Title

Instrumental variable analysis with censored data in the presence of many weak instruments: Application to the effect of being sentenced to prison on time to employment

Journal

The Annals of Applied Statistics, 12(4)

ISSN

1932-6157

Authors

Ertefaie, Ashkan
Nguyen, Anh
Harding, David J
et al.

Publication Date

2018

DOI

10.1214/18-aoas1174

Peer reviewed

Instrumental Variable Analysis with Censored Data in the Presence of Many Weak Instruments: Application to the Effect of Being Sentenced to Prison on Time to Employment

Ashkan Ertefaie [1], Anh Nguyen[†2], David Harding[†3], Jeffrey Morenoff [*2], Wei Yang[4]
[1]Department of Biostatistics and Computational Biology, University of Rochester;
[2]Department of Sociology, University of Michigan
[3]Department of Sociology, University of California at Berkeley
[4]Department of Biostatistics Epidemiology, University of Pennsylvania

March 2, 2018

Abstract

This article discusses an instrumental variable approach for analyzing censored data that includes many instruments that are weakly associated with the endogenous variable. We study the effect of imprisonment on time to employment using an administrative data on all individuals sentenced for felony in Michigan in the years 2003-2006. Despite the large body of research on the effect of prison on employment, this is still a controversial topic, especially since some of the studies could have been affected by unmeasured confounding. We take advantage of a natural experiment based on the random assignment of judges to felony cases and construct a vector of instruments based on judges' ID that can avoid the confounding bias. However, some of the constructed instruments are weakly associated with the sentence type, i.e., the endogenous variable, which can potentially lead to misleading results. Using a dimension reduction technique, we propose a novel semi-parametric estimation procedure in a survival context that is robust to the presence of many weak instruments. Specifically, we construct a test statistic based on the structural failure time model and provide inference by inverting the testing procedure. Under some assumptions, the optimal choice of the test statistic has also been derived. Analyses show a significant negative impact of imprisonment on time to employment which is consistent with some of the previous results. Our simulation studies highlight the importance

[*]*National Science Foundation (SES1061018), with additional support from center grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development to the Population Studies Centers at the University of Michigan (R24 HD041028) and at UC Berkeley (R24 HD073964)*

of accounting for weak instruments in the analyses in terms of both bias and inflated type-I error rates.

Keywords: Instrumental variables, Survival time, Test statistics, Two-stage least squares, Unmeasured confounders.

1 Introduction

We consider instrumental variable (IV) analyses with censored data in settings that include many weak instruments– weakly associated with the treatment variable. The key advantage of IV method is that it allows relaxation of the ”no unmeasured confounders” assumption under some conditions (Wright, 1934; Haavelmo, 1943; Angrist et al., 1996; Abadie, 2003; Inoue & Solon, 2010; Ertefaie et al., 2017a). However, IV analyses may lead to misleading results in the presence of weak instruments. In fact, Bound et al. (1995) showed that in such settings, IV estimates may approach to ordinary least squares estimates which we know are subject to bias because of unmeasured confounding (Staiger & Stock, 1994; Imbens & Rosenbaum, 2005; Small & Rosenbaum, 2008).

In 2013, the incarceration rate in the US was the highest in the world (Mauer, 2001; Austin & Irwin, 2012; Currie, 2013; Travis et al., 2014). The rise of mass incarceration over the last four decades has prompted intense interest among social scientists in the consequences of incarceration for the individuals and families who experience it (Kling, 2006; Alexander, 2012; Turney & Wildeman, 2015; Kilgore, 2015). A large body of research suggests that serving time in prison affects one’s employment (Pager, 2008), relegates workers to the secondary labor market (Western, 2002; Weiman et al., 2007), and affects attachment to the labor market (Apel & Sweeten, 2010). However, there are several studies that find much smaller or nonexistent prison effects (Kling, 2006; Loeffler, 2013). Thus, the evidence on the effect of prison on employment remains inconclusive, especially since some of these studies could have been affected by unmeasured confounding. For example, a judge’s assessment of how likely an offender is to reoffend – and thus the sentencing decision – can be influenced by information available to her that does not get recorded in administrative data (e.g., statements from witnesses) and may also be related to the time to employment, resulting in omitted variable bias/unobserved confounding (Nagin et al., 2009).

Our data includes all individuals sentenced for a felony in Michigan in the years 2003-2006 – over 100,000 individuals assigned to 151 judges. One important feature of this data that encourages IV analysis is the fact that judges are randomly assigned in Michigan within counties. Specifically, criminal cases are assigned to judges by the court clerk when cases are initially filed (at indictment).

Therefore initial charges are filed before the prosecutor knows which judge will be assigned. In this paper, we take advantage of a natural experiment based on the random assignment of judges to felony cases and construct a vector of IVs based on judges' ID. We are interested in the effect of imprisonment on time to employment with quarterly earnings above poverty where the unit of time is in calendar quarter format and the poverty line is defined for a single person of working age under 65. Moreover, the earnings come from jobs recorded by the unemployment insurance system.

The first step in IV analyses is to assess the association between the IV and the sentence type, i.e., prison and non-prison sentences. We fit a random effect model that includes judges' ID as a random intercept, baseline covariates including the counties indicator as fixed effects and the sentence type as a dependent variable (Chamberlain & Imbens, 2004). Our results show that the random intercept component is significant which provides evidence that our instrument judges' ID is associated with the sentence type. However, the estimated random intercepts are relatively small for many judges (Figure 1). Moreover, after controlling for measured covariates, the F-statistic is 27 which is relatively small given the large sample size and the dimension of the vector of instruments (Stock & Yogo, 2005). In fact, Stock and Yogo showed that, with 150 IVs, F-statistic of 84 corresponds to more than 20% error rate for a 0.05 level test where the test level is defined as the maximal size of the Wald test of the estimated treatment effect. These observations suggest the possibility of weak instruments which can lead to an invalid inference by inflating the type-I error rate (Stock et al., 2012) and providing a biased estimate (Bound et al., 1995).

In econometrics, there is a vast literature on estimating the treatment effect in the presence of many weak instruments. Staiger & Stock (1997a) developed asymptotic theory for IV analyses when the number of instruments is assumed to be fixed and the coefficients of the instruments in the treatment model are specified to be in an $n^{-1/2}$ neighborhood of zero. Their results show that when IVs are weak, the two-stage least squares (2SLS) and the limited information maximum likelihood (LIML) estimates are not consistent and the standard errors are underestimated while the bias is less of a problem for LIML than 2SLS (Magdalinos, 1990; Choi & Phillips, 1992; Buse, 1992; Magdalinos, 1994; Bekker, 1994; Bound et al., 1995; Anderson et al., 2010). Another direction to

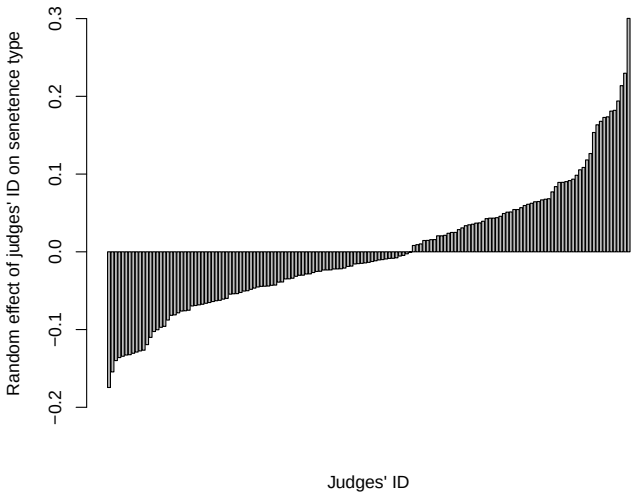

Figure 1: Michigan Sentencing Data. Estimated random intercepts for judges' ID in Michigan in the years 2003-2006.

study the asymptotic behavior of IV estimates has taken by Chao & Swanson (2005) that accounts for both many and weak instruments by allowing the number of instruments to go to infinity as a function of sample size and shrinking the coefficients of the instruments in the treatment model toward zero as the sample size grows (Hahn & Inoue, 2002; Chamberlain & Imbens, 2004; Newey & Windmeijer, 2009; Belloni et al., 2012; Chao et al., 2014; Kaffo & Wang, 2017).

IV methods have been extended to analyses of survival data subject to censoring. Bosco et al. (2010) generalized the 2SLS method to account for censoring by fitting a logistic regression that includes the treatment as dependent variable and the provider-preference based IV as independent variable in the first-stage and included the predicted values obtained by the first-stage in the Cox proportional hazard regression in the second-stage (MacKenzie et al., 2014). In the context of additive hazard models, Li et al. (2015a) developed a closed-form, two-stage treatment effect estimator that relies on assuming linear structural equation models for the hazard function (Tchetgen et al., 2015; Chan, 2016). An alternative two-stage residual inclusion (2SRI) that includes the residual of

the first-stage model in the second stage is proposed by Terza et al. (2008) that can be used in non-linear regression models, e.g., Weibull models. The 2SRI does not account for censoring. However, when instruments are weak none of the aforementioned methods that fit a logistic or a least squares model at the first-stage of the analysis can be applied to settings with many weak IVs (Bound et al., 1995).

In this paper, we propose a pivotal method that adjusts for confounding using IVs to estimate the treatment effect in survival contexts. Specifically, we use a dimension reduction approach to reduce the dimension of the vector of instruments to the dimension of treatment variables and show that our method is robust to the presence of many weak IVs. Similar to Staiger & Stock (1997a), our asymptotic framework assumes a fix number of instruments and lets the coefficients of the instruments in the treatment model go to zero as the sample size goes to infinity (Kleibergen, 2007). This framework is appropriate for settings where the sample size is quite large relative to the number of instruments which is indeed the case in our application (Stock & Yogo, 2005; Hausman & Newey, 2004; Hansen et al., 2008). Our work builds on Kleibergen (2007). In the context of standard linear regression, Kleibergen proposed test statistic that is specifically designed to cases with many weak IVs. He showed that his test statistic is more powerful than the AR statistic proposed by Anderson & Rubin (1949). The IV analysis developed in this article generalizes Kleibergen (2007) to settings with censored data.

2 Framework and Model

2.1 Notation

Suppose that the data is composed of n i.i.d triplet $(T, D, \tilde{Z}^*)$ where D denotes the sentence type and $D \in \{0, 1\}$ for non-prison and prison sentences, respectively. Let $\tilde{Z}^*$ be a single $(L+1)$-valued vector of judge's IDs that is used to form a L mutually independent orthogonal binary instruments Z^*. Also, let T be the time to employment with quarterly earnings above poverty. In the presence of censoring, we only observe $Y = \min(T, C)$ where C is the censoring time. We consider a particular

type of censoring where censored subjects are those who stayed unemployed or are employed with salary below poverty at the planned end of study, i.e., administrative censoring. Thus, the random variable C records the difference between the end of follow-up date and the subject's sentence date. The administrative censoring time C may vary across subjects, but for each subject the value of C is known at the start of the follow-up. For example, all offenders in our study whose sentence date is on July 30, 2005 had the same potential follow-up time c_0. We assume that C is independent of potential outcomes and covariates. Let $\Delta = I(T < C)$ be the censoring indicator and $\mathbf{X}$ be a vector of baseline measured covariates.

