable population, the conditional Gini is the optimal outcome-weighted prediction of an individual's rank advantage. We use the result to construct an R^2-type goodness-of-fit measure for GIR. To ease implementation, the proposed methods are available in a companion `R` package, `GIR`.[2]

We revisit DiNardo et al. (1996) and apply our framework to U.S. CPS wage data from 1979 to 1988 to study how inequality varies with union status, education, and labor-market experience, and show that conditioning on observed characteristics reveals heterogeneity that is not summarized by unconditional group comparisons. Non-union workers exhibit higher conditional wage inequality, and the union-status gap widens substantially with experience. Moreover, after conditioning, we find that college-educated workers indeed have the lowest levels of inequality for all experience profiles compared to the less educated groups, even though it increases with labor-market experience.

Related Literature. Methodologically, we contribute to the recent literature on conditional inequality. The existing literature falls into two main strands: those that target the conditional distribution function to obtain the implied inequality measure and those that target the measure directly. The former relies heavily on parametric assumptions on the conditional distribution function. For example, to understand how experience and education shape the wage distribution and its inequality, Febrer and Mora (2009) specify a parametric conditional hazard model and then compute the implied conditional Gini; Klein et al. (2015) use a Bayesian parametric distributional regression framework and then obtain a conditional Gini by education level, years of experience, and regions. Aaberge et al. (2006) take a more flexible approach and propose to estimate the conditional distribution function semiparametrically before computing inequality measures. More recently, Hurtado (2025) assumes a linear conditional quantile process and decomposes the implied conditional Gini. Similarly, Jokiel-Rokita et al. (2024) parameterize the conditional quantile function and use it to construct inequality measures.

[2] See `https://github.com/uptoaparameter/GIR` for an `R` package which implements the GIR estimator, R^2, and the related inferences.

We make the following point. When the conditional inequality measure is the object of interest, restrictions on a conditional distribution or quantile process need not automatically translate into a simple or parsimonious regression equation for that functional. Since the Gini and other inequality measures are typically complicated functions of the conditional distribution, quantile, or hazard, even a parametric or semiparametric model for these objects can make the parameters and marginal effects on inequality difficult to obtain or interpret. This is the limitation we address. As in linear regression, our formulation focuses the parametric specification on the object the researcher wants to interpret, removing the need to commit to a particular distribution function when inequality is the single feature of interest. However, the strong dependence of the Gini on the shape of the distribution makes conditional mid-rank estimation inevitable. GIR handles it nonparametrically. On the one hand, GIR allows for a simple interpretation of covariate effects on inequality, summarized in a single parameter vector, while, on the other hand, it remains flexible about the distribution generating the data (up to the specification of the conditional Gini).

Within the second strand of the literature, our paper is closely related to the recent work by Ahlin and Jeong (2021), which proposes a conditional inequality measure given pairs of individual characteristics, x_i and x_j, based on the Gini, to study how wealth differences and income inequality differ by financial access. Their measure coincides with the Gini of the conditional distribution whenever $x_i = x_j = x$. They propose two separate linear parametric specifications for a ratio: one for the conditional Gini mean-difference (GMD) numerator, the other for the conditional mean denominator. Both are estimated by OLS.

Their formulation, as ours, avoids specifying the entire conditional distribution, but specifies the conditional mean function to be linear. This adds functional form structure. Our formulation is agnostic about the exact dependence of the conditional mean function on covariates. Moreover, marginal effects in their proposed ratio specification involve both the numerator, denominator, and derivatives, while the parameters in a simple linear GIR specification directly describe inequality gradients. The two specifications are not generally nested, though a nonlinear GIR model can accommodate their specification. Under model misspecification, a ratio of linear projections is generally not the projection of the ratio. A linear GIR estimand has an explicit weighted linear projection interpretation of the true conditional Gini function on covariates. Nevertheless, neither formulation is guaranteed to take values in the unit interval, analogous to linear probability models. Recently, Farcomeni (2026) targets the quantile share ratio as a measure of inequality, which also requires first-stage estimation of conditional ranks. His asymptotic result depends on the assumed rate for that first stage, and when a parametric distribution regression

model is used, the admissible data-generating processes must satisfy the parametric restrictions of both stages. Instead, we establish $\sqrt{n}$-inference for the conditional Gini with a nonparametric series first stage.

Our paper also relates to applied work computing inequality or relative dispersion measures in economics. Applied researchers typically compute the Gini index separately for each survey wave, city, region, country, or demographic category (gender, age, race, birth cohort, earnings percentile, etc.) and then compare its value across the different categories. Some examples include Gottschalk and Smeeding (1997); Bowlus and Robin (2004); Kopczuk et al. (2010). Since the GIR estimator for the saturated group dummy specification reproduces empirical Ginis, the approach is encompassed by GIR. In the case of continuous covariates, imposing a model specification removes the need to choose how to discretize its support to compute empirical Ginis.

Another approach relates estimated subgroup Ginis to a particular treatment and observed characteristics via linear regression (Lu and Yu, 2015; Berkes and Gaetani, 2023). Such regressions can answer meaningful questions about inequality across groups, but do not directly target the effect on the Gini of the distribution. This approach involves aggregating the response variable in a first stage before running OLS of the group measures on covariates. A linear GIR model uses individual-level observations to first estimate individual conditional mid-ranks used to run a weighted linear regression on covariates. Inference accounts for estimation error in the generated dependent variable. Under correct model specification, the former approach measures the effect on the conditional mean of group Ginis, while the latter captures the effect on the Gini of the conditional distribution of the response variable. Under model misspecification, the former approach estimates the best linear approximation to the conditional mean of group Ginis in a least squares sense, while GIR estimates the best linear approximation to the conditional Gini in a weighted least squares sense. The two approaches can coincide in special cases, including saturated group comparisons, but generally differ in the comparison populations, weighting, and restrictions used to construct the conditional inequality relationship.

To account for first-stage estimation, our first-stage analysis builds on the least squares series theory of Belloni et al. (2015), extending the relevant ordinary and weighted-bootstrap linearizations to the threshold-indexed family of conditional mid-distribution regressions. It complements the recent partitioning-based distribution regression theory of Cattaneo et al. (2026). Related results include uniform distribution-regression and bootstrap theory under fixed-dimensional specifications in Chernozhukov et al. (2013) or Chernozhukov

et al. (2024), and series conditional quantile process theory in Belloni et al. (2019). Our analysis supplies the first-stage expansions needed for inference on the GIR estimator.

We do not, however, provide a solution to the Gini decomposition problem (for a recent treatment, see Heikkuri and Schief (2026)), nor do we analyze the effects of covariates on aggregate unconditional inequality à la RIF regression (Firpo et al., 2009). Our focus is on differences in within-group inequality or relative dispersion across covariate profiles.

Outline. The remainder of the paper is organized as follows. Section 2 provides the representation of the conditional Gini used throughout, parametric examples, and a general characterization of parametric models. The following two sections focus on the linear model. Section 3 provides identification results. Section 4 proposes a feasible estimator and contains the main asymptotic results, including the validity of a weighted bootstrap. Section 5 shows the prediction interpretation of the conditional Gini and proposes an R^2-type goodness-of-fit measure. Finally, Sections 6 and 7 contain, respectively, a simulation study and the empirical application. Section 8 concludes.

2 The Setting for Gini Index Regression

In this section, we first review how the Gini is constructed directly from the Lorenz curve in both the unconditional and conditional cases –which will form the basis for our modeling approach– and then proceed with its parameterization, with a specific focus on the simple linear case. We also emphasize the generality of our results, which are valid for continuous, discrete, mixed, or singular distributions. Inference theory covers the first two.

2.1 Conditional Gini Representation

Our starting point is the elementary definition of the Gini of a distribution. Let Y be a random variable with support $\mathscr{Y} \subseteq \mathbb{R}_+ := [0,\infty)$ and cumulative distribution function (CDF) $F(y)$. Let $F^{-1}(\tau) := \inf\{y : F(y) \geq \tau\}$ denote its left-inverse or quantile function and assume that $0 < \mu < \infty$, where $\mu := \mathbb{E}[Y]$. The Lorenz curve is defined by

$$L(u) := \mu^{-1} \int_0^u F^{-1}(s)\, ds \ , \quad 0 \leq u \leq 1$$

(see Gastwirth (1971)) and the unconditional Gini of Y is given by the following expression

$$\mathbb{G}(Y) := 1 - 2\int_0^1 L(p)\, dp = \frac{\mathbb{E}[Y(2M(Y)-1)]}{\mathbb{E}[Y]} \in [0,1) \tag{1}$$

where $M(y) := \big(F(y^-) + F(y)\big)/2$ is the mid-distribution function of Y, with $F(y^-) := P(Y < y) = \lim_{z \uparrow y} F(z)$. Note that $\mathbb{E}[M(Y)] = 1/2$. The concept of the mid-distribution function is not new (see Parzen (2004)). The Gini formula above is general: it is valid for any non-negative random variable with strictly positive and finite mean, whether continuous, discrete, singular, or mixed; with or without density. If $F(y)$ is continuous, then $M(y) = F(y)$. While covariance representations of the Gini based on $F(y)$ are standard for continuous distributions, and mid-distribution adjustments have been proposed for discrete distributions, we provide a unified proof of (1) in Appendix A, showing that the mid-distribution representation is valid for arbitrary distributions, along with its equivalence to the independent-copies or Gini mean-difference (GMD), and covariance representations. The proof we provide thus extends known continuous and discrete representations to arbitrary distributions. To our knowledge, the exact unified statement proven in Lemma 4 of Appendix A, covering continuous, discrete, mixed, and singular distributions, has not previously been recorded in this form.

Note that the re-scaled centered mid-rank satisfies $2M(y) - 1 = P(Y < y) - P(Y > y) =: R(y)$, which we refer to as "rank advantage", that is, the proportion of individuals whose outcomes are strictly below y minus the proportion whose outcomes are strictly above y. The Gini then averages the rank advantage of individuals, weighting them proportionally to their outcome levels: $\mathbb{G}(Y) = \mathbb{E}[w(Y)R(Y)]$, where $w(Y) := Y/\mathbb{E}[Y] \geq 0$ and $\mathbb{E}[w(Y)] = 1$.

Likewise, we can construct the Gini index in the conditional case. Let $X = (X_1, \ldots, X_p)'$ denote a $p \times 1$ vector of covariates with law P_X and let $F(y|x)$ denote the conditional CDF of Y given $X = x$. Let $F^{-1}(\tau|x)$ denote its quantile function defined as its left-inverse, and

$$\mu(u|x) := \int_0^u F^{-1}(v|x)\, dv \ , \quad 0 \leq u \leq 1$$

Assume that $0 < \mu(1|x) < \infty$ for P_X-almost every x and define the conditional Lorenz curve

$$L(u|x) := \frac{\mu(u|x)}{\mu(1|x)} = \frac{\int_0^u F^{-1}(v|x)\, dv}{\int_0^1 F^{-1}(v|x)\, dv} \ , \quad 0 \leq u \leq 1$$

(see Aaberge et al. (2006)). The conditional Gini of Y given $X = x$ is defined analogously

$$G(x) := 1 - 2\int_0^1 L(p|x)\, dp = \frac{\mathbb{E}[Y(2M(Y|X = x) - 1)|X = x]}{\mathbb{E}[Y|X = x]} \in [0, 1) \tag{2}$$

where $M(y|X = x) = M(y|x) := \big(F(y^-|x) + F(y|x)\big)/2$ is the conditional mid-distribution function of Y given $X = x$, with $F(y^-|x) := P\big(Y < y|X = x\big) = \lim_{z \uparrow y} F(z|x)$. Again, the

formula is valid for any type of response variable as long as it is non-negative and admits a conditional distribution with strictly positive and finite conditional mean. In particular, if $F(y|x)$ is continuous in y for P_X-almost every x, then $M(y|x) = F(y|x)$. The derivation of (2), which is the conditional analog of (1), is also included in Appendix A, along with its equivalent conditional versions of the independent-copies and covariance representations.

$G(x)$ will be referred to as the Conditional Gini Function (CGF), and will typically be treated as a random object by evaluating it at the random variable X, i.e., $G \circ X$, written

$$\mathbb{G}(Y|X) = \frac{\mathbb{E}[Y(2M(Y|X)-1)|X]}{\mathbb{E}[Y|X]} \in [0,1) \tag{3}$$

Expressions (2) and (3) are the starting point for the parameterization of the conditional Gini. Throughout the paper, we thus focus on non-negative outcomes, $Y \geq 0$ *a.s.*, satisfying both $0 < \mathbb{E}[Y] < \infty$ and $0 < \mathbb{E}[Y|X] < \infty$ *a.s.* The asymptotic theory developed in Section 4.4 covers both continuous and discrete responses as well as general covariates.

Analogous to the unconditional case, the re-scaled centered conditional mid-rank satisfies $2M(y|x) - 1 = P(Y < y|X = x) - P(Y > y|X = x) =: R(y|x)$, which we refer to as "conditional rank advantage": the proportion of individuals in group $X = x$ whose outcomes are strictly below y minus the proportion whose outcomes are strictly above y in the same group.

2.2 The Parametric Conditional Gini

We now introduce the parameterization of the Gini of the conditional distribution of Y given $X = x$. Let the mapping $m : \mathscr{X} \times \Theta \to \mathbb{R}$, defined by $(x, \theta) \mapsto m(x;\theta)$ be a parametric model for the conditional Gini, where $\mathscr{X} \subseteq \mathbb{R}^p$ denotes the support of X, $\Theta \subseteq \mathbb{R}^q$ is the parameter space and $\theta \in \Theta$ an unknown parameter of interest. To begin our discussion, we state the following assumption regarding the specification of the model and present some examples of general parametric models.

Assumption 1. *(Correct Specification) There exists some true value of the parameter* $\theta_0 \in \Theta$ *such that* $\mathbb{G}(Y|X = x) = G(x) = m(x;\theta_0) \in [0,1)\ \forall\, x \in \mathscr{X}$.

Under Assumption 1, the CGF becomes

$$\mathbb{G}(Y|X) = m(X;\theta_0) \tag{4}$$

2.3 Examples of Parametric Conditional Ginis

For illustration purposes, we provide some examples of well-known (parametric) conditional distributions, some typically used to model inequality, for which the Gini can be expressed in closed form as a function of their parameters, and show that the same parametric model can have different underlying distributions. The following non-exhaustive list of examples is key to simulating data with the desired functional form for the conditional Gini, key to conducting Monte Carlo experiments. Note that, in the following examples, the conditional mean is allowed to be highly nonlinear.

Example 2.1. *Let $Y|X=x \sim P(c(x), \alpha(x;\theta_0))$, where $c(x)>0$ and $\alpha(x;\theta_0)$ are the minimum and shape parameters of the conditional Pareto distribution, respectively. Let $\alpha(x;\theta_0) = \frac{1+m(x;\theta_0)}{2m(x;\theta_0)} > 1$. It can be shown that $\mathbb{G}(Y|X=x) = m(x;\theta_0) \in (0,1)$.*

□

Example 2.2. *Let $Y|X=x \sim LogN\big(\mu(x), \sigma(x;\theta_0)\big)$, where $\mu(x) \in \mathbb{R}$ and $\sigma(x;\theta_0)>0$ are the location and shape parameters of the conditional Log-Normal distribution, respectively. Let $\sigma(x;\theta_0) = \sqrt{2}\,\Phi^{-1}\left(\frac{m(x;\theta_0)+1}{2}\right)$, where $\Phi^{-1}(\cdot)$ denotes the standard normal quantile function. It can be shown that $\mathbb{G}(Y|X=x) = m(x;\theta_0) \in (0,1)$.*

□

Example 2.3. *Let $Y|X=x \sim LL\big(\alpha(x), \beta(x;\theta_0)\big)$, where $\alpha(x)>0$ and $\beta(x;\theta_0)>1$ are the scale and shape parameters of the conditional Log-Logistic distribution, respectively. Let $\beta(x;\theta_0) = 1/m(x;\theta_0) > 1$. It can be shown that $\mathbb{G}(Y|X=x) = m(x;\theta_0) \in (0,1)$.*

□

Example 2.4. *Let $Y|X=x \sim We(\lambda(x), k(x;\theta_0))$, where $\lambda(x)>0$ and $k(x;\theta_0)>0$ are the scale and shape parameters of the conditional Weibull distribution, respectively. Let $k(x;\theta_0) = \frac{-1}{\log_2(1-m(x;\theta_0))}$. It can be shown that $\mathbb{G}(Y|X=x) = m(x;\theta_0) \in (0,1)$.*

□

Example 2.5. *Let $Y|X=x \sim Exp(\lambda(x;\theta_0))$, where $\lambda(x;\theta_0)>0$ is the rate parameter of the conditional Exponential distribution. It can be shown that $\mathbb{G}(Y|X=x) = 0.5$, so the conditional exponential distribution cannot generate non-constant CGFs.*

□

Example 2.6. *Let $Y|X=x \sim Beta\big(\alpha(x;\theta_0), \beta(x;\theta_0)\big)$, where $\alpha(x;\theta_0)>0$ and $\beta(x;\theta_0)>0$ are the shape parameters of the conditional Beta distribution. Let $\alpha(x;\theta_0) = \frac{1-m(x;\theta_0)}{2m(x;\theta_0)}$ and $\beta(x;\theta_0) = \beta = 1$. It can be shown that $\mathbb{G}(Y|X=x) = m(x;\theta_0) \in (0,1)$.*

□

Example 2.7. *Let $Y|X = x \sim U[a(x;\theta_0), b(x;\theta_0)]$, where $a(x;\theta_0) < b(x;\theta_0)$ are the minimum and maximum parameters of the conditional Uniform distribution. Let $b(x;\theta_0) = \frac{a(x;\theta_0)(1+3m(x;\theta_0))}{1-3m(x;\theta_0)}$ and $a(x;\theta_0) > 0$. It can be shown that $\mathbb{G}(Y|X = x) = m(x;\theta_0) \in (0, 1/3)$.*

□

Examples 2.1 and 2.5 warn against routinely replacing the outcome by its logarithm if the object of interest is the Gini of the conditional distribution of Y. Note that if $Y|X = x \sim P(c(x), \alpha(x;\theta_0))$ with $c(x) = c = 1$ and $\alpha(x:\theta_0) = \frac{1+m(x;\theta_0)}{2m(x;\theta_0)} > 1$, it can be shown that $\ln(Y)|X = x \sim Exp(\alpha(x))$, so that $\mathbb{G}(Y|X = x) = m(x;\theta_0)$ and $\mathbb{G}(\ln(Y)|X = x) = 0.5$ for P_X-almost every x. Thus, the conditional Gini of the log-outcome is constant even when the conditional Gini of the outcome in levels varies with covariates. An analysis based solely on the former would not reveal the covariate-related variation present in the latter.

The Gini index is not invariant to monotonic transformations. The logarithm changes the measure in a highly distribution-dependent manner. Moreover, unlike the Gini of Y, the Gini of $\ln(Y)$ is generally sensitive to the units in which the original outcome is measured: rescaling Y induces an additive shift in $\ln(Y)$, to which the Gini is not invariant.

2.4 Characterization of Gini Index Regression Models

Unlike in Mean or Quantile Regression, no characterization of a parametric model in terms of an equation and a conditional moment restriction on an error term has been exploited for the conditional Gini. The Gini lacks the convenient linearity property of expectations or the location-scale equivariance of quantiles[3]. It is also, *a priori*, not obvious how to define the error term in the context of a parametric model for the conditional Gini. As an illustration, consider the following simple example.

Example 2.8. *Let $Y = m(X;\theta_0) + e$, where $X \perp e$, $e \sim U[-1, 1]$, and $m(x;\theta_0) > 1 \ \forall x \in \mathscr{X}$. This implies that $Y|X = x \sim U[m(x;\theta_0) - 1, m(x;\theta_0) + 1] \ \forall x \in \mathscr{X}$. We know that*

1. $\mathbb{E}[Y|X] = m(X;\theta_0)$
2. $\mathbb{Q}_{Y|X}(u|X) = m(X;\theta_0) + F_e^{-1}(u), \quad \forall u \in [0, 1]$

However, using (2) and (3), and after some calculations, it can be shown that the conditional Gini is given by $\mathbb{G}(Y|X) = \frac{1}{3m(X;\theta_0)} \in \left(0, \frac{1}{3}\right)$.

□

[3] However, the Gini is scale invariant, that is, it satisfies $\mathbb{G}(cY) = \mathbb{G}(Y)$ for any real number $c > 0$. Likewise, for a strictly positive function $c(x) > 0$, $\mathbb{G}(c(x)Y|X = x) = \mathbb{G}(Y|X = x)$ for P_X-almost every x.

Hence, we propose the following characterization of the CGF. Note that from the definition in (3), under its parameterization in (4), and under Assumption 1 we can write

$$\mathbb{G}(Y|X) = \frac{\mathbb{E}[Y(2M(Y|X)-1)|X]}{\mathbb{E}[Y|X]} = m(X;\theta_0) \tag{5}$$

which suggests the following simple characterization in terms of an equation with an additive error term and a conditional moment restriction[4]. Define the error term

$$U := U(\theta_0) = Y(2M(Y|X)-1) - Y\,m(X;\theta_0)$$

Proposition 1. *(Characterization of Parametric GIR Models) Let $Y \geq 0$ be a random variable with $0 < \mathbb{E}[Y|X] < \infty$ a.s., and let $m(x;\theta)$ be a parametric model, $\theta \in \Theta$. Then*

$$Y(2M(Y|X)-1) = m(X;\theta_0)Y + U \ , \ \mathbb{E}[U|X] = 0 \iff \mathbb{G}(Y|X) = m(X;\theta_0) \tag{6}$$

The model is misspecified if $\nexists\theta \in \Theta$ such that $\mathbb{E}[U(\theta)|X] = 0$ a.s.

Proof. The result follows directly from (5).

□

This simple definition of an additive error term U and the conditional moment restriction $\mathbb{E}[U|X] = 0$ form the basis of all the theoretical results in this paper. Although simple, this characterization has not been realized before in the literature. We add an important example to conclude the discussion of Assumption 1 and the characterization of the model.

Example 2.9. *(Heterogeneous Gini Treatment Effects). Consider the potential outcomes framework with binary treatment and continuous response $Y = Y(0) + (Y(1) - Y(0)) \cdot D$, where $Y(d) \geq 0$ a.s. denotes the potential outcome of individual i under the treatment value $D = d \in \{0,1\}$. Define the Conditional Gini Treatment Effect (CGTE) by the difference between the Gini of the conditional distributions of $Y(1)$ and that of $Y(0)$ given a vector of characteristics $X = x$, i.e., $\tau(x) = \mathbb{G}(Y(1)|X = x) - \mathbb{G}(Y(0)|X = x)$. Under an unconfoundedness or selection on observables assumption, $(Y(0), Y(1)) \perp D \mid X$, and $0 < P(D = 1|X = x) < 1$ for every $x \in \mathscr{X}$, we can identify –as we do with the Conditional Average Treatment Effect (CATE) and Conditional Quantile Treatment Effects (CQTEs)– the CGTE using a correctly specified*

[4] Without parameterization, equation (5) holds replacing $m(X;\theta_0)$ by $G(X)$ (nonparametric characterization).

linear Gini Index Regression model $m_d(x;\theta_0) := m(x,d;\theta_0)$*:*

$$
\begin{aligned}
&Y(2F(Y|X,D)-1) = Y\,m(X,D;\theta_0) + U \\
&\quad\Longrightarrow \mathbb{E}[Y(2F(Y|X,D)-1)|X=x,D=d] = \mathbb{E}[Y\,m(X,D;\theta_0)+U|X=x,D=d] \\
&\quad\Longrightarrow \mathbb{E}[Y(d)(2F(Y(d)|X=x,D=d)-1)|X=x,D=d] = \mathbb{E}[Y(d)|X=x,D=d]\,m(x,d;\theta_0) \\
&\quad\Longrightarrow \mathbb{G}(Y(d)|X=x,D=d) = m_d(x;\theta_0) = \mathbb{G}(Y|X=x,D=d)
\end{aligned}
$$

Under the conditional independence assumption we can write

$$
\begin{aligned}
\tau(x) &= \mathbb{G}(Y(1)|X=x) - \mathbb{G}(Y(0)|X=x) = \mathbb{G}(Y(1)|X=x,D=1) - \mathbb{G}(Y(0)|X=x,D=0) \\
&= m_1(x;\theta_0) - m_0(x;\theta_0)
\end{aligned}
$$

□

3 The Linear Gini Index Regression (GIR) Model

In this section, we provide a detailed understanding of the linear GIR model. We first provide a simple identification result, analogous to the Ordinary Least Squares (OLS) case, of the parameter of interest in terms of moments of the joint distribution of Y and X. Second, we develop a nonparametric identification result in the case of model misspecification, with an interpretation in terms of a best linear approximation that is analogous to the Quantile Regression (QR) case in Angrist et al. (2006). Third, we derive a simple Omitted Variable Bias (OVB) formula.

Assumption 2. *(Linearity) The parametric model for the conditional Gini is linear in parameters, i.e.,* $m(x;\theta) = x'\theta$*.*

Note that a linear and correctly specified model would satisfy $x'\theta_0 \in [0,1)\ \forall x \in \mathscr{X}$. This is also the case in the classical (correctly specified) Linear Probability Model (LPM). We also note that if, for example, the outcome variable Y is taken to be income, a linear GIR model still allows for nonlinearities in the income process. An important case of Assumptions 1 and 2 being true is analogous to the cases of conditional mean and conditional quantiles. This is the case of mutually exclusive and exhaustive categories. It is well known that in the presence of a finite number of mutually exclusive categories or groups, a linear model for the Conditional Expectation Function (CEF) or the Conditional Quantile Function (CQF) that includes all categories, e.g., as binary variables, is correctly specified, and that each

coefficient is the population mean or quantile of that particular group[5] (saturated model).

This fact is important since it is standard practice in applied work to compute empirical Ginis by groups, e.g., in each year that a particular survey is conducted, city, region, country, or by gender, age, race, birth cohort, percentile earners, etc; and possibly make inference on the differences between them. See, for example, Bhattacharya (2007) testing for region and year Gini-differences in India using an inference method for multistage survey data; Kopczuk et al. (2010) for a Gini-based study of earnings inequality in the US; Krueger and Perri (2006) which examines Gini indices by years to identify trends in inequality for posterior structural modeling; or Berkes and Gaetani (2023) which computes Gini indices to measure economic segregation by city in the US. Other examples are Bowlus and Robin (2004), which computes Ginis by educational levels, years of experience, etc., at different points in time to study the evolution of inequality in expected and realized lifetime earnings; or Gottschalk and Smeeding (1997), which computes Gini indices by countries and addresses potential differences.

This procedure makes it difficult to account for within-group differences across the comparison groups, particularly in the presence of continuous covariates. The linear Gini Index Regression model provides an alternative to estimate and conduct inference on group Ginis, which may become particularly advantageous when the researcher wants to control for within-group differences in characteristics when making comparisons that involve continuous covariates, without having to discretize its support.

An important example to conclude the discussion of the linearity assumption, illustrating the previous points, is the following.

Example 3.1. *(Gini Treatment Effects in a Randomized Control Trial). Consider the potential outcomes framework with binary treatment and continuous response* $Y = Y(0) + (Y(1) - Y(0)) \cdot D$*, where* $Y(d) \geq 0$ *a.s. denotes the potential outcome of individual* i *under the treatment value* $D = d \in \{0,1\}$*. Define the Gini Treatment Effect (GTE) by the difference between the Gini of the distribution of* $Y(1)$ *and that of* $Y(0)$*, i.e.,* $\tau = \mathbb{G}(Y(1)) - \mathbb{G}(Y(0))$*. In the case of a successful treatment randomization or Randomized Control Trial (RCT),* $D \perp (Y(0), Y(1))$*, we can identify –as we do with the Average Treatment Effect (ATE) and Quantile Treatment Effects (QTEs) using OLS and QR, respectively– the GTE using the cor-*

[5] Alternatively, if one adds an intercept and excludes one category, then the coefficients represent the deviation with respect to the excluded or baseline category.

rectly specified linear Gini Index Regression model $m(D;\theta_0) = \theta_{01} + \theta_{02}D$*, since*

$$\begin{aligned} Y(2F(Y|D)-1) &= Y(\theta_{01}+\theta_{02}D) + U \\ &\implies \mathbb{E}[Y(2F(Y|D)-1)|D=d] = \mathbb{E}[Y(\theta_{01}+\theta_{02}D)+U|D=d] \\ &\implies \mathbb{E}[Y(d)(2F(Y(d)|D=d)-1)|D=d] = \mathbb{E}[Y(d)|D=d](\theta_{01}+\theta_{02}d) \\ &\implies \mathbb{G}(Y(d)|D=d) = \theta_{01}+\theta_{02}d = \mathbb{G}(Y|D=d) \end{aligned}$$

Under the independence assumption we can write

$$\tau = \mathbb{G}(Y(1)) - \mathbb{G}(Y(0)) = \mathbb{G}(Y|D=1) - \mathbb{G}(Y|D=0) = \theta_{02}$$

□

3.1 Identification

We show how, under a standard rank assumption, we can uniquely write the parameter of interest as a function of moments of the joint distribution of Y and X in a linear projection fashion. First note that under Assumptions 1 and 2 we can write equation (6) using only one orthogonality condition as follows

$$Y(2M(Y|X)-1) = YX'\theta_0 + U \ , \ \mathbb{E}[XU] = 0 \tag{7}$$

Consider, similar to the OLS case, the following identifying assumption.

Assumption 3. *(Identification)* $rank\left(\mathbb{E}\left[YXX'\right]\right) = p$.

In words, the finite matrix $\mathbb{E}[YXX']$ exists and is nonsingular. If $Y > 0$ *a.s.*, one can easily show that $rank\left(\mathbb{E}[XX']\right) = p$ if and only if $rank\left(\mathbb{E}[YXX']\right) = p$, so the assumption is neither stronger nor weaker than the standard rank assumption in linear regression[6], given moment existence. We are then ready to state the first result of this section.

Proposition 2. *(Identification in the Linear GIR Model) Let Assumptions 1-3 hold,* $\mathbb{E}\|X\|^2 < \infty$*, and* $\mathbb{E}\left[Y^2\right] < \infty$*. Then, we can identify* θ_0 *defined in (7) by*

$$\theta_0 = \mathbb{E}\left[YXX'\right]^{-1}\mathbb{E}[XY(2M(Y|X)-1)]$$

Proof. As in the OLS case, pre-multiplying by X both sides of equation (7), taking expectations, applying the moment condition in equation (7), and premultiplying by the inverse of

[6] However, if $Y \geq 0$ *a.s.* then one can guarantee $rank\left(\mathbb{E}[YXX']\right) = p \implies rank\left(\mathbb{E}[XX']\right) = p$, but not conversely in general.

the matrix in Assumption 3 yields the above result. The existence of moments is guaranteed by Assumption 3 and the Cauchy-Schwarz inequality.

□

Note that by Proposition 2, the parameter of interest can be written as the coefficient of the linear projection of $\sqrt{Y}(2M(Y|X)-1)$ on $X\sqrt{Y}$. Alternatively, it is the coefficient of the weighted linear projection of the conditional rank advantage $2M(Y|X)-1$ on covariates, with projection weights $w(Y) = Y/\mathbb{E}[Y]$. However, as for the LPM, it is difficult to expect in applications that under the linearity assumption, the fitted values will remain inside the unit interval or that marginal effects will be constant. This raises questions about the correct specification of the linear model. Our next result addresses this concern.

3.2 Identification under Model Misspecification

Suppose that a researcher works under Assumption 2 in the sense that she specifies a linear (in parameters) model for the conditional Gini, but the true model is some unknown, potentially nonlinear function. Hence, Assumption 1 may fail. That is, the researcher incorrectly specifies the linear GIR model

$$Y(2M(Y|X)-1) = YX'\gamma_0 + V \ , \ \mathbb{E}[XV] = 0 \tag{8}$$

where γ_0 is the researcher's estimand or pseudo-true parameter. Under the same rank and moment conditions as in Proposition 2, the researcher estimates γ_0 as if it were θ_0, i.e., we can express this pseudo-true parameter as

$$\gamma_0 = \mathbb{E}\left[YXX'\right]^{-1}\mathbb{E}\left[XY(2M(Y|X)-1)\right] \tag{9}$$

The natural question to ask is what is the researcher actually estimating under misspecification. Denote the CDF of X by $F_X(x) := P_X((-\infty, x])$. Theorem 1 provides a nonparametric identification result.

Theorem 1. *(Identification in the Linear GIR Model under Misspecification) Let Assumption 3 hold,* $0 < \mathbb{E}[Y|X] < \infty$ *a.s.,* $0 < \mathbb{E}[Y]$, $\mathbb{E}\|X\|^2 < \infty$, *and* $\mathbb{E}\left[Y^2\right] < \infty$. *Moreover, let* γ_0 *be defined as in equation (8). Then,* γ_0 *is the best linear approximation to the Conditional Gini Function* $\mathbb{G}(Y|X)$ *in a weighted least squares sense, i.e., it is the solution to the weighted least squares problem*

$$\gamma_0 \ = \ \underset{\gamma \in \mathbb{R}^p}{\operatorname{argmin}} \ \mathbb{E}\left[\left(\mathbb{G}(Y|X) - X'\gamma\right)^2 \omega(X)\right]$$

where the weights are given by $\omega(X) = \frac{\mathbb{E}[Y|X]}{\mathbb{E}[Y]} \geq 0$ *and satisfy* $\int_{\mathscr{X}} \omega(x)\, dF_X(x) = 1$.

Proof. By definition,

$$\begin{aligned}\gamma_0 &= \mathbb{E}\left[YXX'\right]^{-1}\mathbb{E}\left[XY\left(2M(Y|X)-1\right)\right] = \mathbb{E}\left[XX'\mathbb{E}[Y|X]\right]^{-1}\mathbb{E}\left[X\mathbb{E}\left[Y\left(2M(Y|X)-1\right)|X\right]\right] \\ &= \mathbb{E}\left[XX'\mathbb{E}[Y|X]\right]^{-1}\mathbb{E}\left[X\mathbb{E}[Y|X]\mathbb{G}(Y|X)\right] = \mathbb{E}\left[XX'\omega(X)\right]^{-1}\mathbb{E}\left[X\mathbb{G}(Y|X)\omega(X)\right] \\ &= \underset{\gamma\in\mathbb{R}^p}{\operatorname{argmin}}\ \mathbb{E}\left[\left(\mathbb{G}(Y|X)-X'\gamma\right)^2\omega(X)\right]\end{aligned}$$

where we used the law of iterated expectations and the definition of CGF.

□

The result establishes that the linear GIR estimand minimizes the weighted mean squared specification error[7]. The result is similar to the misspecification result of Angrist et al. (2006) for QR, where the QR estimand is the best linear approximation to the conditional quantile function in a weighted least squares sense[8]. In this case, $\omega(X)$ assigns more weight to the specification errors corresponding to a given $X = x$ such that $\mathbb{E}[Y|X = x]$ is large compared to the unconditional mean $\mathbb{E}[Y]$, e.g., if we think of Y as income, specification errors are penalized more for groups of richer individuals. Using the data from the empirical application in Section 7 (CPS data for the 1979 and 1988 waves), we illustrate that a potentially misspecified linear GIR model (using just a constant and years of experience as regressors) may tend to predict the Gini more accurately for groups of individuals with higher wages, and less accurately for poorer groups. See Figure 1 below.

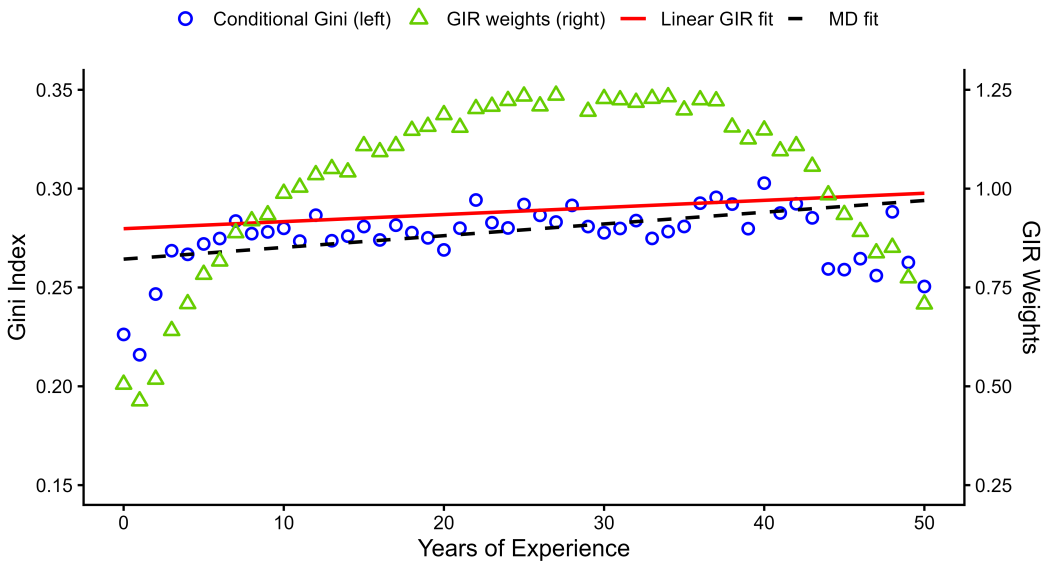
In Figure 1, the blue circles correspond to the nonparametric estimate of the conditional Gini function –wage Gini by experience group–, the red line is the estimated linear GIR model fit, and the green triangles represent the misspecification weights –also nonparametrically estimated. Note that GIR predictions tend to be more accurate for groups with higher weight, i.e., higher group-wage mean relative to the mean wage level. Finally, following Angrist et al. (2006), the dashed black line corresponds to the Minimum Distance (MD) estimator solving the empirical version of (10), which does not weight specification errors.

$$\lambda_0 = \underset{\lambda\in\mathbb{R}^p}{\operatorname{argmin}}\ \mathbb{E}\left[\left(\mathbb{G}(Y|X)-X'\lambda\right)^2\right] \tag{10}$$

[7] This could also be seen as a mean squared specification error under the distribution $\tilde{F}_X(x)$, i.e., $\mathbb{E}_{\tilde{F}}\left[\left(\mathbb{G}(Y|X)-X'\gamma\right)^2\right]$, where $\omega(x)$ is the Radon-Nikodym derivative of $\tilde{P}_X$ with respect to F_X.

[8] See theorems 1 and 2 in Angrist et al. (2006).

Figure 1: Misspecified GIR Weighting Scheme



Theorem 1 allows us to write

$$\mathbb{G}(Y|X) = X'\gamma_0 + g(Y|X), \quad \text{where} \quad \mathbb{E}[Xg(Y|X)\omega(X)] = 0$$

and $g(Y|X)$ can be seen as the errors created by a weighted linear projection of $\mathbb{G}(Y|X)$ on X. This result ensures that the researcher using a (misspecified) linear GIR model is still estimating a best linear approximation to the true conditional Gini of Y given X, as with OLS or QR.

3.3 Omitted Variable Bias

There is, however, one caveat attached to the result of Theorem 1. Unlike in OLS or Angrist et al. (2006) for QR, the above misspecification result cannot be used to derive an Omitted Variable Bias (OVB) formula. We show that omitting or adding variables changes the estimand through two channels: the familiar projection channel, present in OLS and QR, and a change in the conditional population used to rank outcomes; so that for inequality, the choice of controls does more than determine the regression specification: it determines how the outcomes are compared.

