Synthetic Instrumental Variable Method Utilizing the Dual Tendency Condition[*]

Ratbek Dzhumashev[†]and Ainura Tursunalieva[‡]

March 21, 2025

Abstract

This paper introduces a novel Synthetic Instrumental Variable (SIV) method that constructs valid instruments using only existing data, addressing key challenges in traditional instrumental variable approaches. We demonstrate that any valid instrument can be represented as a linear combination of coplanar vectors spanned by the outcome and endogenous variables in a reduced form. Based on these coplanar vectors, we develop a "dual tendency" (DT) condition that provides a moment-based criterion for identifying valid SIVs, bridging the gap between unobservable orthogonality conditions and observable data characteristics. The SIV method is robust to address potential heteroscedasticity. Our method can also determine the true sign of $cov(\mathbf{x}, \mathbf{u})$, often assumed a priori in empirical work. Using simulated data and empirical applications, we demonstrate the effectiveness of our approach. The SIV method offers advantages over traditional IV approaches by mitigating issues of weak or invalid instruments and reducing reliance on scarce external instruments. This approach has broad implications for improving causal inference in various fields, such as economics, epidemiology, and policy evaluation.

Key words: *IV, OLS, endogeneity, generated instruments, synthetic instruments, causal inference*
JEL Code: C13, C18

1 Introduction

Endogeneity poses a significant challenge in econometric analysis, affecting the reliability of causal inferences in economic and social science research. This problem arises when the error term in a regression model correlates with the explanatory variables, leading to estimates that are inconsistent and biased. Ignoring endogeneity can lead to misguided policy recommendations, flawed theoretical insights, and a distorted understanding of complex economic phenomena. To advance economic knowledge and create effective policies, endogeneity must be addressed. Credible causal inference depends on it. Finding reliable solutions to the endogeneity problem is still a major challenge in contemporary econometrics, despite methodological advancements (Imbens, 2024).

Instrumental variable (IV) methods have long been the cornerstone of econometric approaches to addressing endogeneity. These methods rely on finding variables that are correlated with the endogenous regressors but uncorrelated with the error term. However, the application of IV methods is fraught with challenges: Firstly, finding valid external instruments is often an arduous task. It is almost always difficult to find a variable that satisfies all of the required restrictions imposed on IVs (Angrist and Krueger, 2001; Hausman, 2001). Secondly, even when potential instruments are identified, they frequently suffer from the weak instrument problem. When instruments only weakly correlated with endogenous regressors

[*]The authors thank Arthur Lewbel, Richard Butler, Jeff Wooldridge, and Guido Imbens for valuable comments on the paper.

[†]Department of Economics, Monash University, Ratbek.Dzhumashev@monash.edu

[‡]Data61, CSIRO ainura.tursunalieva@data61.csiro.au

can lead to large inconsistencies in IV estimates, sometimes exceeding the bias in OLS estimates they were meant to correct (Chernozhukov and Hansen, 2008; Stock et al., 2002). This issue is exacerbated in finite samples, where weak instruments can produce misleading inferences and confidence intervals. Lastly, the use of multiple instruments weakens the reliability of IV estimations (Bound et al., 1995; Stock et al., 2002; Wooldridge, 2013). These challenges underscore the need for new approaches that can overcome the limitations of traditional IV methods while maintaining their power to address endogeneity.

To overcome these challenges, we propose the Synthetic Instrumental Variable (SIV) method, a new approach for constructing valid instruments directly from the data without relying on external variables. The foundation of our method is the coplanarity among the outcome, the endogenous regressor, and the error term vectors. We utilize this coplanarity to demonstrate that any valid instrument can be represented as a linear combination of vectors within the plane formed by the outcome and the endogenous regressor. By leveraging this geometric insight, we can generate instruments from existing data, thus eliminating the need for external instruments, which are often difficult to obtain. To ensure the validity of our synthetic instruments, we introduce the "dual tendency" (DT) condition– a moment-based criterion that identifies valid SIVs within the coplanar plane. This condition effectively links the unobservable orthogonality requirement to observable features in the data, providing both theoretical rigor and a practical means of verification. Another significant advantage of our method is its ability to accurately identify the true sign of $\mathrm{cov}(\mathbf{x},\mathbf{u})$, a parameter that is typically assumed a priori in empirical research. This capability enhances the applicability and robustness of our approach across various empirical contexts, making it a powerful alternative to traditional instrumental variable techniques.