We use potential outcome framework to present our causal model and the required assumptions (Neyman, 1923; Rubin, 1978). Let D^{z^*} denote the potential sentence type when assigned to a particular judge, i.e., $Z^* = z^*$, and $T^{z^*,d}$ denote the counterfactual time to employment if $Z^* = z^*$ and $D = d$. Also, define $Y^{z^*,d} = \min(T^{z^*,d}, C)$ as the counterfactual length of follow-up time and $\Delta^{z^*,d} = I(T^{z^*,d} < C)$ as the counterfactual censoring indicator.

2.2 Assumptions

Our proposed method requires that the following assumptions hold for every L:

Assumption 1. Stable Unit Treatment Value Assumption (Rubin, 1978): Let $\underline{Z}^*$ and $\underline{D}$ denote the $n \times L$ matrix of instruments and n-dimensional vector of sentence type, respectively.

a. If $z_i^* = z_i'^*$, then $D_i^{\underline{\mathbf{z}}^*} = D_i^{\underline{\mathbf{z}}'^*}$ for all subjects.

b. If $z_i^* = z_i'^*$ and $d_i = d_i'$, then $T_i^{\underline{\mathbf{z}}^*,\underline{\mathbf{d}}} = T_i^{\underline{\mathbf{z}}'^*,\underline{\mathbf{d}}'}$ for all subjects.

The SUTVA implies that the sentence type status of any individual does not affect the sentence type and the time to employment of other subjects. Under this assumption we can write $(T_i^{\underline{\mathbf{z}}^*,\underline{\mathbf{d}}}, \Delta_i^{\underline{\mathbf{z}}^*,\underline{\mathbf{d}}}, Y_i^{\underline{\mathbf{z}}^*,\underline{\mathbf{d}}}, D_i^{\underline{\mathbf{z}}^*})$ as $(T_i^{z_i^*,d_i}, \Delta_i^{z_i^*,d_i}, Y_i^{z_i^*,d_i}, D_i^{z_i^*})$, respectively for subject i.

Assumption 2. Consistency: Individuals' observed time to employment T_i is the counterfactual time to employment under the sentence type D and instrument Z^*, i.e., $T_i = T_i^{Z^*,D}$.

Assumption 3. Z^* is associated with D conditional on the vector of measured covariate $\mathbf{X}$.

Assumption 4. Z^* is uncorrelated with unmeasured confounders conditional on $\mathbf{X}$. More specifically, $Z^* \perp\!\!\!\perp (T^{z^*,d}, D^{z^*})|\mathbf{X}$.

Assumption 5. Z^* affects the time to employment only through the sentence type, i.e., $T_i^{z^*,d} = T_i^{z'^*,d}$ for all z^*, z'^*, d and all individuals. So we can write $T_i^{z^*,d} = T_i^d$.

Assumptions 1 & 2 link the the potential outcome and the observed data. We believe that these two assumptions are plausible. However, a case can be made for violations. For example, a reason for violation of SUTVA is that felons who are sentenced to prison may dissuade their friends to commit crimes. We refer to an instrument as a valid IV when assumptions 3, 4, and 5 are satisfied. Using the random effect model discussed in the introduction, we have verified that assumption 3 holds despite the possibility of weak association. We discuss the plausibility of assumptions 4 & 5 in detail in Section 5.2.

2.3 Preliminaries

We first introduce the Anderson-Rubin (AR) statistic (Anderson & Rubin, 1949) and the KJ-statistics (Kleibergen, 2007) in standard linear regression settings. Assuming that all the time to employment were observed and followed an accelerated failure time (AFT) model, one could use the existing AR or KJ- statistic to overcome the challenges arising in the presence of many weak IVs. Under these assumptions, $\log(T) = \beta_0 D + \epsilon$ where ϵ has a mean zero normal distribution with variance σ^2. It is assumed that $\text{cov}(\epsilon, \epsilon_d) \neq 0$ where $\epsilon_d = D - (\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')D$ (Kleibergen, 2007). For simplicity of notation, baseline covariates $\mathbf{X}$ are excluded from both models. Then,

$$AR(\beta_0) = (\log(T) - \beta_0 D)\mathbf{Z}[\mathbf{Z}'\sigma^2\mathbf{Z}]^{-1}\mathbf{Z}'(\log(T) - \beta_0 D)',$$

and

$$K(\beta_0) = (\log(T) - \beta_0 D)\mathbf{Z}\Pi(\beta_0)[\Pi(\beta_0)'\mathbf{Z}'\sigma^2\mathbf{Z}\Pi(\beta_0)]^{-1}\Pi(\beta_0)'\mathbf{Z}'(\log(T) - \beta_0 D)',$$
$$J(\beta_0) = (\log(T) - \beta_0 D)\mathbf{Z}\zeta(\beta_0)_\perp[\zeta(\beta_0)'_\perp\mathbf{Z}'\sigma^2\mathbf{Z}\zeta(\beta_0)_\perp]^{-1}\zeta(\beta_0)'_\perp\mathbf{Z}'(\log(T) - \beta_0 D)',$$

where $\Pi(\beta_0) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'[D - (\log(T) - \beta_0 D)\frac{\text{cov}(\epsilon,\epsilon_d)}{\sigma^2}]$, and $\zeta(\beta_0)_\perp = (\mathbf{Z}'\mathbf{Z})^{-1}\Pi(\beta_0)_\perp$ is the orthonormal complement of $\zeta(\beta_0)$, i.e., $\zeta(\beta_0)'_\perp\zeta(\beta_0)_\perp = I_{k-1}$ and $\zeta(\beta_0)_\perp\zeta(\beta_0) = 0$. Under the null hypothesis $H_0 : \beta = \beta_0$, Anderson & Rubin (1949) showed that the AR converges in distribution to $\chi^2(L)$, and Kleibergen (2007) showed that K- and J- statistics converge in distribution to $\chi^2(1)$ and $\chi^2(L-1)$, respectively. Our contribution is to generalize these statistics to settings with censored data.

2.4 Failure time model

Following Robins & Tsiatis (1991), consider the following structural failure time model

$$T^0 = T^d \exp[\beta_0 d], \tag{1}$$

where T^d is the counterfactual time to employment if $D = d$. Thus, $\exp[\beta_0 d]$ can be interpreted as the factor by which sentence type d shortens or accelerates the time to employment (Joffe, 2001). This model assumes rank preservation of the subjects's time to employment meaning that if it takes longer time for subject i to find a job compared with subject j when both are sentenced to non-prison sentences, it would also take longer time to find a job for subject i than subject j if both sentenced to prison (Mark & Robins, 1993; Hernán et al., 2005; Solomon et al., 2014).

Under assumptions 1& 2 the structural model (1) can be linked to the observed data as

$$T(\beta_0) \equiv T^0 = T \exp[\beta_0 D].$$

We use the notation $T(\beta_0)$ to acknowledge the dependence on β_0. The counterfactual time to

employment T^0 is independent of the vector of instruments given that assumptions 4 and 5 hold. However, we do not observe the time to employment T for all of the subjects because some cannot find a job at the end of the follow-up. Intuitively, one may define $Y^0 = \min(T^0, C^0)$ where $C^0 = C\exp[\beta_0 D]$ and follow a standard G-estimation method (Robins, 1993, 1997). Unfortunately, this is not a proper way of handling censored data because even when β_0 is the true causal effect, Y^0 is no longer independent of D since C^0 is a function of the sentence type (Hernán et al., 2005).

Artificial censoring is a method to handle censored data in causal inference. The idea is to restrict the analysis to subjects whose employment time would have been observed regardless of their sentence type. Let $\mathcal{D}$ be a bounded set that is the support of D. Define the minimum potential censoring time as $C^\dagger(\beta) = C\min\{\exp(\beta d); d \in \mathcal{D}\}$ which reduces to $C\min\{1, \exp(\beta)\}$ when D is binary. Unlike C^0, $C^\dagger(\beta)$ is not a function of the sentence type. Define a new censoring indicator $\Delta^\dagger(\beta) = I(U(\beta) < C^\dagger(\beta))$ where $U(\beta) = \min(Y(\beta), C^\dagger(\beta))$ and $Y(\beta) = Y\exp[\beta D]$. Notice when $\beta = 0$, both censoring indicators are identical and when $\beta \neq 0$, $\Delta = 0$ implies $\Delta^\dagger(\beta) = 0$ but not the other way around. Thus, we are artificially censoring some subjects whose time to employment is observed to preserve the exchangeability result

$$\{U(\beta), \Delta^\dagger(\beta)\} \perp \mathbf{Z}^*|\mathbf{X}. \tag{2}$$

The independence result holds because $(U(\beta), \Delta^\dagger(\beta))$ are functions of $(U(\beta), C^\dagger(\beta))$ that is independent of $\mathbf{Z}^*$ under assumptions 4 & 5. For the simplicity of notation, assuming an additive model for the association between $\mathbf{Z}^*$ and $\mathbf{X}$, we rewrite (2) as

$$\{U(\beta), \Delta^\dagger(\beta)\} \perp \mathbf{Z}. \tag{3}$$

where $\mathbf{Z} = \mathbf{Z}^* - P_{\mathbf{X}}\mathbf{Z}^*$ with $P_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Thus, assuming that $P_{\mathbf{X}}$ is well-defined, $\mathbf{Z}$ is the orthogonal projection of $\mathbf{Z}^*$ onto the space spanned by covariates $\mathbf{X}$. In the following, we refer to $\mathbf{Z}$ as our vector of instruments and assume that all the variables are centered including the artificial censoring variable $\Delta^\dagger(\beta)$.

3 Estimation in the presence of many weak IVs

3.1 Testing procedures

The independence result (3) plays a central role in our proposed inferential procedure. Based on the AR statistic (Anderson & Rubin, 1949), a test statistic that explores the correlation between $\Delta^\dagger(\beta_0)$ and $\mathbf{Z}$ can be formed as

$$AR(\beta_0) = \Delta^\dagger(\beta_0)\mathbf{Z}[\mathbf{Z}' s_{\Delta\Delta}\mathbf{Z}]^{-1}\mathbf{Z}'\Delta^\dagger(\beta_0)', \tag{4}$$

where $M_{\mathbf{Z}} = I - P_{\mathbf{Z}}$ and $s_{\Delta\Delta} = \Delta^\dagger(\beta_0)' M_{\mathbf{Z}} \Delta^\dagger(\beta_0)/(N-L)$. Under H_0, the AR statistic converges to $\chi^2(L)$. In contrast with the t-test based on the 2SLS IV analyses, the Anderson-Rubin test has correct size regardless of the strength of the IVs. This is mainly because the denominator of the 2SLS estimators is a function of $\mathrm{cov}(\mathbf{Z}, D)$, and thus, weak IVs result in unstable estimators. This is not the case in (4) and the strength of IVs can only affect the power of AR test statistic (Staiger & Stock, 1997b; Hahn et al., 2004; Jiang et al., 2014).