Partition the covariate vector as $X = (X_1', X_2')'$. Suppose that the true CGF is of the form

$$\mathbb{G}(Y|X) = X_1'\theta_{01} + R(X) \tag{11}$$

which can be characterized by

$$Y\left(2M(Y|X)-1\right) = Y\left(X_1'\theta_{01} + R(X)\right) + U \;,\; \mathbb{E}[U|X] = 0 \tag{12}$$

where $R : \mathscr{X} \mapsto \mathbb{R}$ is some possibly nonlinear function of X, i.e., the part of the CGF not explained by the linear function of X_1, and θ_{01} is the parameter of interest. However, the researcher omits $R(X)$ in the estimation step, either because X_2 or $R(X)$ are unobserved, unknown, or simply omitted[9]. By Proposition 2 and Theorem 1 we know that the linear model $X_1'\gamma_{01}$ provides the best linear approximation to $\mathbb{G}(Y|X_1)$ in a WLS sense, i.e.,

$$\gamma_{01} = \mathbb{E}\left[Y X_1 X_1'\right]^{-1}\mathbb{E}\left[X_1 Y\left(2M(Y|X_1)-1\right)\right] = \mathbb{E}\left[X_1 X_1'\omega(X_1)\right]^{-1}\mathbb{E}\left[X_1\mathbb{G}(Y|X_1)\omega(X_1)\right]$$

However, it is not the best linear approximation to $\mathbb{G}(Y|X)$, only to $\mathbb{G}(Y|X_1)$. To see this, consider the following two parameters

$$\delta_{01} \;=\; \underset{\delta\,\in\,\mathbb{R}^{p_1}}{\arg\min}\; \mathbb{E}\left[\left(\mathbb{G}(Y|X) - X_1'\delta\right)^2\omega(X_1)\right] \quad \text{and} \quad \alpha_{01} \;=\; \underset{\alpha\,\in\,\mathbb{R}^{p_1}}{\arg\min}\; \mathbb{E}\left[\left(\mathbb{G}(Y|X) - X_1'\alpha\right)^2\omega(X)\right]$$

These two parameters are unfeasible to estimate if the researcher does not observe X_2 because only X_1 is available. This is because estimation of either δ_{01} or α_{01} would require a preliminary estimate of the conditional mid-distribution function of Y given both X_1 and X_2. In particular, $X_1'\alpha_{01}$ would be the best linear approximation to $\mathbb{G}(Y|X_1, X_2)$ in a WLS sense.

Hence, the key conceptual point is that the "short" regression estimand is tied to the CGF based on the reduced information set, whereas the "long" regression is formulated in terms of the CGF based on the full information set. As in OLS, and under the weighted-approximation representation of Angrist et al. (2006) for QR, a short-regression coefficient can be represented as a projection of the full-information conditional mean or conditional quantile function onto the retained set of regressors. GIR has an additional feature: the short-regression moment itself is constructed from the reduced-information conditional mid-rank $M(Y|X_1)$. Rewriting this moment in terms of the full-information, conditional mid-rank $M(Y|X_1, X_2)$ generates the second term in our next result.

Proposition 3. *(Omitted Variable Bias in the Linear GIR Model) Let the CGF be as defined in equations (11) and (12), suppose Assumption 3 holds, let* $\mathbb{E}[Y^2] < \infty$, $\mathbb{E}\|X\|^2 < \infty$, *and*

[9] Measurement error in the explanatory variables can be treated as an omitted variable.

$\mathbb{E}[\|Y X_1 R(X)\|] < \infty$. *The omitted variable bias incurred by omitting* $R(X)$ *is*

$$\gamma_{01} - \theta_{01} = \mathbb{E}\left[Y X_1 X_1'\right]^{-1} \mathbb{E}\left[Y X_1 R(X)\right] + 2\mathbb{E}\left[Y X_1 X_1'\right]^{-1} \mathbb{E}\left[Y X_1 \left(M(Y|X_1) - M(Y|X)\right)\right]$$

Furthermore, if $R(X) = X_2'\theta_{02}$ *then the expression simplifies to*

$$\gamma_{01} - \theta_{01} = \mathbb{E}\left[Y X_1 X_1'\right]^{-1} \mathbb{E}\left[Y X_1 X_2'\right] \theta_{02} + 2\mathbb{E}\left[Y X_1 X_1'\right]^{-1} \mathbb{E}\left[Y X_1 \left(M(Y|X_1) - M(Y|X)\right)\right]$$

Proof. See Appendix C.

□

Proposition 3 shows two sources of bias due to a lack of information. The first is due to the misspecification of the functional form of the CGF and is analogous to the OLS and QR cases. The second source of bias is due to the lack of information about X_2, which makes it impossible to make $M(Y|X)$ estimable without further assumptions. Suppose that X_2 is empty or that $Y \perp X_2 \mid X_1$, so that $M(Y|X) = M(Y|X_1)$. Then the second bias term disappears. The only source of bias would thus be due to some nonlinearity in X_1 if X_2 is empty –in which case γ_0 can be obtained by the weighted MSE minimization problem in Theorem 1 and is the best linear approximation to the CGF in a WLS sense– or to some nonlinearity in both X_1 and X_2 if X_2 is nonempty.

After discussing the lack of information case, one may ask what happens when the researcher adds extra regressors that are in fact not in the conditional Gini index of interest, even though the extra regressor is not statistically independent of Y (given X_1). Formally,

(i) $M(Y|X_1, X_2) \neq M(Y|X_1)$

(ii) $\mathbb{G}(Y|X_1, X_2) = X_1'\theta_{01}$

Note that the "irrelevant regressor" condition (ii) does not necessarily imply independence nor conditional independence: in many examples given in the previous section, the Gini depends only on one of the parameters of the distribution so that one of the conditioning variables may only affect the parameter of the conditional distribution that does not pin down its Gini. Suppose that conditions (i) and (ii) hold and the researcher estimates a linear model for the conditional Gini index $X'\gamma_0 = X_1'\gamma_{01} + X_2'\gamma_{02}$ so that

$$\gamma_0 = \mathbb{E}\left[Y X X'\right]^{-1} \mathbb{E}\left[X Y \left(2M(Y|X) - 1\right)\right]$$

By Theorem 1, we know that γ_0 provides the best linear approximation in a WLS sense.

Ideally, extra information is not a problem, i.e., $\gamma_{01} = \theta_{01}$ and $\gamma_{02} = 0$. The following lemma on irrelevant regressors verifies this intuition.

Lemma 1. *(Irrelevant Regressors) Let γ_0 be defined as in (8) and consider the partitions $X = (X_1', X_2')'$ and $\gamma_0 = (\gamma_{01}', \gamma_{02}')'$. Let Assumption 3, $\mathbb{E}[Y^2] < \infty$, $\mathbb{E}\|X\|^2 < \infty$, and the irrelevant regressor condition (ii) hold. Then, $\gamma_{01} = \theta_{01}$ and $\gamma_{02} = 0$.*

Proof. See Appendix C.

□

4 Estimation and Inference Theory for the Linear GIR Model

In this section, we discuss the estimation of the parameter of interest in the linear GIR model in (7). We propose a two-step feasible GIR estimator based on the plug-in principle and establish its asymptotic properties, along with a bootstrap validity result to facilitate inference. Assume that we observe a sample $\{Y_i, X_i\}_{i=1}^n$. A feasible GIR estimator is given by the following statistic

$$\hat{\theta}_n = \left(\frac{1}{n} \sum_{i=1}^n X_i X_i' Y_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i Y_i \left(2\hat{M}_n(Y_i | X_i) - 1 \right) \right) \tag{13}$$

where $\hat{M}_n(y|x)$ denotes an estimator of the conditional mid-distribution function at (y, x).

4.1 First Stage Estimation

The main complication is that a preliminary estimate $\hat{M}_n(y_i|x_i)$ for the conditional mid-distribution function – henceforth, CMDF – of Y given X is needed at each pair of observation values (y_i, x_i) in the sample. One of the main advantages of GIR is that it allows the researcher to remain somewhat agnostic about the conditional distribution generating the data –up to the specification of its conditional Gini. If the conditional distribution is fully specified, the parameters of the conditional distribution could be estimated by maximum or quasi-maximum likelihood methods, and then the Gini of the conditional distribution could be computed based on these estimates for a fixed value of covariates. $M(\cdot|\cdot)$ should therefore be treated as an infinite-dimensional nuisance to be estimated flexibly; otherwise the parametric modeling of the conditional Gini alone does not offer a robustness gain, rendering the proposed approach unjustified. Hence, an ideal approach would use a fully nonparametric first-stage. We thus propose a nonparametric first-stage series-based estimator. We refer to this estimator as the GIR estimator.

Least Squares Series Approach. Sieve estimators are common in econometric applications when the exact form of a conditional expectation is unknown and must be flexibly estimated (see Chen (2007)). Series methods[10] are very easy to implement and approximate the unknown function based on K basis functions, where $K := K(n)$ is chosen as a deterministic function of n and allowed to grow with the sample size to balance the bias-variance trade-off. Large sample properties of series estimators are well-known (see Andrews (1991); Newey (1997); Belloni et al. (2015); Chen and Christensen (2015), etc.), and Newey (1994b) shows $\sqrt{n}$-consistency and asymptotic normality of two-step estimators with a nonparametric first-stage. A series estimator approximates the target function $M := M(y|x)$ by linear forms $x \mapsto M_K(y|x) := q_K(x)'\beta_K(y)$ (surrogate function) at each y, where $x \mapsto q_K(x) := (q_{K1}(x), \ldots, q_{KK}(x))'$ is a vector of approximating functions that can change with n, e.g., polynomial series, splines, wavelets, etc. The method is particularly attractive in the presence of continuous regressors, since no discretization of the support of X is needed[11] and it directly approximates M. Finally, in the case of mutually exclusive dummies as regressors, the choice $q_K(x) = x$ yields a correctly specified linear model at each $y \in \mathscr{Y}$, so no approximation is needed and the GIR estimator coincides with the sample Gini of each group, defined by the corresponding dummy (see Lemma 2).

First Stage Estimation. We now give the expression for the first-stage conditional mid-distribution estimator based on Least-Squares (LS) Series. The target function is a conditional expectation function: $M(y|x) = \mathbb{E}[\mathbb{1}_{\{Y<y\}} + 0.5\mathbb{1}_{\{Y=y\}}|X = x]$, $y \in \mathscr{Y}$. We treat the first stage as a continuum of regressions indexed by $y \in \mathscr{Y}$. Define $\mathscr{Y}\mathscr{X} := \{(y,x) : y \in \mathscr{Y}, x \in \mathscr{X}\}$ and $\psi_y(t) := \mathbb{1}_{\{t<y\}} + 0.5\mathbb{1}_{\{t=y\}}$. The series-based estimator uses the standard OLS estimator,

$$\hat{\beta}_K(y) \;=\; \underset{b \in \mathbb{R}^K}{\arg\min} \sum_{i=1}^{n} \left(\psi_y(y_i) - q_K(x_i)'b\right)^2 , \quad y \in \mathscr{Y} \tag{14}$$

which is then used to compute the CMDF estimator by

$$\hat{M}_n(y|x) = q_K(x)'\hat{\beta}_K(y) , \quad (y,x) \in \mathscr{Y}\mathscr{X} \tag{15}$$

Note that since K is a deterministic function of n, we index the series-based CMDF estimator by n, and use K to denote the dimension of both $q(x)$ and its coefficient vector.

[10] Series estimators are defined as sieve M-estimators with concave criterion functions and finite-dimensional linear sieve spaces (Chen (2007)).

[11] Unlike in the estimation of the Gini by groups, where one has to discretize the support of X to define the group. GIR allows estimation at a particular point $X = x$ of the continuous support (no discretization).

4.2 GIR Estimation

Once the CMDF has been estimated, fitted values can be computed for each individual in the sample, and the GIR can be easily obtained, since it is just a linear projection coefficient. The following simple algorithm summarizes how the GIR estimator can be computed (the given basis dimension choice is formally justified in the next section).

Algorithm 1. *(GIR Estimator) (i) Obtain the OLS-Series estimates $\hat{\beta}_K(y)$ for all $y \in \{y_1, \ldots, y_n\}$ using (14) for a choice of $q_K(\cdot)$, e.g., splines, and a choice of K, e.g., $K(n) = \lfloor n^{0.3} \rfloor$. (ii) Obtain estimates $\hat{M}_n(y_i|x_i)$ of the CMDF for all $(y,x) \in \{(y_1,x_1),\ldots,(y_n,x_n)\}$ using (15). (iii) Obtain the GIR estimator $\hat{\theta}_n$ via an OLS regression of $\sqrt{y_i}\left(2\hat{M}_n(y_i|x_i) - 1\right)$ on $\sqrt{y_i}x_i$.*

4.3 Numerical Equivalence and Scale Invariance of the GIR Estimator

In this section, we establish two important properties of the linear GIR estimator. First, we show that, for a sample $\{Y_i, X_i\}_{i=1}^n$, each coefficient of the linear GIR estimator of Y on a $p \times 1$ vector of mutually exclusive and exhaustive dummies X, without intercept, and no first-stage basis transformations, $q_K(x) = x$ (so $K = p$), is numerically equivalent to the empirical-distribution, or plug-in, sample Gini of the corresponding group of Y defined by each dummy. This is an exact finite-sample identity. Analogous finite-sample identities also hold for OLS and QR with the same saturated group specification: their coefficients are subgroup sample means and quantiles, respectively. Second, reflecting the scale-invariance property of the Gini, that is, $\mathbb{G}(cY) = \mathbb{G}(Y)$ for every $c > 0$ (which also holds in the conditional case), we show that the proposed linear GIR estimator (13)-(15) is scale invariant with respect to Y.

For the first result, let p be the number of mutually exclusive and exhaustive groups and let $D_i \in \{1,\ldots,p\}$ be the group identifier of observation $i \in \{1,\ldots,n\}$. Define the $p \times 1$ vector of group-indicator regressors $X_i := \left(X_{i1},\ldots,X_{ip}\right)'$, where $X_{ig} = \mathbb{1}_{\{D_i=g\}}$, $g \in \{1,\ldots,p\}$. Then $\sum_{i=1}^n X_{ig} = n_g$ for all $g \in \{1,\ldots,p\}$, $\sum_{g=1}^p X_{ig} = 1$, and $X_{ig}X_{ih} = 0$ for all $h \neq g$. Moreover, for group g, define the set $I_g := \{i : D_i = g\}$; the sample size of the group $n_g := \text{card}(I_g)$, where $\text{card}(A)$ denotes the cardinality of set A; and the group sample average $\bar{Y}_g := S_g / n_g$, where $S_g := \sum_{i \in I_g} Y_i$.

Lemma 2. *(Numerical Equivalence) Let $Y_i \geq 0$ $\forall i \in \{1,\ldots,n\}$, $n_g > 0$, and $\bar{Y}_g > 0$ $\forall g \in \{1,\ldots,p\}$. The linear GIR estimator $\hat{\theta}_n$ defined by (13), (14), and (15) using mutually exclusive*

group-indicators X_{ig} as regressors without intercept, and $q_K(x) = x$ satisfies, for $g \in \{1, \ldots, p\}$

$$\hat{\theta}_{ng} = \frac{\sum_{i \in I_g} \sum_{j \in I_g} |Y_i - Y_j|}{2n_g^2 \bar{Y}_g}$$

Thus, each coefficient of the GIR estimator is numerically equivalent to the corresponding plug-in sample Gini of group g.

Proof. See Appendix C.

□

The typical finite-sample correction can be easily implemented by multiplying each GIR estimator coefficient by $n_g/(n_g - 1)$, $n_g > 1$. Moreover, since the group indicators span the function of interest on the finite group support, no basis transformations q_K are needed; it is sufficient to regress $\psi_y(Y_i)$ on the vector of group dummies in the first stage. Since the expression on the right-hand side is equivalent to the order-statistic version of the sample Gini, equivalence to the latter expression also follows.

The following result regarding the scale-invariance property of the GIR estimator with respect to the response variable holds generally, irrespective of the nature of the regressors, and reflects the scale-invariance property of the index, $\mathbb{G}(cY|X = x) = \mathbb{G}(Y|X = x)$ for every $c > 0$ and all $x \in \mathcal{X}$ for which the conditional Gini is defined.

Lemma 3. *(Scale Invariance of the GIR estimator) Let $c > 0$ and define $\tilde{Y}_i := cY_i$, $\forall i \in \{1, \ldots, n\}$. Let $\tilde{\theta}_n$ be the GIR estimator of $\tilde{Y}_i$ on X_i, and let $\hat{\theta}_n$ be the GIR estimator of Y_i on X_i as defined by (13), (14), and (15), both using the same choice of q_K. Suppose as well that the matrices inverted in both stages are invertible. Then, $\tilde{\theta}_n = \hat{\theta}_n$, that is, the GIR estimator is scale-invariant with respect to the response variable, Y_i.*

Proof. See Appendix C.

□

4.4 Inference Theory

This section contains the main results regarding the asymptotic properties of the GIR estimator and the validity of the bootstrap, which rely on uniform series asymptotics. We first state some baseline GIR assumptions.

Assumption 4. *(Baseline GIR Conditions)*

(i) (Sampling) We observe an i.i.d. sample $\{Y_i, X_i\}_{i=1}^n$.

(ii) (Identification) $rank\big(\mathbb{E}\left[XX'Y\right]\big) = p$.

These are just standard sampling and identification conditions. Before stating the series-specific assumptions, we introduce some new definitions and notation. Let $\mathbb{R}_+ := [0,\infty)$, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, and $\Lambda^s(\mathscr{X})$ denote the class of real-valued functions $h : \mathscr{X} \to \mathbb{R}$ that are r-times continuously differentiable and satisfy a Hölder s-smoothness condition, endowed with the norm $\|h\|_{\Lambda^s}$ (see Appendix D for the exact definitions). Finally, define the Gram matrix $Q_K := \mathbb{E}[q_K(X)q_K(X)']$, the mean-square projection operator $(P_K h)(x) = q_K(x)'Q_K^{-1}\mathbb{E}[q_K(X)h(X)]$ and the operator norm $\|P_K\|_{\infty\to\infty} := \sup_{\|h\|_\infty \le 1} \|P_K h\|_\infty$. We use the notation $b_{1n} \lesssim b_{2n}$ to denote $b_{1n} \le c b_{2n}$ for some $c \in \mathbb{R}_+$ that does not depend on n, and two sequences b_{1n} and b_{2n}. For simplicity of exposition, the assumptions are stated for continuous, untransformed regressors, all used as arguments of the first-stage basis. With a fixed finite number of discrete covariate cells of positive probability, the analogous extension proceeds by applying the sieve approximation separately within each cell, under the corresponding within-cell assumptions.

Assumption 5. *(LS-Series Conditions)*

(i) (Density and Support) $\mathscr{Y}$ *is either a compact interval in* $\mathbb{R}_+$ *or a finite subset of* $\mathbb{R}_+$. *In the former case, the conditional density function exists and is uniformly bounded.* $\mathscr{X} = \prod_{j=1}^p [a_j, b_j]$, $a_j < b_j$. *The region* $\mathscr{Y}\mathscr{X}$ *is a compact subset of* $\mathbb{R}^{p+1}$.

(ii) (Smoothness) $\exists\, r \in \mathbb{N}_0$ *and* $\gamma \in (0,1]$ *such that* $\sup_{y\in\mathscr{Y}} \|f_{Y|X}(y|\cdot)\|_{\Lambda^s} < \infty$ *for* $s = r + \gamma$, *where* $f_{Y|X}(y|x)$ *denotes the conditional density or probability mass function.*

(iii) (Basis) Let $q_K(x) \in \mathbb{R}^K$ *be a tensor-product*[12] *basis from either: Legendre polynomial series, spline/B-spline series of order* s_0 *(degree* $s_0 - 1$*) with equally-spaced knots, or (CDV) wavelet series of order* s_0. *The eigenvalues of the Gram matrix* Q_K *are bounded above and away from zero uniformly over* K.

(iv) (Rates) Let $\bar{s} := s$ *for Legendre polynomial series and* $\bar{s} := \min\{s, s_0\}$ *for spline/B-spline or (CDV) wavelet series. Let* $l_K := 1 + \|P_K\|_{\infty\to\infty}$ *and* $\xi_K := \sup_{x\in\mathscr{X}} \|q_K(x)\|$. *As* $n \to \infty$ *we have that*

$$(a)\, K\xi_K\sqrt{\frac{\log K}{n}} \longrightarrow 0,\ \ (b)\, l_K K^{-\bar{s}/p} \longrightarrow 0,\ \ \textit{and}\ \ (c)\, \sqrt{n}K^{-2\bar{s}/p} \longrightarrow 0.$$

[12] We define tensor products as in Example 3.6 of Belloni et al. (2015): if the basic regressors are $X_i = \left(X_{1i}, \ldots, X_{pi}\right)'$, we can create p series for each and take all interactions of functions from these p series and collect them into q_K. If each series for a basic regressor has J terms, q_K has dimension $K = J^p$.

Assumption 5-(i) allows for a continuous or discrete response variable and imposes a boundedness condition on the conditional density. Assumption 5-(ii) is a Hölder s-smoothness condition on the conditional density/mass, standard in the sieve literature so that the function of interest can be well approximated by linear sieves, with the additional requirement that it holds uniformly over $y \in \mathscr{Y}$, so the family of maps $\{x \mapsto M(y|x) : y \in \mathscr{Y}\}$ lies in a common Hölder ball. It is natural in the CMDF setup and is analogous to the 'uniform over a class of functions' viewpoint in Belloni et al. (2015). The assumption enables the approximation constants to be uniform over $y \in \mathscr{Y}$. If X contains an intercept or fixed smooth transformations in addition to the underlying basic covariates, the rectangular-support and covariate-smoothness conditions are understood with respect to the basic, untransformed, covariates over which the tensor-product basis is constructed. For example, if $X = \left(C, C^2\right)'$, the first-stage basis is constructed as a function of C, and the support and smoothness conditions apply to C, not to $\left(C, C^2\right)'$. The stability of eigenvalues in assumption 5-(iii) is also standard in the sieve literature: it imposes the restriction that $q_{1K}, \ldots, q_{KK}$ are not too co-linear, where $q_K(x) := (q_{1K}(x), \ldots, q_{KK}(x))' \in \mathbb{R}^K$. A simple sufficient condition is given in Proposition 2.1 in Belloni et al. (2015).

Finally, Assumption 5-(iv) gives the rate conditions for the growth of the basis vector dimension, K. For splines, B-splines and wavelets, $\xi_K \lesssim \sqrt{K}$, and for Legendre polynomials $\xi_K \lesssim K$ (see examples 3.1-3.6 in Belloni et al. (2015)). Condition (b) yields a supremum-norm bound (and therefore a mean-square norm bound) of the series approximation error for Hölder smooth classes: Propositions 3.1 and 3.2 in Belloni et al. (2015), with $c_K \lesssim K^{-\bar{s}/p}$ for smooth functions, establish that the so-called Lebesgue factor l_K satisfies $l_K = 1 + \|P_K\|_{\infty \to \infty} \lesssim 1 + \xi_K$. Sharper bounds from Approximation Theory can be used if we further assume a density for the regressors that is bounded above and away from zero: $l_K = O(1)$ for splines and wavelets. In general, $l_K = O\left(\sqrt{K}\right)$ for splines and wavelets, and $l_K = O(K)$ for Legendre polynomials. Condition (b), together with Assumption 5-(ii), delivers a uniform-approximation bias condition for the family $\{x \mapsto M(y|x) : y \in \mathscr{Y}\}$. The rate condition (c) is weaker than the condition $\sqrt{n}K^{-\bar{s}/p} \to 0$ in Newey (1997), required for the asymptotic normality of functionals of conditional expectations (either linear or nonlinear), and coincides with the comparable bias component of the $\sqrt{n}\delta_{gn}^2 \to 0$ condition in Newey (1994b) (see Theorems 9 and 10) for two-step series estimators (among other conditions). The rate conditions (a) and (c) make the stochastic and approximation remainders asymptotically negligible and yield an asymptotically linear representation of the GIR numerator that includes the first-stage estimation effect. Condition (a) is rather conservative, though not an outlier in the literature, e.g., for splines and wavelets,

where $\xi_K \lesssim \sqrt{K}$, it is stronger than Newey's $K^2/n \to 0$ requirement for linear functionals of conditional expectations, but weaker than $K^4/n \to 0$ for nonlinear functionals. For power series, where $\xi_K \lesssim K$, it is stronger than Newey's $K^3/n \to 0$ for linear functionals, but weaker than the $K^6/n \to 0$ required for nonlinear functionals.

We now give examples where the rate conditions (a)-(c) in Assumption 5-(iv) can be implemented for a tensor product basis of dimension K. If $\bar{\rho} := \bar{s}/p > 3/4$ then one can choose a sequence $K = K(n)$, e.g., $K \asymp n^{\kappa}$ with $\kappa \in (1/4\bar{\rho}, 1/3)$ for splines, B-splines and wavelet series[13] (assuming a bounded density for X, above and away from 0). For Legendre polynomials, rates $(a)-(c)$ can be simultaneously satisfied with $K \asymp n^{\kappa}$, $\kappa \in (1/4\bar{\rho}, 1/4)$ for a smoothness requirement of $\bar{\rho} > 1$.

Throughout this section, we denote the estimand of interest by

$$\theta_0 = \mathbb{E}\left[XX'Y\right]^{-1}\mathbb{E}\left[XY\left(2M(Y|X)-1\right)\right]$$

regardless of whether the GIR model is correctly specified or not, since asymptotic theory applies to both cases and Theorem 1 gives the estimand's interpretation under model misspecification[14]. Inference results are derived letting $K \to \infty$ as $n \to \infty$. Let $\rightsquigarrow$ denote weak convergence. We are now well-equipped to state the main results of this section.

Theorem 2. *(Asymptotic Normality of the GIR Estimator) Suppose that Assumptions 4 and 5 hold. Then as* $(n, K) \to \infty$

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \rightsquigarrow W$$

where W *is a zero-mean Gaussian random vector in* $\mathbb{R}^p$.

Proof. See Appendix D. □

The result shows that the GIR estimator satisfies a Central Limit Theorem (CLT). The limiting distribution W has a complicated variance-covariance matrix because of the non-parametric estimation effect of the first stage. Due to the complex form of the asymptotic distribution, we next provide a bootstrap validity result and a computationally convenient bootstrap algorithm to facilitate inference. In particular, we propose the use of the Bayesian bootstrap, which is a particular case of the weighted bootstrap.

[13] For tensor-product CDV wavelet series with resolution level J_n, typically $K \asymp 2^{pJ_n}$, so one must choose J_n such that $2^{pJ_n} \asymp n^{\kappa}$. For p-dimensional tensor-product splines of order s_0, choose knots J_n such that $K(n) = \prod_{l=1}^{p}(J_l + s_0) \asymp n^{\kappa}$ (see Chen (2007); Belloni et al. (2015)).

[14] Note that in the misspecified case, the estimand θ_0 used in this section coincides with the pseudo-true parameter γ_0 defined in Section 3.2.

We extend the relevant weighted-bootstrap linearization in Belloni et al. (2015) to cover the case of the CMDF, treated as a continuum of regressions indexed by $y \in \mathscr{Y}$ with response variable $\psi_y(Y_i)$, making the results uniform over $(y,x) \in \mathscr{Y}\mathscr{X}$, not only over $x \in \mathscr{X}$. For that purpose, we introduce the following convenient normalizations before stating the bootstrap-specific assumption: let $p_K(x) := Q_K^{-1/2} q_K(x)$ be the normalized basis vector, $\alpha_K(x) := p_K(x)/\|p_K(x)\|$, which is well-defined for our choices of basis, and

$$\xi_K^L := \sup_{x,x' \in \mathscr{X}: x \neq x'} \frac{\|\alpha_K(x) - \alpha_K(x')\|}{\|x - x'\|}$$

These normalizations simplify the proofs and are without loss of generality under the previous eigenvalue assumption on the Gram matrix. The bootstrap assumption is the following.

Assumption 6. *(Bayesian Bootstrap) Let $(w_1, \ldots, w_n)'$ be a vector of i.i.d. random variables independent of the data. For each n, define $\bar{w}_n = n^{-1}\sum_{i=1}^n w_i$ and $w_{ni} := w_i/\bar{w}_n$ for $i \in \{1, \ldots, n\}$, where*

(i) (Weights) $\{w_i\} \sim$ i.i.d. Exp(1).

(ii) (First Stage Remainder) For $r_{K,n}^b := \left[\frac{\xi_K^2 \log^3(en)}{n}\right]^{1/2} \{\sqrt{\log(en)} + \sqrt{K} l_K K^{-\bar{s}/p}\}$, we have that $\xi_K r_{K,n}^b \longrightarrow 0$ as $(K,n) \to \infty$. Also, $\log\left(1 + \xi_K^L\right) \lesssim \log(eK)$.

Assumption 6-(i) concerns the weighted bootstrap weights. This weighted bootstrap scheme with exponential weights is standard and is known as the Bayesian bootstrap (see example 3.7.9 in Van der Vaart and Wellner (2023)). Assumption (ii) controls the stochastic and approximation remainders in the first-stage bootstrap linearization, uniformly over $(y,x) \in \mathscr{Y}\mathscr{X}$. Its form is based on the weighted least-squares series argument of Belloni et al. (2015) to establish bootstrap validity, including the logarithmic inflation generated by exponential weights. Their uniform linearization theory concerns a single conditional-mean regression and uniformity over its evaluation point x. It does not directly provide a joint expansion for the family of regressions indexed by $y \in \mathscr{Y}$. Propositions 4 and 5 in Appendix E establish the required uniform-in-$(y,x) \in \mathscr{Y}\mathscr{X}$ extension for the CMDF first-stage.

Regarding the choice of basis dimension $K := K(n)$ that satisfies both Assumption 5-(iv) and 6-(ii), one can show that, for example, the aforementioned polynomial series dimension $K(n) \asymp n^\kappa$ still satisfies the rate conditions required for inference. If $\bar{\rho} := \bar{s}/p > 3/4$ then one can choose a sequence $K \asymp n^\kappa$ with $\kappa \in (1/4\bar{\rho}, 1/3)$ for splines, B-splines

and wavelet series, as discussed before. For Legendre polynomials, the previous given smoothness condition is slightly strengthened: if $7/6 < \bar{\rho} < 3/2$ one may take $K \asymp n^{\kappa}$ with $\kappa \in (1/4\bar{\rho}, 1/(7-2\bar{\rho}))$, and if $\bar{\rho} \geq 3/2$ then one may take $K \asymp n^{\kappa}$ with $\kappa \in (1/4\bar{\rho}, 1/4)$. Legendre polynomials now need $\bar{\rho} > 7/6$ instead of the previous $\bar{\rho} > 1$. The data-generating processes in the simulation study in Section 6 are sufficiently smooth, and the first-stage estimators use $K = \lfloor n^{0.3} \rfloor$ for splines and $K = \lfloor n^{0.245} \rfloor$ for Legendre polynomials.

We can easily obtain a bootstrap draw of the GIR estimator as

$$\hat{\theta}_n^* = \left(\frac{1}{n} \sum_{i=1}^{n} w_{ni} x_i x_i' y_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} w_{ni} x_i y_i \left(2\hat{M}_n^* \left(y_i | x_i \right) - 1 \right) \right) \tag{16}$$

where $\hat{M}_n^* \left(y_i | x_i \right)$ denotes the bootstrap draw of the CMDF estimator at (y_i, x_i). The bootstrap draw of the series-based CMDF estimator can be obtained by

$$\hat{M}_n^*(y|x) \;=\; q_K(x)' \hat{\beta}_K^*(y) \tag{17}$$

where $\hat{\beta}_K^*(y)$ denotes the bootstrap draw of the series estimator obtained as the solution to the weighted least squares problem

$$\hat{\beta}_K^*(y) \;=\; \underset{b \in \mathbb{R}^K}{\operatorname{argmin}} \; \sum_{i=1}^{n} w_{ni} \left(\psi_y(y_i) - q_K(x_i)' b \right)^2 \tag{18}$$

We say that the bootstrap law of $T_n^* := \sqrt{n} \left(\hat{\theta}_n^* - \hat{\theta}_n \right)$ consistently estimates the law of the random element W and write $T_n^* \underset{P}{\rightsquigarrow} W$ if

$$\sup_{h \in BL_1(\mathbb{R}^p)} \left| \mathbb{E}_w \left[h(T_n^*) \right] - \mathbb{E}\left[h(W) \right] \right| \underset{P}{\longrightarrow} 0$$

where $BL_1(\mathbb{R}^p)$ denotes the space of functions on $\mathbb{R}^p$ with Lipschitz norm at most 1 and $\underset{P}{\longrightarrow}$ denotes convergence in probability (all the relevant concepts are formally defined in Appendix E). The next result establishes the validity of the proposed bootstrap scheme.

Theorem 3. *(Validity of the Bootstrap) Let Assumptions 4, 5, and 6 hold. Then, the Bayesian bootstrap consistently estimates the limit law W of the GIR estimator, that is, as $(n, K) \to \infty$*

$$\sqrt{n} \left(\hat{\theta}_n^* - \hat{\theta}_n \right) \underset{P}{\rightsquigarrow} W$$

Proof. See Appendix E.

□

The following simple algorithm summarizes how $(1-\alpha)$ confidence intervals can be computed using the Bayesian bootstrap[15].

Algorithm 2. *(Bayesian Bootstrap for the GIR Estimator)* (i) Draw independent realizations $\{e_i\}_{i=1}^n \sim Exp(1)$ and compute $w_{ni} = e_i/\bar{e}_n$. (ii) Obtain the bootstrap draws $\{\hat{M}_n^*(y_i|x_i)\}_{i=1}^n$ of the CMDF estimator $\{\hat{M}_n(y_i|x_i)\}_{i=1}^n$ by (17) and (18). (iii) Compute the bootstrap GIR estimator $\hat{\theta}_n^*$ by (16). (iv) Repeat steps (i)-(iii) B times and get $\{\hat{\theta}_{nb}^*\}_{b=1}^B$ to compute the $\frac{\alpha}{2}$ and $(1-\frac{\alpha}{2})$ sample quantiles of $\left\{\hat{\theta}_{nb,j}^* - \hat{\theta}_{n,j} : 1 \leq b \leq B\right\}$ for each coordinate $j \in \{1,\ldots,p\}$, denoted by $t_{\frac{\alpha}{2},j}^b$ and $t_{1-\frac{\alpha}{2},j}^b$, respectively, and where B is a large number. (v) Compute the $(1-\alpha)$ bootstrap confidence intervals as $\left(\hat{\theta}_{n,j} - t_{1-\frac{\alpha}{2},j}^b\,,\, \hat{\theta}_{n,j} - t_{\frac{\alpha}{2},j}^b\right)$.

5 Conditional Gini via Loss Minimization and Goodness of Fit for Gini Index Regression Models

It is well known that the conditional expectation function (CEF) minimizes the mean squared error (MSE) or expected quadratic loss and that the conditional quantile function (CQF) minimizes the expected check loss function[16]. In turn, these loss functions have been used for prediction evaluation and to construct goodness-of-fit measures. In this section, we show that the conditional Gini function (CGF) –regardless of its functional form, in contrast with the previous results involving a parametric linear model– is the solution to an expected loss minimization problem.

Intuitively, we show that given the characteristics defining a comparable population, the conditional Gini is the optimal outcome-weighted prediction of an individual's centered position within that population. We then use the loss function to propose an R^2-type goodness-of-fit measure for GIR models, similar in spirit to the ones used in OLS and QR.

5.1 A Loss Function for the Gini

We start by defining the loss function $L^G : \mathscr{Y}\mathscr{X}\mathscr{A} \to \mathbb{R}_+$ given by

$$L^G(y,x,a) = y\left((2M(y|x)-1)-a\right)^2 \tag{19}$$

[15] The usual first-order coverage interpretation concerns coordinates with nondegenerate limiting distributions.

[16] In fact, the CEF minimizes the expected value of the class of Bregman functions, and the CQF the expected value of the class of Generalized Piecewise Linear functions, see Savage (1971), Gneiting (2011a), and Gneiting (2011b).

where $a \in \mathscr{A} \subseteq \mathbb{R}$ and $L^G = 0$ if $a = (2M(y|x) - 1)$. We coin L^G the Conditional Gini Loss Function and show in Appendix F that the conditional Gini minimizes L^G in a conditional expected loss minimization problem for P_X-almost every x.

We are now ready to show that the CGF solves the unconditional expected loss minimization problem analogously to the CEF and CQF minimizing the expected quadratic and check loss functions, respectively. We do this by showing that the CGF is an optimal predictor of the conditional rank advantage $(2M(Y|X) - 1)$ in a weighted least squares sense with weights $w(Y) = Y/\mathbb{E}[Y]$.

Theorem 4. *(Conditional Gini Function via Expected Loss Minimization) Let $Y \geq 0$ a.s. be a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ with $0 < \mathbb{E}[Y] < \infty$. Let X be a measurable function from $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\Psi, \mathscr{G})$ and $0 < \mathbb{E}[Y|X] < \infty$ a.s. Let $h(X)$ be a predictor, i.e., $h \in \aleph$, where $\aleph$ is the set of Borel functions $h : \mathscr{X} \to \mathbb{R}$ such that $\mathbb{E}\left[w(Y)h^2(X)\right] < \infty$. Then the Conditional Gini Function optimally predicts the conditional rank advantage $(2M(Y|X) - 1)$ in a weighted least squares sense*

$$\mathbb{G}(Y|X) = \underset{h \in \aleph}{\operatorname{argmin}} \, \mathbb{E}\left[w(Y)\{(2M(Y|X) - 1) - h(X)\}^2\right] \tag{20}$$

where $w(Y) = Y/\mathbb{E}[Y]$ are the weights and $\int_{\mathscr{Y}} w(y)dF(y) = 1$. On the other hand, the unconditional Gini optimally predicts the unconditional rank advantage $(2M(Y) - 1)$ in the same weighted least squares sense.

Proof. See Appendix F. □

The result shows that the CGF minimizes L^G and provides a simple interpretation for the conditional Gini problem in terms of prediction. For example, if $F(y|x)$ is the income distribution for the subpopulation with characteristics x from which we take a random draw, the Gini optimally predicts the recipient's position (rank advantage) in the distribution by weighting individuals in proportion to their outcomes. The Gini of any distribution (continuous, discrete, mixed, etc.) can thus be recovered by taking the difference between the outcome-weighted expected mid-rank and the (unweighted) expected rank, which is always 0.5, and scaling it by 2, that is, $2\{\mathbb{E}_w[M(Y)] - \mathbb{E}[M(Y)]\}$. Theorem 4 is new.[17]

Note that the expectation in the minimization problem is taken with respect to the joint distribution of Y and X. If, for example, $\aleph$ is restricted to be the class of linear measurable functions indexed by $\theta \in \Theta \subseteq \mathbb{R}^p$ such that $h(x;\theta) = x'\theta \in \aleph$, and minimization in (20) takes

[17] Note that, in the continuous case, $(2F(Y|X) - 1)|X = x \sim \mathrm{U}[-1, 1]$ and $(2F(Y) - 1) \sim \mathrm{U}[-1, 1]$.

place over the parameter space Θ, the first-order condition coincides with the moment condition in (7), i.e., $\mathbb{E}[XU] = 0$. Likewise, for a general parametric model $h(x;\theta) = m(x;\theta)$ and under a smoothness assumption in θ, the first-order condition with respect to θ coincides with the classical orthogonality condition $\mathbb{E}[m_\theta(X;\theta_0)U] = 0$ where $m_\theta(x;\theta_0) = \partial_\theta m(x;\theta)|_{\theta=\theta_0}$ is the score of the model evaluated at the true parameter value.

Theorem 4 thus suggests the following alternative characterization of GIR models

$$(2M(Y|X) - 1) = m(X;\theta_0) + e \ , \ \mathbb{E}[w(Y)e|X] = 0 \tag{21}$$

where e and U are related through $Ye = U$. Note that L^G is a weighted MSE loss function used to evaluate predictions for $(2M(Y|X) - 1)$ instead of predictions for Y. In Appendix B, we show that L^G is related –through a simple location shift that does not change the minimizer– to the family of loss functions introduced in Gneiting (2011a) that yield ratios of expectations as optimal predictors for Y.