To further strengthen our method, we develop a robust extension that accounts for heteroscedasticity in the structural and first-stage error terms. This ensures validity across a broader range of data-generating processes, increasing its practical utility. In summary, our SIV approach provides a powerful alternative to traditional IV methods by: Mitigating weak or invalid instrument problems, reducing reliance on external instruments, and offering a data-driven strategy for instrument construction. These advancements make SIV a valuable tool for causal inference, expanding the scope of valid instrument selection in econometric analysis.

In terms of finding a valid IV, the SIV method operates on a fundamentally different principle than traditional IV approaches. The method begins by recognizing that in a reduced form, the outcome variable ($\mathbf{y}$), endogenous regressor ($\mathbf{x}$), and structural error term ($\mathbf{u}$) are coplanar. We demonstrate that for any valid instrument, $\mathbf{z}$, its orthogonal projection onto plane $\mathscr{W}$ spanned by $\mathbf{x}$ and $\mathbf{y}$ can be expressed in the form $\mathbf{z_0} = \mathbf{x} + k\delta\mathbf{r}$, where $\mathbf{r}$ is orthogonal to $\mathbf{x}$ and lies on plane $\mathscr{W}$. We use parameter $k = -1$ if $cov(\mathbf{x},u) > 0$, and $k = 1$ if $cov(\mathbf{x},u) < 0$, to capture the orientation of the valid IV with respect to the unobservable error term.[1]

This geometric insight allows us to construct any IV on the plane $\mathscr{W}$ as $\mathbf{s} = \mathbf{x} + k\delta\mathbf{r}$. The question is whether one can determine a valid IV among these synthetic vectors. We show that the above-introduced "dual tendency" (DT) condition identifies a valid synthetic instrumental variable (SIV) in the set of vectors constructed in such a way. In its baseline form, the DT condition asserts that if the vector $\mathbf{u}$ is homoscedastic, it implies that the first-stage error term, $\mathbf{e}$, associated with a valid IV, is also homoscedastic. Therefore, a valid SIV, $\mathbf{s}^*$, within the plane spanned by the vectors $\mathbf{x}$ and $\mathbf{y}$ must satisfy two conditions simultane-

[1] In certain instances, researchers might have strong opinions about the correlation sign of $\mathbf{x}$ and $\mathbf{u}$ (DiTraglia and García-Jimeno, 2021; Moon and Schorfheide, 2009). However, since $\mathbf{u}$ is unobservable, empirical researchers generally lack a rigorous method to ascertain this sign. Nonetheless, we demonstrate that the SIV method can help determine the sign of $cor(\mathbf{x},u)$, as will be covered in more detail below.

ously: $E(\mathbf{s}^{*\prime} \cdot \mathbf{u}) = 0$ and $E(\mathbf{s}^{*\prime} \cdot \mathbf{e}'\mathbf{e}) = 0$. This implies that if we synthesize such an SIV, $\mathbf{s}^*$, that satisfies $E(\mathbf{s}^{*\prime} \cdot \mathbf{e}'\mathbf{e}) = 0$, then according to the DT condition, $E(\mathbf{s}^{*\prime} \cdot \mathbf{u}) = 0$ also must hold. To implement this method, we iteratively adjust δ using the given values of $\mathbf{x}$, $\mathbf{r}$, and k, to identify the instrument $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ that satisfies the dual tendency (DT) condition, where δ_0 represents the optimal parameter value ensuring compliance with the DT condition.

In practice, we will only be using sample data; therefore, the optimal SIV, denoted as $\mathbf{s}^*$, is determined solely based on that sample. This issue can be remedied using the bootstrapping method, which allows the observed sample to estimate the population distribution and establish asymptotic properties of the SIV estimates. Specifically, we can infer the distributional characteristics of the optimal SIV $\mathbf{s}^*$ by applying the bootstrapping technique to the dataset. This involves constructing confidence intervals and calculating the expected value of the regression parameter using the SIVs determined for each bootstrap sample.

The baseline DT condition provides valuable guidance for identifying the true effects of the endogenous regressor. However, in practice, it may be weakened by issues such as heteroscedasticity in the error term. Using the following approach, we modify the baseline DT condition to account for the intrinsic heteroscedasticity of the error term. Building on the intuition from the Generalized Least Squares (GLS) estimator method, the spherical properties of the disturbances can be restored by transforming the original variables as follows:

$$\mathbf{P}'\mathbf{x} = \mathbf{P}'\mathbf{s}\gamma + \mathbf{P}'\mathbf{e}, \quad \text{where } \mathbf{P}'\mathbf{P} = \mathbf{H}^{-1}.$$

Here, $\mathbf{H}$ represents the covariance structure of the disturbances, but its full content is generally unknown.