A deficiency of the AR statistic is that it has low power because the degrees of freedom parameter of its limiting distribution is equal to the number of instruments L. We propose the following two chi-square test statistics to overcome this deficiency,

$$\begin{aligned} K(\beta_0) &= \Delta^\dagger(\beta_0)\mathbf{Z}\Pi(\beta_0)[\Pi(\beta_0)'\mathbf{Z}' s_{\Delta\Delta}\mathbf{Z}\Pi(\beta_0)]^{-1}\Pi(\beta_0)'\mathbf{Z}'\Delta^\dagger(\beta_0)', \\ J(\beta_0) &= \Delta^\dagger(\beta_0)\mathbf{Z}\zeta(\beta_0)_\perp[\zeta(\beta_0)'_\perp \mathbf{Z}' s_{\Delta\Delta}\mathbf{Z}\zeta(\beta_0)_\perp]^{-1}\zeta(\beta_0)'_\perp \mathbf{Z}'\Delta^\dagger(\beta_0)', \end{aligned} \tag{5}$$

where $\Pi(\beta_0) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'[D - \Delta^\dagger(\beta_0)\frac{s_{\Delta d}}{s_{\Delta\Delta}}]$, and $\zeta(\beta_0)_\perp = (\mathbf{Z}'\mathbf{Z})^{-1}\Pi(\beta_0)_\perp$ is the orthonormal complement of $\zeta(\beta_0)$, i.e., $\zeta(\beta_0)'_\perp\zeta(\beta_0)_\perp = I_{k-1}$ and $\zeta(\beta_0)_\perp\zeta(\beta_0) = 0$. Also, $s_{\Delta d} = \Delta^\dagger(\beta_0)' M_Z D/(N-L)$. Assuming that instruments are valid, under H_0, the K- and J- statistics converge to $\chi^2(1)$ and $\chi^2(L-1)$, respectively. An interesting feature of these two statistics is that they are independent because the projection operators $\mathbf{Z}\Pi(\beta_0)[\Pi(\beta_0)'\mathbf{Z}' s_{\Delta\Delta}\mathbf{Z}\Pi(\beta_0)]^{-1}\Pi(\beta_0)'\mathbf{Z}'$ and $\mathbf{Z}\zeta(\beta_0)_\perp[\zeta(\beta_0)'_\perp\mathbf{Z}' s_{\Delta\Delta}\mathbf{Z}\zeta(\beta_0)_\perp]^{-1}\zeta(\beta_0)'_\perp\mathbf{Z}'$ are orthogonal. This implies that the K- and J-

statistics add-up to the AR statistic. The K- and J- statistics are the generalizations of the test statistics in Kleibergen (2007) to survival outcome settings. Proposition 1 states the limiting behavior of the K- and J- statistics for different limiting sequences of the association parameter between the treatment and instruments.

Proposition 1 *Let γ be the association parameter between the treatment indicator and instruments. Assuming that Assumptions 4 & 5 hold, under the null hypothesis $H_0 : \beta = \beta_0$, and when*

a. the instruments are strong such that $\gamma = O(1)$,

b. the instruments are weak such that $\gamma = n^{-1/2}H$ where H is a vector of constants,

the K- and J- statistics converge in distribution to $\chi^2(1)$ and $\chi^2(L-1)$, respectively.

The proof is presented in Section S3 of the supplementary material. Proposition 1 shows that the limiting distribution of the K- and J- statistics are robust to the specification of the association parameter between the treatment and instruments.

Assuming that the IV is valid, these statistics are testing whether, for a given β, the artificial censoring indicator $\Delta^{\dagger}(\beta)$ is uncorrelated with the vector of instrument $\mathbf{Z}$. Specifically, $K(\beta_0)$ uses a dimension reduction tool and instead of directly looking at the correlation between $\Delta^{\dagger}(\beta)$ and vector of instruments, it explores the correlation between $\Delta^{\dagger}(\beta)$ and a projection of a function of the sentence type variable onto the space spanned by instruments. By construction when there is no unmeasured confounding, $\Delta^{\dagger}(\beta)$ is independent of the sentence type. However, in the presence of unmeasured confounding which is the main subject of this paper, this independence does not hold anymore. In fact, the artificial censoring indicator may be a function of unmeasured confounders. For example, it is plausible that individuals with more serious crimes are more likely to have $\Delta^{\dagger}(\beta) = 0$ because it may be more difficult for them to find a job and if such characteristics is not captured by measured covariates, then even under the null hypothesis, $\Delta^{\dagger}(\beta)$ will depend on the sentence type through the unmeasured confounders.

The key idea in $K(\beta)$ is to asymptotically retrieve the independence condition under H_0 by projecting a new variable $D - \Delta^{\dagger}(\beta_0)\frac{s_{\Delta d}}{s_{\Delta\Delta}}$ onto the space spanned by instruments where $D -$

$\Delta^{\dagger}(\beta_0)\frac{s_{\Delta d}}{s_{\Delta\Delta}}$ can be viewed as the residual of the least squares regression of D on $\Delta^{\dagger}(\beta_0)$. Thus, the K-statistic tests if under the null hypothesis, there is any path from instruments that goes through the sentence type to the artificial censoring indicator. The J-statistic serves a different purpose. The test statistic $J(\beta)$ is constructed using the orthogonal complement $\Pi(\beta_0)_{\perp}$ and tests given that the null hypothesis holds, if there are any other paths that goes from instruments to the artificial censoring indicator. For example, when there is a direct path from $\mathbf{Z}$ to the time to employment, i.e., Assumption 5 is violated, $J(\beta_0)$ may reject the null hypothesis $H_0 : \beta = \beta_0$ even when H_0 is true (Kleibergen, 2007, Sec. 3).

The proposed K- and J- statistics have more power than AR because their limiting chi-square distributions have degrees of freedom of less than L. Particularly, the K-statistic is using only one degrees of freedom which is due to the dimension reduction process that replaces $\mathbf{Z}$ with $\mathbf{Z}\Pi(\beta_0)$ that is a $n\times 1$ vector. One may think that estimation based on the K-statistic alone would outperform other estimation procedures based on any combinations of the K- and J- statistics due to the small degrees of freedom. However, Kleibergen (2007) showed that although the K-statistic is a powerful test, in some cases, the power drops significantly around the inflexion points and local maxima of the statistic. He overcame this deficiency by using the J-statistic as a pretest for the K-statistic. Thus, a testing procedure of size α, first tests the null hypothesis using the J-statistic with level α_J and if rejected, performs the test using the K-statistic with level α_K such that $(1-\alpha) = (1-\alpha_J)(1-\alpha_K)$, i.e., $\alpha \approx \alpha_J + \alpha_K$. From now on, we refer to this testing procedure as $\text{KJ}_{a,b}$ statistic where $a = \alpha_J \times 100$ and $b = \alpha_K \times 100$. Note that the $\text{KJ}_{0,5}$ is equivalent to the K-statistic.

We generalized the dimension reduction technique proposed by Kleibergen (2007) to settings with censored data. Another potential method to reduce the dimension of the vector of instruments is to utilize the idea in the provider preference based IVs and construct a "harshness" score for each judge which leads to a one-dimensional instrument (Brookhart & Schneeweiss, 2007; Brookhart et al., 2006). However, the harshness score is not known and must be estimated using the available data by, for example, estimating the proportion of felons sentenced to prison by each judge. To provide valid inferences, this method requires data splitting so that the harshness is estimated using

the first portion of the data and the prison effect is estimated using the second portion of the data (Bound et al., 1995; Ertefaie et al., 2017b). One drawback of the sample splitting is the reduction in the sample size and thus, reducing the statistical power. Our dimension reduction method does not suffer from this drawback and utilizes the entire data. Hernán & Robins (2006) discussed the other issues related to the provider preference based IVs in detail (Li et al., 2015b).

Confidence intervals for β can be constructed by inverting the testing procedures. Accordingly, when the AR or KJ statistic with $\alpha_J = 0$ is used, point estimates $\hat{\beta}$ are obtained as a parameter value β for which the test results in the highest p-value. The point estimate of the KJ statistic with $\alpha_J > 0$ is the one with highest K-statistic p-value among those that are not rejected by the J-statistic. However, when the treatment effect is heterogeneous, AR and KJ statistic with $\alpha_J > 0$ may result in an empty confidence intervals that is discussed in Section S1 of the Supplementary Material (Kadane & Anderson, 1977; Small, 2007; Davidson & MacKinnon, 2014).

3.2 More powerful testing procedures

The testing procedures proposed in the previous section are functions of the observed time to employments only through the artificial censoring indicator. Intuitively, the power of the tests can be potentially improved by incorporating more information of the observed time to employments than just a binary artificial censoring indicator. However, it may require imposing some parametric models on the time to employments.

In the following proposition, we consider a more general form of the test statistics $K(\beta_0)$, $J(\beta_0)$ and $AR(\beta_0)$ where the censoring indicator is replace by a function $g(\Delta^\dagger, \beta)$. We then derive the optimal choice of this function.