5.2 Goodness of Fit for Gini Index Regression

It is natural to use the optimized value of an objective function defining a regression estimator for a parameterized statistical functional as a goodness-of-fit measure. The standard example is the classical R^2 used in OLS regression analysis in the context of the conditional mean, and later extended to the QR setting in Koenker and Machado (1999). Consider the partition of the $p \times 1$ vector of regressors $X = \left(X_1', X_2'\right)'$ and of the $p \times 1$ parameter vector $\theta = \left(\theta_1', \theta_2'\right)' \in \mathbb{R}^{p_1+p_2}$. For a linear model, the OLS- and QR- R^2 measures are defined, respectively, as

$$R^2 = 1 - \frac{\hat{S}}{\tilde{S}} \quad \text{and} \quad R^2(\tau) = 1 - \frac{\hat{S}(\tau)}{\tilde{S}(\tau)}$$

where $\hat{S} := \min_\theta \sum_i (y_i - x_i'\theta)^2$ and $\tilde{S} := \min_{\theta_1} \sum_i (y_i - x_{1i}'\theta_1)^2$ denote the sum of squared residuals from the unrestricted and restricted ($\theta_2 = 0$) models, respectively. The R^2 that is typically reported takes the restricted model to have only an intercept ($x_{1i} = 1 \ \forall i$), that is, the MSE evaluated at the unconditional mean of Y. Likewise, for a particular τ-quantile, $\hat{S}(\tau) := \min_\theta \sum_i \rho_\tau(y_i - x_i'\theta)$ and $\tilde{S}(\tau) := \min_{\theta_1} \sum_i \rho_\tau(y_i - x_{1i}'\theta_1)$ denote the weighted sum of absolute residuals from the unrestricted and restricted models, where $\rho_\tau(u) = u(\tau - \mathbb{1}_{\{u<0\}})$ is the check loss function. Again, a "classical" $R^2(\tau)$ measure would also use $x_{1i} = 1 \ \forall i$ as the restricted model, that is, the intercept-only model.

For GIR, we may proceed in the same manner in light of the result of Theorem 4. However, unlike in OLS or QR –where the parameter that minimizes the expected quadratic (check) loss function using a constant model is the unconditional mean (quantile)– the minimizer of the conditional Gini loss function with an intercept-only model does not coincide with the unconditional version of the statistical functional[18].

We define the GIR-R^2 for a general GIR model $m(\cdot)$ by

$$R^2_G = 1 - \frac{\hat{S}_G}{\tilde{S}_G} \tag{22}$$

where $\hat{S}_G := \min_{\theta\in\Theta} \sum_i \hat{L}^G(y_i, x_i, m(x_i;\theta))$ and $\tilde{S}_G := \min_{\theta\in\tilde{\Theta}} \sum_i \hat{L}^G(y_i, x_i, m(x_{1i};\theta))$, where the minimization takes place over the unrestricted and restricted parameter spaces, respectively $\Theta, \tilde{\Theta} \subseteq \mathbb{R}^q$, and $\hat{L}^G$ denotes the estimated loss using the preliminary estimate $\hat{M}_n(y|x)$. The "classical" version of R^2_G uses the constant-only model $m(1;\theta) = \theta_1$ to compute $\tilde{S}_G$. As is the case with nonlinear least squares or OLS with no intercept in the unrestricted model, R^2_G cannot be guaranteed to lie within the unit interval for an arbitrary nonlinear $m(\cdot)$.

Focusing on the linear case with an intercept, $m(x;\theta) = \theta_1 + x_2'\theta_2$, the classical version of R^2_G is computed as

$$R^2_G = 1 - \frac{\sum_i \hat{L}^G(y_i, x_i, x_i'\hat{\theta}_n)}{\sum_i \hat{L}^G(y_i, x_i, \hat{\theta}_{1n})} = 1 - \frac{\sum_i y_i\{(2\hat{M}_n(y_i|x_i) - 1) - x_i'\hat{\theta}_n\}^2}{\sum_i y_i\{(2\hat{M}_n(y_i|x_i) - 1) - \hat{\theta}_{1n}\}^2} \tag{23}$$

where $\hat{\theta}_n$ and $\hat{\theta}_{1n}$ are the unrestricted and restricted ($\theta_2 = 0$) linear GIR estimators, respectively[19]. Since the loss function is non-negative and the unrestricted linear model nests the intercept-only model, R^2_G is always between zero and one. The following algorithm summarizes how to compute R^2_G for linear GIR models.

Algorithm 3. *(GIR-R^2) (i) Obtain estimates $\hat{M}_n(y_i|x_i)$ of the conditional mid-distribution function for all $(y,x) \in \{(y_1,x_1),\ldots,(y_n,x_n)\}$ using (15). (ii) Obtain the GIR estimator $\hat{\theta}_n$ via an OLS regression of $\sqrt{y_i}\left(2\hat{M}_n(y_i|x_i) - 1\right)$ on $\sqrt{y_i}x_i$ and generate the predictions $x_i'\hat{\theta}_n$ for*

[18] It would be true under independence $M(y|x) = M(y)$, since the unconditional Gini minimizes a slightly different loss function: (19) with the unconditional mid-distribution instead of the conditional.

[19] It can be shown that

$$\hat{\theta}_{1n} = \frac{\mathbb{E}_n[y_i(2\hat{M}_n(y_i|x_i) - 1)]}{\mathbb{E}_n[y_i]} = \mathbb{E}_n[\hat{w}(y_i)(2\hat{M}_n(y_i|x_i) - 1)] \neq \hat{G}(y_i) \text{ where } \hat{w}(y_i) = y_i/\mathbb{E}_n[y_i]$$

and $\hat{G}(y_i)$ is the sample Gini of Y.

all $x \in \{x_1, \ldots, x_n\}$. *(iii) Obtain the restricted, constant-only, GIR estimator* $\hat{\theta}_{1n}$ *via an OLS regression of* $\sqrt{y_i}\left(2\hat{M}_n(y_i|x_i) - 1\right)$ *on* $\sqrt{y_i}$. *(iv) Compute the GIR-*R^2 *as in (23).*

In terms of interpretation, R^2_G could be seen as the fraction of explained weighted variance of the conditional centered mid-rank, where the weights are given as in Theorem 4. Define the weighted sum of squared residuals $wSSR := \sum_i \hat{w}(y_i)\{(2\hat{M}_n(y_i|x_i) - 1) - x_i'\hat{\theta}_n\}^2$, the weighted total sum of squares $wTSS := \sum_i \hat{w}(y_i)\{(2\hat{M}_n(y_i|x_i) - 1) - \mathbb{E}_n[\hat{w}(y_i)(2\hat{M}_n(y_i|x_i) - 1)]\}^2$, and $wESS := \sum_i \hat{w}(y_i)\{x_i'\hat{\theta}_n - \mathbb{E}_n[\hat{w}(y_i)(2\hat{M}_n(y_i|x_i) - 1)]\}^2$, the weighted explained sum of squares. One can show that $wTSS = wSSR + wESS$ by exploiting the linear GIR first-order condition (in-sample) $\mathbb{E}_n[\hat{w}(y_i)x_i\hat{e}_i] = 0$, where x_i includes a constant, $\hat{w}(y_i) = y_i/\mathbb{E}_n[y_i]$, and $\hat{e}_i := (2\hat{M}_n(y_i|x_i) - 1) - x_i'\hat{\theta}_n$. Then:

$$R^2_G = 1 - \frac{wSSR}{wTSS} = \frac{wESS}{wTSS} \tag{24}$$

Finally, note that R^2_G is not equivalent to the classical OLS-R^2 of the second-stage OLS regression of $\sqrt{y_i}\left(2\hat{M}_n(y_i|x_i) - 1\right)$ on $\sqrt{y_i}x_i$, typically reported in linear regression software. We compute R^2_G in the empirical application (see Table 4 in Section 7) and analyze how it increases with the number of variables included in each model specification.

6 Simulation Study

In this section, we study the finite sample performance of the linear GIR estimator and the proposed bootstrap scheme in a Monte Carlo (MC) simulation. Since the MC experiment involves bootstrap estimators, we use the "Warp-Speed" method proposed in Giacomini et al. (2013), which significantly eases computational costs. This method considers a single bootstrap draw from each simulated MC sample to generate the bootstrap estimators later used to compute the quantiles with which to assess confidence interval (CI) coverage. For the experiment, we consider the following data-generating processes (DGPs).

- **DGP1.** $\mathbb{G}(Y|X) = X_1\theta_{01} + X_2\theta_{02} + X_3\theta_{03}$, where $Y|X = x \sim Beta\big(\alpha(x), \beta(x)\big)$, with $\beta(x) = \beta = 1$, and $\theta_0 = (0.1, 0.2, 0.3)'$. Regressors X_k are mutually exclusive dummies indicating on which third of the support of $Z \sim U[1,4]$ a realization z is.

- **DGP2.** $\mathbb{G}(Y|X) = X_1\theta_{01} + X_2\theta_{02} + X_3\theta_{03}$, where $Y|X = x \sim Beta\big(\alpha(x), \beta(x)\big)$ with $\beta(x) = \beta = 1$, $X_1 \sim U[0.1, 1] \perp X_k \sim Ber(0.5)$ for $k \in \{2,3\}$, and $\theta_0 = (0.075, 0.1, 0.125)'$.

- **DGP3.** $\mathbb{G}(Y|X) = \theta_{00} + X_1\theta_{01} + X_2\theta_{02}$, where $\mathscr{Y} = \{1,3,5\}$ with conditional probability mass function $P(Y=1|X=x) = P(Y=5|X=x) = \left(1-\sqrt{1-3x'\theta_0}\right)/2$, $0 < x'\theta_0 < 1/3$ $\forall x \in \mathscr{X}$, $X_1 \sim U[0,1] \perp X_2 \sim Ber(0.5)$, and $\theta_0 = (0.225, -0.075, 0.075)'$.

For DGP1 and DGP2, the choices of parameters $\theta = (\theta_{01}, \theta_{02}, \theta_{03})'$ and of the support $\mathscr{X}$ of the regressors $X = (X_1, X_2, X_3)'$ ensure that the linear conditional Gini lies in the appropriate range. The functional forms of the parameters of the conditional distributions, $\alpha(x)$, are as given in Section 2 to guarantee a linear conditional Gini. All simulations use R = 10,000 Monte Carlo replications, and all first stages estimate the conditional mid-distribution function nonparametrically. For DGP1, since the dummies are mutually exclusive, the choice $q_K(x) = x$ yields a correctly specified first stage. For DGP2 and DGP3, since they contain continuous regressors, we use the least-squares series method with both spline and Legendre polynomial bases. All simulations assess the finite sample bias of the GIR estimator and coverage provided by the Bayesian bootstrap 95% CIs.

Table 1: GIR Monte Carlo Summary, Group-Dummy Design (DGP1).

n	**500**			**1000**			**4000**		
Coefficient	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**
GIR (LS)									
θ_{01}	0.942	-0.0006	0.008	0.951	-0.0003	0.005	0.956	0.0000	0.003
θ_{02}	0.943	-0.0009	0.014	0.944	-0.0003	0.010	0.952	-0.0001	0.005
θ_{03}	0.945	-0.0013	0.018	0.954	-0.0007	0.013	0.948	-0.0002	0.006
Oracle									
θ_{01}	0.948	0.0000	0.042	0.954	-0.0002	0.029	0.949	0.0000	0.015
θ_{02}	0.951	-0.0016	0.040	0.949	-0.0001	0.028	0.953	-0.0001	0.014
θ_{03}	0.942	-0.0016	0.039	0.952	-0.0004	0.027	0.948	0.0000	0.014

Note: Warp-Speed simulation results of DGP1 for R = 10,000 Monte Carlo replications. Displayed coverage, bias, and Monte Carlo standard deviation results use the true Gini parameter values. “GIR (LS)" refers to the results obtained using the estimated CMDF via least-squares with $q_K(x) = x$ (correct specification), whereas “Oracle" refers to those obtained using the true CMDF.

DGP1 illustrates the case of mutually exclusive and exhaustive categories as regressors, where both first and second stages are correctly specified, and the GIR estimator exactly coincides with the sample Gini of each group (common in applied work). Table 1 contains the main results, showing bootstrap CI coverage for the true parameter, along with the bias and MC standard deviation of the GIR estimator. Following the Warp-Speed method, Bayesian bootstrap quantiles are then constructed from the resulting centered bootstrap

draws across MC replications, rather than from B draws within each replication. We compare the GIR estimation results with those obtained by using the oracle GIR estimator, that is, the one that uses the true CMDF as if it were known by the econometrician. The numerical equivalence of GIR estimates to the empirical group Ginis is also assessed. The observed agreement, at floating-point precision, is an overall maximum difference, across all DGP replications, below 1.3×10^{-15} between GIR estimates and empirical Ginis computed using the standard sorted-observation formula.

Table 2: GIR Monte Carlo Summary, Continuous Response Design (DGP2).

n	**1000**			**2000**			**7000**		
Coefficient	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**
GIR (Sp.)									
θ_{01}	0.942	-0.0016	0.008	0.943	-0.0010	0.006	0.950	-0.0005	0.003
θ_{02}	0.946	-0.0018	0.008	0.942	-0.0010	0.006	0.954	-0.0005	0.003
θ_{03}	0.941	-0.0023	0.009	0.947	-0.0015	0.006	0.947	-0.0006	0.003
GIR (Lg.)									
θ_{01}	0.947	-0.0018	0.009	0.948	-0.0008	0.006	0.951	-0.0003	0.003
θ_{02}	0.948	-0.0007	0.008	0.943	-0.0006	0.006	0.954	-0.0003	0.003
θ_{03}	0.946	-0.0010	0.009	0.947	-0.0010	0.006	0.950	-0.0004	0.003
Oracle									
θ_{01}	0.951	0.0002	0.043	0.956	0.0001	0.030	0.953	-0.0001	0.016
θ_{02}	0.955	-0.0001	0.032	0.949	-0.0006	0.022	0.948	0.0000	0.012
θ_{03}	0.950	-0.0004	0.031	0.952	0.0002	0.022	0.946	0.0001	0.012

Note: Warp-Speed simulation results of DGP2 for $R = 10{,}000$ Monte Carlo replications. Displayed coverage, bias, and Monte Carlo standard deviation results use the true Gini parameter values. “GIR (Sp.)" refers to the results obtained using the estimated CMDF via least-squares series with spline basis, whereas “GIR (Lg.)" refers to the results obtained using the estimated CMDF via least-squares series with Legendre basis. “Oracle" refers to those obtained using the true CMDF. The DGP satisfies $\bar{\rho} > 7/6$ and the first-stage estimators use $K = \lfloor n^{0.3} \rfloor$ for splines and $K = \lfloor n^{0.245} \rfloor$ for Legendre polynomials in each discrete-cell.

The simulation results for DGP2, with continuous response and both continuous and binary regressors, are reported in Table 2. First-stage estimation with mixed regressors applies the proposed series approximation to the continuous coordinate in each cell of discrete covariate values, applying the rule $K_g = \lfloor n_g^{0.3} \rfloor$ for splines and $K_g = \lfloor n_g^{0.245} \rfloor$ for Legendre polynomials, where K_g and n_g denote the basis length and sample size in cell g, respectively. Oracle estimation results are also reported. Finally, DGP3 considers the discrete-response case, where Y follows a three-point conditional distribution, and the

model uses both continuous and discrete regressors. Results in Table 3 show bootstrap confidence interval coverage for the true parameter values, bias, and MC standard deviation of the GIR estimator.

Table 3: GIR Monte Carlo Summary, Discrete Response Design (DGP3).

n	**1000**			**2000**			**7000**		
Coefficient	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**
GIR (Sp.)									
θ_{00}	0.950	-0.0025	0.015	0.946	-0.0017	0.011	0.952	-0.0007	0.006
θ_{01}	0.945	0.0010	0.023	0.951	0.0006	0.016	0.952	0.0003	0.009
θ_{02}	0.953	-0.0009	0.013	0.946	-0.0004	0.009	0.952	-0.0003	0.005
GIR (Lg.)									
θ_{00}	0.951	-0.0016	0.015	0.948	-0.0012	0.011	0.952	-0.0004	0.006
θ_{01}	0.948	0.0006	0.023	0.952	0.0004	0.016	0.950	0.0002	0.009
θ_{02}	0.954	-0.0006	0.013	0.949	-0.0002	0.009	0.955	-0.0002	0.005
Oracle									
θ_{00}	0.954	-0.0002	0.034	0.954	-0.0004	0.024	0.948	0.0000	0.013
θ_{01}	0.952	0.0000	0.052	0.952	0.0002	0.037	0.952	0.0002	0.020
θ_{02}	0.951	-0.0002	0.030	0.949	0.0004	0.021	0.947	-0.0002	0.011

Note: Warp-Speed simulation results of DGP3 for $R = 10{,}000$ Monte Carlo replications. Displayed coverage, bias, and Monte Carlo standard deviation results use the true Gini parameter values. “GIR (Sp.)" refers to the results obtained using the estimated CMDF via least-squares series with spline basis, whereas “GIR (Lg.)" refers to the results obtained using the estimated CMDF via least-squares series with Legendre basis. “Oracle" refers to those obtained using the true CMDF. The DGP satisfies $\bar{\rho} > 7/6$ and the first-stage estimators use $K = \lfloor n^{0.3} \rfloor$ for splines and $K = \lfloor n^{0.245} \rfloor$ for Legendre polynomials in each discrete cell.

The results are consistent with the theoretical findings. The GIR estimator in the binary design performs quite well and is free of first-stage approximation error, since the first stage is correctly specified due to the saturated specification. In the continuous response and mixed regressor case (continuous and binary), GIR also performs well in terms of coverage and bias decay, showing that the approximation of the CMDF around the point of interest is accurate enough. As expected, approximation bias is larger in smaller samples compared to the oracle estimator, but this difference is negligible in large samples. Finally, regarding the discrete-response design, results show that the theoretical machinery developed in the preceding sections naturally accommodates discrete responses through the CMDF term. An interesting point to mention is that the MC standard deviation of the feasible estimator

is lower than that of the oracle. This seems counterintuitive, since the oracle is using more information and is thus expected to be less noisy than the feasible one, which adds an extra term in the asymptotic variance of the estimator due to the first stage estimation of the CMDF (as shown in the proof of Theorem 2). However, it is likely that within the simulation settings, the covariance term between the Gaussian random vectors showing up in the asymptotic distribution is negative, which reduces the overall variance of the feasible estimator. In Appendix G, we verify this possibility by conducting an extra MC experiment in which the first stage estimate is computed using an independent sample from the one used for the second stage. The results show that with independent samples, the MC standard deviation of the estimates is higher than that of the oracle estimator, which we interpret as evidence in support of our hypothesis.

7 Empirical Application

In this section, we examine how wage inequality in the United States evolved during the 1980s, focusing on the 1979–1988 period originally analyzed by DiNardo et al. (1996) (henceforth DFL) and by Chernozhukov et al. (2013) (henceforth CFM). Although our results cannot be directly compared with those of CFM, who study the unconditional Gini of the distribution of log wages, whereas we study the conditional Gini of the distribution of wages in levels[20], analyzing the same period allows us to implement the theoretical framework developed in Sections 3-5 and to provide a complementary perspective on the evolution of inequality.

The empirical analysis uses the Outgoing Rotation Groups of the Current Population Survey (CPS) for 1979 and 1988, chosen to capture the steep increase in wage dispersion during the 1980s. We use hourly wages (expressed in 1979 dollars) and construct a rich set of covariates: union status; four education categories (collapsing all sub–high-school levels into a single category) interacted with experience; a cubic polynomial in experience; and occupation, industry, race, wave, Standard Metropolitan Statistical Area (SMSA), and part-time indicators. As in DFL and CFM, each observation is weighted by the CPS sampling weights multiplied by hours worked. Consistent with their main analysis, we restrict attention to men (96,144 observations). In the first stage, we explicitly account for minimum-wage censoring by treating wages below the statutory minimum as left-censored observations. The conditional distribution function is then recovered from censoring-adjusted, series least-squares estimates. We estimate the first-stage regression coefficients on a grid of out-

[20] See the discussion at the end of Section 2.3.

Table 4: GIR Estimates in Gini points (100×Gini) for Models 1–3.

Variable	**Model 1**	**Model 2**	**Model 3**
Union	-4.70**	-1.80**	0.72
	(−4.98, −4.31)	(−2.59, −0.63)	(−0.23, 2.40)
Wave	1.09**	1.11**	1.03**
	(0.22, 1.15)	(0.36, 1.16)	(0.28, 1.07)
S.M.S.A	2.14**	2.13**	2.16**
	(1.71, 2.60)	(1.73, 2.60)	(1.75, 2.63)
Part-time	6.74**	6.53**	6.49**
	(6.01, 7.94)	(5.64, 7.67)	(5.61, 7.61)
Non-white	-0.46	-0.42	-1.29
	(−1.25, 0.32)	(−1.22, 0.36)	(−2.80, 0.011)
High School	-1.24**	-4.39**	-4.38**
	(−1.70, −0.68)	(−5.35, −3.34)	(−5.33, −3.33)
Some College	-1.45**	-4.23**	-4.28**
	(−1.99, −0.63)	(−5.46, −3.01)	(−5.50, −3.11)
College	-3.82**	-5.43**	-5.64**
	(−4.36, −2.98)	(−6.88, −4.35)	(−7.09, −4.56)
Experience	0.30**	0.35**	0.34**
	(0.26, 0.39)	(0.18, 0.52)	(0.17, 0.52)
Experience^2	-0.0042**	-0.011**	-0.011**
	(−0.0060, −0.0028)	(−0.018, −0.0035)	(−0.018, −0.0034)
Experience^3		0.00012**	0.00012**
		(0.000021, 0.00023)	(0.000018, 0.00023)
Experience × High School		0.14**	0.14**
		(0.099, 0.17)	(0.099, 0.17)
Experience × Some College		0.12**	0.12**
		(0.091, 0.19)	(0.088, 0.18)
Experience × College		0.061**	0.060**
		(0.028, 0.13)	(0.026, 0.13)
Union × Experience		-0.13**	-0.13**
		(−0.18, −0.087)	(−0.18, −0.086)
Experience × Nonwhite			0.023
			(−0.052, 0.073)
Union × Nonwhite			1.59*
			(−0.23, 3.14)
Fixed Effects:			
Industry and Occupation	Yes	Yes	Yes
Union × Occupation	No	No	Yes
$R^2_G \times 100$	84.51	84.51	84.52

Note: In parentheses and below the coefficient estimates are the 95% bootstrap confidence intervals based on Algorithm 2 with $B = 100$. '**', and '*' denote significance at the 5% and 10% levels, respectively.

come thresholds, corresponding to selected quantile points of the wage distribution, and obtain $M(y|x)$ at intermediate points by interpolation, which eases computational costs significantly. The main results use a spline basis (as in the simulations) in the least-squares series estimation of the first stage, but we report very similar results in Appendix H using a Legendre polynomial basis instead (see Table 10). We consider a linear specification for the conditional Gini, leading to the following model:

$$Y(2M(Y|X)-1) = Y b(X)'\gamma + \epsilon \ , \ \mathbb{E}[b(X)\epsilon] = 0 \tag{25}$$

Here Y denotes hourly wages and $M(Y|X)$ is the conditional mid-distribution function of Y given covariates X, γ is a $p \times 1$ vector of coefficients of interest, and $b(x)$ is a transformation of the covariates used to model the conditional Gini that includes all covariates entering in levels, along with the relevant interaction terms used to model heterogeneity.

Table 4 reports the GIR estimates, which describe how the conditional Gini varies with the individual characteristics of interest and R^2 of the model. Figures 3 and 4 illustrate the evolution of inequality in years of experience, by union status and level of education, for an individual whose covariate vector is set to the modal characteristics (most frequently observed) of the sample, constructed using the GIR estimates from Model 3. We will focus mainly on the effect of union, experience, education, and how their interactions affect wage inequality when we control for other factors. Comparing Gini indices by groups does not account for differences in the characteristics of individuals within the group. The proposed analysis makes these comparisons while controlling for observable group characteristics.

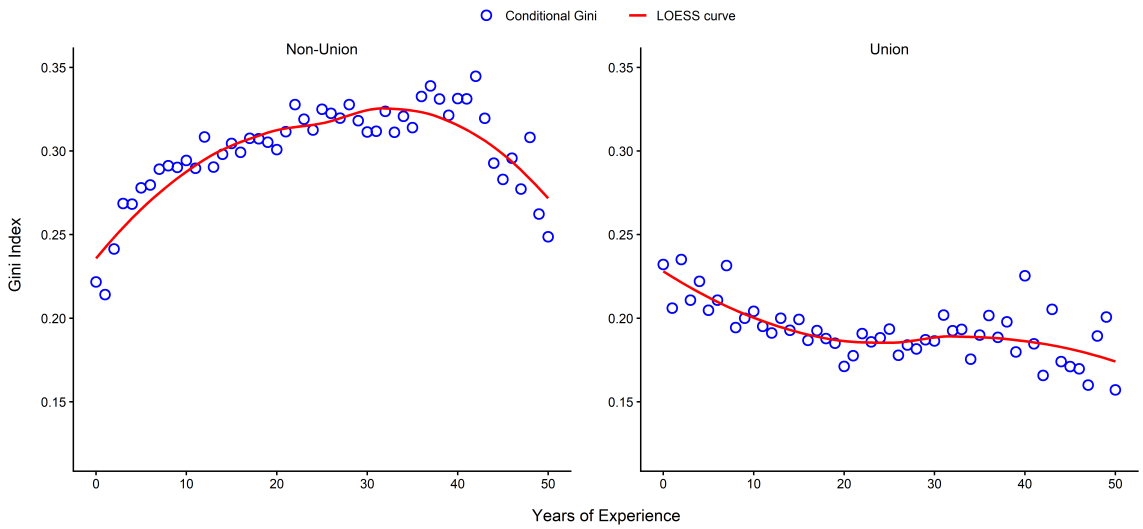
Table 5: Gini of Wages by Groups

Variable	**Gini Index**
Non-Unionized	0.325
Unionized	0.222
Below High School	0.277
High School	0.272
Some College	0.289
College	0.288
Total	0.313

One of the key interests in the labor economics literature is to study the impact of unions on wages and its distribution. Deunionization has been considered a key driver of inequality

(Freeman (1991); Card (1992); DiNardo et al. (1996); Farber et al. (2021)). As shown in Table 5, where we compare Gini indices across unionized and non-unionized workers, we find that inequality for non-unionized workers is about 10 Gini points (defined as 100×Gini) higher than that of unionized workers. However, this does not account for differences in characteristics between unionized and non-unionized individuals. Additionally, it is not informative about potential heterogeneous differences in union status inequality by observable characteristics. One of those characteristics is experience. Figure 2 helps us visualize this heterogeneity. We can observe that the differences in inequality between union and non-union workers at lower levels of experience are moderate. However, with the increase in experience, the predicted inequality gap widens considerably, as inequality for non-unionized workers is trending upwards, while for unionized workers it is slightly trending downwards. Therefore, studying the impact of unions without considering this interaction does not provide a complete picture.

Figure 2: Gini of Wages by Union Status and Years of Experience



When analyzing the conditional predicted difference of unionization on inequality in Table 4, we find a statistically significant difference across the first two model specifications, and in the third model its impact is captured by the interaction variables. The extended specifications in Models 2 and 3 provide a clearer picture of how the impact of unionization differs across individual and firm characteristics.

Additionally, we find that the effect of unionization on the Gini coefficient varies with experience. In the most comprehensive specification, unionization is associated with lower wage inequality, with the magnitude of this reduction increasing among more experienced workers. The estimated reduction ranges from 1.3 Gini points for workers with 10 years

of experience to 5.2 Gini points for workers with 40 years of experience. This pattern is illustrated in Figure 3, which shows little difference in inequality between unionized and non-unionized workers at lower levels of experience, but a substantially wider gap as experience increases. These findings are consistent with the broader literature documenting that declining unionization, or deunionization, has contributed to rising wage inequality.

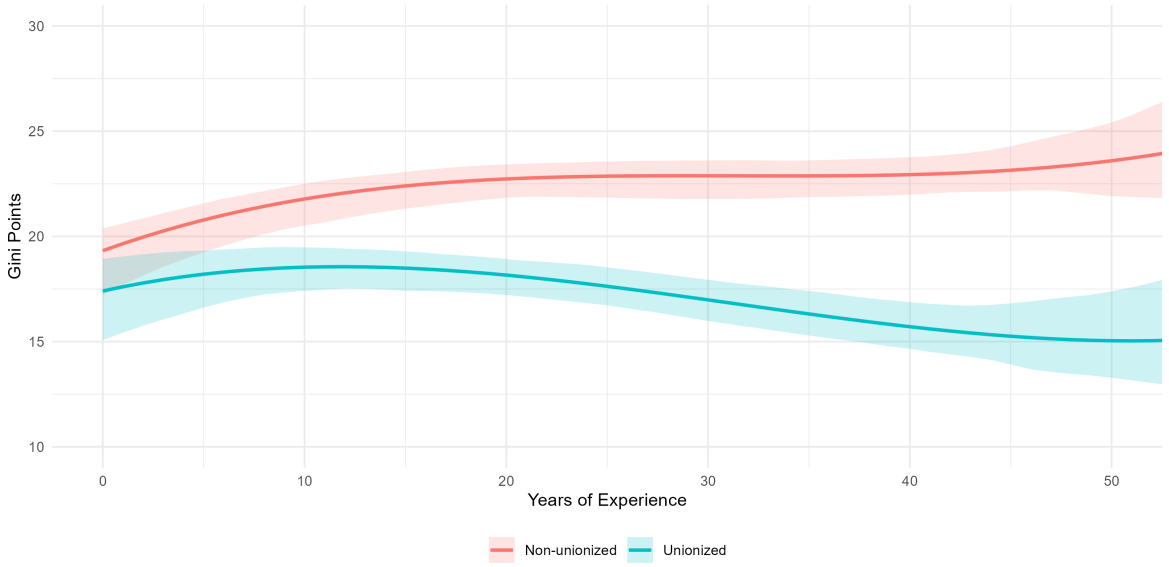
This finding further reinforces the point that inequality cannot be adequately studied using only two or three variables. The differences in other characteristics between the considered groups matter. For instance, when considering only unionization and experience, Figure 2 suggests that unionization is associated with reduced inequality among more experienced workers. However, once additional covariates are accounted for, the relationship becomes more nuanced. Our findings suggest that efforts to reduce inequality may be more effective if policies prioritized increasing unionization rates among more experienced workers. Moreover, we find that unionization might not lead to lower inequality among people of color compared to white workers.

In addition to examining the effects of unionization, labor economists have documented a pronounced rise in income inequality among highly educated workers, particularly college graduates. This rise in within-group inequality has been a major driver of the overall increase in wage dispersion. A central explanation in the literature is the skill-biased technical change (SBTC) hypothesis, which argues that technological advances, especially the rapid expansion of information technology, have disproportionately increased the relative demand for certain high-skill tasks compared with others. Consequently, wages among college-educated workers have diverged as technology rewards specific skill sets more strongly. See, for example (Juhn et al. (1993); Lemieux (2006); Autor et al. (2008); Lemieux (2011)).

Hence, it is important to examine how inequality varies across education groups, and to assess this both unconditionally and after controlling for other relevant covariates, while also exploring how inequality heterogeneity evolves with experience within each education group. When we look at the unconditional inequality across education categories, we find that the differences are modest (see Table 5): inequality among college-educated workers is only about 1 Gini point higher than among individuals who did not complete high school.

Model 1 in Table 4 reports estimated differences in inequality across education groups, comparing workers in the same industry and occupation and with similar experience, race, and union status. We find that inequality is lower among workers with a high school diploma, some college, or a college degree relative to workers without high school diploma.

Figure 3: Gini Estimates by Union Status and Experience, given Controls



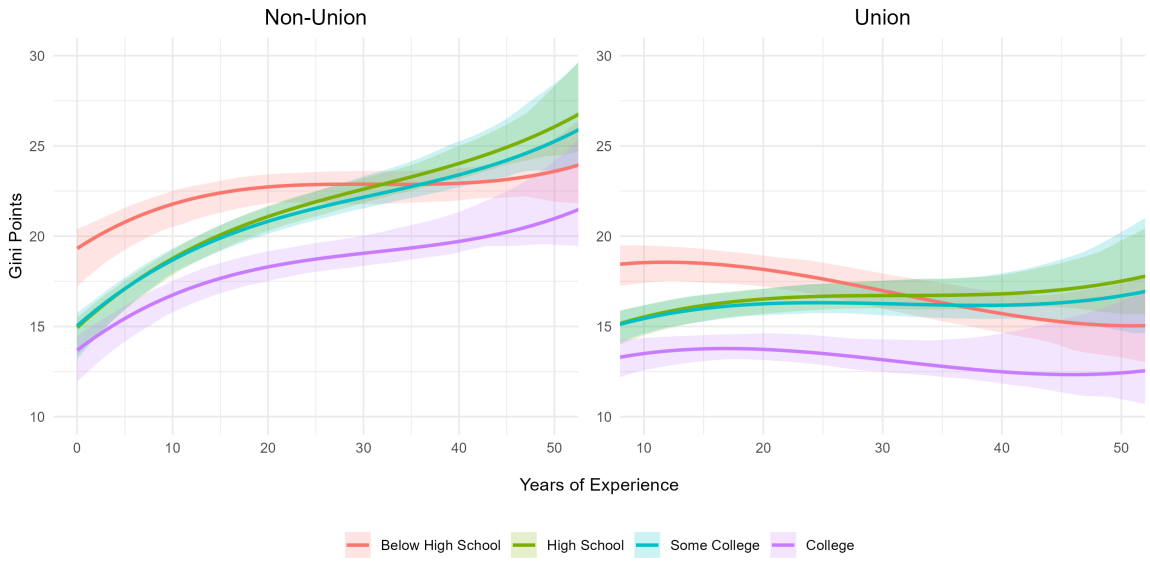
Note: The shaded regions represent the 95% bootstrap confidence intervals.

Models 2 and 3 yield a particularly interesting insight. The larger negative coefficients for the some college and college educated groups indicate that individuals in these categories, when working in the same industry, occupation and year, having the same union status, and racial group, earn relatively similar wages at the start of their careers. This pattern is also visible in Figure 4, and is consistent with the SBTC hypothesis. Industry and occupation dummies proxy for major choice, job type, and task content. This suggests that a fraction of the initial wage inequality among new college graduates is explained by sorting into different occupations and industries (Johnson, 1997; Card and DiNardo, 2002).

In Figure 4, we observe that among non-unionized workers, inequality diverges with experience. More experienced workers with a high school degree, some college education, or a college degree exhibit substantially higher wage inequality than their less experienced counterparts. This suggests that, conditional on the other observed characteristics and job-related factors considered in the model, inequality among workers with at least a high school education is partly driven by heterogeneous returns to experience. In other words, workers with similar education levels and comparable characteristics may begin with relatively similar wage distributions, but their ability to convert accumulated experience into higher earnings differs substantially, thereby contributing to rising inequality over the life cycle.

However, when we turn to unionized workers, we do not observe a similarly stark increase

Figure 4: Gini Estimates by Union Status, Education, and Experience, given Controls



Note: The shaded regions represent the 95% bootstrap confidence intervals.

in inequality among workers with at least a high school degree. Moreover, inequality appears to decline with experience among workers without a high school degree. This suggests that unionization may blunt the inequality-enhancing effect of experience, and in some cases may even reverse it. A plausible explanation is that collective bargaining compresses wage dispersion by standardizing wage schedules and limiting the extent to which individual returns to experience vary across workers (Freeman (1980)). Thus, while experience generates increasing wage dispersion among comparable non-unionized workers, unionization appears to dampen this channel of inequality.

Finally, examining the goodness-of-fit of our specifications, we observe consistently high and similar values of R^2_G in Table 4 for all models (model 3 nests model 2, and model 2 nests model 1). To gain some insight on the magnitude of the change in R^2_G between specifications, we informally compare the three models using the change in R^2_G to compute the standard F-test from linear regression: comparing Models 1 and 2, we get an F-statistic of 2.2, while comparing Models 1 and 3 we get an F-statistic of 5.8. The statistic increases to 10.3 when the comparison involves Models 2 and 3. The change in R^2_G is the largest when model 3 is used as benchmark (unrestricted model). This suggests model 3 to be the preferred specification[21].

[21] We note that the reported F-type quantities are used only as descriptive indices of fit and are not intended to have the usual F or χ^2 distribution.

8 Conclusion

In this paper, we introduce a novel regression framework to understand concentration, relative dispersion, or inequality differences between subpopulations that a simple difference in means or in an inequality measure cannot capture and that emerge when controlling for other characteristics or confounding factors. The approach is based on modeling the Gini of a conditional distribution directly and providing an identifying moment condition without specifying its mean or distribution, directly or indirectly, e.g., via conditional quantiles. We refer to this method as Gini Index Regression (GIR) and place particular emphasis on the advantages of the linear model, for which we establish the main theoretical results. In the same spirit as ordinary least squares and quantile regression, this new regression method enables simple estimation, implementation, interpretation, and inference on inequality differences or marginal effects between population subgroups.

Extensions of GIR that consider alternative (nonlinear) parametric specifications and estimation methods are the subject of ongoing research. In addition, an important avenue for future work is to incorporate time dependence into the model, thereby making the methodology suitable for time-series and dynamic settings.

Appendix

A General Unconditional and Conditional Gini Formulae

In this section, we state and prove two lemmas regarding the validity of the given expressions for the Gini of an unconditional distribution (1) and its conditional version (2), key to the results in this paper, for continuous, discrete, mixed, or singular distributions. We also prove their equivalences to the popular Gini mean-difference (GMD) and covariance representations.

Lemma 4. *(Unconditional Gini) Let Y be a random variable satisfying $Y \geq 0$ a.s., with CDF $F(y)$ and $0 < \mathbb{E}[Y] < \infty$. The Gini of Y satisfies*

$$\mathbb{G}(Y) = \frac{\mathbb{E}[Y(2M(Y)-1)]}{\mathbb{E}[Y]} = \frac{\mathbb{E}[|Y_1 - Y_2|]}{\mathbb{E}[Y_1 + Y_2]} = 2\frac{\mathbb{C}ov(Y, M(Y))}{\mathbb{E}[Y]} \in [0,1)$$

where Y_1 and Y_2 are independent copies of Y and $M(y) = \left(F(y) + F(y^-)\right)/2$ is the mid-distribution function of Y, where $F(y^-) = P(Y < y) = \lim_{z \uparrow y} F(z)$ and $\mathbb{C}ov(\cdot,\cdot)$ denotes the covariance operator.

Proof. We first prove the first equality. Let $L(p)$ be the (unconditional) Lorenz curve defined in Section 2.1 and let $\mu = \mathbb{E}[Y]$. By definition of Gini,

$$\begin{aligned}\mathbb{G}(Y) &= 1 - 2\int_0^1 L(p)\,dp = 1 - \frac{2}{\mu}\int_0^1\int_0^1 \mathbb{1}_{\{s\le p\}}F^{-1}(s)\,ds\,dp = 1 - \frac{2}{\mu}\int_0^1\int_0^1 \mathbb{1}_{\{s\le p\}}\,dp\,F^{-1}(s)\,ds \\ &= \frac{1}{\mu}\int_0^1 F^{-1}(s)\,ds - \frac{2}{\mu}\int_0^1 (1-s)F^{-1}(s)\,ds = \frac{1}{\mu}\int_0^1 (2s-1)F^{-1}(s)\,ds \\ &= \frac{1}{\mu}\int_0^1 2sF^{-1}(s)\,ds - 1\end{aligned}$$

where the last equality in the first line follows from Tonelli's theorem, since $(s,p) \mapsto \mathbb{1}_{\{s\le p\}}F^{-1}(s)$ is non-negative ($Y \ge 0$ *a.s.*) and measurable, and the first equality in the second line uses $\mu > 0$ and the quantile (left-inverse) representation of the mean, which holds for any real-valued, integrable random variable. Finiteness follows from $\mu < \infty$ and $0 \le \int_0^1 2sF^{-1}(s)\,ds \le 2\int_0^1 F^{-1}(s)\,ds = 2\mu < \infty$.