The feasible GLS (FGLS) method allows us to derive a consistent estimator for $\mathbf{H}$, denoted as $\hat{\mathbf{H}} = \mathbf{H}(\hat{\theta})$. This implies that we can obtain two estimates of the first-stage error term: one from OLS ($\hat{\mathbf{e}}$) and one from FGLS ($\hat{\mathbf{e}}_\mathbf{g}$). In this context, measuring the degree to which the SIV contributes into the heteroscedasticity in the first-stage errors is crucial as it affects the difference between the estimated heteroscedasticities using $\hat{\mathbf{e}}$ and $\hat{\mathbf{e}}_\mathbf{g}$.

If the SIV deviates from the true IV, heteroscedasticity in the first-stage errors may arise from two sources: (1) the inherent heteroscedasticity associated with the true IV and (2) the misalignment between the SIV and the true IV. Intuitively, when the SIV perfectly aligns with the true IV, the variance of the first-stage errors remains independent of the SIV's divergence from the true IV. Thus, any residual heteroscedasticity should be attributable solely to the true IV. This insight suggests that at the DT condition point, the difference in heteroscedasticity between the two error estimates and should be minimized. We formally establish this result as the robust DT condition, demonstrating its effectiveness in mitigating heteroscedasticity and ensuring a more reliable identification of valid instruments.

The robust DT condition can be established as follows. First, we measure the degree of the difference in heteroscedasticity between the two error estimates as:

$$\Delta = \hat{\mathbf{e}}_\mathbf{g}\hat{\mathbf{e}}_\mathbf{g}'|\mathbf{s} - \hat{\mathbf{e}}\hat{\mathbf{e}}'|\mathbf{s}.$$

Denoting by $\mathbf{D}_\Delta$, the function that maps δ onto Δ, we solve for $\delta_0 = \arg\min_\delta(\mathbf{D}_\Delta)$ and identify the SIV $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ that satisfies $E(\mathbf{s}^{*\prime} \cdot \mathbf{u}) = 0$. This adjustment allows us to use the DT condition in the presence of heteroscedasticity, improving its robustness in practical applications.

The SIV method offers several advantages over traditional instrumental variable (IV) techniques: 1. No need for external instruments: Instruments are constructed directly from the data, removing the search for valid external variables. 2. Stronger instrument relevance: Ensures relevance by design, reducing concerns about weak instruments. 3. No need for multiple instruments: Synthesizes a single optimal instrument to avoid overidentification bias. 4. Data-driven endogeneity sign identification: Empirically determines the sign of $\mathrm{cov}(\mathbf{x}, u)$ without relying on assumptions. 5. Robust to heteroscedasticity: Accounts for structural error heteroscedasticity for broader applicability. These features make SIV a powerful alternative for addressing endogeneity in econometric analysis. It is important to note that approaches not using external instrumental variables (IVs) have already been proposed in the literature, as several methods for estimation have been developed. For instance, there are approaches based on higher moments, such as those by Lewbel (1997) and Erickson and Whited (2002). Other methods utilize heteroscedastic covariance restrictions, including contributions from Rigobon (2003), Lewbel (2012), and Klein and Vella (2010). Additionally, the Latent Instrumental Variables (LIV) method introduced by Ebbes et al. (2005) offers a distinct approach. In the LIV framework, the endogenous explanatory variable is separated into an exogenous component and an endogenous error term, where the exogenous part is represented by an unobserved discrete variable. The model parameters are identified and estimated using maximum likelihood methods.

Several alternative approaches exist for constructing synthetic instrumental variables (IVs) based on specific structural constraints. For instance, Gallo and Páez (2013) propose using an eigenvector spatial filter as an instrument, where eigenvectors derived from a transformed weights matrix capture underlying spatial patterns. This method relies on the correlations between the synthetic IV and the endogenous regressor to achieve identification. Similarly, Tang et al. (2024) demonstrates that causal effects can be identified through sparse causation, even in the presence of unmeasured confounding. Their approach constructs synthetic instruments from observed exposures. However, parameter identification requires that the sparsity assumption holds and that the effects of unmeasured confounding can be isolated. In another contribution, Vives-i Bastida and Gulek (2023) use the term the Synthetic IV (SIV). However, their approach to "synthetic" focuses on the use of synthetic controls to debias data within the framework of IV-DiD (instrumental variables with difference-in-differences) (Abadie, 2021), rather than on directly constructing synthetic IVs. It is worth mentioning that some researchers suggest tackling the endogeneity issues by employing an instrument-free estimation method that relies on the Gaussian copula function to capture the dependence between endogenous regressors and composite errors (Haschka, 2024; Park and Gupta, 2012). Haschka (2022) argues that such an approach works only if the joint distribution of endogenous regressors and the composite error are not multivariate normal.