Proposition 2 *Suppose T has a density function $f_T(t)$. Then the density function of the treatment-free survival time T_0 is $f_{T_0}(t) = \exp(\beta D) f_T(t \exp\{\beta D\})$. Assuming that the score function corresponding to the survival likelihood can be decomposed as $S_\beta(\xi) = D g^{opt}(\Delta^\dagger(\beta), U(\beta), \xi)$ where ξ is the set of nuisance parameters, then the function $g^{opt}(\Delta^\dagger(\beta), U(\beta), \hat{\xi})$ is the optimal choice of $g(\Delta^\dagger, \beta)$ where $\hat{\xi}$ is the estimated nuisance parameters.*

Proof Because the map from T to $T^0 = T\exp[\beta D]$ is one to one with strictly positive Jacobian determinant $\partial T^0/\partial\beta = \exp(\beta D)$, we have $f_T(t) = \exp(\beta D) f_{T^0}(t\exp\{\beta D\})$. Define a score function $S_\beta(\xi) = \partial\mathcal{L}/\partial\beta = Dg(\Delta^\dagger(\beta), U(\beta), \xi)$ where $\mathcal{L}$ is the corresponding log-likelihood. Then following Tsiatis (2007), the efficient score function can be constructed as $S_\beta^{eff}(\xi) = S_\beta(\xi) - \mathbb{E}[S_\beta(\xi)|\mathbf{X}] = \Big(D - \mathbb{E}[D|\mathbf{X}]\Big)g(\Delta^\dagger(\beta), U(\beta), \xi)$. The unknown vector of nuisance parameters ξ can be replaced by a consistent estimator $\hat{\xi}$. Now, because we are interested in the null hypothesis of no association between D and T_0, assuming that IVs are valid, i.e., the IVs are associated with the outcome only through their association with the treatment, we can replace D with a function of instruments $q(\mathbf{Z})$ and define $S_\beta^{eff^\dagger}(\hat{\xi}) = \Big(q(\mathbf{Z}) - \mathbb{E}[q(\mathbf{Z})|\mathbf{X}]\Big)g(\Delta^\dagger(\beta), U(\beta), \hat{\xi})$. However, by definition, our instruments $\mathbf{Z} = \mathbf{Z}^* - P_{\mathbf{X}}\mathbf{Z}^*$, where $P_X = \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$, are orthogonal to the space spanned by $\mathbf{X}$. Thus $\mathbb{E}[q(\mathbf{Z})|\mathbf{X}] = 0$. In AR statistic, $q(\mathbf{Z}) = \mathbf{Z}$ and in the K- and J- statistics, $q(\mathbf{Z}) = \mathbf{Z}\Pi(\beta)$. Thus, for example, the optimal version of the AR statistic is given by

$$AR^{eff}(\beta_0) = g^{opt}(\Delta^\dagger(\beta), U(\beta), \hat{\xi})\mathbf{Z}[\mathbf{Z}' s_{\Delta\Delta}\mathbf{Z}]^{-1}\mathbf{Z}' g^{opt}(\Delta^\dagger(\beta), U(\beta), \hat{\xi})',$$

where $s_{\Delta\Delta} = \frac{1}{N-L} g^{opt}(\Delta^\dagger(\beta), U(\beta), \hat{\xi})' M_{\mathbf{Z}} g^{opt}(\Delta^\dagger(\beta), U(\beta), \hat{\xi})$ and $M_{\mathbf{Z}} = I - P_{\mathbf{Z}}$. The optimal version of the K and J statistics can be derived similarly.

To clarify the proposition, for example, assume that T is exponentially distributed with mean $1/\lambda$. Then the density function of the treatment-free survival time T^0 is $f_{T^0}(t) = \lambda\exp(\beta D)\exp(-t\lambda\exp\{\beta D\})$ and the log-likelihood function for $U(\beta)$ is

$$\mathcal{L} = \beta D\Delta^\dagger(\beta) + \Delta^\dagger(\beta)\log(\lambda_0) - \lambda_0 Y\exp(\beta D).$$

Taking derivatives of $\mathcal{L}$ with respect to β gives the score function

$$S_\beta(\lambda) = \frac{\partial\mathcal{L}}{\partial\beta} = D\Delta^\dagger(\beta) - \lambda DU(\beta).$$

After replacing λ with its maximum likelihood estimator $\hat{\lambda}$, we have $S_\beta(\hat{\lambda}) = \frac{\partial\mathcal{L}}{\partial\beta} = D\left(\Delta^\dagger(\beta) - U(\beta)\frac{\sum\Delta^\dagger(\beta)}{\sum U(\beta)}\right)$.

Thus, $g^{opt}(\Delta^{\dagger}(\beta), U(\beta), \hat{\lambda}) = \left(\Delta^{\dagger}(\beta) - U(\beta)\frac{\sum \Delta^{\dagger}(\beta)}{\sum U(\beta)}\right)$.

The asymptotic distributions of our test statistics are agnostic to the number of covariates. This is because we are not projecting out the effect of covariates from the outcome model. However, one can gain efficiency by including the covariates into the analysis and estimate $\hat{\lambda}$ using the covariates, i.e., $\hat{\lambda}(\mathbf{x})$, and plugging it in $g^{opt}(\Delta^{\dagger}(\beta), U(\beta), \hat{\lambda}(\mathbf{x}))$. In this situation, the denominator of the variance estimator must be modified, and

$$s_{\Delta\Delta} = \frac{1}{N - L - p} g^{opt}(\Delta^{\dagger}(\beta), U(\beta), \hat{\lambda}(\mathbf{x}))' M_{\mathbf{Z}} g^{opt}(\Delta^{\dagger}(\beta), U(\beta), \hat{\lambda}(\mathbf{x})),$$

where p is the dimension of the vector of covariates. See Section S4 of the supplementary material.

4 Simulation studies

We further evaluate in simulation studies the performance of the proposed method under different survival time distributions. Within each scenario, we also vary the strength and number of instruments. The treatment variable is generated from

$$D \sim Binomial\left(\frac{\exp\{\gamma\mathbf{Z} + \eta\mathbf{X} + U\}}{1 + \exp\{\gamma\mathbf{Z} + \eta\mathbf{X} + U\}}\right), \tag{6}$$

where $\eta = (0.5, 0.5)$, $\mathbf{X} = (X_1, X_2)$ and U generated from a standard normal distribution with with a diagonal covariance matrix, and $\mathbf{Z}$ is a $k \in \{5, 50\}$ dimensional vector of IVs that are generated independently from a standard normal distribution. We assume that $\mathbf{X}$ are U are measured and unmeasured confounders, respectively. The parameter $\gamma = (\gamma_1, 0, 0, ..., 0)$ reflects the strength of IVs with $\gamma_1 = 2.0$, and 1.0. Thus, in our simulation, only the first IV is valid. We have tuned the censoring mechanism such that we get about 30% censoring in all the scenarios. We implemented five methods:

- KJ represents the estimator obtained by inverting the test statistic in (5). The KJ^{eff} corresponds to the more powerful version of the KJ statistic discussed in Proposition 2.

- AR represents the estimator obtained by inverting the test statistic in (4). The AR^{eff} corresponds to the more powerful version of the AR statistic discussed in Proposition 2.
- 2SLS correspond to a two-stage IV method in which the second stage fits a parametric AFT model that includes predicted treatment values obtained in the first stage model that regresses the treatment on $\mathbf{Z}$.
- LIML is the limited information maximum likelihood estimator that is obtained using the `ivmodel` R package where we replace the outcome with the artificial censoring indicator.
- Fuller represents an estimator proposed by Fuller (1977) that is obtained using the `ivmodel` R package where we replace the outcome with the artificial censoring indicator.
- 2SLS^{arti} is a G-estimation based estimator that involves artificial censoring and assumes a parametric AFT model for the survival times.
- Regression ignores the presence of unmeasured confounding and fits an AFT model using treatment D as an independent variable.

4.1 Simulation study: Exponential

Consider exponential distribution for the survival time,

$$Y \sim exponential\left(5\exp\{2D + 0.5X_1 + 0.5X_2 + U\}\right).$$

We also assume that the first IV, i.e., the only valid IV, is associated with $\mathbf{X}$ such that $Z_1 = X_1 + X_2 + \epsilon$ where ϵ is a standard normal random variable. We generate 500 datasets of size 1000 according to this model.

In the KJ^{eff} and AR^{eff}, we consider $g^{opt}(\Delta^{\dagger}(\beta), U(\beta), \hat{\lambda}) = \left(\Delta^{\dagger}(\beta) - U(\beta)\hat{\lambda}(\mathbf{x})\right)$ where for any given β, $\hat{\lambda}(\mathbf{x}) = \exp\{\hat{\eta}\mathbf{X}\}$ with $\hat{\eta} = \arg\min_{\eta}\left\{\beta D\Delta^{\dagger}(\beta) + \Delta^{\dagger}(\beta)\eta\mathbf{X} - Y\exp(\beta D + \eta\mathbf{X})\right\}$. In the 2SLS and 2SLS^{arti}, we assume exponential model for the outcome, i.e., no model misspecification. Both the KJ and AR statistics lead to an unbiased estimators and KJ^{eff} outperforms all the

other methods. When the strength of IV is moderate, i.e., $\gamma_1 = 2.0$, the LIML and Fuller estimators perform well but their confidence intervals are slightly undercovered when L=50. Moreover, when the dimension of the vector of instruments is small, i.e., L=5, the latter methods lead to confidence intervals that are 10%-20% wider than the one obtained by the KJ^{eff}. Although the estimator 2SLS^{arti} with L=5 and $\gamma_1 = 1.0$ is unbiased, the coverage rate of the confidence interval is slightly below the nominal rate and as the number of instruments increases to L=50, the coverage rate decreases drastically and the estimator reveals significant bias. The LIML and Fuller estimators also suffer from low coverage rates when the strength of the IV is reduced and L=50.

When there are L=50 instruments and the only valid IV is weak, i.e., $\gamma_1 = 1.0$, KJ^{eff} is the only unbiased approach that provides two sided confidence intervals with a valid coverage rate. The other test based procedures KJ, AR^{eff} and AR fail to provide an upper bound for the confidence intervals and the LIML and the Fuller lead to undercovered confidence intervals. Figure 3 provides more detail about the power of KJ^{eff} (solid line), KJ (dashed line), AR^{eff} (dotted line) and AR (dashed-dotted line) statistics. The plot with L=50 and $\gamma_1 = 1.0$ shows low power on the right side of the true parameter value. Figure S2 in the supplementary material compares the power plot of the LIML with KJ^{eff} and AR^{eff}. The LIML reveals some type-I error rate inflation when $L = 50$ and $\gamma_1 = 1$. The Fuller estimators have similar behavior to LIML, and thus, omitted from Figure S2.

We also studies the effect of the sample size to the instrument dimension ratio on the estimates by fixing the $L = 50$ and increasing the sample size from 500 to 10,000. Figure 2 shows that for smaller sample sizes the 2SLS estimator (solid line) has a notable bias and as the sample size increases the bias reduces. Moreover, the KJ statistic (dotted line) has the smallest absolute bias and the bias is slightly larger for the AR estimate (dashed line).When the instruments are stronger with $\gamma_1 = 2$, the 2SLS estimator is less bias but the KJ and the AR methods outperform the 2SLS uniformly for all the sample sizes considered. The LIML and Fuller estimators reveal similar behavior as the KJ method and are omitted.

Table 1: Simulation study: The survival outcome is generated using a Exponential distribution with 30% censoring rate. L: number of instruments; $\gamma = (\gamma_1, 0, 0, ..., 0)$: strength of the instruments; CI: confidence interval; Covg: coverage of confidence intervals. True $\beta_0 = 2$.