Note that for $0 < v < 1$, $f_V(v) = 2v$ is the density function of a continuous random variable V with CDF $F_V(v) = v^2$, $0 \le v \le 1$, which is the CDF of the maximum between two independent, standard uniform random variables, i.e., we can take $V = \max\{U_1, U_2\}$, where U_1 and U_2 are independent copies of $U \sim \mathrm{U}[0,1]$, since $P(V \le v) = P(U_1 \le v, U_2 \le v) = P(U_1 \le v)P(U_2 \le v) = v^2$. Hence, by monotonicity of F^{-1}, we can write

$$\int_0^1 2sF^{-1}(s)\,ds = \mathbb{E}[F^{-1}(V)] = \mathbb{E}[\max\{F^{-1}(U_1), F^{-1}(U_2)\}] = \mathbb{E}[\max\{Y_1, Y_2\}] = \int_0^\infty y\,d\left\{F^2(y)\right\}$$

where the third equality follows from the standard fact $F^{-1}(U) \sim F$ (which holds for any CDF) so $Y_i := F^{-1}(U_i)$, $i \in \{1,2\}$, and the last equality from the fact that $F^2(y)$ is the CDF of $\max\{Y_1, Y_2\}$, where Y_1 and Y_2 are independent copies of Y. We thus have the key expression

$$\mathbb{G}(Y) = \frac{1}{\mu}\int_0^\infty y\,d\left\{F^2(y)\right\} - 1$$

which is finite because $\mathbb{E}[\max\{Y_1, Y_2\}] \le \mathbb{E}[Y_1 + Y_2] = 2\mu < \infty$. Moreover, write $\max\{Y_1, Y_2\} = Y_1\mathbb{1}_{\{Y_2 \le Y_1\}} + Y_2\mathbb{1}_{\{Y_2 > Y_1\}}$ so that by the towering property of expectations and independence

$$\begin{aligned}\mathbb{E}[\max\{Y_1, Y_2\}] &= \mathbb{E}[Y_1 P(Y_2 \le Y_1 | Y_1)] + \mathbb{E}[Y_2 P(Y_1 < Y_2 | Y_2)] = \mathbb{E}[Y_1 F(Y_1)] + \mathbb{E}[Y_2 F(Y_2^-)] \\ &= \mathbb{E}[Y(F(Y) + F(Y^-))]\end{aligned}$$

where the last equality follows by Y_1 and Y_2 being independent copies of Y. Thus,

$$\mathbb{G}(Y) = \frac{1}{\mu}\int_0^\infty y\, d\{F^2(y)\} - 1 = \frac{\mathbb{E}[Y(F(Y) + F(Y^-))] - \mu}{\mu} = \frac{\mathbb{E}[Y(2M(Y) - 1)]}{\mathbb{E}[Y]}$$

which proves the first equality. For the second equality, note that we can write $\max\{Y_1, Y_2\} = (Y_1 + Y_2 + |Y_1 - Y_2|)/2$, so $\mathbb{E}[\max\{Y_1, Y_2\}] = \mu + \mathbb{E}[|Y_1 - Y_2|]/2$, which gives

$$\mathbb{G}(Y) = \frac{1}{\mu}\int_0^\infty y\, d\{F^2(y)\} - 1 = \frac{1}{\mu}\left(\mu + \frac{\mathbb{E}[|Y_1 - Y_2|]}{2}\right) - 1 = \frac{\mathbb{E}[|Y_1 - Y_2|]}{\mathbb{E}[Y_1 + Y_2]}$$

The last equivalence is easily proven by realizing that $\mathbb{E}[M(Y)] = 1/2$ since

$$2\mathbb{E}[M(Y)] = \mathbb{E}[F(Y)] + \mathbb{E}[F(Y^-)] = P(Y_2 \leq Y_1) + P(Y_2 > Y_1) = 1$$

where the second equality follows since Y_1 and Y_2 are independent copies. We can then write $\mathbb{E}[Y(2M(Y) - 1)] = 2\mathbb{E}[Y(M(Y) - \mathbb{E}[M(Y)])] = 2\mathbb{C}ov(Y, M(Y))$. Dividing by $\mu > 0$ proves the last representation. The finiteness of the covariance follows from $0 \leq M(Y) \leq 1$ and $\mu < \infty$.

Finally, we show that $0 \leq \mathbb{G}(Y) < 1$. Note that $\mathbb{G}(Y) = 0$ if and only if $\mathbb{E}[|Y_1 - Y_2|] = 0$, i.e., whenever $P(Y_1 = Y_2) = 1$. Since Y_1 and Y_2 are independent copies, this occurs if and only if Y is degenerate; otherwise $\mathbb{G}(Y) > 0$. Moreover, note that $|Y_1 - Y_2| \leq Y_1 + Y_2$ and taking expectations gives $\mathbb{G}(Y) \leq 1$. For strict inequality, note that $Y_1 + Y_2 - |Y_1 - Y_2| = 2\min\{Y_1, Y_2\}$, since $Y \geq 0$. Thus, $\mathbb{G}(Y) = 1$ would require $\mathbb{E}[\min\{Y_1, Y_2\}] = 0$, but since $\mathbb{E}[Y] > 0$ we have $P(Y > 0) > 0$. Independence gives $P(Y_1 > 0, Y_2 > 0) > 0$. Therefore, $\mathbb{E}[\min\{Y_1, Y_2\}] > 0$, so $\mathbb{G}(Y) < 1$. Hence $0 \leq \mathbb{G}(Y) < 1$.

□

The literature has many equivalent representations of the Gini under different assumptions and presented differently depending on whether the distribution is continuous or discrete. The mid-distribution function formula has been suggested as an adjustment to the Gini calculation in the case of discrete variables (see chapter 2.4 of Yitzhaki and Schechtman (2013)), and as a method to improve the accuracy of Gini calculations based on weighted or grouped (instead of individual) observations (Lerman and Yitzhaki, 1989). In the discrete case, Ferreira and Garín (1997) formally prove the equivalence between the Gini obtained directly from the Lorenz curve and the mid-distribution function formula (see Lemma 3 therein). The covariance formula is proven for continuous variables, using $F(Y)$ instead of $M(Y)$, in Lerman and Yitzhaki (1984). The index range part of the statement has been

proven under the same assumptions in Vila and Saulo (2026) (Remark 2.1 and subsequent derivation). The lemma above states that the mid-distribution function formula is, in fact, valid for arbitrary distributions.

Lemma 5. *(Conditional Gini) Let Y be a random variable satisfying $Y \geq 0$ a.s. and let X be a $p \times 1$ random vector with law P_X. Let $M(y|x)$ be the CMDF of Y given $X = x$ with $0 < \mathbb{E}[Y|X = x] < \infty$ for P_X-almost every x. The conditional Gini of Y given $X = x$ satisfies*

$$\mathbb{G}(Y|X = x) = \frac{\mathbb{E}[Y(2M(Y|X = x) - 1)|X = x]}{\mathbb{E}[Y|X = x]} = \frac{\mathbb{E}[|Y_1 - Y_2||X = x]}{\mathbb{E}[Y_1 + Y_2|X = x]}$$

$$= 2\frac{\mathbb{C}ov(Y, M(Y|X = x)|X = x)}{\mathbb{E}[Y|X = x]} \in [0, 1)$$

for P_X-almost every x, where Y_1 and Y_2 are conditionally independent copies of Y given $X = x$, that is, $P(Y_1 \leq y_1, Y_2 \leq y_2|X = x) = P(Y_1 \leq y_1|X = x)P(Y_2 \leq y_2|X = x) = F(y_1|x)F(y_2|x)$; and $M(y|X = x) := M(y|x) = \big(F(y|x) + F(y^-|x)\big)/2$ is the conditional mid-distribution function of Y given $X = x$, where $F(y^-|x) = P(Y < y|X = x) = \lim_{z \uparrow y} F(z|x)$ and $\mathbb{C}ov(\cdot, \cdot|X = x)$ denotes the conditional covariance given $X = x$.

Proof. The proof follows by fixing x and applying Lemma 4 to the distribution $F(y|x)$. □

B Loss Functions for Ratios of Expectations and the Conditional Gini Function

In this section, we show how the loss function L^G –used to predict $(2M(Y|X) - 1)$– in (19) relates to the family of loss functions for ratios of expectations –used to predict Y– derived in Gneiting (2011a). Gneiting derives a family of loss functions that yield a ratio of expectations as the solution to an expected loss minimization problem. Let $L : A \times A \to \mathbb{R}$ be a loss function defined by $(y, a) \mapsto L(y, a)$ used to evaluate predictions for $Y \geq 0$ *a.s.* given some point forecast $a \in A \subseteq \mathbb{R}$, so that $L(y, a) = 0$ if $y = a$.

Theorem 8 in Gneiting (2011a) shows that statistical functionals[22] of the form

$$T(F) = \frac{\mathbb{E}_F[r(Y)]}{\mathbb{E}_F[s(Y)]} \tag{26}$$

[22] Formally, these are the mappings $T : \mathscr{F} \to \mathscr{P}(D)$ defined by $F \mapsto T(F) \subseteq D$, where $\mathscr{F}$ is the class of probability measures on $D \subseteq \mathbb{R}$ and $\mathscr{P}(D)$ is the power set of D.

minimize a class of loss functions, where $r:\mathbb{R}\to\mathbb{R}$ and $s:\mathbb{R}\to(0,\infty)$ are two measurable functions, in an expected minimization framework. That class of loss functions is given by

$$L^R(y,a) = s(y)(\phi(y)-\phi(a)) - \phi'(a)\big(r(y)-as(y)\big) + \phi'(y)\big(r(y)-ys(y)\big) \tag{27}$$

where ϕ is some convex function with subgradient ϕ'. Gneiting shows that

$$T(F) \;=\; \underset{a\in A}{\operatorname{argmin}}\; \mathbb{E}_F\left[L^R(Y,a)\right] \tag{28}$$

for all probability distributions[23] F on $A\subseteq\mathbb{R}_+$ (uniquely if ϕ is strictly convex).

Noting that the unconditional Gini of a distribution can be expressed as the ratio of expectations in (1), and setting $r(y)=y(2M(y)-1)$, $s(y)=y$, and $\phi(u)=u^2$ (strictly convex)[24] in the loss function (27), we can obtain the unconditional Gini $T(F)=\mathbb{G}(Y)$ by solving the problem in (28) –and requiring the existence of the necessary moments. This is just a particular case of the result in Gneiting (2011a), treating M as a known function. We can also extend it to the case of the conditional mid-distribution function of Y given $X=x$, $M(y|x)$, by defining the following loss function

$$\tilde{L}^G(y,x,a) = ya^2 - 2r(y,x)a + 2yr(y,x) - y^3 \tag{29}$$

where $r(y,x)=y(2M(y|x)-1)$ and $\tilde{L}^G=0$ if $y=a$. This loss function is a particular case of L^R for fixed x, $s(y)=y$, and $\phi(u)=u^2$. Define $s(x):=\mathbb{E}[Y|X=x]$ and $t(x):=\mathbb{E}[Y(2M(Y|x)-1)|X=x]$. Appendix F shows that the conditional Gini minimizes $\tilde{L}^G$ in a conditional expected loss minimization problem for P_X-almost every x.

We can also show that the CGF solves the unconditional expected loss minimization problem with loss function $\tilde{L}^G$.

Theorem 5. *(Conditional Gini Function via Expected Loss Minimization) Let $Y\geq 0$ a.s. be a random variable on $(\Omega,\mathscr{F},\mathbb{P})$ with $0<\mathbb{E}[Y]<\infty$, $0<\mathbb{E}[Y|X]<\infty$ a.s., $\mathbb{E}\left[Y^3\right]<\infty$ and let X be a measurable function from $(\Omega,\mathscr{F},\mathbb{P})$ to $(\Psi,\mathscr{G})$. Let $h(X)$ be a predictor, i.e., $h\in\aleph$, where $\aleph$ is the set of Borel functions $h:\mathscr{X}\to\mathbb{R}$ such that $\mathbb{E}\left[Yh^2(X)\right]<\infty$, and $\tilde{L}^G$ is the loss*

[23] Such that under F the moments $\mathbb{E}_F[r(Y)]$, $\mathbb{E}_F[s(Y)]>0$, $\mathbb{E}_F[r(Y)\phi'(Y)]$, $\mathbb{E}_F[s(Y)\phi(Y)]$, and $\mathbb{E}_F[Ys(Y)\phi'(Y)]$ exist and are finite.

[24] The choice of ϕ does not matter in terms of the functional minimizing the loss function, but it does determine uniqueness and can require the existence of more complicated moments. For example, the choice $\phi(u)=u^2$ in the class of Bregman loss functions yields the MSE and only requires the existence of second moments. See Gneiting (2011a).

function defined in (29). Then

$$\mathbb{G}(Y|X) = \underset{h \in \aleph}{\arg\min}\ \mathbb{E}\left[\tilde{L}^G(Y, X, h(X))\right] \ \ a.s. \tag{30}$$

Proof. See Appendix F.

□

If $\aleph$ is restricted to be the class of linear measurable functions indexed by $\theta \in \Theta \subseteq \mathbb{R}^p$ such that $h_\theta(x) = x'\theta \in \aleph$, and minimization in (30) takes place over the parameter space Θ, the first-order condition coincides with the moment condition in (7), i.e., $\mathbb{E}[XU] = 0$. This coincidence motivates the choice of ϕ in the definition of $\tilde{L}^G$ (along with the moment requirements). Likewise, for a general parametric model $h_\theta(x) = m(x;\theta)$ smooth enough in θ, the first-order condition with respect to θ coincides with the classical orthogonality condition $\mathbb{E}[m_\theta(X;\theta_0)U] = 0$ where $m_\theta(x;\theta_0) = \partial_\theta m(x;\theta)|_{\theta=\theta_0}$ is the score of the model evaluated at the true parameter value.

The loss function $\tilde{L}^G$ and its expected value can be negative even though it is minimized at the Gini of the conditional distribution of Y. Shifting the values of the loss function by exactly the amount of the loss evaluated at the pointwise minimizer over a in (29), that is, by the amount $\tilde{L}^G_{\min}(y,x) = -r(y,x)^2/y + 2yr(y,x) - y^3$, we get the loss function L^G defined in (19). This shift does not change the minimizer $\mathbb{G}(Y|X)$. Heuristically, L^G evaluates predictions for $(2M(y|x) - 1)$ and $\tilde{L}^G$ evaluates predictions for Y. Moreover, Theorem 5 requires a finite third moment for Y, while Theorem 4 only requires a finite first moment. This shows how the GIR loss function in (19) relates to the loss functions for ratios of expectations derived in Gneiting (2011a).

C Proof of Proposition 3 and Lemmas 1 – 3

C.1 *Proof of Proposition 3*

The estimand is

$$\begin{aligned}
\gamma_{01} &= \mathbb{E}\left[Y X_1 X_1'\right]^{-1} \mathbb{E}\left[X_1 Y (2M(Y|X_1) - 1)\right] \\
&= \mathbb{E}\left[X_1 X_1' Y\right]^{-1} \mathbb{E}\left[X_1 Y (2M(Y|X) - 1)\right] + 2\mathbb{E}\left[X_1 X_1' Y\right]^{-1} \mathbb{E}\left[X_1 Y (M(Y|X_1) - M(Y|X))\right] \\
&= \theta_{01} + \mathbb{E}\left[X_1 X_1' Y\right]^{-1} \mathbb{E}\left[X_1 Y R(X)\right] + 2\mathbb{E}\left[X_1 X_1' Y\right]^{-1} \mathbb{E}\left[X_1 Y (M(Y|X_1) - M(Y|X))\right]
\end{aligned}$$

where the last equality uses the equation and conditional moment restriction in (12).

□

C.2 *Proof of Lemma 1*

Let $A_0 := \mathbb{E}[YXX']$ be the GIR $p \times p$ design matrix and $\theta_0 := \left(\theta_{01}', 0\right)' \in \mathbb{R}^p$. Condition (ii) gives $\mathbb{G}(Y|X) = X_1'\theta_{01}$. By the definition of conditional Gini and (ii), we have

$$\mathbb{E}[XY(2M(Y|X)-1)] = \mathbb{E}[\mathbb{E}[Y|X]X\mathbb{G}(Y|X)] = \mathbb{E}[YXX_1']\theta_{01}$$

On the other hand, simple block matrix multiplication gives $A_0\theta_0 = \mathbb{E}[YXX_1']\theta_{01}$, so that $\mathbb{E}[XY(2M(Y|X)-1)] = A_0\theta_0$. By nonsingularity of A_0 in Assumption 3,

$$\gamma_0 = A_0^{-1}\mathbb{E}[XY(2M(Y|X)-1)] = A_0^{-1}A_0\theta_0 = \theta_0$$

Hence, $\gamma_{01} = \theta_{01}$ and $\gamma_{02} = 0$.

□

C.3 Proof of Lemma 2

The first-stage OLS solution at the threshold y is the p-dimensional coefficient vector

$$\hat{\beta}(y) = \left(\sum_{l=1}^{n} X_l X_l'\right)^{-1}\left(\sum_{l=1}^{n} X_l \psi_y(Y_l)\right)$$

Since the regressors $X_l = \left(X_{l1}, \ldots, X_{lp}\right)'$ are mutually exclusive dummies $X_{lg} = \mathbb{1}_{\{D_l = g\}}$, where $g \in \{1, \ldots, p\}$, we have $\sum_{l=1}^{n} X_l X_l' = \mathrm{diag}(n_1, \ldots, n_p)$ and $\sum_{l=1}^{n} X_{lg}\psi_y(Y_l) = \sum_{l \in I_g}\psi_y(Y_l)$, since $X_{lg} = 1$ whenever $l \in I_g$. The g-th coefficient is thus

$$\hat{\beta}_g(y) = n_g^{-1}\sum_{l \in I_g}\psi_y(Y_l) =: \hat{M}_g(y)$$

which is exactly the sample mid-distribution function of Y within group g. Hence, we have $\hat{M}(Y_i|X_i) = X_i'\hat{\beta}(Y_i) = \hat{M}_g(Y_i)$, $i \in I_g$. Regarding the second stage, note that $\sum_{i=1}^{n} X_i X_i' Y_i = \mathrm{diag}(S_1, \ldots, S_p)$, where $S_g = \sum_{i \in I_g} Y_i$, and the numerator's g-th component is given by $\sum_{i=1}^{n} X_{ig} Y_i \left(2\hat{M}_g(Y_i) - 1\right) = \sum_{i \in I_g} Y_i \left(2\hat{M}_g(Y_i) - 1\right)$. Therefore, the g-th coefficient of the

GIR estimator satisfies

$$\hat{\theta}_{ng} = \frac{\sum_{i\in I_g} Y_i \left(2\hat{M}_g(Y_i) - 1\right)}{S_g} = \frac{n_g^{-1}\sum_{i\in I_g} Y_i \left(2\hat{M}_g(Y_i) - 1\right)}{\bar{Y}_g}$$

Note that $2\hat{M}_g(Y_i) - 1 = n_g^{-1}\sum_{j\in I_g}\left(2\mathbb{1}_{\{Y_j<Y_i\}} + \mathbb{1}_{\{Y_j=Y_i\}} - 1\right) = n_g^{-1}\sum_{j\in I_g}\text{sgn}(Y_i - Y_j)$, so the numerator can be written as $n_g^{-1}\sum_{i\in I_g} Y_i \left(2\hat{M}_g(Y_i) - 1\right) = n_g^{-2}\sum_{i\in I_g}\sum_{j\in I_g} Y_i \,\text{sgn}(Y_i - Y_j) =: A_g$. Finally, note that

$$\begin{aligned}
2A_g &= n_g^{-2}\sum_{i\in I_g}\sum_{j\in I_g} Y_i \,\text{sgn}(Y_i - Y_j) + n_g^{-2}\sum_{j\in I_g}\sum_{i\in I_g} Y_j \,\text{sgn}(Y_j - Y_i) \\
&= n_g^{-2}\sum_{i\in I_g}\sum_{j\in I_g} Y_i \,\text{sgn}(Y_i - Y_j) - n_g^{-2}\sum_{j\in I_g}\sum_{i\in I_g} Y_j \,\text{sgn}(Y_i - Y_j) \\
&= n_g^{-2}\sum_{i\in I_g}\sum_{j\in I_g} (Y_i - Y_j)\,\text{sgn}(Y_i - Y_j) = n_g^{-2}\sum_{i\in I_g}\sum_{j\in I_g} |Y_i - Y_j|
\end{aligned}$$

where $\text{sgn}(u) = \mathbb{1}_{\{u>0\}} - \mathbb{1}_{\{u<0\}}$ is the sign function. Hence, for $g \in \{1,\dots,p\}$

$$\hat{\theta}_{ng} = \frac{\sum_{i\in I_g}\sum_{j\in I_g} |Y_i - Y_j|}{2n_g^2\bar{Y}_g}$$

□

C.4 Proof of Lemma 3

Let $c \in \mathbb{R}$ such that $c > 0$. The first stage remains unchanged since

$$\psi_{cy}(cY_i) = \mathbb{1}_{\{cY_i<cy\}} + \frac{1}{2}\mathbb{1}_{\{cY_i=cy\}} = \mathbb{1}_{\{Y_i<y\}} + \frac{1}{2}\mathbb{1}_{\{Y_i=y\}} = \psi_y(Y_i)$$

So that $\tilde{\beta}_K(cy) = \hat{\beta}_K(y)$, where $\tilde{\beta}_K(cy)$ is the first stage coefficient estimated from $\{\tilde{Y}_i, X_i\}_{i=1}^n$ with corresponding fitted values denoted by $\tilde{M}_n(\tilde{Y}_i|X_i)$, and $\tilde{M}_n(cY_i|X_i) = q_K(X_i)'\tilde{\beta}_K(cY_i) = q_K(X_i)'\hat{\beta}_K(Y_i) = \hat{M}_n(Y_i|X_i)$. Finally,

$$\begin{aligned}
\tilde{\theta}_n &= \left(\sum_{i=1}^n X_i X_i' cY_i\right)^{-1}\left(\sum_{i=1}^n X_i cY_i \left(2\tilde{M}_n(cY_i|X_i) - 1\right)\right) \\
&= \left(\sum_{i=1}^n X_i X_i' Y_i\right)^{-1}\left(\sum_{i=1}^n X_i Y_i \left(2\hat{M}_n(Y_i|X_i) - 1\right)\right) = \hat{\theta}_n
\end{aligned}$$

□

D Proof of Theorem 2

Before proving the theorem, we first clarify the notation and definitions to be used throughout the proof.

D.1 *Definitions and Notation*

Let $Z_i = (Y_i, X_i)$, $i \in \{1, \dots, n\}$ be i.i.d., with law P and $Y_i \in \mathscr{Y}$, where $\mathscr{Y}$ is either a compact interval $\mathscr{Y} = [\underline{y}, \bar{y}]$, where $0 \leq \underline{y} < \bar{y} < \infty$ (continuous regime); or a finite set $\mathscr{Y} = \{y_1, \dots, y_J\}$, where $0 \leq y_1 < \cdots < y_J < \infty$ and $J < \infty$ (discrete regime), following Assumption 5-(i). Normalize the support of X to the unit rectangle $X_i \in \mathscr{X} := [0,1]^p$ and define $\mathscr{Y}\mathscr{X} := \mathscr{Y} \times \mathscr{X}$, a compact subset of $\mathbb{R}^{p+1}$. In either regime, let $f(y|x)$ denote the Radon-Nikodym derivative of the conditional law of Y given $X = x$ with respect to the measure ν, so that, for every measurable set $A \subseteq \mathscr{Y}$, $P(Y \in A | X = x) = \int_A f(t|x)\, d\nu(t)$, where ν is the Lebesgue measure on $\mathscr{Y}$ in the continuous regime and the counting measure on $\mathscr{Y}$ in the discrete regime. Notice that $\nu(\mathscr{Y}) < \infty$ in both regimes. Assumption 5-(i) is understood to require $0 \leq f(y|x) \leq C_f < \infty$ uniformly over $(y,x) \in \mathscr{Y}\mathscr{X}$. For $y, t \in \mathscr{Y}$, define $\psi_y(t) := \mathbb{1}_{\{t<y\}} + \frac{1}{2}\mathbb{1}_{\{t=y\}}$, $F(y|x) := P(Y \leq y | X = x)$, $F(y^-|X = x) := P(Y < y | X = x)$, and the conditional mid-distribution function (CMDF):

$$M(y|x) := \mathbb{E}[\psi_y(Y)|X = x] = F(y^-|X = x) + \frac{1}{2}P(Y = y|X = x) = \int_{\mathscr{Y}} \psi_y(t) f(t|x)\, d\nu(t)$$

In the continuous regime $M(y|x) = F(y|x)$, and in the discrete regime $M(y_j|x) = \sum_{l<j} P(Y = y_l|X = x) + \frac{1}{2}P(Y = y_j|X = x)$. In either regime, $0 \leq \psi_y(t) \leq 1$ and $0 \leq M(y|x) \leq 1$. Every integral with respect to ν is understood as an ordinary Lebesgue integral in the continuous regime and a finite sum in the discrete regime.

We use standard empirical processes notation. Let P denote the law of $Z := (Y, X')' \in \mathbb{R}^{p+1}$ and P_X the law of X. For a measurable function f on $\mathscr{Y}\mathscr{X}$ (either scalar-, vector-, or matrix-valued), we write

$$P_n f = P_n[f] := \int f(z)\, dP_n = \frac{1}{n}\sum_{i=1}^n f(Z_i) \qquad Pf = P[f] := \int f(z)\, dP = \mathbb{E}[f(Z)]$$

And thus $L_2(P)$ denotes the space of square integrable functions with respect to the law P. For two sequences of positive numbers b_{1n} and b_{2n}, the notation $b_{1n} \asymp b_{2n}$ means that the ratio b_{1n}/b_{2n} is bounded below and above by positive constants that are independent of n. Similarly, the notation $b_{1n} \lesssim b_{2n}$ denotes $b_{1n} \leq c b_{2n}$ for some constant $c > 0$ that

does not depend on n. Euclidean vector norms and the induced matrix operator norm are both denoted by $\|\cdot\|$; $\|A\|_F = \{\text{tr}(A'A)\}^{1/2}$ is the Frobenius norm; $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote the minimum and maximum eigenvalues of A, respectively. Finally, $\|\cdot\|_\infty$ denotes the supremum norm. All logarithms, denoted by log, are natural logarithms (base e).

Choose, from Assumption 5-(iii), a K-dimensional sieve tensor-product basis $q_K(x) := (q_{1K}(x),\ldots,q_{KK}(x))' \in \mathbb{R}^K$, and define the series estimator $\hat{M}_n(y|x) = q_K(x)'\hat{\beta}_K(y)$, where $\hat{\beta}_K(y) = \hat{Q}_K^{-1} P_n[q_K \psi_y]$ with $\hat{Q}_K = P_n[q_K q_K']$ being the sample Gram matrix. Define also the population projection $M_K(y|x) = q_K(x)'\beta_K(y)$, where $\beta_K(y) = Q_K^{-1}\mathbb{E}[q_K(X)\psi_y(Y)]$ and $Q_K := \mathbb{E}[q_K(X)q_K(X)']$ denotes the population Gram matrix. Throughout, we use $\theta_0 = \mathbb{E}\left[YXX'\right]^{-1}\mathbb{E}[XY(2M(Y|X)-1)]$ for the estimand and denote the estimator by

$$\hat{\theta}_n = \left(\frac{1}{n}\sum_{i=1}^n Y_i X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\left(2\hat{M}_n(Y_i|X_i)-1\right)\right)$$

Define $a(y,x) := xy$, $\phi(y,x) := xy\{2M(y|x)-1\}$, and $\hat{\phi}(y,x) := xy\{2\hat{M}_n(y|x)-1\}$ for the GIR estimand and estimator numerators, so $a,\phi,\hat{\phi} \in \mathbb{R}^p$. Moreover, fix $v \in \mathbb{R}^p$ and define $a_v(y,x) := yv'x$, $\phi_v(y,x) := a_v(y,x)\{2M(y|x)-1\}$, and $\hat{\phi}_v(y,x) = a_v(y,x)\{2\hat{M}_n(y|x)-1\}$.

Let $\mathscr{H}_K := \text{span}\{q_{1K},\ldots,q_{KK}\}$ be the sieve space and let the population $L_2(P_X)$-projection be defined by the map $P_K : L_2(P_X) \to \mathscr{H}_K$ such that $(P_K h)(x) = q_K(x)'Q_K^{-1}\mathbb{E}[q_K(X)h(X)]$. For every fixed $v \in \mathbb{R}^p$, define $\delta_v(y,x) := y(v'x)f(y|x)$, $\alpha_{v,K}(y,x) := [P_K\delta_v(y,\cdot)](x)$ written $\alpha_{v,K}(y,x) = q_K(x)'\eta_{v,K}(y)$, and $M_K(y|x) := [P_K M(y|\cdot)](x)$. Define the norm for the projection operator $\|P_K\|_{\infty\to\infty} := \sup_{\|h\|_\infty \le 1}\|P_K h\|_\infty$. Finally, we define the specific notion of smoothness that is used, which is standard in the series literature (Chen, 2007).

Let $g : \mathscr{X} \to \mathbb{R}$ and let $0 < \gamma \le 1$. A real-valued function g on $\mathscr{X}$ is said to satisfy a Hölder condition with exponent γ if there exists a positive number c such that $|g(x)-g(y)| \le c\|x-y\|_2^\gamma \ \forall x,y \in \mathscr{X}$. Given a p-tuple of non-negative integers $\alpha = (\alpha_1,\ldots,\alpha_p)$, let $[\alpha] = \alpha_1 + \cdots + \alpha_p$ and let D^α denote the differential operator defined by

$$D^\alpha := \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1}\ldots\partial x_p^{\alpha_p}}$$

Let r be a non-negative integer and set $s = r+\gamma$. A real-valued function g on $\mathscr{X}$ is said to be s-smooth if it is r-times continuously differentiable on $\mathscr{X}$, i.e., $g \in C^r(\mathscr{X})$, and $D^\alpha g$ satisfies a Hölder condition with exponent γ for all α with $[\alpha] = r$.

Denote the class of s-smooth real-valued functions on $\mathscr{X}$ by $\Lambda^s(\mathscr{X})$ (Hölder class). Define

a Hölder ball of finite radius

$$\Lambda_c^s(\mathscr{X}) = \left\{ g \in C^r(\mathscr{X}) : \sup_{[\alpha]\le r} \|D^\alpha g\|_\infty \le c, \sup_{[\alpha]=r} [D^\alpha g]_\gamma \le c \right\}$$

where

$$[D^\alpha g]_\gamma = \sup_{x\neq x'} \frac{|D^\alpha g(x) - D^\alpha g(x')|}{\|x - x'\|_2^\gamma}$$

A norm on $\Lambda^s(\mathscr{X})$ is $\|g\|_{\Lambda^s} := \sum_{[\alpha]\le r} \|D^\alpha g\|_\infty + \sum_{[\alpha]=r} [D^\alpha g]_\gamma$.

D.2 *Implied Rates, Smoothness, and Uniform Approximation Bounds*

We now state and prove a lemma including the main implications of Assumption 5 that will be used routinely throughout the proof. Some items in the lemma are well-known results in the sieve literature or in approximation theory, but are stated for completeness.

Lemma 6. *(Implications of Assumptions 4 and 5) Under Assumptions 4 and 5 we have:*

1. $\exists\, d_B, D_B \in \mathbb{R}$ *such that* $0 < d_B \le \lambda_{min}(Q_K) \le \lambda_{max}(Q_K) \le D_B < \infty$.

2. $\mathbb{E}\|q_K(X)\|^2 = \mathrm{tr}(Q_K) \asymp K$.

3. *The sample Gram matrix satisfies*

$$\|\hat{Q}_K - Q_K\| = O_p\left(\xi_K \sqrt{\frac{\log K}{n}}\right), \qquad \|\hat{Q}_K^{-1}\| = O_p(1)$$

   Moreover,

$$\|\hat{Q}_K^{-1} - Q_K^{-1}\| = O_p\left(\xi_K \sqrt{\frac{\log K}{n}}\right) = o_p(1).$$

4. *For every* $g \in \Lambda^s(\mathscr{X})$*, we have that*

$$\inf_{h\in\mathscr{H}_K} \|g - h\|_\infty \lesssim K^{-\bar{s}/p} \|g\|_{\Lambda^s} \;\; \textit{and} \;\; \inf_{h\in\mathscr{H}_K} \|g - h\|_{L_2(P_X)} \lesssim K^{-\bar{s}/p} \|g\|_{\Lambda^s}.$$

5. *Uniform Hölder smoothness:*

$$\sup_{y\in\mathscr{Y}} \|M(y|\cdot)\|_{\Lambda^s} < \infty, \qquad \sup_{y\in\mathscr{Y}} \|\delta_v(y,\cdot)\|_{\Lambda^s} \lesssim \|v\| < \infty.$$

6. *Uniform approximation bounds:*

(i) $\sup_{y\in\mathscr{Y}} \|M(y|\cdot) - M_K(y|\cdot)\|_\infty \lesssim l_K K^{-\bar{s}/p}$.

(ii) $\sup_{y\in\mathscr{Y}} \|M(y|\cdot) - M_K(y|\cdot)\|_{L_2(P_X)} \lesssim K^{-\bar{s}/p}$.

(iii) $\sup_{y\in\mathscr{Y}} \|\delta_v(y,\cdot) - \alpha_{v,K}(y,\cdot)\|_{L_2(P_X)} \lesssim \|v\| K^{-\bar{s}/p}$.

and $\exists K_0 \in \mathbb{N}$ *such that* $\forall K \geq K_0$ *we have that*

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |M_K(y|x) - M(y|x)| < 1, \qquad |M_K(y|x)| \leq 2 \qquad \forall (y,x) \in \mathscr{Y}\mathscr{X}.$$

Proof. We prove the Lemma statement by statement.

1. Immediate by Assumption 5-(iii).

2. $\mathbb{E}\|q_K(X)\|^2 = \text{tr}(Q_K) = \sum_{j=1}^K \lambda_j(Q_K)$, where λ_j denotes the $j-th$ eigenvalue of Q_K. By (1) we have $K d_B \leq \sum_{j=1}^K \lambda_j(Q_K) \leq K D_B$. Hence, $\mathbb{E}\|q_K(X)\|^2 \asymp K$.

3. Apply Lemma 6.2 in Belloni et al. (2015) (which is a version of Rudelson's LLN for matrices) with $Q_i = q_K(X_i)q_K(X_i)'$ so that $Q = \mathbb{E}[Q_i] = Q_K$, $n^{-1}\sum_{i=1}^n Q_i = \hat{Q}_K$, and $M = \xi_K^2$. Since 5-(i) and 5-(iii) imply $\|Q_K\| = O(1)$, and 5-(iv)-(a) gives $\xi_K\sqrt{\log(K)/n} = o(1)$, the Lemma yields

$$\mathbb{E}\|\hat{Q}_K - Q_K\| \lesssim \xi_K^2 \frac{\log K}{n} + \xi_K\sqrt{\frac{\log K}{n}} \lesssim \xi_K\sqrt{\frac{\log K}{n}}$$

and Markov's inequality gives

$$\|\hat{Q}_K - Q_K\| = O_p\left(\xi_K\sqrt{\frac{\log(K)}{n}}\right)$$

which by 5-(iv) $\|\hat{Q}_K - Q_K\| = o_p(1)$. For the second statement, define the event $E_n := \left\{\|\hat{Q}_K - Q_K\| \leq \frac{d_B}{2}\right\}$. Since $\|\hat{Q}_K - Q_K\| = o_p(1)$ we have that $\mathbb{P}(E_n) \to 1$ as $n \to \infty$. Note that for any $u \in \mathbb{R}^K$ with $\|u\| = 1$ we have that

$$\begin{aligned} u'\hat{Q}_K u &= u'Q_K u + u'(\hat{Q}_K - Q_K)u \geq \inf_{\|u\|=1} u'Q_K u - |u'(\hat{Q}_K - Q_K)u| \\ &\geq \lambda_{min}(Q_K) - \|\hat{Q}_K - Q_K\| \geq d_B - \|\hat{Q}_K - Q_K\| \\ &\geq \frac{d_B}{2} \quad \text{with probability approaching one (wpa 1).} \end{aligned}$$

which taking infimum over the unit-norm vector u implies $\lambda_{min}(\hat{Q}_K) \geq d_B/2$ with probability tending to one (so the inverse exists wpa 1). Hence $\|\hat{Q}_K^{-1}\| = \lambda_{max}(\hat{Q}_K^{-1}) =$

$\frac{1}{\lambda_{min}(\hat{Q}_K)} \leq 2/d_B$ with probability tending to one. This implies $\|\hat{Q}_K^{-1}\| = O_p(1)$. Then,

$$\begin{aligned}\|\hat{Q}_K^{-1} - Q_K^{-1}\| &= \|\hat{Q}_K^{-1}(Q_K - \hat{Q}_K)Q_K^{-1}\| \leq \|\hat{Q}_K^{-1}\|\|\hat{Q}_K - Q_K\|\|Q_K^{-1}\| \\ &\lesssim \|\hat{Q}_K^{-1}\|\|\hat{Q}_K - Q_K\| = O_p\left(\xi_K\sqrt{\frac{\log(K)}{n}}\right)\end{aligned}$$

4. Item 4 follows from standard approximation results for orthogonal polynomial, spline/B-spline, and wavelet bases; see Chen (2007). Rates are given for the sup-norm of the approximation error, but $\|\cdot\|_{L_2(P_X)} \leq \|\cdot\|_\infty$.

5. Because $M(y|x) = \int_{\mathscr{Y}} \psi_y(t) f(t|x)\, d\nu(t)$ for every multi-index α with $[\alpha] \leq r$, $D_x^\alpha M(y|x) = \int_{\mathscr{Y}} \psi_y(t) D_x^\alpha f(t|x)\, d\nu(t)$, In the continuous regime, differentiation under the integral sign is allowed since 5-(ii) provides a ν-integrable uniform bound for $D_x^\alpha f(t|x)$ on $\mathscr{Y}$. In the discrete regime, the integral is a finite sum, and differentiation is term-by-term. Since $|\psi_y(t)| \leq 1$,

$$\sup_{y\in\mathscr{Y}} \|D_x^\alpha M(y|\cdot)\|_\infty \leq \nu(\mathscr{Y}) \sup_{t\in\mathscr{Y}} \|D_x^\alpha f(t\,|\,\cdot)\|_\infty$$

and similarly, when $[\alpha] = r$

$$\begin{aligned}|D_x^\alpha M(y|x) - D_x^\alpha M(y|x')| &\leq \int_{\mathscr{Y}} |D_x^\alpha f(t|x) - D_x^\alpha f(t|x')|\, d\nu(t) \\ [D_x^\alpha M(y|\cdot)]_\gamma &\leq \int_{\mathscr{Y}} [D_x^\alpha f(t|\cdot)]_\gamma\, d\nu(t) \leq \nu(\mathscr{Y}) \sup_{t\in\mathscr{Y}} [D_x^\alpha f(t|\cdot)]_\gamma < \infty \text{ by 5-}(ii)\end{aligned}$$

So $\sup_{y\in\mathscr{Y}} \|M(y|\cdot)\|_{\Lambda^s} < \infty$. For fixed v, the map $x \mapsto v'x$ is smooth and bounded on $[0,1]^p$. Since Hölder classes are stable under multiplication by bounded smooth functions, we also have that $\sup_{y\in\mathscr{Y}} \|\delta_v(y,\cdot)\|_{\Lambda^s} \lesssim \|v\| < \infty$.