All these mentioned methods are useful when access to suitable IVs is limited. However, these methods may lead to less accurate estimates due to reliance on higher-order moments or applying only to certain types of models as highlighted by Lewbel (2012). The primary limitation of the mentioned approaches is their reliance on additional constraints associated with the structural and first-stage error term dependencies.[2] In contrast, our method does not necessitate assumptions beyond the usual linear regression assumptions, offering greater generality in addressing IV estimation without external instruments.

The method developed in this paper may also be related to the design-based approaches to IVs, as our synthetic IV is also computed using a formula. Such a design-based approach leverages knowledge of the formula construction $f(\cdot)$ for an instrument to correct for potential endogeneity (confounding). In a few economic studies, instruments are created by applying a known formula to a combination of exogenous shocks and predetermined variables.[3] There is emerging literature on econometric tools for this setting, which

[2] Notably, the moments condition used for identification purposes in this paper, does not assume heteroscedastic covariance restriction as in Lewbel (2012), Klein and Vella (2010), and Rigobon (2003). In addition, the assumption of normal distribution of errors as in Ebbes et al. (2005) is not necessary in our case as our method allows for non-parametric estimation.

[3] Bartik (1991) and Blanchard and Katz (1992) pioneered the use of such instruments to measure labor demand elasticities.

leverage the assignment process of the exogenous shocks and the structure of the formula for identification. For example, Adão et al. (2019) studied inference in shift-share regression designs, whereas Borusyak et al. (2021) developed the approach to identification and consistency in this setting, and Borusyak and Hull (2023) extended this approach to general design-based instruments. A review of the use of the design-based ('formula') instrument approaches is given by Borusyak et al. (2024). Unlike design-based methods, our approach does not require additional exogenous shocks or variables to create synthetic instrumental variables.

The rest of the paper is organized as follows. Section 2 explains the DT condition for valid IVs. Section 3 discusses the SIV method. In Section 4, we describe the application of the SIV method to artificial and empirical data. Finally, Section 5 presents conclusions.

2 The properties of coplanar IVs

2.1 A coplanar instrumental variable in a regression model

Assume we have the classical endogeneity problem with $E(\mathbf{u}|\mathbf{x}) \neq 0$, in a simple model given as a triangular system[4]

$$\mathbf{y} = \beta\mathbf{x} + \mathbf{u}, \tag{1}$$

$$\mathbf{x} = \gamma\mathbf{z} + \mathbf{e} \tag{2}$$

Let $\mathscr{H}$ be a separable Hilbert space where the inner product of two vectors $\mathbf{v}$ and $\mathbf{w}$ is denoted by $(\mathbf{v} \cdot \mathbf{w})$.[5] All the vectors involved in (1) and (2) are determined in $\mathscr{H}$, where $\mathbf{y} \in \mathscr{H}$ is an $N \times 1$ vector of the outcome (dependent) variable, $\mathbf{x} \in \mathscr{H}$ is the endogenous regressor given by a $N \times 1$ vector, and $\mathbf{u} \in \mathscr{H}$ is a $N \times 1$ vector of unobservable error (disturbance) term.

An instrumental variable $\mathbf{z}$, an $N \times 1$ vector, is used in the first-stage equation (2). Note that any model with additional matrix $\mathbf{V} \in \mathscr{H}$ of predetermined or exogenous regressors, including a vector of ones, can be reduced to this form by defining $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{z}$ as residuals from the orthogonal projection onto the closed linear subspace spanned by $\mathbf{V}$. Specifically, let $\mathbf{P_V}$ be the orthogonal projection matrix onto span$(\mathbf{V})$, and define: $\mathbf{y} = (\mathbf{I} - \mathbf{P_V})\tilde{\mathbf{y}}$, $\mathbf{x} = (\mathbf{I} - \mathbf{P}_{\tilde{\mathbf{V}}})\tilde{\mathbf{x}}$, $\mathbf{z} = (\mathbf{I} - \mathbf{P_V})\tilde{\mathbf{z}}$, where $\mathbf{y}$, $\tilde{\mathbf{x}}$, and $\tilde{\mathbf{z}}$ denotes the original vectors in this extended case, and $\mathbf{I}$ is the identity matrix on $\mathscr{H}$. By construction, the residual vectors $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{z}$ are orthogonal to span$(\mathbf{V})$, effectively partialling out the exogenous regressors.