Methods	L=5				L=50			
	Estimate	95% CI	Length of 95% CI	Covg.	Estimate	95% CI	Length of 95% CI	Covg.
$\gamma_1 = 2.0$								
$KJ_{1,4}^{eff}$	1.99	(0.90,3.10)	2.20	0.94	2.03	(0.77,3.25)	2.48	0.95
$KJ_{1,4}$	2.01	(0.88,3.28)	2.40	0.94	1.96	(0.73,3.45)	2.72	0.95
$KJ_{0,5}^{eff}$	1.97	(0.95,3.04)	2.09	0.95	1.95	(0.85,3.20)	2.35	0.95
$KJ_{0,5}$	1.95	(0.87,3.22)	2.35	0.95	1.96	(0.77,3.41)	2.64	0.94
AR^{eff}	1.95	(0.69,3.36)	2.67	0.93	2.00	(-0.15,4.25)	4.40	0.96
AR	1.94	(0.65,3.64)	2.99	0.94	2.05	(-0.26,5.42)	5.68	0.95
$2SLS^{arti}$	1.92	(1.01,3.02)	2.01	0.94	1.45	(0.67,2.46)	1.79	0.69
LIML	1.95	(0.87,3.22)	2.35	0.96	2.08	(0.86,3.21)	2.35	0.92
Fuller	1.97	(0.85,3.17)	2.32	0.95	1.90	(0.85,3.20)	2.35	0.92
2SLS	3.31	(2.51,4.11)	1.60	0.00	3.60	(2.72,4.48)	1.76	0.00
Regression	5.52	(5.08,5.96)	0.88	0.00	5.50	(5.06,5.94)	0.88	0.00
$\gamma_1 = 1.0$								
$KJ_{1,4}^{eff}$	2.00	(0.32,3.76)	3.44	0.95	2.03	(-0.44,5.12)	5.56	0.95
$KJ_{1,4}$	2.01	(0.26,4.88)	4.62	0.96	1.96	(-0.46,$+\infty$)	$+\infty$	0.98
$KJ_{0,5}^{eff}$	2.04	(0.34,3.74)	3.40	0.96	2.11	(-0.42,5.10)	5.52	0.96
$KJ_{0,5}$	2.03	(0.28,4.82)	4.54	0.95	2.12	(-0.44,$+\infty$)	$+\infty$	0.97
AR^{eff}	2.02	(-0.14,4.36)	4.50	0.96	2.10	(-1.90,$+\infty$)	$+\infty$	0.97
AR	2.04	(-0.22,5.32)	5.54	0.95	2.18	(-2.00,$+\infty$)	$+\infty$	0.97
$2SLS^{arti}$	1.90	(0.40,3.26)	2.82	0.87	1.16	(-0.24,2.24)	2.44	0.36
LIML	2.05	(0.25,4.41)	4.16	0.95	1.98	(0.06,7.90)	7.84	0.89
Fuller	2.04	(0.24,4.30)	4.06	0.94	1.91	(0.05,7.85)	7.80	0.90
2SLS	2.64	(1.76,3.52)	1.76	0.41	3.14	(2.30,3.98)	1.68	0.21
Regression	5.88	(5.48,6.28)	0.80	0.00	5.88	(5.48,6.28)	0.80	0.00

4.2 Simulation study: Weibull

Similar to to Section 4.1, we assume that $\eta = (0.5, 0.5)$ but consider a Weibull distribution for the survival time,

$$Y \sim Weibull\left(shape = 0.5, scale = 5\exp\{\beta D + X_1 + X_2 + 2U\}\right).$$

Our goal is to study how model misspecification affects the results. Specifically, in the 2SLS and $2SLS^{arti}$, we postulate exponential model for the outcome and in the KJ^{eff} and AR^{eff}, and con-

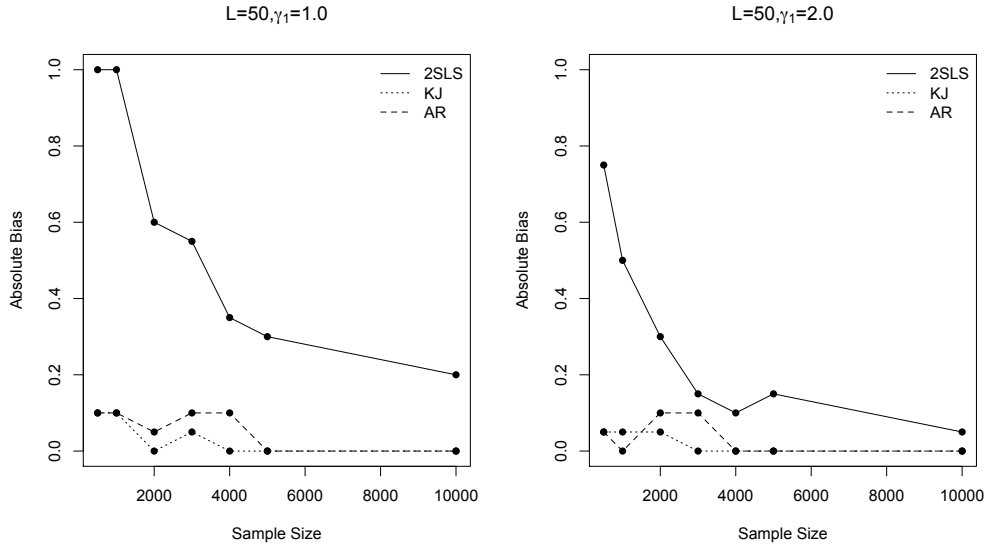

Figure 2: Simulation study: The survival outcome is generated using a Exponential distribution with 30% censoring rate and $L = 50$ instruments. $\gamma = (\gamma_1, 0, 0, ..., 0)$: strength of the instruments. Plots show the absolute bias of the KJ (dotted line), the AR (dashed line) and the 2SLS (solid line). Sample sizes are from 500 to 10,000.

sider $g^{opt}(\Delta^{\dagger}(\beta), U(\beta), \hat{\lambda}) = \left(\Delta^{\dagger}(\beta) - U(\beta)\frac{\sum \Delta^{\dagger}(\beta)}{\sum U(\beta)}\right)$ which was derived in Section 3.2. In Table 2, we have also reported the coverage of the corresponding confidence intervals to reflect the effect of model misspecification. Our previous results in Table 1 suggest that the 2SLS^{arti} provides unbiased effect estimates when there are small number of instruments, i.e., L=5. However, in Table 2, the coverage rate corresponding to the constructed confidence interval for the 2SLS^{arti} shows that type-I error rate is significantly inflated due to the model misspecification and/or the presence of weak IVs. The LIML and the Fuller estimators are substantially less efficient and when $L = 5$, the corresponding confidence intervals are up to 35% wider than the one obtained by the KJ^{eff}. Moreover, The LIML and the Fuller fail to provide finite confidence intervals. Figure 4 displays the power of KJ^{eff} (solid line), KJ (dashed line), AR^{eff} (dotted line) and AR (dashed-dotted line) statistics. See also Figure S3 in the supplementary material.

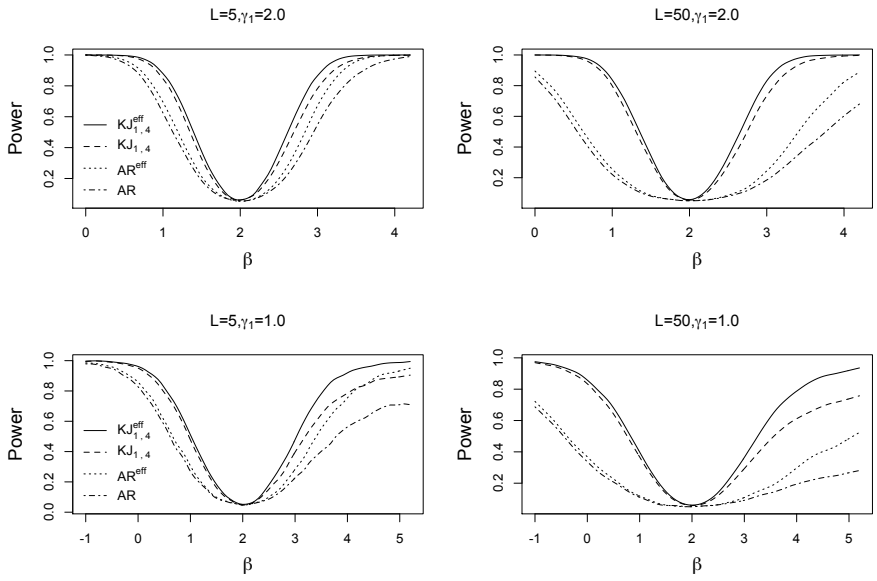

Figure 3: Simulation study: The survival outcome is generated using a Exponential distribution with 30% censoring rate. Power plots of efficient KJ (solid line), KJ (dashed line), efficient AR (dotted line) and AR (dotted-dashed line). L: number of instruments; $\gamma = (\gamma_1, 0, 0, ..., 0)$: strength of the instruments.

4.3 Simulation results summary

We studied the effect of number of instruments and the strength of instruments on different estimators. Our results showed that the g-estimation based the 2SLS estimator, i.e., 2SLS^{arti}, is unbiased only when there is a small number of instruments and as the number of instruments increases this estimator shows significant bias. However, the proposed test based estimators KJ^{eff}, KJ, AR^{eff}, AR, LIML, and Fuller remain unbiased regardless of the number of IVs. Estimators that are obtained by the KJ methods are more efficient and the corresponding confidence intervals are shorter than the ones obtained by AR, LIML, and Fuller statistics. For example, in Table 1 for L=50 and $\gamma_1 = 2$, the confidence interval obtained by the AR^{eff} is 1.87 wider than the one obtained by the KJ^{eff}. The importance of the KJ statistic and particularly the KJ^{eff} becomes more evident when

Table 2: Simulation study: Weibull. The survival outcome is generated using a Weibull distribution with 30% censoring rate. L:number of instruments; $\gamma = (\gamma_1, 0, 0, ..., 0)$: strength of the instruments; CI: confidence interval; Covg: coverage of confidence intervals. True $\beta_0 = 2$.