6. Since by item 5, $M(y|\cdot) \in \Lambda^s(\mathscr{X})$ for every $y \in \mathscr{Y}$, and $M_K \in \mathscr{H}_K$, by definition of orthogonal $L_2(P_X)$-projection and item 4, we have that

$$\|M(y|\cdot) - M_K(y|\cdot)\|_{L_2(P_X)} \leq \inf_{h\in\mathscr{H}_K} \|M(y|\cdot) - h\|_{L_2(P_X)} \lesssim \|M(y|\cdot)\|_{\Lambda^s} K^{-\bar{s}/p}$$

By the same arguments and since $\alpha_{v,K}(y,\cdot)$ is the $L_2(P_X)$-projection of $\delta_v(y,\cdot)$ onto $\mathscr{H}_K$, we also have

$$\|\delta_v(y,\cdot) - \alpha_{v,K}(y,\cdot)\|_{L_2(P_X)} \leq \inf_{h\in\mathscr{H}_K} \|\delta_v(y,\cdot) - h\|_{L_2(P_X)} \lesssim \|\delta_v(y,\cdot)\|_{\Lambda^s} K^{-\bar{s}/p}$$

By item 5, $\sup_{y\in\mathscr{Y}}\|M(y|\cdot)\|_{\Lambda^s}<\infty$, so $\{M(y|\cdot):y\in\mathscr{Y}\}$ lies in a single Hölder ball, so the constant can be taken uniformly over y. The same arguments apply to δ_v so that:

$$\begin{aligned}\sup_{y\in\mathscr{Y}}\|M(y|\cdot)-M_K(y|\cdot)\|_{L_2(P_X)} &\le \sup_{y\in\mathscr{Y}}\inf_{h\in\mathscr{H}_K}\|M(y|\cdot)-h\|_{L_2(P_X)}\\ &\lesssim \sup_{y\in\mathscr{Y}}\|M(y|\cdot)\|_{\Lambda^s}K^{-\bar{s}/p}\\ \sup_{y\in\mathscr{Y}}\|\delta_v(y,\cdot)-\alpha_{v,K}(y,\cdot)\|_{L_2(P_X)} &\le \sup_{y\in\mathscr{Y}}\inf_{h\in\mathscr{H}_K}\|\delta_v(y,\cdot)-h\|_{L_2(P_X)}\\ &\lesssim \sup_{y\in\mathscr{Y}}\|\delta_v(y,\cdot)\|_{\Lambda^s}K^{-\bar{s}/p}\lesssim\|v\|K^{-\bar{s}/p}\end{aligned}$$

We now prove the sup-norm bound for $M-M_K$. First note that $M_K(y|\cdot)=P_KM(y|\cdot)$ and $P_Kh=h$ for all $h\in\mathscr{H}_K$, so that $M(y|\cdot)-M_K(y|\cdot)=(I-P_K)M(y|\cdot)=(I-P_K)(M(y|\cdot)-h)+(I-P_K)h=(I-P_K)(M(y|\cdot)-h)$, where I denotes the identity operator. Taking sup-norm yields $\forall h\in\mathscr{H}_K$

$$\|M(y|\cdot)-M_K(y|\cdot)\|_\infty\le\|I-P_K\|_{\infty\to\infty}\|M(y|\cdot)-h\|_\infty\le l_K\|M(y|\cdot)-h\|_\infty$$

Hence

$$\begin{aligned}\|M(y|\cdot)-M_K(y|\cdot)\|_\infty &\le l_K\inf_{h\in\mathscr{H}_K}\|M(y|\cdot)-h\|_\infty\lesssim l_K\|M(y|\cdot)\|_{\Lambda^s}K^{-\bar{s}/p}\\ \sup_{y\in\mathscr{Y}}\|M(y|\cdot)-M_K(y|\cdot)\|_\infty &\lesssim l_K\sup_{y\in\mathscr{Y}}\|M(y|\cdot)\|_{\Lambda^s}K^{-\bar{s}/p}\lesssim l_KK^{-\bar{s}/p}\end{aligned}$$

Under Assumption 5-(iv) $l_KK^{-\bar{s}/p}\to0$ we have that

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}\left|M(y|x)-M_K(y|x)\right|\to0$$

so by definition of convergence, $\exists K_0$ such that $\forall K\ge K_0$

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}\left|M(y|x)-M_K(y|x)\right|<1$$

and since $M(y|x)$ is a CMDF we have that $\forall(y,x)\in\mathscr{Y}\mathscr{X}$ and $\forall K\ge K_0$

$$|M_K(y|x)|\le\left|M_K(y|x)-M(y|x)\right|+|M(y|x)|<2.$$

□

D.3 *Proof of Theorem 2 – Asymptotic Normality*

Define $A_n := n^{-1}\sum_{i=1}^n X_i X_i' Y_i$, $B_n := n^{-1}\sum_{i=1}^n X_i Y_i (2\hat{M}_n(Y_i|X_i) - 1)$, $A_0 := \mathbb{E}[XX'Y]$, and $B_0 := \mathbb{E}[XY(2M(Y|X) - 1)]$ so that $\hat{\theta}_n = A_n^{-1} B_n$ and $\theta_0 = A_0^{-1} B_0$. Write

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = A_n^{-1}\left\{\sqrt{n}\,(B_n - B_0) - \sqrt{n}\,(A_n - A_0)\,\theta_0\right\}$$

We show that a Central Limit Theorem (CLT) applies to the terms in brackets, and then use Slutsky's Lemma to obtain convergence of the GIR estimator to a zero-mean Gaussian random vector W. To invoke the Cramér-Wold device, fix $v \in \mathbb{R}^p$.

Step 1 ($v'\sqrt{n}\,(B_n - B_0)$ satisfies a CLT). Using the previous definitions, we can rewrite the term as follows

$$\begin{aligned} v'\sqrt{n}\,(B_n - B_0) &= \sqrt{n}(P_n\hat{\phi}_v - P\phi_v) \\ &= \sqrt{n}(P_n - P)(\hat{\phi}_v - \phi_v) + \sqrt{n}(P_n - P)\phi_v + \sqrt{n}P(\hat{\phi}_v - \phi_v) \\ &=: T_{1,v} + T_{2,v} + T_{3,v} \end{aligned}$$

We break down this step into several sub-steps. The goal is to show that $T_{1,v} = o_p(1)$, and that $T_{2,v}$ and $T_{3,v}$ satisfy a CLT.

Step 1.1 ($T_{2,v}$ satisfies a CLT). Note that $T_{2,v} = n^{-1/2}\sum_{i=1}^n v'\{X_i Y_i (2M(Y_i|X_i) - 1) - P\phi\}$, where the term inside the sum has zero mean and $\left|\phi_v(Y,X)\right|^2 = \left|Yv'X\{2M(Y|X) - 1\}\right|^2 \lesssim \|v\|^2 < \infty$ so $\phi_v \in L_2(P)$ and the standard CLT for i.i.d. data applies.

Step 1.2 ($T_{3,v}$ satisfies a CLT). Write

$$\begin{aligned} T_{3,v} &= \sqrt{n}P(\hat{\phi}_v - \phi_v) = 2\sqrt{n}P[a_v(\hat{M}_n - M)] \\ &= 2\sqrt{n}P[a_v(\hat{M}_n - M_K)] + 2\sqrt{n}P[a_v(M_K - M)] =: 2T_{3.1,v} + 2T_{3.2,v} \end{aligned}$$

Note that

$$P[a_v h] = \int_{\mathscr{Y}\mathscr{X}} a_v(y,x)h(y,x)f(y|x)\,dP_X(x)d\nu(y) = \int_{\mathscr{Y}} \mathbb{E}[\delta_v(y,X)h(y,X)]\,d\nu(y)$$

for any integrable h. From Newey (1994a), this verifies (in Newey's notation) for a single projection $y \in \mathscr{Y}$: $\mathbb{E}[D(Z,\tilde{g})] = \mathbb{E}[\delta(X)\tilde{g}(X)]$, where the linearization term $D(Z,\tilde{g}) = 2Y(X'v)\tilde{g}(Y,X)$, $\tilde{g} = h$, and $\delta(X) = 2\delta_v(y,X)$ is the Riesz representer (for fixed y). Integration with respect to ν over $y \in \mathscr{Y}$ combines the family of first-stage projections indexed by $y \in \mathscr{Y}$, which is finite in the discrete regime. Hence, we have the approximation error

term satisfying $T_{3.2,v} = \sqrt{n}\int_{\mathscr{Y}} \mathbb{E}[\delta_v(y,X)(M_K(y|X) - M(y|X))]\, d\nu(y)$. Since $\alpha_{v,K}(y,\cdot) \in \mathscr{H}_K$ and $M_K(y|\cdot) - M(y|\cdot)$ is orthogonal to $\mathscr{H}_K$,

$$T_{3.2,v} = \sqrt{n}\int_{\mathscr{Y}} \mathbb{E}[(\delta_v(y,X) - \alpha_{v,K}(y,X))(M_K(y|X) - M(y|X))]\, d\nu(y)$$

By the triangle and Cauchy-Schwarz inequalities, along with the approximation error bounds from Lemma 6-6

$$\begin{aligned}
\left|T_{3.2,v}\right| &\leq \sqrt{n}\int_{\mathscr{Y}} \|\delta_v(y,X) - \alpha_{v,K}(y,X)\|_{L_2(P_X)} \|M_K(y|X) - M(y|X)\|_{L_2(P_X)} d\nu(y) \\
&\lesssim \sqrt{n}\nu(\mathscr{Y}) \sup_{y\in\mathscr{Y}} \|\delta_v(y,X) - \alpha_{v,K}(y,X)\|_{L_2(P_X)} \sup_{y\in\mathscr{Y}} \|M_K(y|X) - M(y|X)\|_{L_2(P_X)} \\
&\lesssim \|v\| \sqrt{n} K^{-2\bar{s}/p}
\end{aligned}$$

So $\left|T_{3.2,v}\right| \lesssim \sqrt{n}K^{-2\bar{s}/p} \longrightarrow 0$ by Assumption 5-(iv)-(c). This is the product-bias condition discussed in section 6 of Newey (1994a), a feature which is a consequence of the orthogonality of least-squares projection errors. Define the projection residuals $U_y(Z) := (\psi_y(Y) - M_K(y|X))$, which satisfy projection orthogonality $\mathbb{E}[U_y(Z)q_K(X)] = 0$ for each $y \in \mathscr{Y}$. For the estimation error term $T_{3.1,v}$, first note that

$$\hat{Q}_K\left(\hat{\beta}_K(y) - \beta_K(y)\right) = P_n[q_K\psi_y(Y)] - P_n[q_K M_K] = P_n[q_K U_y]$$

Since $\hat{M}_n(y|\cdot) - M_K(y|\cdot), \alpha_{v,K}(y,\cdot) \in \mathscr{H}_K$ and since $\alpha_{v,K}(y,\cdot)$ is the orthogonal projection of $\delta_v(y,\cdot)$ onto $\mathscr{H}_K$ we have that

$$\begin{aligned}
T_{3.1,v} &= \sqrt{n}\int_{\mathscr{Y}} \mathbb{E}[\delta_v(y,X)(\hat{M}_n(y|X) - M_K(y|X))]\, d\nu(y) \\
&= \sqrt{n}\int_{\mathscr{Y}} \mathbb{E}[\alpha_{v,K}(y,X)(\hat{M}_n(y|X) - M_K(y|X))]\, d\nu(y) \\
&= \sqrt{n}\int_{\mathscr{Y}} \eta_{v,K}(y)'\mathbb{E}[q_K(X)(q_K(X)'\hat{\beta}_K(y) - q_K(X)'\beta_K(y))]\, d\nu(y) \\
&= \sqrt{n}\int_{\mathscr{Y}} \eta_{v,K}(y)' Q_K\left(\hat{\beta}_K(y) - \beta_K(y)\right) d\nu(y) \\
&= \sqrt{n}\int_{\mathscr{Y}} \eta_{v,K}(y)'(Q_K - \hat{Q}_K)\left(\hat{\beta}_K(y) - \beta_K(y)\right) + \eta_{v,K}(y)' P_n[q_K U_y]\, d\nu(y) \\
&= \int_{\mathscr{Y}} \eta_{v,K}(y)'\sqrt{n}(Q_K - \hat{Q}_K)\left(\hat{\beta}_K(y) - \beta_K(y)\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \alpha_{v,K}(y,X_i)U_y(Z_i)\, d\nu(y) \\
&=: T^a_{3.1,v} + T^b_{3.1,v}
\end{aligned}$$

By the triangle and Cauchy-Schwarz inequalities for the first term, we have

$$\begin{aligned}
\left|T^a_{3.1,v}\right| &\leq \sqrt{n}\int_{\mathscr{Y}} \left|\eta_{v,K}(y)'(Q_K - \hat{Q}_K)\left(\hat{\beta}_K(y) - \beta_K(y)\right)\right| d\nu(y) \\
&\leq \sqrt{n}\|Q_K - \hat{Q}_K\| \int_{\mathscr{Y}} \|\eta_{v,K}(y)\|\|\hat{\beta}_K(y) - \beta_K(y)\| d\nu(y) \\
&\leq \sqrt{n}\|Q_K - \hat{Q}_K\| \left(\int_{\mathscr{Y}} \|\eta_{v,K}(y)\|^2 d\nu(y)\right)^{1/2} \left(\int_{\mathscr{Y}} \|\hat{\beta}_K(y) - \beta_K(y)\|^2 d\nu(y)\right)^{1/2} \\
&\lesssim \|v\| O_p\left(\sqrt{K}\xi_K\sqrt{\frac{\log(K)}{n}}\right)
\end{aligned}$$

So $\left|T^a_{3.1,v}\right| \underset{P}{\longrightarrow} 0$ by Assumption 5-(iv)-(a). The last step uses the following two arguments. First,

$$d_B\|\eta_{v,K}(y)\|^2 \leq \mathbb{E}[\alpha_{v,K}(y,X)^2] = \eta_{v,K}(y)'Q_K\eta_{v,K}(y) \leq D_B\|\eta_{v,K}(y)\|^2$$

by item 1 of Lemma 6, so that $\mathbb{E}[\alpha_{v,K}(y,X)^2] \asymp \|\eta_{v,K}(y)\|^2$ with $\mathbb{E}[\alpha_{v,K}(y,X)^2] \leq \mathbb{E}[\delta_v(y,X)^2]$ since by definition of orthogonal projection $\mathbb{E}[(\delta_v(y,X) - \alpha_{v,K}(y,X))^2] = \mathbb{E}[\delta_v(y,X)^2] - \mathbb{E}[\alpha_{v,K}(y,X)^2]$. This gives, along with Assumption 5-(i)

$$\int_{\mathscr{Y}} \|\eta_{v,K}(y)\|^2 d\nu(y) \asymp \int_{\mathscr{Y}} \mathbb{E}[\alpha_{v,K}(y,X)^2] d\nu(y) \leq \int_{\mathscr{Y}} \mathbb{E}[\delta_v(y,X)^2] d\nu(y) \lesssim \|v\|^2 < \infty \quad (*)$$

Second, using the OLS normal equations

$$\int_{\mathscr{Y}} \|\hat{\beta}_K(y) - \beta_K(y)\|^2 d\nu(y) \leq \int_{\mathscr{Y}} \left\|\hat{Q}_K^{-1}P_n[q_K U_y]\right\|^2 d\nu(y) \leq \|\hat{Q}_K^{-1}\|^2 \int_{\mathscr{Y}} \left\|P_n[q_K U_y]\right\|^2 d\nu(y)$$

where $\|\hat{Q}_K^{-1}\| = O_p(1)$ by item 3 in Lemma 6 and

$$\mathbb{E}\|P_n[q_K U_y]\|^2 = n^{-1}\mathbb{E}\left[U_y(Z)^2\|q_K(X)\|^2\right] = n^{-1}\mathbb{E}\left[\mathbb{E}[U_y(Z)^2|X]\|q_K(X)\|^2\right]$$

where

$$\begin{aligned}
\mathbb{E}[U_y(Z)^2|X] &= \mathbb{V}ar[\psi_y(Y)|X] + \mathbb{E}[(M(y|X) - M_K(y|X))^2|X] \\
&\leq \frac{1}{4} + \sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |M(y|x) - M_K(y|x)|^2
\end{aligned}$$

By definition of convergence and item 6-(i) in Lemma 6, $\exists K_0 \in \mathbb{N}$ such that $\forall K \geq K_0$ $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |M(y|x) - M_K(y|x)| < 1$ so that $\forall K \geq K_0$ and by item 1 in Lemma 6

$$\mathbb{E}\|P_n[q_K U_y]\|^2 \leq 2n^{-1}\mathbb{E}\|q_K(X)\|^2 = 2n^{-1}\operatorname{tr}(Q_K) = 2n^{-1}\sum_{j=1}^{K}\lambda_j(Q_K) \lesssim \frac{K}{n} \quad (**)$$

Markov's inequality and $\nu(\mathscr{Y}) < \infty$ using the above give

$$\left(\int_{\mathscr{Y}} \|\hat{\beta}_K(y) - \beta_K(y)\|^2 \, d\nu(y)\right)^{1/2} = O_p(1) O_p\left(\sqrt{\frac{K}{n}}\right) \quad (***)$$

Combining $\sqrt{n}\|Q_K - \hat{Q}_K\| = O_p\left(\xi_K \sqrt{\log K}\right)$ from item 3 in Lemma 6 with $(*)$ and $(***)$ gives the stated $\left|T^a_{3.1,\nu}\right| \xrightarrow[P]{} 0$ under Assumption 5-$(i\nu)$-(a). For the term $T^b_{3.1,\nu}$, we prove the applicability of a CLT. Define first

$$\Xi_{\nu,K}(Z_i) := \int_{\mathscr{Y}} \alpha_{\nu,K}(y, X_i)(\psi_y(Y_i) - M_K(y|X_i)) \, d\nu(y)$$
$$\Xi_{\nu}(Z_i) := \int_{\mathscr{Y}} \delta_{\nu}(y, X_i)(\psi_y(Y_i) - M(y|X_i)) \, d\nu(y)$$

and write $\Xi_{\nu,K}(Z_i) - \Xi_{\nu}(Z_i) =: R_{1,\nu}(Z_i) + R_{2,\nu}(Z_i)$, where

$$R_{1,\nu}(Z_i) := \int_{\mathscr{Y}} (\alpha_{\nu,K}(y, X_i) - \delta_{\nu}(y, X_i))(\psi_y(Y_i) - M_K(y|X_i)) \, d\nu(y)$$
$$R_{2,\nu}(Z_i) := \int_{\mathscr{Y}} \delta_{\nu}(y, X_i)(M(y|X_i) - M_K(y|X_i)) \, d\nu(y)$$

and notice that by item 6 in Lemma 6 $\forall K \geq K_0 \left|\psi_y(Y_i) - M_K(y|X_i)\right| \leq 3$ and thus, by the Cauchy-Schwarz inequality, $\nu(\mathscr{Y}) < \infty$, and item 6 in Lemma 6

$$\begin{aligned}
\mathbb{E}[R_{1,\nu}(Z)^2] &= \mathbb{E}\left[\left\{\int_{\mathscr{Y}} (\alpha_{\nu,K} - \delta_{\nu})(y, X) U_y(Z) \, d\nu(y)\right\}^2\right] \\
&\leq \mathbb{E}\left[\int_{\mathscr{Y}} (\alpha_{\nu,K} - \delta_{\nu})^2(y, X) \, d\nu(y) \int_{\mathscr{Y}} U_y(Z)^2 \, d\nu(y)\right] \\
&\leq 9\nu(\mathscr{Y}) \int_{\mathscr{Y}} \mathbb{E}\left[(\alpha_{\nu,K}(y, X) - \delta_{\nu}(y, X))^2\right] d\nu(y) \lesssim \sup_{y \in \mathscr{Y}} \left\|\alpha_{\nu,K}(y, \cdot) - \delta_{\nu}(y, \cdot)\right\|^2_{L_2(P_X)} \\
&\lesssim \|\nu\|^2 K^{-2\bar{s}/p} \longrightarrow 0
\end{aligned}$$

Likewise, by Cauchy-Schwarz, boundedness of δ_{ν} due to Assumption 5-(i), $\nu(\mathscr{Y}) < \infty$, and $\|\nu\| < \infty$

$$\begin{aligned}
\mathbb{E}[R_{2,\nu}(Z)^2] &\leq \mathbb{E}\left[\int_{\mathscr{Y}} \delta^2_{\nu}(y, X) \, d\nu(y) \int_{\mathscr{Y}} (M(y|X) - M_K(y|X))^2 \, d\nu(y)\right] \\
&\lesssim \int_{\mathscr{Y}} \mathbb{E}\left[(M_K(y|X) - M(y|X))^2\right] d\nu(y) \lesssim \sup_{y \in \mathscr{Y}} \left\|M_K(y|\cdot) - M(y|\cdot)\right\|^2_{L_2(P_X)} \\
&\lesssim K^{-2\bar{s}/p} \longrightarrow 0
\end{aligned}$$

so that $\Xi_{\nu,K}(Z_i) \longrightarrow \Xi_{\nu}(Z_i)$ in $L_2(P)$ and $\mathbb{E}[\Xi_{\nu,K}(Z)] = \mathbb{E}[\Xi_{\nu}(Z)] = 0$, by projection orthogo-

nality and conditional centering, respectively. This implies

$$\mathbb{E}\left[\left\{n^{-1/2}\sum_{i=1}^{n}\left(\Xi_{v,K}(Z_i)-\Xi_v(Z_i)\right)\right\}^2\right]=\mathbb{E}\left[\left(\Xi_{v,K}(Z)-\Xi_v(Z)\right)^2\right]\longrightarrow 0$$

Hence,

$$\begin{aligned}n^{-1/2}\sum_{i=1}^{n}\Xi_{v,K}(Z_i)&=n^{-1/2}\sum_{i=1}^{n}\Xi_v(Z_i)+n^{-1/2}\sum_{i=1}^{n}(\Xi_{v,K}(Z_i)-\Xi_v(Z_i))\\&=n^{-1/2}\sum_{i=1}^{n}\Xi_v(Z_i)+o_p(1)\end{aligned}$$

So the standard CLT for i.i.d. data applies to $2T^b_{3.1,v}$. Note that $\delta_v(y,X)\{\psi_y(Y)-M(y|X)\}$ is the adjustment term in Newey (1994a) for a single mean-square projection. Since we have a family of projections indexed by $y\in\mathscr{Y}$, the integral over $y\in\mathscr{Y}$ appears. This last step verifies assumption 5.3 in Newey (1994a) by showing the existence of some $\alpha(z)$ such that $\sqrt{n}\int D(z,\hat{h}-h_0)dF_0=\sum_{i=1}^n\alpha(z_i)/\sqrt{n}+o_p(1)$ in Newey's notation, where $D(z,\hat{h}-h)=2x'vy\{\hat{h}-h\}$ is the linearization term, $\hat{h}=\hat{M}(y|x)$, $h_0=M(y|x)$, $\alpha(z)=2\Xi_v(z)$, and F_0 is the true distribution of Z.

Step 1.3 ($T_{1,v}$ is $o_p(1)$). Write

$$\begin{aligned}T_{1,v}&=2\sqrt{n}(P_n-P)[a_v(\hat{M}_n-M)]\\&=2\sqrt{n}(P_n-P)[a_v(\hat{M}_n-M_K)]+2\sqrt{n}(P_n-P)[a_v(M_K-M)]=:T_{1.1,v}+T_{1.2,v}\end{aligned}$$

where by Assumption 5-(i) and item 6 in Lemma 6

$$\begin{aligned}\mathbb{E}[T^2_{1.2,v}]\leq 4P[a_v^2(M_K-M)^2]&\lesssim\int_{\mathscr{Y}}\mathbb{E}[(M_K(y|X)-M(y|X))^2]\,dv(y)\\&\lesssim\sup_{y\in\mathscr{Y}}\left\|M_K(y|\cdot)-M(y|\cdot)\right\|^2_{L_2(P_X)}\lesssim K^{-2\bar{s}/p}\longrightarrow 0\end{aligned}$$

which implies that $T_{1.2,v}=o_p(1)$. We now handle the last term $T_{1.1,v}$. First note that $\hat{\beta}_K(y)=\beta_K(y)+\hat{Q}_K^{-1}P_n[q_KU_y]=\beta_K(y)+(\hat{Q}_K^{-1}-Q_K^{-1})P_n[q_KU_y]+Q_K^{-1}P_n[q_KU_y]$, which implies that $\hat{M}_n(y|x)=\tilde{M}(y|x)+q_K(x)'(\hat{Q}_K^{-1}-Q_K^{-1})P_n[q_KU_y]$, where we define
$\tilde{M}(y|x):=M_K(y|x)+q_K(x)'Q_K^{-1}P_n[q_KU_y]$. We can now write

$$T_{1.1,v}=2\sqrt{n}(P_n-P)[a_v(\hat{M}_n-\tilde{M})]+2\sqrt{n}(P_n-P)[a_v(\tilde{M}-M_K)]=:2T^a_{1.1,v}+2T^b_{1.1,v}$$

Define the following objects $D_K:=\hat{Q}_K^{-1}-Q_K^{-1}$, $\Delta_m(y,x):=\hat{M}_n(y|x)-\tilde{M}(y|x)=q_K(x)'(\hat{Q}_K^{-1}-$

$Q_K^{-1})P_n[q_K U_y]$, $a_i := a_v(Y_i, X_i) = Y_i X_i' v$, $q_i := q_K(X_i)$, and

$$c_i := n^{-1} \sum_{j=1}^{n} q_j(\psi_{Y_i}(Y_j) - M_K(Y_i|X_j))$$

For the first term, by the Cauchy-Schwarz inequality and since $\Delta_m(y,\cdot) \in \mathscr{H}_K$,

$$\begin{aligned}
\left|T^a_{1.1,v}\right| &\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left|a_i q_i' D_K c_i\right| + \sqrt{n} \int_{\mathscr{Y}} \left|\mathbb{E}[\delta_v(y,X)\Delta_m(y,X)]\right| d\nu(y) \\
&\leq \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \|a_i q_i' D_K\|^2\right)^{1/2} \left(\sum_{i=1}^{n} \|c_i\|^2\right)^{1/2} \\
&\quad + \sqrt{n} \int_{\mathscr{Y}} \left|\eta_{v,K}(y)' Q_K D_K P_n[q_K U_y]\right| d\nu(y) \\
&\leq \frac{\|D_K\|}{\sqrt{n}} \left(\sum_{i=1}^{n} a_i^2 \|q_i\|^2\right)^{1/2} \left(\sum_{i=1}^{n} \|c_i\|^2\right)^{1/2} \\
&\quad + \sqrt{n}\|Q_K\|\|D_K\| \int_{\mathscr{Y}} \|\eta_{v,K}(y)\| \|P_n[q_K U_y]\| d\nu(y) \\
&\lesssim \frac{\|D_K\|}{\sqrt{n}} \left(\sum_{i=1}^{n} a_i^2 \|q_i\|^2\right)^{1/2} \left(\sum_{i=1}^{n} \|c_i\|^2\right)^{1/2} \\
&\quad + \sqrt{n}\|D_K\| \left(\int_{\mathscr{Y}} \|\eta_{v,K}(y)\|^2 d\nu(y)\right)^{1/2} \left(\int_{\mathscr{Y}} \|P_n[q_K U_y]\|^2 d\nu(y)\right)^{1/2} \\
&= O_p\left(K\xi_K \sqrt{\frac{\log K}{n}}\right) + O_p\left(\sqrt{K}\xi_K \sqrt{\frac{\log K}{n}}\right)
\end{aligned}$$

where we used items 1 and 3 in Lemma 6, $(*)$, $(**)$, and

$$\mathbb{E}[\delta_v(y,X)\Delta_m(y,X)] = \mathbb{E}[\alpha_{v,K}(y,X)\Delta_m(y,X)] = \eta_{v,K}(y)' Q_K D_K P_n[q_K U_y]$$

The bound for $\sum_i \|c_i\|^2$ can be obtained by showing that $\sum_{i=1}^n \mathbb{E}\|c_i\|^2 \lesssim K$ and then invoking Markov's inequality. To show this, define $U_{Y_i}(Z_j) := \psi_{Y_i}(Y_j) - M_K(Y_i|X_j)$, $V_{ij} := q_j U_{Y_i}(Z_j)$ and write $c_i = n^{-1}\sum_{j=1}^n V_{ij}$ so that $\mathbb{E}\|c_i\|^2 = n^{-2}\sum_{j=1}^n \sum_{k=1}^n \mathbb{E}[V_{ij}' V_{ik}]$, where all the terms $j \neq k$ vanish due to independence and projection orthogonality after applying the towering property, conditioning on Z_i. So that $\forall K \geq K_0$

$$\mathbb{E}\|c_i\|^2 = n^{-2}\mathbb{E}\|V_{ii}\|^2 + n^{-2}\sum_{j\neq i} \mathbb{E}\|V_{ij}\|^2 \lesssim \frac{K}{n^2} + \frac{n-1}{n^2}K \lesssim \frac{K}{n}$$

which summing over i and using the fact that c_i's are identically distributed shows that $\sum_{i=1}^n \mathbb{E}\|c_i\|^2 \lesssim K$, and by Markov's inequality we get that $\sum_{i=1}^n \|c_i\|^2 = O_p(K)$. Hence, by

Assumption 5-(iv)-(a) we have that $\left|T^a_{1.1,v}\right| = o_p(1)$.

For the final term, $T^b_{1.1,v}$, note that it is a degenerate V-statistic since

$$\tilde{M}(Y_i|X_i) - M_K(Y_i|X_i) = q_K(X_i)' Q_K^{-1} \left(n^{-1} \sum_{j=1}^{n} q_j U_{Y_i}(Z_j) \right)$$

Write $T^b_{1.1,v} = n^{-3/2} \sum_{i,j=1}^{n} h_{K,v}(Z_i, Z_j)$, with $h_{K,v}(z_1, z_2) := k_{K,v}(z_1, z_2) - P_1 k_{K,v}(\cdot, z_2)$ and

$$k_{K,v}(z_1, z_2) := a_v(z_1)\, q_K(x_1)' Q_K^{-1} q_K(x_2) U_{y_1}(z_2)$$

where $z_1 = (y_1, x_1)$, $z_2 = (y_2, x_2)$, P_1 and P_2 denote the laws of Z_1 and Z_2, respectively, and $U_{y_1}(z_2) = \psi_{y_1}(y_2) - M_K(y_1|x_2)$. The V-statistic is canonical of order 2 since $P_1 h_{K,v}(\cdot, z_2) = 0$ and by projection orthogonality $P_2 h_{K,v}(z_1, \cdot) = 0$. Moreover, note that

$$\begin{aligned}
\mathbb{E}[k^2_{K,v}(Z_1, Z_2)] &= \mathbb{E}[a_v(Z_1) q_K(X_1)' Q_K^{-1} \mathbb{E}[q_K(X_2) q_K(X_2)' U^2_{Y_1}(Z_2)|Z_1] Q_K^{-1} q_K(X_1) a_v(Z_1)] \\
&\lesssim \mathbb{E}[q_K(X_1)' Q_K^{-1} \mathbb{E}[q_K(X_2) q_K(X_2)'|Z_1] Q_K^{-1} q_K(X_1)] \quad \forall K \geq K_0 \\
&= \mathbb{E}[q_K(X_1)' Q_K^{-1} q_K(X_1)] = K
\end{aligned}$$

We now split $T^b_{1.1,v} = D_n + U_n$, where $D_n := n^{-3/2} \sum_{i=1}^{n} h_{ii}$ and $U_n = n^{-3/2} \sum_{i \neq j}^{n} h_{ij}$, with $h_{ii} := h_{K,v}(Z_i, Z_i)$ and $h_{ij} := h_{K,v}(Z_i, Z_j)$. For U_n, since the V-statistic is canonical, using the i.i.d. assumption and Cauchy-Schwarz, we get

$$\begin{aligned}
\mathbb{E}[U_n^2] &= n^{-3} \sum_{i \neq j} \sum_{l \neq k} \mathbb{E}[h_{ij} h_{lk}] = n^{-3} \sum_{i \neq j} \mathbb{E}[h_{ij}^2] + n^{-3} \sum_{i \neq j} \mathbb{E}[h_{ij} h_{ji}] \\
&\leq n^{-3} \sum_{i \neq j} \mathbb{E}[h_{ij}^2] + n^{-3} \sum_{i \neq j} \left(\mathbb{E}[h_{ij}^2] \mathbb{E}[h_{ji}^2] \right)^{1/2} = 2n^{-3} \sum_{i \neq j} \mathbb{E}[h_{ij}^2] \\
&= 2 \frac{n(n-1)}{n^3} \mathbb{E}[h_{12}^2] \leq 2 \frac{n(n-1)}{n^3} \mathbb{E}[k_{12}^2] \lesssim K/n
\end{aligned}$$

So that $U_n = O_p\left(\sqrt{\frac{K}{n}}\right)$, and thus $U_n = o_p(1)$ by Assumption 5-(iv)-(a). For D_n, write

$$\mathbb{E}[D_n^2] = \frac{1}{n} \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} h_{ii} \right)^2 \right] \leq \frac{1}{n} \mathbb{E}[h_{ii}^2] \leq \frac{2}{n} \mathbb{E}\left[k^2_{K,v}(Z_i, Z_i) + \left(P_1[k_{K,v}(\cdot, Z)] \right)^2 \right]$$

and note that $\mathbb{E}\left(P_1[k_{K,v}(\cdot, Z)]\right)^2 \leq \mathbb{E}[P_1 k^2(\cdot, Z)] = \mathbb{E}[k^2_{K,v}(Z_1, Z_2)] \lesssim K$ by Jensen's inequality and conditioning on Z. Also, since by Assumption 5-(iv) and $\xi_K = \sup_{x \in \mathcal{X}} \|q_K(x)\|$ (by

definition), we have that $|k_{K,v}(Z,Z)| \lesssim q_K(X)'Q_K^{-1}q_K(X) \leq \|q_K(X)\|^2\|Q_K^{-1}\| \lesssim \xi_K^2$. So

$$\mathbb{E}[k_{K,v}^2(Z,Z)] \lesssim \mathbb{E}\left[(q_K(X)'Q_K^{-1}q_K(X))^2\right] \leq \|Q_K^{-1}\|K\xi_K^2 \lesssim K\xi_K^2$$

which implies that $\mathbb{E}[D_n^2] \lesssim K/n + \xi_K^2 K/n$, where $\xi_K^2 \geq \mathbb{E}\|q_K(X)\|^2 = \mathrm{tr}(Q_K) \asymp K$, so $\mathbb{E}[D_n^2] \lesssim \xi_K^2 K/n$. By Assumption 5-$(iv)$ we get that $\forall K \geq 3$ and as $n \to \infty$

$$0 \leq \xi_K^2 K/n \leq \xi_K^2 K^2 \log(K)/n = \left(K\xi_K\sqrt{\frac{\log(K)}{n}}\right)^2 \longrightarrow 0$$

This shows that $T_{1.1,v}^b = o_p(1)$ and thus $T_{1,v} = o_p(1)$. This step verifies the stochastic equicontinuity condition, assumption 5.2 in Newey (1994a).

Combining Steps 1.1-1.3 we get $v'\sqrt{n}(B_n - B_0) = n^{-1/2}\sum_{i=1}^n v'(\phi(Z_i) - P\phi + 2\Xi(Z_i)) + o_p(1)$ where $v'\Xi = \Xi_v$, which satisfies a CLT since $\left|(\phi_v(Z_i) - P\phi_v + 2\Xi_v(Z_i))\right|^2 < \infty$ by the compact support assumption and the boundedness of $f(y|x)$ and the CMDF.

Step 2 (The term $v'\sqrt{n}(A_n - A_0)\theta_0$ satisfies a CLT). Note that we can write

$$\sqrt{n}v'(A_n - A_0)\theta_0 = n^{-1/2}\sum_{i=1}^n v'\left(X_iX_i'Y_i - A_0\right)\theta_0$$

where the term in the sum has zero-mean and by the compact support assumption we have that $\left|Y(X'v)(X'\theta_0)\right|^2 \lesssim \|v\|^2\|\theta_0\|^2 < \infty$, so the standard CLT for i.i.d. data applies.

Step 3 (Weak Convergence). Combining Steps 1 and 2 we get the representation

$$\sqrt{n}v'(B_n - B_0 - A_n\theta_0 + A_0\theta_0) = n^{-1/2}\sum_{i=1}^n v'\left(\phi(Z_i) - P\phi + 2\Xi(Z_i) - \left(X_iX_i'Y_i - A_0\right)\theta_0\right) + o_p(1)$$

where $\Xi := \left(\Xi_{e_1}, \ldots, \Xi_{e_p}\right)'$ and $\{e_1, \ldots, e_p\}$ are the canonical basis vectors of $\mathbb{R}^p$, so $\Xi_v = v'\Xi$. So by the standard CLT for i.i.d. data, we have that in $\mathbb{R}$

$$\sqrt{n}v'(B_n - B_0) - \sqrt{n}v'(A_n - A_0)\theta_0 \rightsquigarrow v'\ddot{W}$$

where $v'\ddot{W}$ is a zero-mean Gaussian random variable with the obvious variance. By the Cramér-Wold device, we have that in $\mathbb{R}^p$

$$\sqrt{n}(B_n - B_0) - \sqrt{n}(A_n - A_0)\theta_0 \rightsquigarrow \ddot{W}$$

where $\ddot{W}$ is a zero-mean Gaussian random vector. Note that $P\phi = B_0 = A_0\theta_0$ so

$$\phi(Z_i) - P\phi + 2\Xi(Z_i) - \left(X_i X_i' Y_i - A_0\right)\theta_0 = XU(\theta_0) + 2\Xi(Z_i) =: \chi(Z_i)$$

where $U(\theta_0)$ is the error term defined in Proposition 1 (or equation (7)). The variance of $\ddot{W}$ is then given by $\mathbb{V}ar[\ddot{W}] = \mathbb{E}[\chi(Z)\chi(Z)']$.

Step 4 (Weak Convergence of Series-GIR). By standard arguments under Assumption 4 we have that $A_n^{-1} = A_0^{-1} + o_p(1)$. Furthermore, by Slutsky's Lemma we can conclude that in $\mathbb{R}^p$, $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \rightsquigarrow W := A_0^{-1}\ddot{W}$, where W is a zero-mean Gaussian random vector with variance $\Sigma := \mathbb{V}ar[W] = A_0^{-1}\mathbb{E}[\chi(Z)\chi(Z)']A_0^{-1}$, i.e., $W \sim N(0,\Sigma)$.

□

E Proof of Theorem 3

Before proving the theorem, we first clarify the notation and definitions to be used throughout the proof. The same definitions and notation as defined in Section D.1 and in the asymptotic normality proof of Section D.3 will be used. Section E.1 adds the bootstrap-specific notation and definitions.

E.1 *Additional Bootstrap Definitions and Notation*

We first specify the relevant probability spaces and define the related stochastic orders in order to rigorously state the theoretical results for the bootstrap. We follow a setup similar to that in Cheng and Huang (2010).

Probability Spaces. The bootstrap weights w_i's are defined on the probability space $(\Omega_W, \mathcal{W}, P_W)$. The random variable Z, from which the data is observed, is defined on the probability space $(\mathcal{Z}, \mathcal{A}, P)$. We view the random vector Z_i as the i-th coordinate projection from the canonical probability space $(\mathcal{Z}^\infty, \mathcal{A}^\infty, P^\infty)$ onto the i-th copy of $\mathcal{Z}$. We write P^∞ as P_Z thereafter and note that under P_Z, the random vectors $Z_1, Z_2, \ldots$ are i.i.d. with common law P. For the joint randomness (due to the observed data and the weights) involved in the bootstrap, the product probability space is defined as $(\mathcal{Z}^\infty, \mathcal{A}^\infty, P^\infty) \times (\Omega_W, \mathcal{W}, P_W)$ and we denote the resulting measure by $P_{ZW} := P_Z \times P_W$, where we assume $\mathcal{W}$ to be independent of $\mathcal{A}^\infty$ under P_{ZW}. Let $\mathcal{D}_n := \sigma(Z_1, \ldots, Z_n)$ be the σ-field generated by the observed data. For an event A on the bootstrap product space, define $P_w(A) := P_{ZW}(A|\mathcal{D}_n)$ to be the conditional probability given the data (conditional bootstrap probability). We can now define the corresponding expectation operators.