Conventionally, the IV method involves using an instrumental variable $\mathbf{z} \in \mathscr{H}$, an $N \times 1$ vector that satisfies the following two key conditions:

1. **Exogeneity:** $\mathbf{z}$ is exogenous implying that it is asymptotically uncorrelated with the error $\mathbf{u}$, i.e., $E(\mathbf{u}|\mathbf{z}) = \mathbf{0}$. This condition ensures that the instrument is not correlated with the unobserved factors influencing the outcome variable.

The review of this approach is presented by Goldsmith-Pinkham et al. (2020). Recent applications of this method explore public economics (e.g., Diamond, 2016; Saiz, 2010), macroeconomics (e.g., Jaravel, 2019; Nakamura and Steinsson, 2014; Oberfield and Raval, 2021), finance (e.g., Greenstone et al., 2020; Xu, 2022), immigration (e.g., Card, 2009; Peri et al., 2016), and trade (e.g., Autor et al., 2013; Hummels et al., 2014).

[4] We use the following notation conventions: Scalar variables will be denoted with an italic lowercase letter, such as β or x_i; a column vector of scalar values will be denoted by a boldface lowercase letter, such as $\mathbf{u}$. A boldface uppercase letter will denote a matrix, such as $\mathbf{V}$. Spaces and planes will be denoted using calligraphic letters such as $\mathscr{W}$.

[5] In this context, a Hilbert space is a complete vector space equipped with an inner product that induces a distance metric. A separable Hilbert space is a Hilbert space that has a countable dense subset. It can be thought of as an infinite-dimensional generalization of Euclidean space, allowing us to work with more abstract mathematical objects while retaining many of the geometric properties familiar from finite-dimensional spaces.

2. **Relevance:** $\mathbf{z}$ is relevant meaning that it is partially and sufficiently strongly correlated with the endogenous regressor $\mathbf{x}$. Formally, this implies that $\text{cov}(\mathbf{x},\mathbf{z}) \neq 0$, or equivalently, $E(\mathbf{x}|\mathbf{z}) \neq \mathbf{0}$. The relevance condition ensures that the instrument can effectively predict variations in the endogenous regressor.[6]

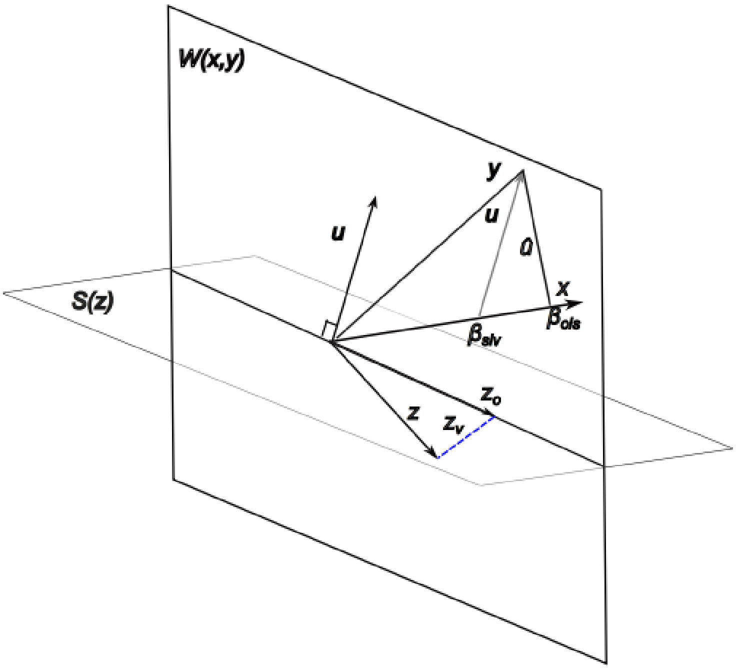


Figure 1: Geometric representation of instrumental variable (IV) and regression planes. The figure illustrates the relationship between the outcome variable $\mathbf{y}$, endogenous regressor $\mathbf{x}$, error term $\mathbf{u}$, and instrumental variable $\mathbf{z}$ in a three-dimensional space. $\mathscr{W}(\mathbf{x},\mathbf{y})$ represents the plane spanned by $\mathbf{x}$ and $\mathbf{y}$, while $\mathscr{S}(\mathbf{z})$ is orthogonal to $\mathbf{u}$. Vector $\mathbf{z_0}$ is the component of $\mathbf{z}$ that is coplanar with $\mathbf{x}$ and $\mathbf{y}$.