Methods	L=5				L=50			
	Estimate	95% CI	Length of 95% CI	Covg.	Estimate	95% CI	Length of 95% CI	Covg.
$\gamma_1 = 2.0$								
$KJ_{1,4}^{eff}$	1.99	(0.60,3.50)	2.90	0.96	2.00	(0.40,3.78)	3.38	0.95
$KJ_{1,4}$	2.01	(0.44,3.80)	3.36	0.96	2.00	(0.32,4.06)	3.74	0.95
$KJ_{0,5}^{eff}$	2.00	(0.66,3.42)	2.67	0.96	2.00	(0.42,3.64)	3.22	0.96
$KJ_{0,5}$	2.02	(0.52,3.72)	3.20	0.96	2.00	(0.34,3.92)	3.58	0.95
AR^{eff}	2.12	(0.24,3.98)	3.74	0.96	2.10	(-0.78,5.72)	6.50	0.95
AR	2.08	(0.12,4.32)	4.20	0.97	2.09	(-0.84,6.32)	5.68	0.94
$2SLS^{arti}$	2.14	(0.84,3.02)	2.18	0.80	1.68	(0.58,2.64)	2.06	0.64
LIML	1.95	(0.45,3.65)	3.20	0.96	1.99	(0.38,3.76)	3.38	0.94
Fuller	1.94	(0.44,3.59)	3.15	0.95	2.00	(0.39,3.75)	3.36	0.95
2SLS	3.34	(2.54,4.14)	1.60	0.01	3.61	(2.81,4.41)	1.60	0.01
Regression	5.48	(5.32,5.64)	0.32	0.00	5.51	(5.35,5.67)	0.32	0.00
$\gamma_1 = 1.0$								
$KJ_{1,4}^{eff}$	2.00	(-0.36,4.98)	5.34	0.97	1.98	(-1.22,7.91)	9.13	0.94
$KJ_{1,4}$	1.94	(-0.48,6.51)	6.99	0.96	1.96	(-1.49,+∞)	$+\infty$	0.97
$KJ_{0,5}^{eff}$	2.01	(-0.34,4.72)	5.06	0.97	1.98	(-1.20,7.73)	8.93	0.95
$KJ_{0,5}$	1.93	(-0.40,6.37)	6.77	0.96	1.96	(-1.44,+∞)	$+\infty$	0.98
AR^{eff}	1.80	(-0.90,6.70)	7.60	0.96	2.00	(-2.81,+∞)	$+\infty$	0.96
AR	2.20	(-1.18,+∞)	$+\infty$	0.97	2.00	(-3.23,+∞)	$+\infty$	0.97
$2SLS^{arti}$	2.04	(0.02,3.68)	3.62	0.81	1.04	(-0.48,2.50)	2.94	0.47
LIML	1.94	(-0.55,6.95)	7.50	0.95	1.95	(-0.80,+∞)	+∞	0.96
Fuller	1.96	(-0.55,6.89)	7.44	0.95	1.95	(-0.84,+∞)	+∞	0.97
2SLS	2.47	(1.25,3.69)	2.44	0.36	3.61	(2.81,4.41)	1.60	0.01
Regression	5.99	(5.37,6.61)	1.24	0.00	5.51	(5.35,5.67)	0.32	0.00

there are many weak IVs, e.g., L=50 and $\gamma_1 = 1$, where AR, LIML, and Fuller based estimators often fail to provide an upper bound for their corresponding confidence intervals (see for example Table 2). We have also studied the effect of model misspecification on different estimators where the true survival times were generated from a Weibull distribution and the postulated model was exponential. This model misspecification results in invalid confidence intervals for the 2SLS based estimator $2SLS^{arti}$ (Table 2). However, the proposed estimators are robust to model misspecification and the inference remains valid. We have also implemented the bias-corrected 2SLS estimator (Hausman & Newey, 2004) and the results showed that its performance was inferior to all the other

estimators considered except the 2SLS and the Regression estimators (results are omitted).

In Sections S4 of the supplementary materials, we have investigated the performance of our methods in the presence of many covariates in the model $p = 20, 100$. The results show that the type-I error rates are slightly inflated for the LIML and the Fuller estimators but the KJ^{eff} performs well. Moreover, when there are many covariates, the LIML and the Fuller have higher and lower statistical power in the left and right sides of the true value, respectively, compared with the KJ^{eff} (Table S2). The KJ statistic also performs well regarding the type-I error rate but as expected has slightly lower power than the LIML and the Fuller statistics. In Sections S5 of the supplementary materials, we have studied the impact of increasing the number of instruments to $L = 100, 250$ on our estimators. Figure S1 shows that while the KJ outperforms all the other methods when L=100 and $\gamma_1 = 1, 2$, it reveals some type-I error rate inflation when L=250 and this is exacerbated when the instrument is weak, i.e., $\gamma_1 = 1$. The LIML performs poorly in all the scenarios and seems to be unreliable due to the dramatic type-I error rate inflation. The Fuller estimator reveals similar behavior as the LIML and is omitted. Our results suggest that, in such extreme cases, i.e., L=250, the AR statistic is more robust compared with the KJ and LIML estimators.

In general, when IVs are not strong, the coverage rate of the 2SLS^{arti} estimator is below the nominal rate and increasing the number of IVs decreases the coverage rate drastically (Stock et al., 2012). Although the LIML and Fuller also suffer when there are many IVs, they seem to be more robust than the 2SLS. Overall, the KJ^{eff} outperforms all the other estimators in all the different scenarios considered in this section. It is robust to model misspecification and provides valid inference in the presence of many weak IVs.

5 Application

5.1 Overview

Our dataset includes 111,000 sentenced for a felony in Michigan between 2003-2006. We are interested in the total effect of incarceration on time to employment with quarterly earnings above

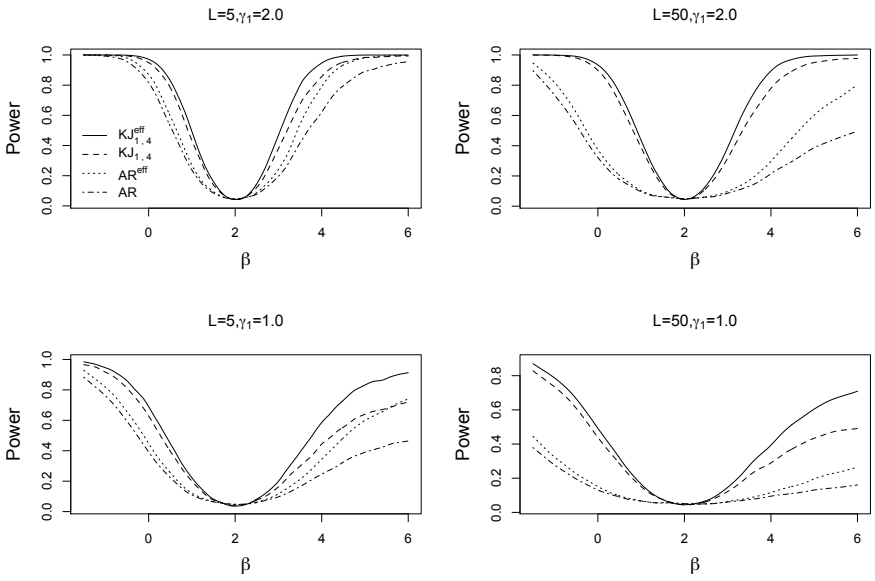

Figure 4: Simulation study with covariates: The survival outcome is generated using a Weibull distribution with 30% censoring rate. Power plots of efficient KJ (solid line), KJ (dashed line), efficient AR (dotted line) and AR (dotted-dashed line). L:number of instruments; $\gamma = (\gamma_1, 0, 0, ..., 0)$: strength of the instruments.

poverty, and thus, the time is measured from the sentence date. At the end of the study follow-up, 45% of felons were unemployed and considered as censored observations. We define a binary treatment variable which is one if sentenced to prison and zero otherwise. Using judges' ID, we create a 150 dimensional vector of mutually orthogonal binary instruments ($\mathbf{Z}^*$).

5.2 Validity of IV assumptions

The validity of a candidate instrument relies the core assumptions A.3-5. We assessed the association between the IV and the sentence type by fitting a random effect model discussed in Section 1. Our results showed that the random intercept component is significant which provides evidence that A.3 is satisfied. However, as shown in Figure 1, the estimated random intercepts are relatively

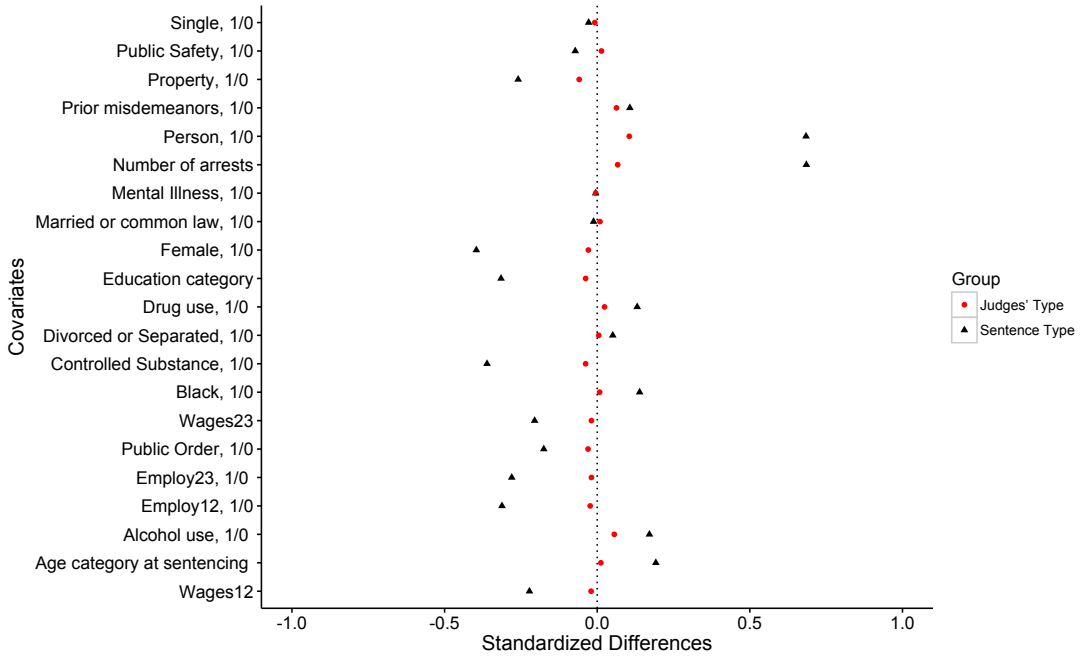

Figure 5: Michigan Sentencing data. Covariate imbalance across the sentence types, i.e., treatment, and the judges' type, i.e., function of candidate IVs. Standardized difference defined as the absolute value of the difference in means divided by the pooled standard deviation.

small for many judges which suggests the presence of weak IVs.

While we cannot empirically verify that judge assignment is random with respect to unobserved variables, we can check that the covariates we observe are uncorrelated with judge assignment. In order to assess the degree of covariate balance across judge types, we created two categories of judge "harshness" based on whether the estimated random effect was above or below the median. The dichotomization helps us to assess whether the covariate imbalances are reduced across the different judges types compared to the sentence types, i.e., treatment groups. This procedure provides insight about the validity of A.4. Specifically, imbalance in measured confounders across categories of the IV makes assumption A.4 less plausible because, for example, if the measured covariates are a proxy of the unmeasured confounders, an association between the measured confounders and the

IV suggests that there will be an association between the IV and unmeasured confounders. Figure 5 shows the standardized differences in means, that are, the values of the differences in means divided by the pooled standard deviation. This standardized differences suggest that there is a significant improvement in terms of the covariate imbalance across the judges types compared to sentence type.