As in Section D, $\mathbb{E}[\cdot]$ denotes the expectation with respect to P, $\mathbb{E}_W[\cdot]$ denotes the expectation with respect to P_W, $\mathbb{E}_Z[\cdot]$ denotes the expectation with respect to P_Z, $\mathbb{E}_{ZW}[\cdot]$ denotes the expectation with respect to P_{ZW}, and $\mathbb{E}_w[\cdot] \equiv \mathbb{E}_{ZW}[\cdot \mid \mathscr{D}_n]$ denotes the (conditional) expectation given the observed data, i.e., with respect to the conditional probability P_w.

Measurability. All the function classes to be considered throughout the proof are taken to be suitably measurable, e.g., pointwise measurability or Suslin measurability are sufficient.

Stochastic Orders. For data-only random variables, $o_{P_Z}(a_n)$ and $O_{P_Z}(a_n)$ denote the usual stochastic orders under P_Z. For random variables on the product space, $o_{P_{ZW}}(a_n)$ and $O_{P_{ZW}}(a_n)$ denote the usual stochastic orders under P_{ZW}. For random variables on the weight space, $o_{P_W}(a_n)$ and $O_{P_W}(a_n)$ denote the usual stochastic orders under P_W. For a measurable bootstrap random vector Δ_n defined on the product probability space and a positive deterministic sequence $a_n > 0$, write $\Delta_n = o_{P_w}(a_n)$ in P_Z-probability if for every $\epsilon > 0$ and every $\delta > 0$,

$$P_Z\left[P_w(\|\Delta_n\| > \epsilon a_n) > \delta\right] \longrightarrow 0$$

Write $\Delta_n = O_{P_w}(a_n)$ in P_Z-probability if for every $\delta > 0$, $\exists M < \infty$ such that

$$P_Z\left[P_w(\|\Delta_n\| > M a_n) > \delta\right] \longrightarrow 0$$

The conditional stochastic $O_{P_w}(a_n)$ and $o_{P_w}(a_n)$ symbols satisfy the usual algebra of stochastic orders.

Notation. For a measurable scalar-, vector-, or matrix-valued function f, we use the standard empirical process notation

$$P_n[f] := n^{-1}\sum_{i=1}^{n} f(L_i) \qquad \mathbb{G}_n(f) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{f(L_i) - \mathbb{E}_L[f(L)]\right\}$$

where $\mathbb{E}_L$ denotes the expectation with respect to the law of the argument of f. The relevant choice is $L_i = Z_i$ for data-only functions and $L_i = (Z_i, w_i)$ for functions depending on both data and weights. Define as well $\bar{w}_n := n^{-1}\sum_{i=1}^{n} w_i$ and, for a function g on $\mathscr{Y}\mathscr{X}$,

$$P_n^* g := \frac{1}{n\bar{w}_n}\sum_{i=1}^{n} w_i g(Z_i) \qquad \widetilde{G}_n^*(g) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(w_i - 1)g(Z_i)$$

$$\mathbb{Z}_n^*(g) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(w_i - \bar{w}_n)g(Z_i) \qquad \hat{G}_n^*(g) := \sqrt{n}(P_n^* - P_n)g = \bar{w}_n^{-1}\mathbb{Z}_n^*(g)$$

For vector- or matrix-valued functions, all expectations and empirical operators act com-

ponentwise. The bootstrap versions of $A_n = P_n\left(XX'Y\right)$ and $B_n = P_n\left(XY\left(2\hat{M}_n(Y|X)-1\right)\right)$ are $A_n^* := P_n^*\left(XX'Y\right)$ and $B_n^* := P_n^*\left(XY\left(2\hat{M}_n^*(Y|X)-1\right)\right)$ so that $\hat{\theta}_n^* = \left(A_n^*\right)^{-1} B_n^*$ is the bootstrap version of the GIR estimator. Let $BL_1(\mathbb{R}^p)$ denote the class of functions

$$BL_1(\mathbb{R}^p) := \left\{ h : \mathbb{R}^p \to \mathbb{R} \;\middle|\; \sup_x |h(x)| \leq 1, \quad |h(x)-h(y)| \leq \|x-y\|, \quad \forall x, y \in \mathbb{R}^p \right\}$$

For a bootstrap statistic R_n^* and some random vector R with law P_R, write $R_n^* \underset{P_Z}{\rightsquigarrow} R$ on $\mathbb{R}^p$ if $\sup_{h \in BL_1(\mathbb{R}^p)} \left|\mathbb{E}_w[h(R_n^*)] - \mathbb{E}_R[h(R)]\right| = o_{P_Z}(1)$, where $\mathbb{E}_R[\cdot]$ denotes the expectation with respect to P_R.

For a probability measure Q, the $L_2(Q)$-norm of a real-valued f is denoted by $\|f\|_{Q,2} := \left(\int |f|^2 \, dQ\right)^{1/2}$. For a semimetric space $(\mathbb{T}, d)$, the covering number $N(\mathbb{T}, d, \epsilon)$ is the minimal number of d-balls of radius ϵ needed to cover $\mathbb{T}$. For a class $\mathscr{F}$ of measurable real-valued functions and a probability measure Q, write $N(\mathscr{F}, L_2(Q), \epsilon)$. A measurable function F is an envelope of the class $\mathscr{F}$ if $|f(u)| \leq F(u)$ for every $f \in \mathscr{F}$ and every u. When writing $\sup_Q N(\mathscr{F}, L_2(Q), \epsilon)$, the supremum is taken over all probability measures Q on the appropriate measurable space.

Recall as well the normalization notation. Following Belloni et al. (2015), define the normalized basis $p_K(x) = Q_K^{-1/2} q_K(x)$ so that $\mathbb{E}[p_K(X)p_K(X)'] = I_K$. Define as well $\alpha_K(x) = p_K(x)/\|p_K(x)\|$, where $\inf_{x \in \mathscr{X}} \|p_K(x)\| > 0$ under the basis choices, as they reproduce constants, and

$$\xi_K^L := \sup_{x, x' \in \mathscr{X} : x \neq x'} \frac{\|\alpha_K(x) - \alpha_K(x')\|}{\|x - x'\|}$$

Moreover, let $\gamma_K(y) = Q_K^{1/2}\beta_K(y)$, $\hat{\gamma}_K(y) := Q_K^{1/2}\hat{\beta}_K(y)$, $\hat{\gamma}_K^*(y) := Q_K^{1/2}\hat{\beta}_K^*(y)$, so that $M_K(y|x) = q_K(x)'\beta_K(y) = p_K(x)'\gamma_K(y)$, $\hat{M}_n(y|x) = q_K(x)'\hat{\beta}_K(y) = p_K(x)'\hat{\gamma}_K(y)$, and $\hat{M}_n^*(y|x) = q_K(x)'\hat{\beta}_K^*(y) = p_K(x)'\hat{\gamma}_K^*(y)$. Finally, let $\hat{Q}_{K,p} := P_n[p_K(X)p_K(X)']$, satisfying $\hat{Q}_{K,p} = Q_K^{-1/2}\hat{Q}_K Q_K^{-1/2}$.

E.2 *Tools*

Lemma 7. *(Transitions Between Probability Orders) Let $a_n > 0$ be a deterministic sequence, Δ_n be a vector-valued random function on $(\mathscr{Z}^\infty, \mathscr{A}^\infty, P^\infty) \times (\Omega_W, \mathscr{W}, P_W)$, Γ_n be a vector-valued $\mathscr{D}_n$-measurable random function on $(\mathscr{Z}^\infty, \mathscr{A}^\infty, P^\infty)$, and Λ_n be a vector-valued random function on $(\Omega_W, \mathscr{W}, P_W)$. The following statements hold.*

(i) If $\Delta_n = o_{P_{ZW}}(a_n)$, then $\Delta_n = o_{P_w}(a_n)$ in P_Z-probability.

(ii) $\Gamma_n = o_{P_Z}(a_n) \iff \Gamma_n = o_{P_{ZW}}(a_n)$ *and* $\Gamma_n = O_{P_Z}(a_n) \iff \Gamma_n = O_{P_{ZW}}(a_n)$.

(iii) $\Gamma_n = o_{P_Z}(a_n) \implies \Gamma_n = o_{P_w}(a_n)$ *in* P_Z*-probability.*

(iv) $\Lambda_n = o_{P_W}(a_n) \iff \Lambda_n = o_{P_w}(a_n)$ *in* P_Z*-probability and* $\Lambda_n = O_{P_W}(a_n) \iff \Lambda_n = O_{P_w}(a_n)$ *in* P_Z*-probability.*

Proof. For part (i), Markov's inequality and iterated expectations give, for every $\epsilon, \delta > 0$:

$$P_Z\left(P_w\left(\|\Delta_n\| > \epsilon a_n\right) > \delta\right) \leq \delta^{-1}\mathbb{E}_Z\left[P_w\left(\|\Delta_n\| > \epsilon a_n\right)\right] = \delta^{-1}P_{ZW}\left(\|\Delta_n\| > \epsilon a_n\right) \longrightarrow 0$$

The statement then follows by definition. For part (ii), since Γ_n depends only on the data,

$$P_{ZW}\left(\|\Gamma_n\| > t\right) = \mathbb{E}_Z\left[\mathbb{E}_w[\mathbb{1}_{\{\|\Gamma_n\|>t\}}]\right] = \mathbb{E}_Z[\mathbb{1}_{\{\|\Gamma_n\|>t\}}] = P_Z(\|\Gamma_n\| > t)$$

where the second equality holds since Γ_n does not depend on the bootstrap weight. The statement then follows by definition. For (iii), since Γ_n does not depend on the weights, $P_w(\|\Gamma_n\| > \epsilon a_n) = \mathbb{1}_{\{\|\Gamma_n\|>\epsilon a_n\}}$ so that, $P_Z\left(P_w\left(\|\Gamma_n\| > \epsilon a_n\right) > \delta\right) = P_Z\left(\|\Gamma_n\| > \epsilon a_n\right) \to 0$, for $\delta \in (0,1)$. For $\delta \geq 1$ the event is empty and the probability is 0. This proves (iii). Finally, note that by independence, for every t, $P_w(\|\Lambda_n\| > t) = P_W(\|\Lambda_n\| > t)$, which is a deterministic sequence so $P_Z(P_w(\|\Lambda_n\| > a_n\epsilon) > \delta) = \mathbb{1}_{\{P_W(\|\Lambda_n\|>a_n\epsilon)>\delta\}}$. Suppose that $\Lambda_n = o_{P_W}(a_n)$. Then, $\forall \epsilon > 0$, $P_w(\|\Lambda_n\| > a_n\epsilon) = P_W(\|\Lambda_n\| > a_n\epsilon) \to 0$, i.e., for all sufficiently large n we have $P_w(\|\Lambda_n\| > a_n\epsilon) \leq \delta$ for every fixed $\delta > 0$. Hence, $P_Z(P_w(\|\Lambda_n\| > a_n\epsilon) > \delta) = \mathbb{1}_{\{P_W(\|\Lambda_n\|>a_n\epsilon)>\delta\}} \to 0$. Conversely, suppose that $\Lambda_n = o_{P_w}(a_n)$ in P_Z-probability. Then, for sufficiently large n, $\mathbb{1}_{\{P_w(\|\Lambda_n\|>a_n\epsilon)>\delta\}} \leq 1/2$ and thus $P_W(\|\Lambda_n\| > a_n\epsilon) \leq \delta$ for large n, i.e., $P_W(\|\Lambda_n\| > a_n\epsilon) \to 0$. For the big-$O$ analog, first suppose that $\Lambda_n = O_{P_W}(a_n)$. Then, for $\delta > 0$, choose $M < \infty$ such that $\limsup_n P_W(\|\Lambda_n\| > Ma_n) < \delta/2$. Then $P_Z(P_w(\|\Lambda_n\| > Ma_n) > \delta) = \mathbb{1}_{\{P_W(\|\Lambda_n\|>Ma_n)>\delta\}} = 0$ for large n. For the reverse implication, suppose that $\Lambda_n = O_{P_w}(a_n)$ in P_Z-probability. Then, for $\delta > 0$, $\exists M < \infty$ such that $P_Z(P_w(\|\Lambda_n\| > Ma_n) > \delta/2) = \mathbb{1}_{\{P_W(\|\Lambda_n\|>Ma_n)>\delta/2\}} \to 0$, i.e., $P_W(\|\Lambda_n\| > Ma_n) \leq \delta/2$ for large enough n, so that $\limsup_n P_W(\|\Lambda_n\| > Ma_n) \leq \delta/2 < \delta$ and thus $\Lambda_n = O_{P_W}(a_n)$ by definition.

□

Remark: In general, the analogous implication of (iii) from an ordinary or joint big-O order to the conditional O_{P_w} order is not true.

Lemma 8. *(Standard Covering Number Properties) Let* $\mathcal{F}$ *and* $\mathcal{G}$ *be two classes of real-valued functions on* $\mathcal{Z}$ *with finite covering numbers in* $L_2(Q)$*. For constants* $a, b \in \mathbb{R}$*, the*

classes $a\mathscr{F} + b\mathscr{G} := \{af + bg : f \in \mathscr{F}, g \in \mathscr{G}\}$ *and* $m \cdot \mathscr{F} := \{(z,w) \mapsto m(w) f(z) : f \in \mathscr{F}\}$ *with* $m : \mathbb{R}_+ \to \mathbb{R}$, *satisfy*

(i) $N(a\mathscr{F} + b\mathscr{G}, L_2(Q), |a|\epsilon_F + |b|\epsilon_G) \leq N(\mathscr{F}, L_2(Q), \epsilon_F) N(\mathscr{G}, L_2(Q), \epsilon_G)$.

(ii) Let $0 < \epsilon \leq 1$, F *be an envelope for* $\mathscr{F}$, *and* $F_m := |m|F$. *If* $\sup_S N\left(\mathscr{F}, L_2(S), \epsilon \|F\|_{S,2}\right) \leq \left(\frac{A}{\epsilon}\right)^V$, *then* $\sup_Q N\left(m\mathscr{F}, L_2(Q), \epsilon \|F_m\|_{Q,2}\right) \leq \left(\frac{A}{\epsilon}\right)^V$, *where* S, Q *range over all probability measures on* $\mathscr{Z}$ *and* $\mathscr{Z} \times \mathbb{R}_+$, *respectively, such that* $s^2_{Q,m} := \int m(w)^2 \, dQ(z,w) < \infty$, $0 < \|F_m\|_{Q,2} < \infty$.

Proof. Let $\{f_j : 1 \leq j \leq M\}$ be an ϵ_F-cover of $\mathscr{F}$ and let $\{g_l : 1 \leq l \leq N\}$ be an ϵ_G-cover of $\mathscr{G}$. The set $\{af_j + bg_l : 1 \leq j \leq M, 1 \leq l \leq N\}$, which has at most NM elements, is an $(|a|\epsilon_F + |b|\epsilon_G)$-cover of $a\mathscr{F} + b\mathscr{G}$, since by the triangle inequality $\|(af + bg) - (af_j + bg_l)\|_{Q,2} \leq |a|\|f - f_j\|_{Q,2} + |b|\|g - g_l\|_{Q,2} \leq |a|\epsilon_F + |b|\epsilon_G$. Taking the minimal covers proves the statement. To prove the last claim, note that if $s^2_{Q,m} = 0$, then $m(w) = 0$ Q-a.s. and the covering number is 1. For $s^2_{Q,m} > 0$ define the probability measure $Q^\dagger_m$ on $\mathscr{Z}$ by $Q^\dagger_m(A) = s^{-2}_{Q,m} \int 1\{z \in A\} m(w)^2 \, dQ(z,w)$. The covering number bound holds uniformly over all probability measures on $\mathscr{Z}$, so take $S = Q^\dagger_m$ and $\epsilon_F = \epsilon \|F\|_{Q^\dagger_m,2}$. The set $\{mf_j : 1 \leq j \leq M\}$ is an $\epsilon\|F_m\|_{Q,2}$-cover of $m\mathscr{F}$ since $\|mf - mf_j\|_{Q,2} = s_{Q,m}\|f - f_j\|_{Q^\dagger_m,2} \leq \epsilon s_{Q,m}\|F\|_{Q^\dagger_m,2} = \epsilon\|F_m\|_{Q,2}$ and thus $N(m\mathscr{F}, L_2(Q), \epsilon\|F_m\|_{Q,2}) \leq N(\mathscr{F}, L_2(Q^\dagger_m), \epsilon\|F\|_{Q^\dagger_m,2}) \leq \left(\frac{A}{\epsilon}\right)^V$.

□

Lemma 9. *(Symmetrization Inequality) Let* $\mathscr{H}$ *be a sigma-field. Conditional on* $\mathscr{H}$, *let the measurable classes* $V_i := \{V_i(t) : t \in T\}$, $i \in \{1, \ldots, n\}$, *be jointly independent random elements conditional on* $\mathscr{H}$ *satisfying* $\mathbb{E}[V_i(t) \mid \mathscr{H}] = 0$ *for every* i, t. *Moreover, let* $\eta_1, \ldots, \eta_n$ *be independent Rademacher random variables (defined on* $(\Omega_\eta, \mathscr{A}_\eta, P_\eta)$ *such that* $P_\eta(\eta_1 = 1) = P_\eta(\eta_1 = -1) = 1/2$*), independent of* $\mathscr{H}$ *and of* $(V_1, \ldots, V_n)$. *Then*

$$\mathbb{E}\left[\sup_{t\in T}\left|\frac{1}{n}\sum_{i=1}^n V_i(t)\right| \mid \mathscr{H}\right] \leq 2\mathbb{E}\left[\mathbb{E}_\eta\left[\sup_{t\in T}\left|\frac{1}{n}\sum_{i=1}^n \eta_i V_i(t)\right|\right] \mid \mathscr{H}\right]$$

Proof. Direct application of Lemma 2.3.1 in Van der Vaart and Wellner (2023) under the conditional law given $\mathscr{H}$.

□

Lemma 10. *(Maximal Inequality) Let* $T \subseteq \mathbb{R}^n$ *be non-empty,* $\|t\|^2_{n,2} := n^{-1}\sum_{i=1}^n t_i^2$, $t \in T$, *and define* $V_t := n^{-1/2}\sum_{i=1}^n \eta_i t_i$, *where* $\eta_1, \ldots, \eta_n$ *are independent Rademacher random variables. Then,*

$$\mathbb{E}_\eta\left[\sup_{t\in T}|V_t|\right] \lesssim \int_0^{2r_T} \sqrt{\log\left(eN(T, \|\cdot\|_{n,2}, \epsilon)\right)} \, d\epsilon$$

where $r_T := \sup_{t\in T} \|t\|_{n,2} < \infty$ *and* $N(T, \|\cdot\|_{n,2}, \epsilon)$ *is finite for every* $\epsilon > 0$.

Proof. Apply Corollary 2.2.5 in Van der Vaart and Wellner (2023) to $T_0 := T \cup \{0\}$, with $V_0 = 0$. The process is sub-Gaussian with respect to $\|\cdot\|_{n,2}$ by Lemma 2.2.8 in Van der Vaart and Wellner (2023). Then note that $\operatorname{diam}(T_0) \le 2r_T$, $N(T_0, \|\cdot\|_{n,2}, \epsilon) \le 2N(T, \|\cdot\|_{n,2}, \epsilon)$ because T is nonempty, $\mathbb{E}_\eta|\cdot| \le \|\cdot\|_{\psi_2}$, and apply the standard packing-covering number inequality, which only changes the universal constant.

□

E.3 *Implied Rates and Entropy of Function Classes*

Lemma 11. *(Bootstrap Processes Transitions Between Probability Orders) Let* $g_n : \mathscr{Y}\mathscr{X} \to \mathbb{R}^d$ *be a measurable function, possibly data-dependent (in which case, measurable with respect to* $\mathscr{D}_n$*), with* $P_n\|g_n\|^2 < \infty$*. Let* $a_n > 0$ *be a deterministic sequence. Under Assumption 6-(*i*):*

(i) $\mathbb{E}_w\|\mathbb{Z}_n^*(g_n)\|^2 \le P_n\|g_n\|^2$.

(ii) If $P_n\|g_n\|^2 = o_{P_Z}(a_n^2)$*, then* $\mathbb{Z}_n^*(g_n) = o_{P_w}(a_n)$ *and* $\hat{\mathbb{G}}_n^*(g_n) = o_{P_w}(a_n)$ *in* P_Z*-probability.*

(iii) If $\exists C < \infty$ *such that* $P_Z\{P_n\|g_n\|^2 \le Ca_n^2\} \to 1$*, then* $\mathbb{Z}_n^*(g_n) = O_{P_w}(a_n)$ *and* $\hat{\mathbb{G}}_n^*(g_n) = O_{P_w}(a_n)$ *in* P_Z*-probability. In particular, for a fixed measurable function* g *with* $P\|g\|^2 < \infty$ *we have* $\mathbb{Z}_n^*(g) = O_{P_w}(1)$ *and* $\hat{\mathbb{G}}_n^*(g) = O_{P_w}(1)$ *in* P_Z*-probability.*

Proof. Under Assumption 6-(i), we have that $\{w_i\} \sim$ *i.i.d. Exp*(1), $i \in \{1,\dots,n\}$, so $\mathbb{E}_W[w_i] = \mathbb{V}ar_W[w_i] = 1$, $\mathbb{E}_W[w_i - \bar{w}_n] = 0$, $\mathbb{E}_W[(w_i - \bar{w}_n)^2] = 1 - n^{-1}$ and $\mathbb{E}_W[(w_i - \bar{w}_n)(w_j - \bar{w}_n)] = -n^{-1}$ for $i \ne j$, which gives, for scalar-valued g_n, $\mathbb{E}_w[\mathbb{Z}_n^*(g_n)^2] = P_n[(g_n - P_n g_n)^2] \le P_n g_n^2 < \infty$. For an $\mathbb{R}^d$-valued function, note that $\mathbb{E}_w\|\mathbb{Z}_n^*(g_n)\|^2 = \sum_{l=1}^d \mathbb{E}_w[\mathbb{Z}_n^*(g_{n,l})^2] \le P_n\|g_n\|^2$. This proves the first statement. For the second claim, Markov's inequality conditional on $\mathscr{D}_n$ and (i) give $P_w\left(\|\mathbb{Z}_n^*(g_n)\| > \epsilon a_n\right) \le \epsilon^{-2}a_n^{-2}P_n\|g_n\|^2 = o_{P_Z}(1)$, so that by definition $P_Z\left(P_w\left(\|\mathbb{Z}_n^*(g_n)\| > \epsilon a_n\right) > \delta\right) \to 0$. Regarding $\hat{\mathbb{G}}_n^*(g_n)$, the claim follows from the relationship $\hat{\mathbb{G}}_n^*(g_n) = \bar{w}_n^{-1}\mathbb{Z}_n^*(g_n)$ and the fact that $\bar{w}_n^{-1} = O_{P_W}(1)$, by the standard law of large numbers, so that by Lemma 7-(iv), $\bar{w}_n^{-1} = O_{P_w}(1)$ in P_Z-probability. The product is thus $o_{P_w}(a_n)$ by standard order algebra. For the last claim, let $\delta > 0$ and choose $M \ge \sqrt{C/\delta}$. Note that Markov's inequality and (i) give $P_w(\|\mathbb{Z}_n^*(g_n)\| > Ma_n) \le P_n\|g_n\|^2/M^2a_n^2 \le C/M^2 \le \delta$, where the second inequality follows with probability P_Z approaching 1 (by hypothesis). Hence $P_Z(P_w(\|\mathbb{Z}_n^*(g_n)\| > Ma_n) > \delta) \longrightarrow 0$ and thus $\mathbb{Z}_n^*(g_n) = O_{P_w}(a_n)$ in P_Z-probability. The claim for $\hat{\mathbb{G}}_n^*(g_n)$ follows from the relationship $\hat{\mathbb{G}}_n^*(g_n) = \bar{w}_n^{-1}\mathbb{Z}_n^*(g_n)$ and the fact that $\bar{w}_n^{-1} = O_{P_W}(1)$, by the standard law of large numbers, so that by Lemma 7-(iv),

$\bar{w}_n^{-1} = O_{P_w}(1)$ in P_Z-probability. The product is thus $O_{P_w}(a_n)$ by standard order algebra. Finally, for fixed g, the standard law of large numbers gives $P_n\|g\|^2 \to P\|g\|^2 < \infty$, so the hypothesis in (iii) holds for some $C < \infty$ and the same arguments apply with $a_n = 1$.

□

Lemma 12. *(Rate Implications of Assumption 6) Let A_K be a deterministic positive sequence. Under Assumptions 5-(iii) and 6-(ii), if $\log(A_K) = O(\log(K))$, then $\log(A_K) = O(\log(n))$. Moreover, Assumption 6-(i) implies $\max_{1\le i\le n} w_i = O_{P_w}(\log(n))$ in P_Z-probability.*

Proof. Note that by Assumption 5-(iii) and Lemma 6, we have $\xi_K^2 = \sup_{x\in\mathscr{X}} \|q_K(x)\|^2 \ge \mathbb{E}\|q_K(X)\|^2 \ge d_B K$. Assumption 6-$(ii)$ gives $\xi_K r_{K,n}^b \to 0$ as $(n,K) \to \infty$, which implies, for the first component of $r_{K,n}^b$, that $0 \le n^{-1/2} K \log^2(en) \lesssim n^{-1/2}\xi_K^2 \log^2(en) \to 0$. By definition of convergence, $\exists n_0 \in \mathbb{N}$ such that $\forall n \ge n_0$, $K \le \sqrt{n}/\log^2(en)$, so $\log(K) \le 0.5\log(n) - 2\log(\log(en)) \lesssim \log(n)$, $\forall n \ge n_0$. Hence, $\log(A_K) = O(\log(K)) = O(\log(n))$. For the last claim, note that by the i.i.d. property of weights and the exponential CDF with rate 1, we have, for $1 < M < \infty$

$$P_W\left(\max_{1\le i\le n} w_i > M\log(n)\right) \le \sum_{i=1}^{n} P_W\left(w_i > M\log(n)\right) = n^{1-M} \longrightarrow 0$$

Lemma 7-(iv) gives $\max_{1\le i\le n} w_i = O_{P_W}(\log(n)) = O_{P_w}(\log(n))$ in P_Z-probability.

□

E.4 *Uniform First-Stage Linearization*

This section contains the extensions of the main uniform linearization and bootstrap results in Belloni et al. (2015) for series asymptotics to hold uniformly over $(y,x) \in \mathscr{Y}\mathscr{X}$ for the case of the conditional mid-distribution function $M(y|x)$. The following results are not direct implications of their theory. We extend Lemma 4.2 and Theorem 4.5-(1) of the mentioned reference in Propositions 4 and 5 of this section.

Proposition 4. *(Uniform Normalized-Basis Linearization) Let Assumptions 4, 5, and 6 hold. Then, $\sqrt{n}\alpha_K(x)'\{\hat{\gamma}_K(y) - \gamma_K(y)\} = \alpha_K(x)'\mathbb{G}_n[p_K U_y] + R_{1n}(y,x)$, where $R_{1n}(y,x)$ obeys*

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |R_{1n}(y,x)| = O_{P_Z}\left(r_{K,n}\right)$$

where $r_{K,n} = \left[\frac{\xi_K^2\log(en)}{n}\right]^{1/2} \cdot \left\{\sqrt{\log(en)} + \sqrt{K}\, l_K K^{-\bar{s}/p}\right\}$.

Proof. Using the normalized model $\psi_y(Y_i) = p_K(X_i)'\gamma_K(y) + U_y(Z_i)$, where $U_y := \varepsilon_y + r_y$ is the composite error term, r_y denotes the approximation error $r_y(X) = M(y|X) - M_K(y|X)$, and ε_y the structural error term $\varepsilon_y(Z) = \psi_y(Y) - M(y|X)$, i.e., $\mathbb{E}[\varepsilon_y(Z)|X] = 0$, $y \in \mathscr{Y}$. Moreover,

$$\gamma_K(y) \in \underset{\gamma \in \mathbb{R}^K}{\arg\min}\, \mathbb{E}\left[\left(\psi_y(Y) - p_K(X)'\gamma\right)^2\right]$$

So that $P[p_K U_y] = 0$ by orthogonality and thus $\sqrt{n}P_n[p_K U_y] = \mathbb{G}_n(p_K U_y)$. The normal equations give

$$\sqrt{n}\alpha_K(x)'\left\{\hat{\gamma}_K(y) - \gamma_K(y)\right\} = \alpha_K(x)'\hat{Q}_{K,p}^{-1}\mathbb{G}_n(p_K U_y) = \alpha_K(x)'\mathbb{G}_n(p_K U_y) + R_{1n}(y,x)$$

where $R_{1n} := R_{1n}^{\varepsilon} + R_{1n}^{r}$ is a remainder term composed of $R_{1n}^{\varepsilon} := \alpha_K(x)'(\hat{Q}_{K,p}^{-1} - I_K)\mathbb{G}_n(p_K\varepsilon_y)$, and $R_{1n}^{r} := \alpha_K(x)'(\hat{Q}_{K,p}^{-1} - I_K)\mathbb{G}_n(p_K r_y)$. We follow the steps in Belloni et al. (2015) to extend their linearization results to the CMDF case.

Step 1 (Rate for R_{1n}^{ε}). We show that $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |R_{1n}^{\varepsilon}(y,x)| = O_{P_Z}\left(\xi_K \log(en)/\sqrt{n}\right)$. Let the independent Rademacher random variables $\eta_1, \ldots, \eta_n$ be independent of the data. Conditional on $X_1, \ldots, X_n$, the indexed summands in R_{1n}^{ε} are independent and have conditional mean zero. By the symmetrization inequality in Lemma 9 we have

$$\mathbb{E}_Z\left[\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |R_{1n}^{\varepsilon}(y,x)| \,\middle|\, X\right] \leq 2\mathbb{E}_Z\left[\mathbb{E}_\eta\left[\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\alpha_K(x)'(\hat{Q}_{K,p}^{-1} - I_K)\mathbb{G}_n(\eta p_K\varepsilon_y)|\right] \,\middle|\, X\right]$$

where conditioning on X is understood as conditioning on $\sigma(X_1, \ldots, X_n)$. To handle the inner expectation, define the class of functions $T := \{t = (t_1, \ldots, t_n) \in \mathbb{R}^n : t_i = \alpha_K(x_0)'(\hat{Q}_{K,p}^{-1} - I_K)p_K(x_i)\varepsilon_y(z_i), (y, x_0) \in \mathscr{Y}\mathscr{X}\}$ and the semimetric $\|t\|_{n,2}^2 = n^{-1}\sum_{i=1}^n t_i^2$. Conditional on the data, $\{n^{-1/2}\sum_{i=1}^n \eta_i t_i\}$ is a Rademacher process. Note that $\|t\|_{n,2}^2 \leq \|\hat{Q}_{K,p}^{-1} - I_K\|^2\|\hat{Q}_{K,p}\| =: \sigma_n^2$, so $\sup_{t\in T}\|t\|_{n,2} \leq \sigma_n$. The maximal inequality in Lemma 10 gives

$$\mathbb{E}_\eta\left[\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\alpha_K(x)'(\hat{Q}_{K,p}^{-1} - I_K)\mathbb{G}_n(\eta p_K\varepsilon_y)|\right] \lesssim \int_0^{2\sigma_n} \sqrt{\log(eN(T, \|\cdot\|_{n,2}, \epsilon))}\, d\epsilon$$

We now bound $N(T, \|\cdot\|_{n,2}, \epsilon)$. Define the vector-valued indexed map $a_{x_0}(x_1, \ldots, x_n) = (a_{x_0}(z_1), \ldots, a_{x_0}(z_n))' \in \mathbb{R}^n$ with $a_{x_0}(z_i) := \alpha_K(x_0)'(\hat{Q}_{K,p}^{-1} - I_K)p_K(x_i)$, and the stacked structural errors $e_y(z_1, \ldots, z_n) = (\varepsilon_y(z_1), \ldots, \varepsilon_y(z_n))' \in \mathbb{R}^n$. The class T is the product class $\{a_{x_0} \odot e_y : (y, x_0) \in \mathscr{Y}\mathscr{X}\}$, for which we construct a cover. The class $\mathscr{A}_n := \{a_{x_0} : x_0 \in \mathscr{X}\}$ with $\|a_{x_0}\|_{n,2} \leq \sigma_n$ satisfies, for $x_0, x_0' \in \mathscr{X}$,

$$\|a_{x_0} - a_{x_0'}\|_{n,2} = \|(\alpha_K(x_0) - \alpha_K(x_0'))'(\hat{Q}_{K,p}^{-1} - I_K)p_K\|_{n,2} \leq \xi_K^L\sigma_n\|x_0 - x_0'\|$$

so it is Lipschitz in the index set and by Theorem 2.7.17 in Van der Vaart and Wellner (2023) and compactness of $\mathscr{X}$, $N(\mathscr{A}_n, \|\cdot\|_{n,2}, \sigma_n\epsilon) \le N(\mathscr{X}, \|\cdot\|, \epsilon/\xi_K^L) \lesssim (\xi_K^L/\epsilon)^p \le ((1+\xi_K^L)/\epsilon)^p$. Define the classes $\mathscr{I}_< := \{u \mapsto \mathbb{1}_{\{u<y\}} : y \in \mathscr{Y}\}$, $\mathscr{I}_\le := \{u \mapsto \mathbb{1}_{\{u\le y\}} : y \in \mathscr{Y}\}$, and $\mathscr{I}_\psi := \{u \mapsto \psi_y(u) : y \in \mathscr{Y}\}$. Since $\psi_y = 0.5\mathbb{1}_{\{\cdot<y\}} + 0.5\mathbb{1}_{\{\cdot\le y\}}$, we have $\mathscr{I}_\psi \subseteq 0.5\mathscr{I}_< + 0.5\mathscr{I}_\le$. In the continuous response regime, both $\mathscr{I}_<$ and $\mathscr{I}_\le$ are well-known VC classes, and Exercise 2.5.4 in Van der Vaart and Wellner (2023) gives $N(\mathscr{I}_<, L_2(Q), \epsilon) + N(\mathscr{I}_\le, L_2(Q), \epsilon) \lesssim (1/\epsilon^2)$, uniformly over probability measures Q, $0 < \epsilon \le 1$. Lemma 8-(i) with $a = b = 0.5$ therefore yields $\sup_Q N(\mathscr{I}_\psi, L_2(Q), \epsilon) \lesssim (1/\epsilon)^4$. Let $\mathscr{M} := \{(u,x) \mapsto M(y|x) : y \in \mathscr{Y}\}$. In the continuous regime, $M(y|x) = F(y|x)$, the class $\mathscr{M}$ satisfies, by boundedness of the conditional density,

$$|F(y|x) - F(y'|x)| \le \int_{y\wedge y'}^{y\vee y'} |f(u|x)|\,du \le C_f|y' - y|$$

for $y, y' \in \mathscr{Y}$, so it is Lipschitz in the index set and Theorem 2.7.17 in Van der Vaart and Wellner (2023) gives $N(\mathscr{M}, \|\cdot\|_{Q,2}, \epsilon) \le N(\mathscr{Y}, |\cdot|, \epsilon/C_f) \lesssim (1/\epsilon)$, where the last inequality uses the compactness assumption. Since $\mathscr{E} := \{z \mapsto \varepsilon_y(z) : y \in \mathscr{Y}\} \subseteq \mathscr{I}_\psi - \mathscr{M}$, lemma 8-$(i)$ then gives $N(\mathscr{E}, \|\cdot\|_{Q,2}, \epsilon) \le N(\mathscr{I}_\psi - \mathscr{M}, \|\cdot\|_{Q,2}, \epsilon) \lesssim (1/\epsilon)^5$, and thus $\sup_Q N(\mathscr{E}, \|\cdot\|_{Q,2}, \epsilon) \lesssim (1/\epsilon)^5$, where the supremum ranges over all probability measures Q on $\mathscr{Y}\mathscr{X}$ and $0 < \epsilon \le 1$. In the (finite) discrete regime, $\mathscr{E} = \{\varepsilon_{y_1}, \ldots, \varepsilon_{y_J}\}$ has at most J elements, so $N(\mathscr{E}, L_2(Q), \epsilon) \le J \lesssim (1/\epsilon)^5$ for $0 < \epsilon \le 1$. Consequently, in either regime $\sup_Q N(\mathscr{E}, L_2(Q), \epsilon) \lesssim (1/\epsilon)^5$. Take the $\sigma_n\epsilon$-cover for $\mathscr{A}_n$, i.e., the set $\{a_{x_0,l} : 1 \le l \le N\}$ and define, for each center $a_{x_0,l}$, the probability measure on $\mathscr{Y}\mathscr{X}$ denoted by $Q_{n,l}(A) := n^{-1}\sum_{i=1}^n (a_{x_0,l}^2(x_i)/\|a_{x_0,l}\|_{n,2}^2)\delta_{Z_i}(A)$, where $\delta_z(A)$ denotes the Dirac measure and $A \subseteq \mathscr{Y}\mathscr{X}$ (if $\|a_{x_0,l}\|_{n,2} = 0$ we require only one center). Then, $N(\mathscr{E}, \|\cdot\|_{Q_{n,l},2}, \epsilon) \lesssim (1/\epsilon)^5$ (which also holds uniformly over $l \in \{1, \ldots, N\}$). Take the ϵ-cover for $\mathscr{E}$, i.e., the set $\{\varepsilon_{yl,j} : 1 \le j \le M_l\}$, and construct the vector $e_{yl,j} := (\varepsilon_{yl,j}(z_1), \ldots, \varepsilon_{yl,j}(z_n))'$. The set $\{a_{x_0,l} \odot e_{yl,j} : 1 \le l \le N, 1 \le j \le M_l\}$, which has at most $\sum_{l=1}^N M_l \le N \max_{1\le l\le N} M_l$ elements, is a $2\sigma_n\epsilon$-cover for T:

$$\begin{aligned}\|a_{x_0} \odot e_y - a_{x_0,l} \odot e_{yl,j}\|_{n,2} &\le \|(a_{x_0} - a_{x_0,l}) \odot e_y\|_{n,2} + \|a_{x_0,l} \odot (e_y - e_{yl,j})\|_{n,2} \\ &\le \|(a_{x_0} - a_{x_0,l})\|_{n,2} + \|a_{x_0,l}\|_{n,2}\|\varepsilon_y - \varepsilon_{yl,j}\|_{Q_{n,l},2} \le 2\sigma_n\epsilon\end{aligned}$$

since $\sup_{y\in\mathscr{Y}} |\varepsilon_y(z)| \le 1$. Taking the minimal cover yields, for $0 < \epsilon \le 1$

$$N(T, \|\cdot\|_{n,2}, 2\sigma_n\epsilon) \le N(\mathscr{A}_n, \|\cdot\|_{n,2}, \sigma_n\epsilon) \sup_l N(\mathscr{E}, \|\cdot\|_{Q_{n,l},2}, \epsilon) \lesssim \left(\frac{1+\xi_K^L}{\epsilon}\right)^p \left(\frac{1}{\epsilon}\right)^5$$

If $\|a_{x_0,l}\|_{n,2} = 0$, the covering number is 1 and the bound still holds. One can choose $A_K = C(1+\xi_K^L)$ for some constant C, and $V = 5 + p$ to get $N(T, \|\cdot\|_{n,2}, \epsilon) \le (2\sigma_n A_K/\epsilon)^V$,