Let us consider the geometry of linear regression given by (1). By definition, vectors are coplanar if one can find no more than two linearly independent vectors in the set of vectors. This implies that the vectors $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{u}$ are coplanar, as they all lie in the same subspace $\mathscr{W} \subseteq \mathscr{H}$, which is the closed linear span of $\mathbf{x}$ and $\mathbf{y}$. Formally, we can express this subspace as

$$\mathscr{W} = \text{span}(\mathbf{x},\mathbf{y}) = \{\alpha\mathbf{x} + \beta\mathbf{y} \mid \alpha,\beta \in \mathbb{R}\} \tag{3}$$

as shown in Figure 1.[7]

On the other hand, the IV vector $\mathbf{z} \in \mathscr{S}$ does not necessarily have to lie in the same subspace of $\mathbf{x}$ and $\mathbf{y}$, $\mathscr{W} = \text{span}(\mathbf{x},\mathbf{y})$. That is $\mathscr{S} \nsubseteq \mathscr{W}$ and $\mathscr{S} \cap \mathscr{W} \neq \emptyset$. Any vector $\mathbf{z}$ that belongs to the orthogonal complement $\mathscr{W}^\perp$ of $\mathscr{W}$ in $\mathscr{H}$, defined as:

$$\mathscr{W}^\perp = \mathbf{z} \in \mathscr{H} \mid \langle \mathbf{z}|\mathbf{w} \rangle = 0, \forall \mathbf{w} \in \mathscr{W}$$

satisfies the exogeneity condition $E(\mathbf{u}|\mathbf{z}) = 0$ by construction, since $\mathbf{u} \in \mathscr{W}$ implies $E(\mathbf{u}|\mathbf{z}) = 0$ for all $\mathbf{z} \in \mathscr{W}^\perp \equiv \mathscr{S}$. Such vectors do not have to be coplanar with $\mathbf{y}$ and $\mathbf{x}$ as any vector on a plane $\mathscr{W}^\perp(\mathbf{z})$, by

[6] As a rule of thumb, an instrument is considered not to be sufficiently strongly correlated with the endogenous variable if its first-stage F-statistics in (2) is less than 10.

[7] See Davidson and MacKinnon (2009, pp.54-56) for a discussion of the geometry of OLS, and a geometric explanation of the IV estimation in Butler (2016).

definition, is orthogonal to $\mathbf{u}$. (see Figure 1). However, if we focus on the orthogonal projection of $\mathbf{z}$ onto plane $\mathscr{W}(\mathbf{x},\mathbf{y})$, we can consider coplanar vectors only. Specifically, in our analysis we can use $\mathbf{z_0} = \mathbf{P_W z}$, the orthogonal projection of $\mathbf{z}$ onto plane $\mathscr{W}(\mathbf{x},\mathbf{y})$ spanned by $\mathbf{y}$ and $\mathbf{x}$, because $E(\mathbf{u}|\mathbf{z_0}) = 0$ implies that $E(\mathbf{u}|\mathbf{z}) = 0$ holds. We formally establish the previous conclusion and state it as the following lemma:

Lemma 2.1 *Let $\mathbf{z_0} = \mathbf{P_W z}$ be the orthogonal projection of IV $\mathbf{z}$ onto plane $\mathscr{W}(\mathbf{x},\mathbf{y})$. Then, $E(\mathbf{u}|\mathbf{z}) = 0$ holds, only if $E(\mathbf{u}|\mathbf{z_0}) = 0$ is true.*

Proof See Appendix A.1.