The exclusion restriction, i.e., A.5, is another core assumption in IV analyses that is not completely testable. In our example, A.5 could be violated if offenders that are assigned to a harsh judge, i.e., judges that have more tendency of sentencing offenders to prison, were more likely to take plea bargain which may eventually plea down to a misdemeanor. Thus it may be easier for the offender to find a job. However, It is conceivable to assume that after controlling for all the measured covariates, the judges' ID affects the time to employment, i.e., outcome, only indirectly through the sentence type which suggests that A.5 is plausible.

5.3 The effect of imprisonment on time to employment with quarterly earnings above poverty

We include all the covariates listed in Figure 5 in our analysis by redefining our IV as the orthogonal projection of $\mathbf{Z}^*$ onto the space spanned by the covariates. Table 3 shows the effect estimates and 95% confidence intervals (CI) obtained by different methods. Overall, our analysis shows that imprisonment has a significant negative effect on time to employment with quarterly earnings above poverty. The point estimates obtained by the KJ and the AR methods are fairly close with the AR estimate being slightly lower. However, there is a drastic difference in terms of the length of the corresponding confidence intervals. Specifically, the $\text{KJ}_{1,4}$ gives the point estimate of -1.81 with 95% CI (-2.08,-1.26) while the point estimate using the AR is -1.70 with 95% CI (-2.58,-0.12). Thus, the length of the confidence interval if the former estimator is 65% shorter than the latter one. This highlights the importance of the dimension reduction technique used in the KJ statistics that leads to a substantial power gain. The point estimates imply that being sentenced to prison multiples the number of quarters to employment by factor of 6 compared with non-prison sentences.

Table 3: Michigan Data. L:number of instruments, i.e., judges, CI: confidence interval

Methods	Estimate	L=150 95% CI	Length of 95% CI
$\text{KJ}_{1,4}^{eff}$	-1.80	(-2.07,-1.32)	0.75
$\text{KJ}_{1,4}$	-1.81	(-2.08,-1.26)	0.82
$\text{KJ}_{0,5}^{eff}$	-1.80	(-2.05,-1.34)	0.71
$\text{KJ}_{0,5}$	-1.81	(-2.05,-1.28)	0.77
AR^{eff}	-1.70	(-2.54,-0.39)	2.17
AR	-1.70	(-2.58,-0.12)	2.46
LIML	-1.80	(-2.06,-1.30)	0.76
Fuller	-1.80	(-2.06,-1.31)	0.75
2SLS^{arti}	-1.97	(-2.15,-1.74)	0.41
Regression	-0.83	(-0.86,-0.80)	0.06

To derive the more powerful test statistics KJ^{eff} and AR^{eff}, we assume exponential model for the outcome and consider $g^{opt}(\Delta^{\dagger}(\beta), U(\beta), \hat{\lambda}) = \left(\Delta^{\dagger}(\beta) - U(\beta)\frac{\sum \Delta^{\dagger}(\beta)}{\sum U(\beta)}\right)$ which was derived in Section 3.2. The 2SLS^{arti} is a G-estimation based estimator that involves artificial censoring and assumes exponential model for the survival times. Also, the Regression approach ignores the presence of unmeasured confounding and, assuming an exponential model, fits an AFT model using treatment D and measured covariates as independent variables. As expected, the KJ^{eff} and AR^{eff} have shorter confidence intervals compared with the KJ and AR. The Regression estimator is bias and seems to underestimate the effect of imprisonment due to both possible unmeasured confounding. The LIML and Fuller estimators are similar to the KJ estimators with confidence intervals that are slightly wider than the KJ^{eff} and shorter than the KJ.

The point estimates obtained by the 2SLS^{arti} and the KJ statistic are fairly close with the 2SLS^{arti} being slightly higher. This might be due to the large ratio of the sample size to instrument dimension, which implies that, in terms of bias, the 2SLS does not suffer by much from the overfitting issue discussed in Bound et al. (1995) (see Figure 2 in the simulation studies section). However, the confidence interval of the 2SLS^{arti} estimator is 50% shorter than the one obtained by the KJ statistic which may be caused by either presence of weak IVs or misspecification of the

postulated survival model. In our analyses, we have 151 judges and after controlling for measured covariates, the F-statistic is only 27 which suggests the possibility of weak instruments (see also Figure 1). Following Stock & Yogo (2005) and our simulation results in Table 2, it is likely that the confidence interval of the 2SLS^{arti} estimator is undercovered, and thus it is not a valid confidence interval.

5.4 Subgroup Analyses

The effect of imprisonment on employment varies across gender and race (Steffensmeier et al., 1998; Pager, 2003, 2008; Decker et al., 2014). Table 4 summarizes the results of our subgroup analysis. Imprisonment have the largest negative effect among White Female offenders. Specifically, the estimated effect using KJ$^{eff}_{0,5}$ among White Male and Female offenders are -1.92 with 95% CI (-2.15,-1.45) and -2.05 with 95% CI (-3.46,0.01), respectively.

The magnitude of the 2SLS^{arti} bias increases as the sample size decreases. For the subgroup of Black Females ($n = 7,867$) which is the smallest in sample size, the KJ$^{eff}_{0,5}$ results in a point estimate of -0.87 with 95% CI (-3.43,2.20) while the 2SLS^{arti} estimate is -1.15 with 95% CI (-2.87,1.10). The performances of the LIML and Fuller estimators are also affected by the smaller sample size in the subgroup of Black Females. Specifically, in this subgroup, the LIML and Fuller estimates are roughly 50% smaller than the KJ estimates while for all the other subgroups the estimates are fairly close to the KJ estimates. The AR and AR^{eff} are only able to provide inference for the largest subgroup with $n = 50,516$ and for the other subgroups, these two test statistic fail because of low power. The KJ$^{eff}_{0,5}$, LIML and Fuller are the only procedures that have enough power to provide valid confidence intervals for all the subgroups regardless of their sample sizes. However, the latter two methods seem to lead to biased estimates for smaller sample sizes. We have investigated this point through our extensive simulation studies in Section 4 and Section S5 of the supplementary material.

Our analyses show that imprisonment has negative effect on time to employment with quarterly earnings above poverty. The effect is significant among Male offenders such that a prison sentence

Table 4: Michigan Data. Subgroup analysis. CI: confidence interval; n: sample size;

	Estimate	95% CI	Length of 95% CI	Estimate	95% CI	Length of 95% CI
	Subgroup: White, Male n=50,516			Subgroup: Black, Male n=39,724		
$KJ_{1,4}^{eff}$	-1.92	(-2.16,-1.40)	0.76	-1.38	(-1.88,-0.48)	1.40
$KJ_{1,4}$	-1.96	(-2.17,-1.32)	0.84	-1.38	(-1.84,-0.28)	1.56
$KJ_{0,5}^{eff}$	-1.92	(-2.15,-1.45)	0.70	-1.38	(-1.84,-0.51)	1.33
$KJ_{0,5}$	-1.96	(-2.15,-1.34)	0.81	-1.38	(-1.79,-0.40)	1.39
AR^{eff}	-1.66	(-2,68,-0.11)	2.57	–	–	–
AR	-1.66	(-2.77,0.10)	2.87	–	–	–
LIML	-1.96	(-2,15,-1.36)	0.79	-1.30	(-1.78,-0.46)	1.32
Full	-1.97	(-2.15,-1.37)	0.78	-1.30	(-1.77,-0.46)	1.31
$2SLS^{arti}$	-1.99	(-2.16,-1.73)	0.43	-1.60	(-1.92,-1.15)	0.77
	Subgroup: White, Female n=11,322			Subgroup: Black, Female n=7,867		
$KJ_{1,4}^{eff}$	-2.05	(-3.47,0.10)	3.57	-0.87	$(-3.44,+\infty)$	–
$KJ_{1,4}$	-2.05	(-3.48,0.42)	3.90	-0.80	$(-3.44,+\infty)$	–
$KJ_{0,5}^{eff}$	-2.05	(-3.46,0.01)	3.40	-0.87	(-3.43,2.20)	5.63
$KJ_{0,5}$	-2.05	(-3.46,0.38)	3.84	-0.80	$(-3.45,+\infty)$	–
AR^{eff}	–	–	–	–	–	–
AR	–	–	–	–	–	–
LIML	-1.99	(-2,91,-0.32)	2.59	-0.40	(-3.30,2.07)	3.37
Full	-1.99	(-2.89,-0.32)	2.57	-0.40	(-3.30,2.08)	3.38
$2SLS^{arti}$	-1.88	(-2.62,-0.56)	2.06	-0.70	(-3.25,1.05)	4.30

multiples number of quarters to the employment by roughly 7 and 4 among White and Black Male offenders, respectively.

6 Discussion

In this paper, we proposed a G-estimation based treatment effect inferential procedure in a survival context that is robust to the presence of many weak instruments. Our method adjusts for confounding using IVs and thus provides an unbiased estimate even when some of the confounders

are unmeasured. In general, one of the limitations of the G-estimation methods is that estimation becomes infeasible when there are several parameters of interest because such methods require a multi-dimensional grid search. Thus, generalization of the proposed method to settings with multiple treatments would be an interesting future work. Notice that selecting subjects based on their received treatment can result in a biased estimate (Swanson et al., 2015; Ertefaie et al., 2016b,a). Hence, performing multiple IV analyses to estimate treatment effects in multi-treatment settings is not possible.

IV analyses rely on some assumptions that cannot be completely tested using observed data. The validity of these assumptions become increasingly important when IVs are weak because estimators in such settings are invariably sensitive even to small departure from the assumptions (Imbens & Rosenbaum, 2005; Small & Rosenbaum, 2008; Baiocchi et al., 2010; Ertefaie et al., 2017b). Thus, developing sensitivity analyses for the proposed procedures is important and can provide support for the validity of the results (Small, 2007; Conley et al., 2012; Kolesár et al., 2015; Kang et al., 2015).

References

ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics* **113**, 231–263.

ALEXANDER, M. (2012). *The new Jim Crow: Mass incarceration in the age of colorblindness*. The New Press.

ANDERSON, T., KUNITOMO, N. & MATSUSHITA, Y. (2010). On the asymptotic optimality of the liml estimator with possibly many instruments. *Journal of Econometrics* **157**, 191–204.

ANDERSON, T. & RUBIN, H. (1949). Estimators of the parameters of a single equation in a complete set of stochastic equations. *The Annals of Mathematical Statistics* **21**.

ANGRIST, J. D., IMBENS, G. W. & RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association* **91**, 444–455.

APEL, R. & SWEETEN, G. (2010). The impact of incarceration on employment during the transition to adulthood. *Social problems* **57**, 448–479.