$0 < \epsilon \leq 2\sigma_n$ satisfying $\log(A_K) \lesssim \log(eK) = O(\log K)$ by Assumption 6-(ii). Lemma 12 then gives $\log(A_K) = O(\log n)$. We also have $\sigma_n = O_{P_Z}(\xi_K\sqrt{\log(K)/n})$ by Lemma 6. Hence

$$\begin{aligned}\int_0^{2\sigma_n}\sqrt{\log(eN(T,\|\cdot\|_{n,2},\epsilon))}\,d\epsilon &\lesssim \int_0^{2\sigma_n}\sqrt{\log(e(2\sigma_n A_K/\epsilon)^V)}\,d\epsilon \\ &= 2\sigma_n\int_0^1\sqrt{\log(e(A_K/u)^V)}\,du \lesssim \sigma_n\left(2\sqrt{\log(eA_K)}+\sqrt{\pi}\right) \\ &= O_{P_Z}\left(\xi_K\frac{\log(en)}{\sqrt{n}}\right)\end{aligned}$$

Conditional Markov's inequality and standard probability algebra then give

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|R_{1n}^{\varepsilon}(y,x)| = O_{P_Z}\left(\xi_K\frac{\log(en)}{\sqrt{n}}\right)$$

Step 2 (Rate for R_{1n}^r). We now show that

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|R_{1n}^{r}(y,x)| = O_{P_Z}\left(\left[\frac{\xi_K^2\log K}{n}\right]^{1/2}\sqrt{K}l_K K^{-\bar{s}/p}\right)$$

First, note that

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|R_{1n}^{r}(y,x)| = \sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|\alpha_K(x)'(\hat{Q}_{K,p}^{-1}-I_K)\mathbb{G}_n(p_K r_y)| \leq \|\hat{Q}_{K,p}^{-1}-I_K\|\sup_{y\in\mathscr{Y}}\|\mathbb{G}_n(p_K r_y)\|$$

where $\|\hat{Q}_{K,p}^{-1}-I_K\| = O_{P_Z}(\xi_K\sqrt{\log(K)/n})$ by Lemma 6. For the second term, note that

$$\begin{aligned}P_K M(y|\cdot)(x) &= q_K(x)'Q_K^{-1}\mathbb{E}[q_K(X)M(y|X)] = p_K(x)'\mathbb{E}[p_K(X)M(y|X)] \\ &= \int_{\mathscr{Y}}p_K(x)'\mathbb{E}[p_K(X)f(t|X)]\psi_y(t)\,d\nu(t) = \int_{\mathscr{Y}}\psi_y(t)P_K f(t|\cdot)(x)\,d\nu(t)\end{aligned}$$

so that $r_y(x) = M(y|x) - M_K(y|x) = \int_{\mathscr{Y}}\psi_y(t)(f(t|x)-P_K f(t|\cdot)(x))\,d\nu(t)$, where $s_{K,t}(x) := f(t|x)-P_K f(t|\cdot)(x)$ is the residual of the projection of f onto $\mathscr{H}_K$. Hence, since $f(t|\cdot)\in\Lambda^s(\mathscr{X})$ uniformly over $t\in\mathscr{Y}$ by Assumption 5-(ii), we have, analogous to item 6-(i) of Lemma 6:

$$\sup_{t\in\mathscr{Y}}\|s_{K,t}\|_\infty = \sup_{t\in\mathscr{Y}}\|(I-P_K)f(t|\cdot)\|_\infty \leq l_K\sup_{t\in\mathscr{Y}}\inf_{h\in\mathscr{H}_K}\|f(t|\cdot)-h\|_\infty \lesssim l_K K^{-\bar{s}/p}$$

and projection orthogonality gives $P[p_K s_{K,t}] = 0$, $t\in\mathscr{Y}$. By these arguments, linearity of

$\mathbb{G}_n$, $|\psi_y(t)| \leq 1$, and Fubini, we have

$$\sup_{y\in\mathscr{Y}} \|\mathbb{G}_n(p_K r_y)\| \leq \sup_{y\in\mathscr{Y}} \int_{\mathscr{Y}} \psi_y(t) \|\mathbb{G}_n(p_K s_{K,t})\| \, d\nu(t) \leq \int_{\mathscr{Y}} \|\mathbb{G}_n(p_K s_{K,t})\| \, d\nu(t)$$

so it is enough to find a rate for $\|\mathbb{G}_n(p_K s_{K,t})\|$ that is uniform over $t \in \mathscr{Y}$. We have

$$\begin{aligned}\mathbb{E}\left\|\mathbb{G}_n(p_K s_{K,t})\right\|^2 &= n^{-1}\mathbb{E}\left\|\sum_{i=1}^n p_K(X_i) s_{K,t}(X_i)\right\|^2 = n^{-1}\sum_{i=1}^n \mathbb{E}\left\|p_K(X_i) s_{K,t}(X_i)\right\|^2 \\ &\leq \sup_{(t,x)\in\mathscr{Y}\mathscr{X}} |s_{K,t}(x)|^2 \mathbb{E}\|p_K(X)\|^2 \lesssim K l_K^2 K^{-2\bar{s}/p}\end{aligned}$$

where the second equality follows by independence and projection orthogonality. By Jensen's inequality, $\mathbb{E}\left\|\mathbb{G}_n(p_K s_{K,t})\right\| \leq \left(\mathbb{E}\left\|\mathbb{G}_n(p_K s_{K,t})\right\|^2\right)^{1/2} \lesssim \sqrt{K} l_K K^{-\bar{s}/p}$ uniformly over $t \in \mathscr{Y}$. Hence, using $\nu(\mathscr{Y}) < \infty$,

$$\mathbb{E}\left[\sup_{y\in\mathscr{Y}} \|\mathbb{G}_n(p_K r_y)\|\right] \leq \int_{\mathscr{Y}} \mathbb{E}\left\|\mathbb{G}_n(p_K s_{K,t})\right\| \, d\nu(t) \lesssim \sup_{t\in\mathscr{Y}} \mathbb{E}\left\|\mathbb{G}_n(p_K s_{K,t})\right\| \lesssim \sqrt{K} l_K K^{-\bar{s}/p}$$

Markov's inequality then gives $\sup_{y\in\mathscr{Y}} \|\mathbb{G}_n(p_K r_y)\| = O_{P_Z}\left(\sqrt{K} l_K K^{-\bar{s}/p}\right)$, and by standard order algebra

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |R_{1n}^r(y,x)| = O_{P_Z}\left(\left[\frac{\xi_K^2 \log K}{n}\right]^{1/2} \sqrt{K} l_K K^{-\bar{s}/p}\right)$$

Combining steps 1 and 2, and using $\log K = O\left(\log(en)\right)$, we can conclude

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |R_{1n}(y,x)| = O_{P_Z}\left(\left[\frac{\xi_K^2 \log(en)}{n}\right]^{1/2} \cdot \left\{\sqrt{\log(en)} + \sqrt{K} l_K K^{-\bar{s}/p}\right\}\right)$$

□

Lemma 13. *(First Stage Uniform Consistency) Let Assumptions 4, 5, and 6 hold. Then, the series estimator of $M(y|x)$ is uniformly consistent, that is,*

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\hat{M}_n(y|x) - M(y|x)| = o_{P_Z}(1)$$

Proof. Write $\hat{M}_n(y|x) - M(y|x) = \left(\hat{M}_n(y|x) - M_K(y|x)\right) + \left(M_K(y|x) - M(y|x)\right)$. By Lemma 6-(6), the approximation part satisfies $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |M_K(y|x) - M(y|x)| \lesssim l_K K^{-\bar{s}/p} = o(1)$ by

the rate assumption, 5-(iv). For the stochastic part, using the normal equations,

$$\begin{aligned}|\hat{M}_n(y|x) - M_K(y|x)| &= \|p_K(x)\||\alpha_K(x)'(\hat{\gamma}_K(y) - \gamma_K(y))| \\ &\leq n^{-1/2}\xi_K\|Q_K^{-1/2}\| \sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\alpha_K(x)'\hat{Q}_{K,p}^{-1}\mathbb{G}_n(p_K U_y)| \\ &\leq n^{-1/2}\xi_K\|Q_K^{-1/2}\|\left(\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\tilde{R}_{1n}^{\varepsilon}(y,x)| + \sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\tilde{R}_{1n}^{r}(y,x)|\right)\end{aligned}$$

where $\tilde{R}_{1n}^{\varepsilon}(y,x_0) := \alpha_K(x_0)'\hat{Q}_{K,p}^{-1}\mathbb{G}_n(p_K\varepsilon_y)$ and $\tilde{R}_{1n}^{r}(y,x_0) := \alpha_K(x_0)'\hat{Q}_{K,p}^{-1}\mathbb{G}_n(p_K r_y)$ are defined analogously to the remainder terms in Proposition 4. By the eigenvalue assumption, 5-(iii), $\|Q_K^{-1/2}\| \leq d_B^{-1/2} = O(1)$. For the supremum terms, we apply the same steps as in the proof of Proposition 4 replacing $\|\hat{Q}_{K,p}^{-1} - I_K\|$ by $\|\hat{Q}_{K,p}^{-1}\|$, which is uniformly bounded in P_Z-probability. We get $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\tilde{R}_{1n}^{\varepsilon}(y,x)| = O_{P_Z}\left(\sqrt{\log K}\right)$ and $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\tilde{R}_{1n}^{r}(y,x)| = O_{P_Z}\left(\sqrt{K}l_K K^{-\bar{s}/p}\right)$. Combining the bounds yields

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\hat{M}_n(y|x) - M_K(y|x)| = O_{P_Z}\left(\xi_K\sqrt{\frac{\log K}{n}} + \xi_K\sqrt{\frac{K}{n}}l_K K^{-\bar{s}/p}\right)$$

where the first term converges to zero by the rates in Assumption 5-(iv), which also imply convergence to zero for the second term since:

$$0 \leq \xi_K\sqrt{\frac{K}{n}}l_K K^{-\bar{s}/p} = \left(K\xi_K\sqrt{\frac{\log K}{n}}\right)\frac{l_K K^{-\bar{s}/p}}{\sqrt{K\log K}} \longrightarrow 0$$

Combining the convergence rates of the stochastic and approximation parts gives

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\hat{M}_n(y|x) - M(y|x)| = o_{P_Z}(1)$$

□

Proposition 5. *(Uniform Bootstrap First-Stage Linearization) Let Assumptions 4, 5, and 6 hold. Then*

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} \left|\sqrt{n}\{\hat{M}_n^*(y|x) - \hat{M}_n(y|x)\} - q_K(x)'Q_K^{-1}\hat{G}_n^*[q_K U_y]\right| = o_{P_w}(1)$$

in P_Z-probability.

Proof. Define the transformed response $\tilde{D}_i(y) := \sqrt{w_i}\psi_y(Y_i)$ and the transformed series vector $\tilde{p}_{K,i} := \sqrt{w_i}p_K(X_i)$. Then, the weighted (Bayesian) bootstrap draw of the normalized

least squares estimator is given by

$$\hat{\gamma}_K^*(y) \in \underset{\gamma \in \mathbb{R}^K}{\arg\min}\, P_n\left[\left(\tilde{D}(y) - \tilde{p}_K'\gamma\right)^2\right]$$

The transformed structural and approximation errors are $\tilde{\varepsilon}_{y,i} := \sqrt{w_i}\varepsilon_y(z_i)$ and $\tilde{r}_{y,i} := \sqrt{w_i}r_y(x_i)$, where $\mathbb{E}[\tilde{\varepsilon}_y(Z)|w,X] = 0$. Then $\tilde{D}_i(y) = \tilde{p}_{K,i}'\gamma_K(y) + \tilde{\varepsilon}_{y,i} + \tilde{r}_{y,i}$ and by independence $\mathbb{E}_{ZW}[\tilde{p}_K(X)\tilde{p}_K(X)'] = I_K$, so that $\gamma_K(y) = \mathbb{E}_{ZW}[\tilde{p}_K(X)\tilde{D}(y)] = \mathbb{E}_Z[p_K(X)\psi_y(Y)]$. The rescaled problem has thus the same population projection coefficient and normalized Gram matrix as the original problem. The normal equations of the rescaled problem give $\hat{\gamma}_K^*(y) - \gamma_K(y) = (\hat{Q}_{K,p}^*)^{-1}P_n[wp_KU_y]$, where $\hat{Q}_{K,p}^* = P_n[wp_Kp_K']$. Thus

$$\begin{aligned}\sqrt{n}\alpha_K(x_0)'\{\hat{\gamma}_K^*(y) - \gamma_K(y)\} &= \alpha_K(x_0)'\mathbb{G}_n(wp_KU_y) + \alpha_K(x_0)'\left\{\left(\hat{Q}_{K,p}^*\right)^{-1} - I_K\right\}\mathbb{G}_n(wp_KU_y)\\ &= \alpha_K(x_0)'\mathbb{G}_n(wp_KU_y) + R_{1n}^*(y,x_0)\end{aligned}$$

On the other hand, Proposition 4 gives

$$\sqrt{n}\alpha_K(x_0)'\{\hat{\gamma}_K(y) - \gamma_K(y)\} = \alpha_K(x_0)'\mathbb{G}_n(p_KU_y) + R_{1n}(y,x_0)$$

Subtracting the two expressions yields

$$\sqrt{n}\alpha_K(x_0)'\{\hat{\gamma}_K^*(y) - \hat{\gamma}_K(y)\} = \alpha_K(x_0)'\widetilde{G}_n^*(p_KU_y) + R_{1n}^b(y,x_0)$$

where $R_{1n}^b := R_{1n}^* - R_{1n}$. We show that $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|R_{1n}^b(y,x)| = O_{P_{ZW}}\left(r_{K,n}^b\right)$, where $r_{K,n}^b$ is the rate defined in Assumption 6-(ii).

To this end, we first find the uniform rate of convergence of the remainder term $R_{1n}^*(y,x)$. Note that Lemma 12 gives, for the re-scaled objects

$$\max_{1\le i\le n}\sup_{x\in\mathscr{X}}\|\tilde{p}_{K,i}(x)\| = O_{P_{ZW}}\left(\xi_K\sqrt{\log n}\right) \quad \max_{1\le i\le n}\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|\tilde{r}_{y,i}(x)| = O_{P_{ZW}}\left(l_KK^{-\bar{s}/p}\sqrt{\log n}\right)$$

$$\max_{1\le i\le n}\sup_{y\in\mathscr{Y}}|\tilde{\varepsilon}_{y,i}| = O_{P_{ZW}}\left(\sqrt{\log n}\right)$$

Using the same steps as in the proof for item 3 of Lemma 6, by Rudelson's LLN (Lemma 6.2 in Belloni et al. (2015)), the above bound for $\|\tilde{p}_K(x)\|$, and the same inverse matrix arguments, we get $\|(\hat{Q}_{K,p}^*)^{-1} - I_K\| = O_{P_{ZW}}\left(\xi_K\log(en)/\sqrt{n}\right)$. Applying the same arguments as in the proof of Proposition 4 to the transformed, re-scaled problem under the joint law P_{ZW}, using the bounds above, and Lemma 8-(ii) so that the metric entropy of the function

classes remains unchanged after the weight multiplication, one gets the uniform rate for R^*_{1n}: Step 1 gives, for the structural error remainder, an order $O_{P_{ZW}}\left(\xi_K \log^2(en)/\sqrt{n}\right)$, since the cover for T is now of radius $\tilde{\sigma}_n\varepsilon := \left(\max_i \sqrt{w_i}\right)\|(\hat{Q}^*_{K,p})^{-1} - I_K\|\,\|\hat{Q}^*_{K,p}\|^{1/2}\epsilon$ because of the bound above for $\tilde{\varepsilon}_y$, and Lemma 8-(ii) guarantees that the covering number bound remains unchanged. Step 2 gives, for the approximation error remainder, an order $O_{P_{ZW}}\left(\xi_K \log^{3/2}(en)\, l_K K^{-\bar{s}/p}\sqrt{K/n}\right)$, by the LLN for the weighted regressors and $\mathbb{E}_{ZW}\|\mathbb{G}_n(w p_K s_{K,t})\|^2 \lesssim K l_K^2 K^{-2\bar{s}/p}$ since $\mathbb{E}_W\left[w^2\right] = 2$. Combining Steps 1 and 2 yields

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |R^*_{1n}(y,x)| = O_{P_{ZW}}\left(\left[\frac{\xi_K^2 \log^3(en)}{n}\right]^{1/2}\left\{\sqrt{\log(en)} + \sqrt{K} l_K K^{-\bar{s}/p}\right\}\right) = O_{P_{ZW}}\left(r^b_{K,n}\right)$$

Moreover, since $r_{K,n} \leq r^b_{K,n}$, we have

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |R^b_{1n}(y,x)| \leq \sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |R^*_{1n}(y,x)| + \sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |R_{1n}(y,x)| = O_{P_{ZW}}\left(r^b_{K,n}\right)$$

which implies

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}} |\sqrt{n}\alpha_K(x)'\{\hat{\gamma}^*_K(y) - \hat{\gamma}_K(y)\} - \alpha_K(x)'\widetilde{G}^*_n(p_K U_y)| = O_{P_{ZW}}\left(r^b_{K,n}\right)$$

We are left to show that the same result is true when replacing $\widetilde{G}^*_n$ by $\hat{G}^*_n$. For that purpose, we exploit the identity $\widetilde{G}^*_n(f) - \hat{G}^*_n(f) = (1-\bar{w}_n^{-1})\widetilde{G}^*_n(f) + \bar{w}_n^{-1}\sqrt{n}(\bar{w}_n - 1)P_n[f]$. Write

$$\begin{aligned}
\sqrt{n}\alpha_K(x_0)'\{\hat{\gamma}^*_K(y) - \hat{\gamma}_K(y)\} &= \alpha_K(x_0)'\hat{G}^*_n(p_K U_y) + (1-\bar{w}_n^{-1})\alpha_K(x_0)'\widetilde{G}^*_n(p_K U_y) \\
&\quad + \bar{w}_n^{-1}\sqrt{n}(\bar{w}_n - 1)\alpha_K(x_0)'P_n[p_K U_y] + R^b_{1n}(y,x_0) \\
&= \alpha_K(x_0)'\hat{G}^*_n(p_K U_y) + (1-\bar{w}_n^{-1})\left(\ddot{R}^\varepsilon_{2,n}(y,x_0) + \ddot{R}^r_{2,n}(y,x_0)\right) \\
&\quad + \bar{w}_n^{-1}\sqrt{n}(\bar{w}_n - 1)n^{-1/2}\left(\ddot{R}^\varepsilon_{1n}(y,x_0) + \ddot{R}^r_{1n}(y,x_0)\right) + R^b_{1n}(y,x_0)
\end{aligned}$$

where $\ddot{R}^\varepsilon_{2,n} := \alpha_K(x_0)'\mathbb{G}_n((w-1)p_K\varepsilon_y)$, $\ddot{R}^r_{2,n} := \alpha_K(x_0)'\mathbb{G}_n((w-1)p_K r_y)$, and both $\ddot{R}^\varepsilon_{1n}$ and $\ddot{R}^r_{1n}$ are defined analogously to R^ε_{1n} and R^r_{1n} from Proposition 4 (replacing $\|\hat{Q}^{-1}_{K,p} - I_K\|$ by $\|I_K\| = 1$). We then apply the same steps as in Proposition 4 to get $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|\ddot{R}^\varepsilon_{1n}(y,x)| = O_{P_Z}(\sqrt{\log K})$ and $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|\ddot{R}^r_{1n}(y,x)| = O_{P_Z}(\sqrt{K} l_K K^{-\bar{s}/p})$. Moreover, we can apply the same steps as in Proposition 4 under P_{ZW} to both $\ddot{R}^\varepsilon_{2,n}$ and $\ddot{R}^r_{2,n}$ in the same way as for $\ddot{R}^\varepsilon_{1n}$ and $\ddot{R}^r_{1n}$ but using Lemma 8-(ii) with $m(w) = w - 1$ so that the metric entropy of the function classes does not change after the multiplication by $w-1$, and Lemma 12 to get the bound $\max_{1\leq i\leq n}|w_i - 1| = O_{P_W}\left(\log(en)\right)$. This gives, in Step 1, the rate $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|\ddot{R}^\varepsilon_{2n}(y,x)| = O_{P_{ZW}}\left(\log(en)\sqrt{\log K}\right)$, since the envelope is now $(\max_i |w_i -$

$1|)\|\hat{Q}_{K,p}\|^{1/2} = O_{P_{ZW}}\left(\log(en)\right)$ and the metric entropy of the function class is unchanged, i.e., we get the order $O\left(\sqrt{\log K}\right)$. Step 2 gives the stochastic order $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|\ddot{R}^r_{2n}(y,x)| = O_{P_{ZW}}\left(\log(en)\sqrt{K}l_K K^{-\bar{s}/p}\right)$. In addition, standard LLN and CLT arguments give $\bar{w}_n^{-1} = O_{P_W}(1)$, $\sqrt{n}(\bar{w}_n - 1) = O_{P_W}(1)$, and $(1-\bar{w}_n^{-1}) = O_{P_W}\left(n^{-1/2}\right)$. We also had, for the last term, $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|R^b_{1n}(y,x)| = O_{P_{ZW}}\left(r^b_{K,n}\right)$. Combining these rates and $\log K = O\left(\log n\right)$ from Lemma 12, we have the following linearization, uniformly over $(y,x)\in\mathscr{Y}\mathscr{X}$

$$\begin{aligned}\sqrt{n}\alpha_K(x_0)'\{\hat{\gamma}^*_K(y) - \hat{\gamma}_K(y)\} &= \alpha_K(x_0)'\hat{G}^*_n(p_K U_y) \\ &\quad + O_{P_{ZW}}\left(\left[\frac{\log^2(en)}{n}\right]^{1/2}\left\{\sqrt{\log(en)} + \sqrt{K}l_K K^{-\bar{s}/p}\right\} + r^b_{K,n}\right) \\ &= \alpha_K(x_0)'\hat{G}^*_n(p_K U_y) + O_{P_{ZW}}\left(r^b_{K,n}\right)\end{aligned}$$

since the first sequence term is bounded by $r^b_{K,n}$, noting that $\log(en) \lesssim \xi_K \log^{3/2}(en)$ since $\xi_K^2 \geq d_B K$. Multiplying both sides by $\|p_K(x)\|$, which is of order $O(\xi_K)$, to undo the normalization yields

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}\left|\sqrt{n}\{\hat{M}^*_n(y|x) - \hat{M}_n(y|x)\} - p_K(x)'\hat{G}^*_n[p_K U_y]\right| = O_{P_{ZW}}\left(\xi_K r^b_{K,n}\right) = o_{P_{ZW}}(1)$$

where the last equality follows by Assumption 6-(ii), which gives $\xi_K r^b_{K,n} \to 0$. Finally, undoing the basis vector normalization $p_K = Q_K^{-1/2} q_K$ and Lemma 7-(i) give

$$\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}\left|\sqrt{n}\{\hat{M}^*_n(y|x) - \hat{M}_n(y|x)\} - q_K(x)'Q_K^{-1}\hat{G}^*_n[q_K U_y]\right| = o_{P_w}(1)$$

in P_Z-probability.

□

E.5 *Proof of Theorem 3 – Validity of the Bootstrap*

The bootstrap statistic of interest is $T^*_n = T_n(Z_1,\ldots,Z_n,w_1,\ldots,w_n) := \sqrt{n}\left(\hat{\theta}^*_n - \hat{\theta}_n\right)$. Consider the bootstrap decomposition:

$$\sqrt{n}\left(\hat{\theta}^*_n - \hat{\theta}_n\right) = \left(A^*_n\right)^{-1}\left\{\sqrt{n}\left(B^*_n - B_n\right) - \sqrt{n}\left(A^*_n - A_n\right)\hat{\theta}_n\right\}$$

Fix $v\in\mathbb{R}^p$. We organize the proof in several steps.

Step 1 (Bootstrap Numerator Linearization). We show that

$$v'\sqrt{n}\left(B^*_n - B_n\right) = \hat{G}^*_n(\phi_v - P\phi_v + 2\Xi_v) + o_{P_w}(1) \quad \text{in } P_Z\text{-probability}$$

Define the bootstrap analog of $\hat{\phi}_v = a_v(y,x)\{2\hat{M}_n(y|x)-1\}$ by $\hat{\phi}^*_v = a_v(y,x)\{2\hat{M}^*_n(y|x)-1\}$, where $a_v(y,x) = y(x'v)$, and write

$$v'\sqrt{n}\left(B^*_n - B_n\right) = \sqrt{n}\left(P^*_n\hat{\phi}^*_v - P_n\hat{\phi}_v\right) =: T^*_{1,v} + T^*_{2,v} + T^*_{3,v}$$

where $T^*_{1,v} := \sqrt{n}\left(P^*_n - P_n\right)(\hat{\phi}^*_v - \hat{\phi}_v)$, $T^*_{2,v} := \sqrt{n}\left(P^*_n - P_n\right)\hat{\phi}_v$, and $T^*_{3,v} := \sqrt{n}P_n(\hat{\phi}^*_v - \hat{\phi}_v)$.

Step 1.1 (Linearization of $T^*_{2,v}$). Note that $T^*_{2,v} := \sqrt{n}\left(P^*_n - P_n\right)\hat{\phi}_v = \hat{G}^*_n(\hat{\phi}_v) = \hat{G}^*_n(\phi_v - P\phi_v) + \hat{G}^*_n(\hat{\phi}_v - \phi_v)$, where $\hat{\phi}_v - \phi_v = 2a_v\left(\hat{M}_n - M\right)$. By compactness and Lemma 13,

$$P_n\|2a_v(\hat{M}_n - M)\|^2 \lesssim P_n(\hat{M}_n - M)^2 \leq \left(\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}\left|\hat{M}_n(y|x) - M(y|x)\right|\right)^2 = o_{P_Z}(1)$$

So by Lemma 11-(ii) we have $\hat{G}^*_n(\hat{\phi}_v - \phi_v) = o_{P_w}(1)$ in P_Z-probability and thus $T^*_{2,v} = \hat{G}^*_n(\phi_v - P\phi_v) + o_{P_w}(1)$ in P_Z-probability.

Step 1.2 ($T^*_{1,v}$ is $o_{P_w}(1)$ in P_Z-probability). Note that $T^*_{1,v} := \sqrt{n}\left(P^*_n - P_n\right)(\hat{\phi}^*_v - \hat{\phi}_v) = 2\hat{G}^*_n(a_v(\hat{M}^*_n - \hat{M}_n))$. For $r^*_n(y,x) := \sqrt{n}\left\{\hat{M}^*_n(y|x) - \hat{M}_n(y|x)\right\} - q_K(x)'Q_K^{-1}\hat{G}^*_n(q_K U_y)$, Proposition 5 gives $\sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|r^*_n(y,x)| = o_{P_w}(1)$ in P_Z-probability. Hence

$$T^*_{1,v} = 2n^{-1/2}\hat{G}^*_n(a_v q_K' Q_K^{-1}\hat{G}^*_n(q_K U_Y)) + 2n^{-1/2}\hat{G}^*_n(a_v r^*_n)$$

where the second term satisfies, by compactness and $P^*_n 1 = P_n 1 = 1$,

$$|2n^{-1/2}\hat{G}^*_n(a_v r^*_n)| = 2|(P^*_n - P_n)a_v r^*_n| \lesssim \sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|r^*_n(y,x)| = o_{P_w}(1) \quad \text{in } P_Z\text{-probability.}$$

We now handle the first term. Note that

$$2n^{-1/2}\hat{G}^*_n(a_v q_K' Q_K^{-1}\hat{G}^*_n(q_K U_Y)) = \left(2\bar{w}_n^{-2}\right)n^{-3/2}\sum_{i=1}^n\sum_{j=1}^n c_i H_{ij} c_j =: S^*_{n,v}$$

where $c_i := w_i - \bar{w}_n$ and $H_{ij} := a_v(Z_i)q_K(X_i)'Q_K^{-1}q_K(X_j)U_{Y_i}(Z_j)$. We can thus write it as a quadratic form $S^*_{n,v} = \left(2n^{-3/2}\bar{w}_n^{-2}\right)c'Hc$, where $H := (H_{ij})$ is an $n\times n$ matrix and $c := (c_1,\ldots,c_n)' \in \mathbb{R}^n$, which can be written as $c = \mathsf{M}_n u$, where $u := (u_1,\ldots,u_n)' \in \mathbb{R}^n$ with $u_i := w_i - 1$, and $\mathsf{M}_n := I_n - n^{-1}\mathbf{1}\mathbf{1}'$ is an $n\times n$ orthogonal projection matrix with $\mathbf{1} := (1,\ldots,1)' \in \mathbb{R}^n$. Hence, $c'Hc = u'\mathsf{M}_n H\mathsf{M}_n u = u'Au$, where $A := \mathsf{M}_n(H' + H)\mathsf{M}_n/2$ is symmetric. This last property of quadratic forms allows us to work with a symmetric matrix $A = (a_{ij})$, since H need not be symmetric. We show that $S^*_{n,v} = o_{P_w}(1)$ in P_Z-probability.

The term $2\bar{w}_n^{-2} = O_{P_W}(1)$ by the standard LLN and CMT arguments. Lemma 7-(iv) thus

gives $2\bar{w}_n^{-2} = O_{P_w}(1)$ in P_Z-probability. It remains to show $n^{-3/2}u'Au = o_{P_w}(1)$ in P_Z-probability. Define $D := \sum_{i=1}^n a_{ii}u_i^2$ and $O := 2\sum_{i<j} a_{ij}u_iu_j$. Note that

$$\mathbb{E}_w[(u'Au)^2] = \mathbb{E}_w[D^2] + \mathbb{E}_w[O^2] + 2\mathbb{E}_w[DO]$$

where $\mathbb{E}_w[DO] = 2\sum_{i=1}^n \sum_{l<k} \mathbb{E}_w[u_i^2 u_l u_k] a_{ii}a_{lk} = 0$ by independence. Moreover, since $\mathbb{E}_W[u^2] = 1$, we have $\mathbb{E}_w[O^2] = 4\sum_{i<j}\sum_{l<k}\mathbb{E}_w[u_iu_ju_lu_k]a_{ij}a_{lk} = 4\sum_{i<j} a_{ij}^2$. For the remaining term, using $\mathbb{E}_W[u^4] = 9$ we get $\mathbb{E}_w[D^2] = \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}_w[u_i^2u_j^2]a_{ii}a_{jj} = 8\sum_{i=1}^n a_{ii}^2 + \mathrm{tr}(A)^2$. This gives

$$\mathbb{E}_w[(u'Au)^2] = 6\sum_{i=1}^n a_{ii}^2 + 2\|A\|_F^2 + \{\mathrm{tr}(A)\}^2 \lesssim \|A\|_F^2 + \{\mathrm{tr}(A)\}^2 \leq \|H\|_F^2 + \{\mathrm{tr}(\mathsf{M}_n H)\}^2$$

where the last inequality uses the fact that the operator norm of the projection matrix is 1 and the cyclical property of the trace operator. For the first term, note that $\forall K \geq K_0$,

$$\begin{aligned}
\mathbb{E}_Z\|H\|_F^2 &= \sum_{i=1}^n \mathbb{E}_Z[H_{ii}^2] + \sum_{i\neq j}^n \mathbb{E}_Z[H_{ij}^2] \\
&\lesssim \sum_{i=1}^n \mathbb{E}_Z[(q_K(X_i)'Q_K^{-1}q_K(X_i))^2] + \sum_{i\neq j}^n \mathbb{E}_Z[q_K(X_j)'Q_K^{-1}q_K(X_i)q_K(X_i)'Q_K^{-1}q_K(X_j)] \\
&\leq n\mathbb{E}_Z[\|q_K(X)\|^4]\|Q_K^{-1}\|^2 + n(n-1)\,\mathrm{tr}(I_K) \\
&\leq n\left(\sup_{x\in\mathscr{X}}\|q_K(x)\|\right)^2 \mathbb{E}_Z\|q_K(X)\|^2\|Q_K^{-1}\|^2 + n(n-1)K \\
&\lesssim n\xi_K^2 K\|Q_K^{-1}\|^2 + n(n-1)K
\end{aligned}$$

where the first inequality follows by compactness of $\mathscr{Y}\mathscr{X}$ and uniform boundedness of U_y for large enough K (Lemma 6-6), the second inequality follows by independence for $i \neq j$, and the third one by definition and Lemma 6-2. Multiplying by n^{-3} and Markov's inequality gives $n^{-3}\|H\|_F^2 = O_{P_Z}(\xi_K^2\frac{K}{n^2} + \frac{K}{n}) = o_{P_Z}(1)$ by the rate assumption 5-(iv), as shown at the end of Step 1.3 of Theorem 2. Regarding the second term, by $(a-b)^2 \leq 2a^2 + 2b^2$ and Cauchy-Schwarz inequality,

$$\{\mathrm{tr}(\mathsf{M}_n H)\}^2 \lesssim \{\mathrm{tr}(H)\}^2 + n^{-2}(\mathbf{1}'H\mathbf{1})^2 \leq \left(\sum_{i=1}^n H_{ii}\right)^2 + n^{-2}\|\mathbf{1}\|^4\|H\|_F^2 \leq n\sum_{i=1}^n H_{ii}^2 + \|H\|_F^2$$

where multiplying by n^{-3} and by the steps above we get $n^{-2}\sum_{i=1}^n H_{ii}^2 = O_{P_Z}(\xi_K^2 K/n) = o_{P_Z}(1)$ and $n^{-3}\|H\|_F^2 = o_{P_Z}(1)$. Hence, $n^{-3}\mathbb{E}_w[(u'Au)^2] = o_{P_Z}(1)$. Markov's inequality gives, for all $\epsilon > 0$, $P_w(n^{-3/2}|u'Au| > \epsilon) \leq \epsilon^{-2}n^{-3}\mathbb{E}_w[(u'Au)^2] = o_{P_Z}(1)$ and thus the term

$n^{-3/2}u'Au = o_{P_w}(1)$ in P_Z-probability by definition. Standard order algebra gives $S^*_{n,v} = o_{P_w}(1)$ in P_Z-probability and thus $T^*_{1,v} = o_{P_w}(1)$ in P_Z-probability.

Step 1.3 (Linearization of $T^*_{3,v}$). Note that $T^*_{3,v} := \sqrt{n}P_n(\hat{\phi}^*_v - \hat{\phi}_v) = 2\sqrt{n}P_n[a_v(\hat{M}^*_n - \hat{M}_n)]$. Using the bootstrap linearization from Proposition 5 we get

$$T^*_{3,v} = 2n^{-1}\sum_{i=1}^n a_v(Z_i)q_K(X_i)'Q_K^{-1}\hat{G}^*_n(q_K U_{Y_i}) + 2P_n[a_v r^*_n]$$

where by compactness of $\mathscr{Y}\mathscr{X}$ and Proposition 5, the second term satisfies $|2P_n[a_v r^*_n]| \lesssim \sup_{(y,x)\in\mathscr{Y}\mathscr{X}}|r^*_n(y,x)| = o_{P_w}(1)$ in P_Z-probability. We now handle the first term to recover the remaining Ξ-component in χ.

Write $T^*_{3,v} = 2\hat{G}^*_n(\widetilde{\Xi}_{v,K,n}) + o_{P_w}(1)$ in P_Z-probability, where we define

$$\widetilde{\Xi}_{v,K,n}(Z_j) := n^{-1}\sum_{i=1}^n a_v(Z_i)q_K(X_i)'Q_K^{-1}q_K(X_j)U_{Y_i}(Z_j) = P_n[k_{K,v}(\cdot,Z_j)]$$

where $k_{K,v}(z_1,z_2)$ is defined as in Step 1.3 of the proof of Theorem 2, i.e., $k_{K,v}(z_1,z_2) = a_v(z_1)q_K(x_1)'Q_K^{-1}q_K(x_2)U_{y_1}(z_2)$, and recall the K-dependent population counterpart

$$\begin{aligned} P_1 k_{K,v}(\cdot,z_2) &= \int k_{K,v}(z_1,z_2)dP(z_1) = \mathbb{E}[a_v(Z)q_K(X)'Q_K^{-1}q_K(x_2)(\psi_Y(y_2) - M_K(Y|x_2))] \\ &= \int_{\mathscr{Y}} \alpha_{v,K}(t,x_2)(\psi_t(y_2) - M_K(t|x_2))d\nu(t) = \Xi_{v,K}(z_2) \end{aligned}$$

where $\alpha_{v,K}(y,X)$ was defined as the projection of $\delta_v(y,X) = y(X'v)f(y|X)$ onto $\mathscr{H}_K$. Recall as well that the proof of Theorem 2 (Step 1.2) showed that $\Xi_{v,K}(z) \to \Xi_v(z)$ in $L_2(P)$, where $\Xi_v(z) = \int_{\mathscr{Y}}\delta_v(t,x)(\psi_t(y) - M(t|x))d\nu(t)$, which implies $\mathbb{E}_Z P_n(\Xi_{v,K} - \Xi_v)^2 = P(\Xi_{v,K} - \Xi_v)^2 \to 0$. The difference between $\widetilde{\Xi}_{v,K,n}$ and Ξ_v can be split into a stochastic difference, denoted by $\Delta_{v,K,n}(z) := \widetilde{\Xi}_{v,K,n} - \Xi_{v,K}$, and an approximation difference $\Xi_{v,K}(z) - \Xi_v(z)$. Write

$$\hat{G}^*_n(\widetilde{\Xi}_{v,K,n}) = \hat{G}^*_n(\Xi_v) + \hat{G}^*_n(\Delta_{v,K,n}) + \hat{G}^*_n(\Xi_{v,K} - \Xi_v)$$

where the latter difference was shown to satisfy $P_n[(\Xi_{v,K} - \Xi_v)^2] = o_{P_Z}(1)$. Lemma 11-$(ii)$ gives $\hat{G}^*_n(\Xi_{v,K} - \Xi_v) = o_{P_w}(1)$ in P_Z-probability. We are thus left to show that $P_n[\Delta^2_{v,K,n}] = o_{P_Z}(1)$ and apply the same Lemma to get the desired result.

Let $\Delta_{v,K,n}(z_2) = \widetilde{\Xi}_{v,K,n}(z_2) - \Xi_{v,K}(z_2) = n^{-1}\sum_{i=1}^n h_{K,v}(Z_i,z_2)$, where, as defined in the proof of the previous theorem, $h_{K,v}(z_1,z_2) = k_{K,v}(z_1,z_2) - P_1 k_{K,v}(\cdot,z_2)$. Likewise, for simplicity, write $h_{ii} := h_{K,v}(Z_i,Z_i)$ and $h_{ij} := h_{K,v}(Z_i,Z_j)$, as well as $h_{11} := h_{K,v}(Z_1,Z_1)$ and $h_{12} :=$

$h_{K,v}(Z_1, Z_2)$, where Z_1 and Z_2 are independent copies. We adopt the same convention for $k_{K,v}$. Note that, conditional on Z_j, the random variables $\{h_{ij} : i \neq j\}$ are mutually independent and centered (and h_{jj} is fixed) so that

$$\mathbb{E}_Z\left[\left(n^{-1}\sum_{i=1}^n h_{ij}\right)^2 \Big| Z_j\right] = n^{-2}\sum_{i=1}^n \mathbb{E}_Z[h_{ij}^2|Z_j] = n^{-2}\left(h_{jj}^2 + \sum_{i\neq j}\mathbb{E}_Z[h_{ij}^2|Z_j]\right)$$

Taking expectations and averaging over j, $\mathbb{E}_Z P_n[\Delta_{v,K,n}^2] = n^{-2}\mathbb{E}_Z[h_{11}^2] + n^{-2}(n-1)\mathbb{E}_Z[h_{12}^2]$. Note that $n^{-2}(n-1)\mathbb{E}_Z[h_{12}^2] \leq n^{-1}\mathbb{E}_Z[\mathbb{V}ar[k_{12}|Z_2]] \leq n^{-1}\mathbb{E}_Z[k_{12}^2]$ and that

$$\begin{aligned} n^{-2}\mathbb{E}_Z[h_{11}^2] &\leq 2n^{-2}\mathbb{E}_Z[k_{11}^2] + 2n^{-2}\mathbb{E}_Z[(P_1 k_{12})^2(Z)] \\ &\leq 2n^{-2}\mathbb{E}_Z[k_{11}^2] + 2n^{-2}\mathbb{E}_Z[k_{12}^2] \leq 2n^{-2}\mathbb{E}_Z[k_{11}^2] + 2n^{-1}\mathbb{E}_Z[k_{12}^2] \\ &\lesssim n^{-2}\mathbb{E}_Z[k_{11}^2] + n^{-1}\mathbb{E}_Z[k_{12}^2] \end{aligned}$$

where the first inequality uses $(a-b)^2 \leq 2a^2 + 2b^2$ and the second Jensen's inequality. We then have $\mathbb{E}_Z P_n[\Delta_{v,K,n}^2] \lesssim n^{-2}\mathbb{E}_Z[k_{11}^2] + n^{-1}\mathbb{E}_Z[k_{12}^2]$. From the proof of Theorem 2 (Step 1.3), we have $\mathbb{E}_Z[k_{11}^2] = O(K\xi_K^2)$ and $\mathbb{E}_Z[k_{12}^2] = O(K)$. Markov's inequality and the rate assumption 5-(iv) give $P_n[\Delta_{v,K,n}^2] = O_{P_Z}(K\xi_K^2/n^2 + K/n) = o_{P_Z}(1)$. Lemma 11-$(ii)$ then gives $\hat{G}_n^*(\Delta_{v,K,n}) = o_{P_w}(1)$ in P_Z-probability. Standard order algebra then gives $T_{3,v}^* = \hat{G}_n^*(2\Xi_v) + o_{P_w}(1)$ in P_Z-probability.