In light of Lemma 2.1, we re-formulate the model (1)-(2) as follows:

$$\mathbf{y} = \beta \mathbf{x} + \mathbf{u}, \tag{4}$$

$$\mathbf{x} = \gamma_0 \mathbf{z}_0 + \mathbf{e_0}. \tag{5}$$

According to Lemma 2.1, the vector $\mathbf{z}$ defined at the intersection of $\mathscr{W}(\mathbf{x},\mathbf{y})$ and $\mathscr{S}(\mathbf{z})$ satisfies the orthogonality condition; thus, $\mathbf{z_0} \perp \mathbf{u} : \mathbf{z} \in \mathscr{W}(\mathbf{x},\mathbf{y}) \cap \mathscr{S}(\mathbf{z})$. Because all vectors required for the IV estimation can be contained in plane $\mathscr{W}(\mathbf{x},\mathbf{y})$, in synthesizing IVs, we limit our focus to plane $\mathscr{W}(\mathbf{x},\mathbf{y})$ only and state the following lemma.

Lemma 2.2 *A projection of a valid instrument onto plane $\mathscr{W}(\mathbf{x},\mathbf{y})$ such that $\mathbf{z_0} \perp \mathbf{u} \in \mathscr{W}(\mathbf{x},\mathbf{y})$, can be expressed as a linear combination of the vectors representing the endogenous regressor, $\mathbf{x}$, and a coplanar vector, $\mathbf{r}$. That is, $\mathbf{z_0} = \zeta \mathbf{x} + \omega \mathbf{r}$, where $\zeta, \omega \in \mathbb{R}$.*

Proof Since $\mathbf{z_0}$, $\mathbf{x}$, and $\mathbf{r}$ belong to the subspace $\mathscr{W} = \text{span}(\mathbf{x},\mathbf{y}) \subseteq \mathscr{H}$, these vectors all lie in the same subspace of the Hilbert space $\mathscr{H}$. By definition, any coplanar vector on the plane can be expressed as a vector sum of two independent coplanar vectors. Therefore, we can express any valid instrument $\mathbf{z_0}$ on a plane $\mathscr{W}(\mathbf{x},\mathbf{y})$ as a linear combination of coplanar vectors $\mathbf{x}$ and $\mathbf{r}$.■

The vectors $\mathbf{x}$, $\mathbf{r}$ and $\mathbf{z_0} \in \mathscr{W}(\mathbf{x},\mathbf{y})$, and thus, they are coplanar. Let us impose the additional correlation conditions on these vectors. The motivation for imposing additional restrictions on the vectors is to simplify the task of synthesizing IVs. By carefully constructing the vectors to satisfy specific conditions, we can facilitate the process of finding suitable IVs that meet the required assumptions and properties. Since we can construct IVs that are positively correlated with the endogenous regressor, we can restrict ourselves to valid IVs that satisfy $corr(\mathbf{x},\mathbf{z_0}) > \mathbf{0}$. In addition, we can consider only such vectors $\mathbf{r} \in \mathscr{W}(\mathbf{x},\mathbf{y})$ that satisfy the following conditions: $0 = corr(\mathbf{x},\mathbf{r}) < corr(\mathbf{x},\mathbf{z_0})$, $0 = corr(\mathbf{x},\mathbf{r}) < corr(\mathbf{r},\mathbf{z_0})$ and $corr(\mathbf{y},\mathbf{r}) > 0$. These assumptions reflected in Figure 2 allow us to state the following lemma.

Lemma 2.3 *On plane $\mathscr{W}(\mathbf{x},\mathbf{y})$ spanned by $\mathbf{x}$ and $\mathbf{y}$, there are noncolinear vectors $\mathbf{x}$, $\mathbf{r}, \mathbf{z_0} \in \mathscr{W}$ that satisfy assumptions: $0 = corr(\mathbf{x},\mathbf{r}) < corr(\mathbf{x},\mathbf{z_0})$, $0 = corr(\mathbf{x},\mathbf{r}) < corr(\mathbf{r},\mathbf{z_0})$, and $corr(\mathbf{y},\mathbf{r}) > 0$. Then, a vector $\mathbf{z_0} \in \mathscr{W}(\mathbf{x},\mathbf{y})$ can be written as a linear combination of $\mathbf{x}$ and $\mathbf{r}$, in the following form:*

$$\mathbf{z_0} = \mathbf{x} + k\delta \mathbf{r}, \tag{6}$$

where $\mathbf{r} : E(\mathbf{r}'\mathbf{x}) = 0$, and $k = (-1) \cdot sign[cov(\mathbf{x},\mathbf{u})]$, with $\mathbf{u}$ being the structural error term.

Proof See Appendix A.2.