AUSTIN, J. & IRWIN, J. (2012). *It's about time: America's imprisonment binge*. Cengage Learning.

BAIOCCHI, M., SMALL, D. S., LORCH, S. & ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association* **105**, 1285–1296.

BEKKER, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society* , 657–681.

BELLONI, A., CHEN, D., CHERNOZHUKOV, V. & HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–2429.

BOSCO, J. L., SILLIMAN, R. A., THWIN, S. S., GEIGER, A. M., BUIST, D. S., PROUT, M. N., YOOD, M. U., HAQUE, R., WEI, F. & LASH, T. L. (2010). A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *Journal of clinical epidemiology* **63**, 64–74.

BOUND, J., JAEGER, D. A. & BAKER, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association* **90**, 443–450.

BROOKHART, M. A. & SCHNEEWEISS, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The international journal of biostatistics* **3**.

BROOKHART, M. A., WANG, P., SOLOMON, D. H. & SCHNEEWEISS, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology (Cambridge, Mass.)* **17**, 268.

BUSE, A. (1992). The bias of instrumental variable estimators. *Econometrica: Journal of the Econometric Society* , 173–180.

CHAMBERLAIN, G. & IMBENS, G. (2004). Random effects estimators with many instrumental variables. *Econometrica* **72**, 295–306.

CHAN, K. C. G. (2016). Instrumental variable additive hazards models with exposure-dependent censoring. *Biometrics* .

CHAO, J. C., HAUSMAN, J. A., NEWEY, W. K., SWANSON, N. R. & WOUTERSEN, T. (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of Econometrics* **178**, 15–21.

CHAO, J. C. & SWANSON, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica* **73**, 1673–1692.

CHOI, I. & PHILLIPS, P. C. (1992). Asymptotic and finite sample distribution theory for iv estimators and tests in partially identified structural equations. *Journal of Econometrics* **51**, 113–150.

CONLEY, T. G., HANSEN, C. B. & ROSSI, P. E. (2012). Plausibly exogenous. *Review of Economics and Statistics* **94**, 260–272.

CURRIE, E. (2013). *Crime and punishment in America*. Macmillan.

DAVIDSON, R. & MACKINNON, J. G. (2014). Confidence sets based on inverting anderson–rubin tests. *The Econometrics Journal* **17**, S39–S58.

DECKER, S. H., SPOHN, C., ORTIZ, N. R. & HEDBERG, E. (2014). Criminal stigma, race, gender and employment: An expanded assessment of the consequences of imprisonment for employment. *Department of Justice. https://www. ncjrs. gov/pdffiles1/nij/grants/244756. pdf* .

ERTEFAIE, A., SMALL, D. S., FLORY, J. & HENNESSY, S. (2016a). Selection bias when using instrumental variable methods to compare two treatments but more than two treatments are available. *The international journal of biostatistics* **12**, 219–232.

ERTEFAIE, A., SMALL, D. S., FLORY, J. & HENNESSY, S. (2016b). A sensitivity analysis to assess bias due to selecting subjects based on treatment received. *Epidemiology* **27**, e5–e7.

ERTEFAIE, A., SMALL, D. S., FLORY, J. H. & HENNESSY, S. (2017a). A tutorial on the use of instrumental variables in pharmacoepidemiology. *Pharmacoepidemiology and Drug Safety* .

ERTEFAIE, A., SMALL, D. S. & ROSENBAUM, P. R. (2017b). Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. *Journal of the American Statistical Association* .

FULLER, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica: Journal of the Econometric Society* , 939–953.

HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society* , 1–12.

HAHN, J., HAUSMAN, J. & KUERSTEINER, G. (2004). Estimation with weak instruments: Accuracy of higher-order bias and mse approximations. *The Econometrics Journal* **7**, 272–306.

HAHN, J. & INOUE, A. (2002). A monte carlo comparison of various asymptotic approximations to the distribution of instrumental variables estimators. *Econometric Reviews* **21**, 309–336.

HANSEN, C., HAUSMAN, J. & NEWEY, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics* **26**, 398–422.

HAUSMAN, C. H. J. & NEWEY, W. K. (2004). Many instruments, weak instruments, and microeonometric practice .

HERNÁN, M. A., COLE, S. R., MARGOLICK, J., COHEN, M. & ROBINS, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and drug safety* **14**, 477–491.

HERNÁN, M. A. & ROBINS, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology* **17**, 360–372.

IMBENS, G. W. & ROSENBAUM, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **168**, 109–126.

INOUE, A. & SOLON, G. (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics* **92**, 557–561.

JIANG, Y., SMALL, D. & ZHANG, N. (2014). Sensitivity analysis and power for instrumental variable studies. *Biometrics* .

JOFFE, M. M. (2001). Administrative and artificial censoring in censored regression models. *Statistics in medicine* **20**, 2287–2304.

KADANE, J. B. & ANDERSON, T. (1977). A comment on the test of overidentifying restrictions. *Econometrica: Journal of the Econometric Society* , 1027–1031.

KAFFO, M. & WANG, W. (2017). On bootstrap validity for specification testing with many weak instruments. *Economics Letters* .

KANG, H., CAI, T. T. & SMALL, D. S. (2015). Robust confidence intervals for causal effects with possibly invalid instruments. *arXiv preprint arXiv:1504.03718* .

KILGORE, J. (2015). *Understanding Mass Incarceration: A People's Guide to the Key Civil Rights Struggle of Our Time*. The New Press.

KLEIBERGEN, F. (2007). Generalizing weak instrument robust iv statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *Journal of Econometrics* **139**, 181–216.

KLING, J. R. (2006). Incarceration length, employment, and earnings. Tech. rep., National Bureau of Economic Research.

KOLESÁR, M., CHETTY, R., FRIEDMAN, J., GLAESER, E. & IMBENS, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics* **33**, 474–484.

LI, J., FINE, J. & BROOKHART, A. (2015a). Instrumental variable additive hazards models. *Biometrics* **71**, 122–130.

LI, Y., LEE, Y., WOLFE, R. A., MORGENSTERN, H., ZHANG, J., PORT, F. K. & ROBINSON, B. M. (2015b). On a preference-based instrumental variable approach in reducing unmeasured confounding-by-indication. *Statistics in medicine* **34**, 1150–1168.

LOEFFLER, C. E. (2013). Does imprisonment alter the life course? evidence on crime and employment from a natural experiment. *Criminology* **51**, 137–166.

MACKENZIE, T. A., TOSTESON, T. D., MORDEN, N. E., STUKEL, T. A. & O'MALLEY, A. J. (2014). Using instrumental variables to estimate a cox's proportional hazards regression subject to additive confounding. *Health Services and Outcomes Research Methodology* **14**, 54–68.

MAGDALINOS, M. A. (1990). The classical principles of testing using instrumental variables estimates. *Journal of Econometrics* **44**, 241–279.

MAGDALINOS, M. A. (1994). Testing instrument admissibility: some refined asymptotic results. *Econometrica: Journal of the Econometric Society* , 373–403.

MARK, S. D. & ROBINS, J. M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statistics in medicine* **12**, 1605–1628.

MAUER, M. (2001). The causes and consequences of prison growth in the united states. *Punishment & Society* **3**, 9–20.

NAGIN, D. S., CULLEN, F. T. & JONSON, C. L. (2009). Imprisonment and reoffending. *Crime and justice* **38**, 115–200.

NEWEY, W. K. & WINDMEIJER, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica* **77**, 687–719.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science* **5**, 465–472.

PAGER, D. (2003). The mark of a criminal record 1. *American journal of sociology* **108**, 937–975.

PAGER, D. (2008). *Marked: Race, crime, and finding work in an era of mass incarceration*. University of Chicago Press.

ROBINS, J. M. (1993). Analytic methods for estimating hiv treatment and cofactor effects. In *Methodological Issues of AIDS Mental Health Research*, D. Ostrow & R. Kessler, eds. New York: Plenum Publishing.

ROBINS, J. M. (1997). Structural nested failure time models. in: Survival analysis. In *The Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds. Andersen, P.K., and Keiding, N, Section Editors.

ROBINS, J. M. & TSIATIS, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics-Theory and Methods* **20**, 2609–2631.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6**, 34–58.

SMALL, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association* **102**, 1049–1058.

SMALL, D. S. & ROSENBAUM, P. R. (2008). War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association* **103**, 924–933.

SOLOMON, B. J., MOK, T., KIM, D.-W., WU, Y.-L., NAKAGAWA, K., MEKHAIL, T., FELIP, E., CAPPUZZO, F., PAOLINI, J., USARI, T. et al. (2014). First-line crizotinib versus chemotherapy in alk-positive lung cancer. *New England Journal of Medicine* **371**, 2167–2177.

STAIGER, D. & STOCK, J. (1997a). Instrumental variables regression with weak instruments. *Econometrica* **65**, 557.

STAIGER, D. & STOCK, J. H. (1997b). Instrumental variables regression with weak instruments. *Econometrica: Journal of the Econometric Society* , 557–586.

STAIGER, D. O. & STOCK, J. H. (1994). Instrumental variables regression with weak instruments.

STEFFENSMEIER, D., ULMER, J. & KRAMER, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology* **36**, 763–798.

STOCK, J. H., WRIGHT, J. H. & YOGO, M. (2012). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* .

STOCK, J. H. & YOGO, M. (2005). Testing for weak instruments in linear iv regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* .

SWANSON, S. A., ROBINS, J. M., MILLER, M. & HERNÁN, M. A. (2015). Selecting on treatment: a pervasive form of bias in instrumental variable analyses. *American journal of epidemiology* **181**, 191–197.

TCHETGEN, E. J. T., WALTER, S., VANSTEELANDT, S., MARTINUSSEN, T. & GLYMOUR, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)* **26**, 402.

TERZA, J. V., BASU, A. & RATHOUZ, P. J. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of health economics* **27**, 531–543.

TRAVIS, J., WESTERN, B. & REDBURN, F. S. (2014). *The growth of incarceration in the United States: Exploring causes and consequences*.

TSIATIS, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

TURNEY, K. & WILDEMAN, C. (2015). Detrimental for some? heterogeneous effects of maternal incarceration on child wellbeing. *Criminology & Public Policy* **14**, 125–156.

WEIMAN, D. F., STOLL, M. A. & BUSHWAY, S. (2007). The regime of mass-incarceration: A labor-market perspective. In *Pp. 29-79 in Barriers to Reentry?: The Labor Market for Released Prisoners in Post-Industrial America, edited by Shawn Bushway, Michael A. Stoll, and David F. Weiman.* New York: Russell Sage Foundation.

WESTERN, B. (2002). The impact of incarceration on wage mobility and inequality. *American Sociological Review* , 526–546.

WRIGHT, S. (1934). The method of path coefficients. *The annals of mathematical statistics* **5**, 161–215.