Combining Steps 1.1-1.3 we get $v'\sqrt{n}\left(B_n^* - B_n\right) = \hat{G}_n^*(\phi_v - P\phi_v + 2\Xi_v) + o_{P_w}(1)$ in P_Z-probability.

Step 2 (Bootstrap Denominator Linearization). We show that

$$v'\sqrt{n}\left(A_n^* - A_n\right)\hat{\theta}_n = \hat{G}_n^*\left(v'\{XX'Y - A_0\}\theta_0\right) + o_{P_w}(1) \quad \text{in } P_Z\text{-probability.}$$

Write $v'\sqrt{n}\left(A_n^* - A_n\right)\hat{\theta}_n = \hat{G}_n^*(v'XX'Y\hat{\theta}_n)$. Since $\hat{G}_n^*$ absorbs constants, we have

$$\hat{G}_n^*(v'XX'Y\hat{\theta}_n) = \hat{G}_n^*\left(v'\{XX'Y - A_0\}\theta_0\right) + \hat{G}_n^*\left(v'\{XX'Y - A_0\}(\hat{\theta}_n - \theta_0)\right)$$

Note that $P_n|v'\{XX'Y - A_0\}(\hat{\theta}_n - \theta_0)|^2 \leq \|v\|^2\|\hat{\theta}_n - \theta_0\|^2 P_n\|XX'Y - A_0\|^2 = o_{P_Z}(1)$ by consistency of $\hat{\theta}_n$ for θ_0 from Theorem 2 and by the standard LLN and compactness of $\mathscr{Y}\mathscr{X}$, which give $P_n\|XX'Y - A_0\|^2 = O_{P_Z}(1)$. Since $\|v\| < \infty$, the product is then $o_{P_Z}(1)$. Lemma 11-(ii) then gives $\hat{G}_n^*\left(v'\{XX'Y - A_0\}(\hat{\theta}_n - \theta_0)\right) = o_{P_w}(1)$ in P_Z-probability. Thus, $v'\sqrt{n}\left(A_n^* - A_n\right)\hat{\theta}_n = \hat{G}_n^*\left(v'\{XX'Y - A_0\}\theta_0\right) + o_{P_w}(1)$ in P_Z-probability.

Step 3 (Bootstrap GIR Linearization) We now show that $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) = A_0^{-1}\hat{G}_n^*(\chi) + o_{P_w}(1)$

in P_Z-probability, where $\chi(Z) = \phi(Z) - P\phi + 2\Xi - \{XX'Y - A_0\}\theta_0 = XU(\theta_0) + 2\Xi(Z)$ with $P\chi = 0$. Steps 1 and 2 then give $v'\sqrt{n}\left(B_n^* - B_n\right) - v'\sqrt{n}\left(A_n^* - A_n\right)\hat{\theta}_n = \hat{G}_n^*(v'\chi) + o_{P_w}(1)$, and since p is fixed, one can obtain the vector statement by applying the scalar expansions for each canonical basis vector, $v = e_i$, $i \in \{1, \ldots, p\}$, and stacking them in a $p \times 1$ vector. Then

$$\sqrt{n}\left(B_n^* - B_n\right) - \sqrt{n}\left(A_n^* - A_n\right)\hat{\theta}_n = \hat{G}_n^*(\chi) + o_{P_w}(1)$$

Write $A_n^* = A_0 + (A_n^* - A_n) + (A_n - A_0)$, where $(A_n^* - A_n) = n^{-1/2}\hat{G}_n^*(XX'Y - A_0) = o_{P_w}(1)$ in P_Z-probability, by the same arguments as in Step 2. The standard LLN gives $(A_n - A_0) = o_{P_Z}(1)$, and since the quantity does not depend on the weights, Lemma 7 gives $(A_n - A_0) = o_{P_w}(1)$ in P_Z-probability. We then have $A_n^* = A_0 + o_{P_w}(1)$ in P_Z-probability. The triangle and Weyl's inequality give $|\lambda_{\min}(A_n^*) - \lambda_{\min}(A_0)| \leq \|A_n^* - A_n\| + \|A_n - A_0\| = o_{P_w}(1)$ in P_Z-probability. So A_n^* is invertible with P_w-probability approaching 1 in P_Z-probability by nonsingularity of A_0 in Assumption 4-(ii), which gives $\lambda_{\min}(A_0) > 0$. Continuous mapping then gives $\left(A_n^*\right)^{-1} = A_0^{-1} + o_{P_w}(1)$ in P_Z-probability. Hence, since by compactness and boundedness of f, and $\nu(\mathscr{Y}) < \infty$, we have $P\|\chi\|^2 < \infty$, Lemma 11-(iii) gives $\hat{G}_n^*(\chi) = O_{P_w}(1)$ in P_Z-probability, and standard order algebra yields

$$\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) = \left(A_n^*\right)^{-1}\left\{\sqrt{n}\left(B_n^* - B_n\right) - \sqrt{n}\left(A_n^* - A_n\right)\hat{\theta}_n\right\} = A_0^{-1}\hat{G}_n^*(\chi) + o_{P_w}(1)$$

in P_Z-probability.

Step 4 (Weak Convergence). Since $P\chi = 0$, $P\|\chi\|^2 < \infty$, $\mathbb{E}_W[w - 1] = 0$, and $\mathbb{E}_W[(w-1)^2] = 1$, by the conditional multiplier CLT in Lemma 2.9.5 of Van der Vaart and Wellner (2023), we have $\widetilde{G}_n^*(\chi) = n^{-1/2}\sum_{i=1}^n (w_i - 1)\chi(Z_i) \underset{P_Z}{\rightsquigarrow} \ddot{W}$, where $\ddot{W} \sim N(0, \mathbb{E}[\chi(Z)\chi(Z)'])$. Moreover, $\hat{G}_n^*(\chi) = \bar{w}_n^{-1}\left(\widetilde{G}_n^*(\chi) - \sqrt{n}(\bar{w}_n - 1)P_n\chi\right)$, where $\bar{w}_n^{-1} = 1 + o_{P_w}(1)$ and $\widetilde{G}_n^*(\chi) = O_{P_w}(1)$, both in P_Z-probability. The latter follows since $\mathbb{E}_w\|\widetilde{G}_n^*(\chi)\|^2 = P_n\|\chi\|^2 \to P\|\chi\|^2 < \infty$ and by conditional Markov's inequality. Moreover, by Markov's inequality

$$P_w\left(\|\sqrt{n}(\bar{w}_n - 1)P_n\chi\| > \epsilon\right) \leq \epsilon^{-2} n\mathbb{E}_W[(\bar{w}_n - 1)^2]\|P_n\chi\|^2 = \epsilon^{-2}\|P_n\chi\|^2 = o_{P_Z}(1)$$

for all $\epsilon > 0$, which gives $\sqrt{n}(\bar{w}_n - 1)P_n\chi = o_{P_w}(1)$ in P_Z-probability. Conditional Slutsky then gives $\hat{G}_n^*(\chi) \underset{P_Z}{\rightsquigarrow} \ddot{W}$.

Note that we have shown $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) = A_0^{-1}\hat{G}_n^*(\chi) + R_n^*$ for some remainder $R_n^* = o_{P_w}(1)$ in P_Z-probability, and where by conditional Slutsky we have $A_0^{-1}\hat{G}_n^*(\chi) \underset{P_Z}{\rightsquigarrow} W$, where $W \sim$

$N(0,\Sigma)$ and $\Sigma = A_0^{-1}\mathbb{E}[\chi(Z)\chi(Z)']A_0^{-1}$. By definition, this means

$$\sup_{h\in BL_1(\mathbb{R}^p)} \left|\mathbb{E}_w\left[h\left(A_0^{-1}\hat{G}_n^*(\chi)\right)\right] - \mathbb{E}[h(W)]\right| = o_{P_Z}(1)$$

where the expectation $\mathbb{E}[h(W)]$ is taken with respect to the distribution of the random element W. We now show $T_n^* = \sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) \underset{P_Z}{\rightsquigarrow} W$. By definition,

$$\begin{aligned}
D_n^* &:= \sup_{h\in BL_1(\mathbb{R}^p)} \left|\mathbb{E}_w\left[h(T_n^*)\right] - \mathbb{E}[h(W)]\right| \\
&\leq \sup_{h\in BL_1(\mathbb{R}^p)} \left|\mathbb{E}_w\left[h(T_n^*)\right] - \mathbb{E}_w\left[h\left(A_0^{-1}\hat{G}_n^*(\chi)\right)\right]\right| + \sup_{h\in BL_1(\mathbb{R}^p)} \left|\mathbb{E}_w\left[h\left(A_0^{-1}\hat{G}_n^*(\chi)\right)\right] - \mathbb{E}[h(W)]\right| \\
&=: D_{1n}^* + D_{2n}^*
\end{aligned}$$

where $D_{2n}^* = o_{P_Z}(1)$ by the CLT argument. For the first term, note that for every $h \in BL_1(\mathbb{R}^p)$ we have $|h(x+r)-h(x)| \leq |h(x+r)|+|h(x)| \leq 2$ by the boundedness property, and $|h(x+r)-h(x)| \leq \|r\|$ by the Lipschitz property. Hence $|\mathbb{E}_w\left[h(T_n^*)\right] - \mathbb{E}_w[h(A_0^{-1}\hat{G}_n^*(\chi))]| \leq \mathbb{E}_w[\|R_n^*\| \wedge 2]$, which does not depend on h, so that $D_{1n}^* \leq \mathbb{E}_w[\|R_n^*\| \wedge 2]$, which we show to be $o_{P_Z}(1)$ by exploiting the fact that R_n^* is $o_{P_w}(1)$ in P_Z-probability. Let $\delta > 0$ and note that $\|R_n^*\| \wedge 2 \leq \delta + 2\mathbb{1}_{\{\|R_n^*\|>\delta\}}$ so $\mathbb{E}_w[\|R_n^*\| \wedge 2] \leq \delta + 2P_w(\|R_n^*\| > \delta)$. Let $\eta \in (0,2]$ and take $\delta = \eta/2$. In the event that $\left\{\mathbb{E}_w[\|R_n^*\| \wedge 2] > \eta\right\}$ we have $\eta < \eta/2 + 2P_w(\|R_n^*\| > \eta/2)$, which implies the event $\left\{P_w(\|R_n^*\| > \eta/2) > \eta/4\right\}$. Therefore, $P_Z(\mathbb{E}_w[\|R_n^*\| \wedge 2] > \eta) \leq P_Z(P_w(\|R_n^*\| > \eta/2) > \eta/4) \to 0$ since $R_n^* = o_{P_w}(1)$ in P_Z-probability. This gives $D_{1n}^* = o_{P_Z}(1)$ and thus $D_n^* = o_{P_Z}(1)$. By definition, we can conclude that $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right) \underset{P_Z}{\rightsquigarrow} W$, so the Bayesian bootstrap consistently estimates the limit law W of the linear GIR estimator.

□

F Proof of Theorems 4 and 5

Before proving Theorem 4, we first show that the conditional Gini minimizes L^G in a conditional expected loss minimization problem for P_X-almost every x (Lemma 14 below). We do the same for the loss function $\tilde{L}^G$ defined in Appendix B before proving Theorem 5 (Lemma 15 below).

F.1 *Conditional Gini as a Conditional Expected Loss Minimizer*

Lemma 14. *(Conditional Gini as a Conditional Expected Loss Minimizer) Let $Y \geq 0$ a.s. be a random variable with $0 < \mathbb{E}[Y] < \infty$ and $0 < \mathbb{E}[Y|X = x] < \infty$ for P_X-almost every x.*

Let $L^G(y,x,a)$ be the Gini Index Loss Function defined in (19). Then, for P_X-almost every x, $\mathbb{E}\left[L^G(Y,x,a)\middle|X=x\right]$ is finite $\forall\, a\in\mathbb{R}$ and uniquely minimized at $a^*(x)=\mathbb{G}(Y|X=x)=G(x)$.

Proof. Let $P_X(x)$ denote the law of X. Note that $\left|L^G\right| \leq y(1+a^2+2|a|)$, so its conditional expectation $\mathbb{E}\left[|L^G(Y,x,a)|\middle|X=x\right]<\infty\ \forall a\in\mathbb{R}$ and for P_X-almost every x, since $0<\mathbb{E}[Y|X=x]<\infty$. For the second claim:

$$\begin{aligned}
\mathbb{E}\left[L^G(Y,x,a)|X=x\right] &= \mathbb{E}\left[Y\{(2M(Y|x)-1)-G(x)+G(x)-a\}^2|X=x\right] \\
&= \mathbb{E}\left[Y\{(2M(Y|x)-1)-G(x)\}^2|X=x\right]+\mathbb{E}\left[Y\,(G(x)-a)^2\,|X=x\right] \\
&\quad +2\mathbb{E}\left[Y\{(2M(Y|x)-1)-G(x)\}|X=x\right](G(x)-a) \\
&= \mathbb{E}\left[L^G(Y,x,G(x))|X=x\right]+\mathbb{E}[Y|X=x]\,(G(x)-a)^2 \\
&\geq \mathbb{E}\left[L^G(Y,x,G(x))|X=x\right]
\end{aligned}$$

with equality if and only if $a=G(x)=\mathbb{G}(Y|X=x)$.

□

F.2 *Proof of Theorem 4 – CGF as an Optimal Predictor*

Note that $\mathbb{E}\left[\left|L^G(Y,X,h(X))\right|\right]\leq\mathbb{E}[Y]+\mathbb{E}[Yh(X)^2]+2\mathbb{E}[Y|h(X)|]<\infty$, since by assumption $\mathbb{E}[Y]<\infty$, $\mathbb{E}\left[Yh(X)^2\right]<\infty$, and by the Cauchy-Schwarz inequality $\mathbb{E}[Y|h(X)|]\leq\mathbb{E}\left[Yh(X)^2\right]^{1/2}\mathbb{E}[Y]^{1/2}<\infty$. When $h(X)=\mathbb{G}(Y|X)$ we have $\mathbb{E}\left[Y\mathbb{G}(Y|X)^2\right]\leq\mathbb{E}[Y]<\infty$, since $|\mathbb{G}(Y|X)|\leq 1$. Finally, by the Law of Iterated Expectations and Lemma 14 we get

$$\mathbb{E}\left[L^G(Y,X,h(X))\right]-\mathbb{E}\left[L^G(Y,X,\mathbb{G}(Y|X))\right]\geq 0$$

with equality if and only if $h(X)=\mathbb{G}(Y|X)$ *a.s.*

Since $\mathbb{E}[L^G(Y,X,h(X))]/\mathbb{E}[Y]=\mathbb{E}\left[w(Y)\{(2M(Y|X)-1)-h(X)\}^2\right]$ the minimizer remains unchanged. For the second claim:

$$\begin{aligned}
\mathbb{E}\left[w(Y)\{(2M(Y)-1)-a\}^2\right] &= \mathbb{E}\left[w(Y)\{(2M(Y)-1)-\mathbb{G}(Y)+\mathbb{G}(Y)-a\}^2\right] \\
&= \mathbb{E}\left[w(Y)\{(2M(Y)-1)-\mathbb{G}(Y)\}^2\right]+\mathbb{E}\left[w(Y)\,(\mathbb{G}(Y)-a)^2\right] \\
&\quad +2\mathbb{E}\left[w(Y)\{(2M(Y)-1)-\mathbb{G}(Y)\}\right](\mathbb{G}(Y)-a) \\
&= \mathbb{E}\left[w(Y)\{(2M(Y)-1)-\mathbb{G}(Y)\}^2\right]+(\mathbb{G}(Y)-a)^2 \\
&\geq \mathbb{E}\left[w(Y)\{(2M(Y)-1)-\mathbb{G}(Y)\}^2\right]
\end{aligned}$$

with equality if and only if $a=\mathbb{G}(Y)$.

□

F.3 *CGF as a Conditional Expected Ratio-Loss Minimizer*

Lemma 15. *(Conditional Gini as a Conditional Expected Loss Minimizer) Let $Y \geq 0$ a.s. be a random variable with $0 < \mathbb{E}[Y] < \infty$, $0 < \mathbb{E}[Y|X = x] < \infty$ for P_X-almost every x, and $\mathbb{E}\left[Y^3\right] < \infty$. Let $\tilde{L}^G(y,x,a)$ be the loss function defined in (29). Then, for P_X-almost every x, $\mathbb{E}\left[\tilde{L}^G(Y,x,a)\middle|X = x\right]$ is finite $\forall\, a \in \mathbb{R}$, and strictly convex in a with unique minimizer $a^*(x) = t(x)/s(x) = \mathbb{G}(Y|X = x) = G(x)$.*

Proof. Let $P_X(x)$ denote the marginal distribution function of X. Note that $\tilde{L}^G(y,a,x) = -y^3 + a^2 y - 2ar(y,x) + 2yr(y,x)$, where $r(y,x) = y(2M(y|x) - 1)$ and thus $|r(y,x)| \leq y$, since $y \geq 0$. Then, $\left|\tilde{L}^G\right| \leq y^3 + a^2 y + 2|a|y + 2y^2$, so $\mathbb{E}\left[|\tilde{L}^G(Y,x,a)|\middle|X = x\right] < \infty\ \forall\, a \in \mathbb{R}$ and for P_X-almost every x, since $\mathbb{E}[Y^3] < \infty$. We now show strict convexity and that $\mathbb{G}(Y|X = x) = G(x)$ is the unique minimizer.

Note that $\mathbb{E}\left[\tilde{L}^G(Y,x,a)\middle|X = x\right] = s(x)a^2 - 2t(x)a + C(x)$, where $s(x) := \mathbb{E}[Y|X = x] > 0$, $t(x) := \mathbb{E}[Y(2M(Y|x) - 1)|X = x]$, and $C(x) := \mathbb{E}\left[2Yr(Y,x) - Y^3\middle|X = x\right]$ is finite and does not depend on a. Strict convexity in a for P_X-almost every x follows directly, since $s(x) > 0$. Noticing that $G(x) = t(x)/s(x)$ is the candidate optimum, one can show that

$$\begin{aligned}
\mathbb{E}\left[\tilde{L}^G(Y,x,a)\middle|X = x\right] - \mathbb{E}\left[\tilde{L}^G(Y,x,G(x))\middle|X = x\right] &= s(x)a^2 - 2t(x)a - s(x)G(x)^2 + 2t(x)G(x) \\
&= s(x)a^2 - 2s(x)G(x)a + G(x)^2 s(x) \\
&= s(x)\,(a - G(x))^2 \;\geq\; 0
\end{aligned}$$

with equality if and only if $a = G(x) = \mathbb{G}(Y|X = x)$.

□

F.4 *Proof of Theorem 5 – CGF as an Optimal Predictor Under Ratio-Loss*

Note that $\mathbb{E}\left[\left|\tilde{L}^G(Y,X,h(X))\right|\right] \leq \mathbb{E}[Y^3] + \mathbb{E}[Yh(X)^2] + 2\mathbb{E}[Y|h(X)|] + 2\mathbb{E}[Y^2] < \infty$, since by assumption $\mathbb{E}\left[Y^3\right] < \infty$, $\mathbb{E}\left[Yh(X)^2\right] < \infty$, and by the Cauchy-Schwarz inequality $\mathbb{E}[Y|h(X)|] \leq \mathbb{E}\left[Yh(X)^2\right]^{1/2}\mathbb{E}[Y]^{1/2} < \infty$. When $h(X) = \mathbb{G}(Y|X)$ we have $\mathbb{E}\left[Y\mathbb{G}(Y|X)^2\right] \leq \mathbb{E}[Y] < \infty$, since $|\mathbb{G}(Y|X)| \leq 1$. Finally, by the Law of Iterated Expectations and Lemma 15:

$$\mathbb{E}\left[\tilde{L}^G(Y,X,h(X))\right] - \mathbb{E}\left[\tilde{L}^G(Y,X,\mathbb{G}(Y|X))\right] = \mathbb{E}\left[\mathbb{E}[Y|X]\,(h(X) - \mathbb{G}(Y|X))^2\right] \;\geq\; 0$$

with equality if and only if $h(X) = \mathbb{G}(Y|X)$ *a.s.*

□

G Simulation Checks

Some simulation checks are collected in Tables 6 and 7 below, assessing the possibility that a negative covariance term in the asymptotic distribution of the GIR estimator in the proof of Theorem 2 drives the standard deviation of the feasible GIR estimator below that of the oracle estimator. This possibility is assessed by computing the first-stage estimator with a random sample of the same DGP but independent of the sample used to compute the second-stage linear projection coefficient. The independent samples make the leading first-stage and second-stage influence-function components independent. Their covariance should therefore be zero in the asymptotic variance and fairly close to zero in finite samples[25]. For this exercise, we consider DGPs 1 and 3 from Section 6. Results show that our hypothesis is highly plausible, since the MC standard deviation of the feasible estimator with independent samples is higher than that of the oracle estimator.

Table 6: Monte Carlo Summary, Independent Samples, Group-Dummy Design (DGP1).

n	**500**			**1000**			**4000**		
Coefficient	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**
GIR (LS)									
θ_{01}	0.951	0.0003	0.063	0.948	0.0001	0.044	0.949	-0.0003	0.022
θ_{02}	0.946	-0.0016	0.062	0.947	0.0002	0.043	0.950	-0.0001	0.022
θ_{03}	0.949	-0.0020	0.061	0.951	0.0000	0.042	0.956	0.0001	0.021
Oracle									
θ_{01}	0.948	0.0000	0.042	0.954	-0.0002	0.029	0.949	0.0000	0.015
θ_{02}	0.951	-0.0016	0.040	0.949	-0.0001	0.028	0.953	-0.0001	0.014
θ_{03}	0.942	-0.0016	0.039	0.952	-0.0004	0.027	0.948	0.0000	0.014

Note: Warp-Speed simulation results of DGP1 for R = 10,000 Monte Carlo replications. Each replication uses two independent samples from DGP1, each of size n. The first sample estimates the CMDF, and the second evaluates that estimate and computes the GIR coefficient. Displayed coverage, bias, and Monte Carlo standard deviation results use the true Gini parameter values. "GIR (LS)" refers to the results obtained using the estimated CMDF via least-squares with $q_K(x) = x$ (correct specification), whereas "Oracle" refers to those obtained using the true CMDF on the same second-stage sample.

[25] We thank Carlos Velasco for suggesting this exercise to assess the negative covariance hypothesis.

Table 7: Monte Carlo Summary, Independent Samples, Discrete Response Design (DGP3).

n	**1000**			**2000**			**7000**		
Coefficient	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**	**Coverage**	**Bias**	**SD**
GIR (Sp.)									
θ_{00}	0.954	0.0003	0.051	0.954	-0.0002	0.035	0.949	-0.0003	0.019
θ_{01}	0.948	-0.0007	0.078	0.954	0.0002	0.054	0.955	0.0006	0.029
θ_{02}	0.947	-0.0002	0.044	0.948	0.0004	0.031	0.955	0.0000	0.017
GIR (Lg.)									
θ_{00}	0.954	0.0003	0.051	0.954	-0.0003	0.035	0.949	-0.0003	0.019
θ_{01}	0.945	-0.0007	0.077	0.953	0.0002	0.054	0.954	0.0006	0.029
θ_{02}	0.948	-0.0002	0.044	0.949	0.0004	0.031	0.956	-0.0001	0.017
Oracle									
θ_{00}	0.954	-0.0002	0.034	0.954	-0.0004	0.024	0.948	0.0000	0.013
θ_{01}	0.952	0.0000	0.052	0.952	0.0002	0.037	0.952	0.0002	0.020
θ_{02}	0.951	-0.0002	0.030	0.949	0.0004	0.021	0.947	-0.0002	0.011

Note: Warp-Speed simulation results of DGP3 for R = 10,000 Monte Carlo replications. Each replication uses two independent samples from DGP3, each of size n. The first sample estimates the CMDF, and the second evaluates that estimate and computes the GIR coefficient. Displayed coverage, bias, and Monte Carlo standard deviation results use the true Gini parameter values. “GIR (Sp.)" refers to the results obtained using the estimated CMDF via least-squares series with spline basis, whereas “GIR (Lg.)" refers to the results obtained using the estimated CMDF via least-squares series with Legendre basis. “Oracle" refers to those obtained using the true CMDF on the same second-stage sample. The first-stage estimators use $K = \lfloor n^{0.3} \rfloor$ for splines and $K = \lfloor n^{0.245} \rfloor$ for Legendre polynomials in each discrete cell.

H Empirical Application: Additional Tables

Table 8: Descriptive Statistics by Waves

	Wave 1		Wave 2		Combined	
Variable	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Wage	7.295	3.650	7.217	4.310	7.232	4.186
Union	0.309	0.462	0.211	0.408	0.230	0.421
S.M.S.A	0.670	0.470	0.791	0.407	0.766	0.423
Part-time	0.051	0.221	0.059	0.237	0.058	0.233
Non-White	0.108	0.311	0.122	0.327	0.119	0.324
High School	0.380	0.485	0.383	0.486	0.383	0.486
Some College	0.184	0.387	0.196	0.397	0.193	0.395
College	0.197	0.397	0.251	0.434	0.240	0.427
Experience	18.337	13.260	18.211	11.867	18.236	12.158

Table 9: Descriptive Statistics by Wave and Union Status

	Wave 1				Wave 2				Combined			
	Union		Non-Union		Union		Non-Union		Union		Non-Union	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Wage	7.924	2.782	7.014	3.946	8.093	3.213	6.983	4.529	8.047	3.104	6.988	4.430
Union	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
S.M.S.A	0.707	0.455	0.653	0.476	0.803	0.397	0.787	0.409	0.778	0.416	0.763	0.425
Part-time	0.033	0.178	0.060	0.237	0.029	0.169	0.068	0.251	0.030	0.171	0.066	0.248
Non-White	0.132	0.338	0.098	0.297	0.145	0.352	0.116	0.320	0.141	0.348	0.113	0.316
High School	0.488	0.500	0.331	0.471	0.483	0.500	0.356	0.479	0.485	0.500	0.352	0.478
Some College	0.170	0.375	0.190	0.393	0.190	0.392	0.197	0.398	0.184	0.388	0.196	0.397
College	0.085	0.278	0.247	0.431	0.167	0.373	0.273	0.446	0.145	0.352	0.269	0.443
Experience	21.204	12.974	17.053	13.186	22.345	11.401	17.107	11.744	22.038	11.855	17.098	12.016

Note: Wave 1 is 1979 (n = 21,483); Wave 2 is 1988 (n = 74,661). The combined sample contains 96,144 men. Wages are observed hourly wages in 1979 dollars (no minimum-wage censoring). Observations are weighted by the product of CPS sampling weights and hours worked. Means and SDs describe the weighted empirical distribution. Combined statistics pool the observations with their original weights. High School denotes 12 years of education, Some College 13–15 years, and College 16 or more years.

Table 10: GIR Estimates in Gini points (100×Gini) for Models 1–3.

Variable	**Model 1**	**Model 2**	**Model 3**
Union	-4.71**	-1.81**	0.72
	(−4.97, −4.29)	(−2.61, −0.74)	(−0.22, 2.38)
Wave	0.85**	0.90**	0.82**
	(0.28, 1.08)	(0.34, 1.14)	(0.26, 1.06)
S.M.S.A	2.09**	2.08**	2.10**
	(1.72, 2.56)	(1.70, 2.53)	(1.72, 2.55)
Part-time	7.03**	6.80**	6.77**
	(6.11, 8.01)	(5.93, 7.76)	(5.88, 7.70)
Non-white	-0.48	-0.44	-1.21
	(−1.24, 0.25)	(−1.19, 0.29)	(−2.60, 0.090)
High School	-1.15**	-4.30**	-4.29**
	(−1.63, −0.66)	(−5.24, −3.29)	(−5.20, −3.30)
Some College	-1.25**	-4.27**	-4.33**
	(−1.72, −0.56)	(−5.49, −3.15)	(−5.53, −3.24)
College	-3.61**	-5.47**	-5.68**
	(−4.12, −2.84)	(−6.81, −4.35)	(−6.95, −4.46)
Experience	0.32**	0.34**	0.34**
	(0.27, 0.38)	(0.18, 0.49)	(0.17, 0.49)
Experience^2	-0.0043**	-0.011**	-0.011**
	(−0.0056, −0.0032)	(−0.017, −0.0036)	(−0.018, −0.0035)
Experience^3		0.00011**	0.00011**
		(0.000015, 0.00020)	(0.000014, 0.00020)
Experience × High School		0.13**	0.14**
		(0.10, 0.17)	(0.10, 0.17)
Experience × Some College		0.13**	0.13**
		(0.095, 0.19)	(0.094, 0.18)
Experience × College		0.073**	0.072**
		(0.036, 0.13)	(0.033, 0.12)
Union × Experience		-0.13**	-0.13**
		(−0.17, −0.087)	(−0.17, −0.087)
Experience × Nonwhite			0.020
			(−0.052, 0.073)
Union × Nonwhite			1.47*
			(−0.049, 2.80)
Fixed Effects:			
Industry and Occupation	Yes	Yes	Yes
Union × Occupation	No	No	Yes
R^2_G	0.849	0.849	0.849

Note: In parentheses and below the coefficient estimates are the 95% bootstrap confidence intervals based on Algorithm 2 with $B = 100$. '**', and '*' denote significance at the 5% and 10% levels, respectively.

References

AABERGE, R., S. BJERVE, AND K. DOKSUM (2006): "Modeling Inequality and Spread in Multiple Regression," *Lecture Notes-Monograph Series*, 49, 120–130. [4, 8]

AHLIN, C., AND H. JEONG (2021): "A Conditional Gini: Measure, Estimation, and Application," *The Journal of Economic Inequality*, 19, 363–384. [5]

ANDREWS, D. W. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59(2), 307–345. [22]

ANGRIST, J., V. CHERNOZHUKOV, AND I. FERNÁNDEZ-VAL (2006): "Quantile Regression under Misspecification, with an Application to the US Wage Structure," *Econometrica*, 74(2), 539–563. [3, 13, 17, 18, 19]

AUTOR, D. H., L. F. KATZ, AND M. S. KEARNEY (2008): "Trends in U.S. Wage Inequality: Revising the Revisionists," *The Review of Economics and Statistics*, 90(2), 300–323. [42]

BELLONI, A., V. CHERNOZHUKOV, D. CHETVERIKOV, AND I. FERNÁNDEZ-VAL (2019): "Conditional Quantile Processes Based on Series or Many Regressors," *Journal of Econometrics*, 213(1), 4–29. [7]

BELLONI, A., V. CHERNOZHUKOV, D. CHETVERIKOV, AND K. KATO (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186(2), 345–366. [4, 6, 22, 25, 26, 27, 28, 56, 69, 73, 74, 79]

BERKES, E., AND R. GAETANI (2023): "Income Segregation and the Rise of the Knowledge Economy," *American Economic Journal: Applied Economics*, 15(2), 69–102. [6, 14]

BHATTACHARYA, D. (2007): "Inference on Inequality from Household Survey Data," *Journal of Econometrics*, 137(2), 674–707. [14]

BOWLUS, A. J., AND J.-M. ROBIN (2004): "Twenty Years of Rising Inequality in US Lifetime Labour Income Values," *The Review of Economic Studies*, 71(3), 709–742. [6, 14]

BRUNELLO, G., M. FORT, AND G. WEBER (2009): "Changes in Compulsory Schooling, Education and the Distribution of Wages in Europe," *The Economic Journal*, 119(536), 516–539. [2]

CARD, D. (1992): "The Effect of Unions on the Distribution of Wages: Redistribution or Relabelling?," Working Paper w4195, NBER. [41]

CARD, D., AND J. DINARDO (2002): "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles," *Journal of Labor Economics*, 20(4), 733–783. [43]

CATTANEO, M. D., Y. FENG, AND B. SHIGIDA (2026): "Uniform Estimation and Inference for Nonparametric Partitioning-Based M-Estimators," *The Annals of Statistics*, 54(4), 1740–1767. [6]

CHEN, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models," *Handbook of Econometrics*, 6, 5549–5632. [22, 27, 54, 57]

CHEN, X., AND T. M. CHRISTENSEN (2015): "Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions," *Journal of Econometrics*, 188(2), 447–465. [22]

CHENG, G., AND J. Z. HUANG (2010): "Bootstrap Consistency for General Semiparametric M-Estimation," *The Annals of Statistics*, 38(5), 2884–2915. [67]

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, J. MEIER, A. VAN VUUREN, AND F. VELLA (2024): "Conditional Rank-Rank Regression," available at arXiv:2407.06387. [6]

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND B. MELLY (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81(6), 2205–2268. [6, 38]

DINARDO, J., N. FORTIN, AND T. LEMIEUX (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64(5), 1001–1044. [4, 38, 41]

FARBER, H. S., D. HERBST, I. KUZIEMKO, AND S. NAIDU (2021): "Unions and Inequality over the Twentieth Century: New Evidence from Survey Data," *The Quarterly Journal of Economics*, 136(3), 1325–1385. [41]

FARCOMENI, A. (2026): "Quantile Share Ratio Regression for the Study of Economic Inequality," *Statistical Papers*, 67(3), 52. [5]

FEBRER, A., AND J. MORA (2009): "Flexible estimation of wage distributions in the presence of covariates," *Computational Statistics & Data Analysis*, 53(6), 2189–2200. [4]

FERREIRA, E., AND A. GARÍN (1997): "Una Nota sobre el Cálculo del Índice de Gini," *Estadística Española*, 39(142), 207–218. [47]

FIRPO, S., N. M. FORTIN, AND T. LEMIEUX (2009): "Unconditional Quantile Regressions," *Econometrica*, 77(3), 953–973. [7]

FREEMAN, R. B. (1980): "Unionism and the Dispersion of Wages," *Industrial and Labor Relations Review*, 34(1), 3–23. [44]

FREEMAN, R. B. (1991): "How Much has De-unionisation Contributed to the Rise in Male Earnings Inequality?," Working Paper w3826, NBER. [41]

FRIEDRICH, B. U. (2022): "Trade Shocks, Firm Hierarchies, and Wage Inequality," *Review of Economics and Statistics*, 104(4), 652–667. [2]

GASTWIRTH, J. L. (1971): "A General Definition of the Lorenz Curve," *Econometrica*, 39(6), 1037–1039. [7]

GIACOMINI, R., D. N. POLITIS, AND H. WHITE (2013): "A Warp-Speed Method for Conducting Monte Carlo Experiments Involving Bootstrap Estimators," *Econometric Theory*, 29(3), 567–589. [34]

GNEITING, T. (2011a): "Making and Evaluating Point Forecasts," *Journal of the American Statistical Association*, 106(494), 746–762. [30, 32, 48, 49, 50]

——— (2011b): "Quantiles as Optimal Point Forecasts," *International Journal of Forecasting*, 27(2), 197–207. [30]

GOTTSCHALK, P., AND T. M. SMEEDING (1997): "Cross-National Comparisons of Earnings and Income Inequality," *Journal of Economic Literature*, 35(2), 633–687. [6, 14]

HEIKKURI, V.-M., AND M. SCHIEF (2026): "Subgroup Decomposition of the Gini Coefficient: A New Solution to An Old Problem," *Econometrica*, 94(1), 169–192. [7]

HURTADO, C. (2025): "Sources of Wage Inequality: Decomposing the Conditional Gini Coefficient," *Applied Economics*, pp. 1–20. [4]

JOHNSON, G. E. (1997): "Changes in Earnings Inequality: The Role of Demand Shifts," *Journal of Economic Perspectives*, 11(2), 41–54. [43]

JOKIEL-ROKITA, A., S. PIĄTEK, AND R. TOPOLNICKI (2024): "Estimation of Conditional Inequality Curves and Measures via Estimating the Conditional Quantile Function," available at arXiv:2412.20228. [4]

JUHN, C., K. M. MURPHY, AND B. PIERCE (1993): "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy*, 101(3), 410–442. [42]

KLEIN, N., T. KNEIB, S. LANG, AND A. SOHN (2015): "Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany," *The Annals of Applied Statistics*, 9(2), 1024–1052. [4]

KOENKER, R., AND J. A. MACHADO (1999): "Goodness of Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, 94(448), 1296–1310. [32]

KOPCZUK, W., E. SAEZ, AND J. SONG (2010): "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data Since 1937," *The Quarterly Journal of Economics*, 125(1), 91–128. [6, 14]

KRUEGER, D., AND F. PERRI (2006): "Does Income Inequality Lead to Consumption Inequality? Evidence and Theory," *The Review of Economic Studies*, 73(1), 163–193. [14]

LEMIEUX, T. (2006): "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?," *American Economic Review*, 96(3), 461–498. [42]

LEMIEUX, T. (2011): "Wage Inequality: A Comparative Perspective," *Australian Bulletin of Labour*, 37(1), 2–32. [42]

LERMAN, R. I., AND S. YITZHAKI (1984): "A Note on the Calculation and Interpretation of the Gini Index," *Economics Letters*, 15(3-4), 363–368. [47]

——— (1989): "Improving the Accuracy of Estimates of Gini Coefficients," *Journal of Econometrics*, 42(1), 43–47. [47]

LU, Y., AND L. YU (2015): "Trade Liberalization and Markup Dispersion: Evidence from China's WTO Accession," *American Economic Journal: Applied Economics*, 7(4), 221–253. [2, 6]

NEWEY, W. K. (1994a): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382. [4, 59, 60, 63, 66]

——— (1994b): "Series Estimation of Regression Functionals," *Econometric Theory*, 10(1), 1–28. [22, 26]

——— (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79(1), 147–168. [22, 26]

PARZEN, E. (2004): "Quantile Probability and Statistical Data Modeling," *Statistical Science*, 19(4), 652–662. [8]

SAVAGE, L. J. (1971): "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, 66(336), 783–801. [30]

VAN DER VAART, A., AND J. WELLNER (2023): *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 2nd edn. [28, 71, 72, 75, 86]

VILA, R., AND H. SAULO (2026): "On the Bias of the Gini Estimator: Poisson and Geometric Cases, a Characterization of the Gamma Family, and Unbiasedness Under Gamma Distributions," *Statistical Methods & Applications*, pp. 1–18. [48]

YITZHAKI, S., AND E. SCHECHTMAN (2013): *The Gini Methodology: A Primer on a Statistical Methodology*, vol. 272. Springer. [3, 47]