3 Synthetic instrumental variable method

Leveraging the result established in Lemma 2.2, we can synthesize an IV $\mathbf{s}^* : \mathbf{s}^* = \mathbf{z_0}$ that lies within the regression subspace $\mathscr{W} = \text{span}(\mathbf{x}, \mathbf{y}) \subseteq \mathscr{H}$, while still satisfying the exogeneity condition with respect to the error term $\mathbf{u} \in \mathscr{W}$. We refer to these IVs as the synthetic IVs. First, to build-up our logic of the SIV estimator, we consider a homoscedastic case. Then, we extend the SIV estimator to heteroscedastic cases, by developing a robust method. In the next subsection, we outline the assumptions required to develop such an SIV method.

3.1 The SIV assumptions

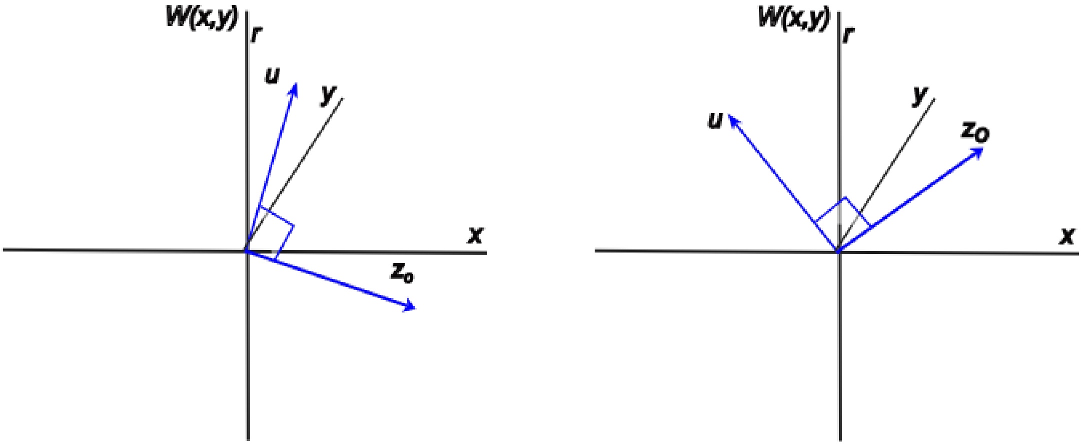
Let us first outline all the assumptions made for the SIV method. To simplify our analysis, we consider only the vectors that satisfy the following assumptions.

A1. Vectors used in SIVs: The vectors $\mathbf{x}$, $\mathbf{r}$ and $\mathbf{s} \in \mathscr{W}(\mathbf{x}, \mathbf{y}) \subseteq \mathscr{H}$, and thus are coplanar. These vectors satisfy Lemma 2.3. Thus, we can express an SIV as in (6) in the form: $\mathbf{s} = \mathbf{x} + k\delta\mathbf{r}$, where $\delta > 0$, and $k = (-1) \cdot sign[cov(\mathbf{x}, \mathbf{u})]$, with $\mathbf{u}$ being the structural error term. Lemma 2.3 implies that, in the construction of an SIV, we can use vector $\mathbf{r} : E(\mathbf{r}'\mathbf{x}) = 0$ determined as $\mathbf{r} = (\mathbf{I} - \mathbf{P_x})\mathbf{y}$, where $\mathbf{P_x} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-\mathbf{1}}\mathbf{x}$ is the orthogonal projection matrix onto $\mathbf{x}$, and $\mathbf{I}$ is the identity matrix of corresponding size.

A2. Synthetic IV: $\mathbf{s}$ is an $N \times 1$ vector and determined by $\mathbf{s} = \mathbf{x} + k\delta\mathbf{r}$. Scalar $\delta \in (0, \bar{\delta})$, where $\bar{\delta} = \arg(\bar{\mathbf{s}} = \mathbf{x} + k\delta\mathbf{r})$ such that $\mathbf{s} \to \bar{\mathbf{s}}$ implies that $E(\mathbf{s}'\mathbf{x}) \to 0$. That is the upper bound for δ is constrained by the orthogonal orientation of $\mathbf{s}$ relative to $\mathbf{x}$ which makes such an IV irrelevant. The SIV is consistent with the exclusion restriction, indicating that the SIV is associated with the outcome variable as a by-product of both $\mathbf{r}$ and $\mathbf{x}$. However, $\mathbf{y}$ cannot directly cause itself through its component represented by $\mathbf{r}$ (Heckman and Pinto, 2024). Thus an SIV defined as $\mathbf{s} = \mathbf{x} + k\delta\mathbf{r}$ can cause $\mathbf{y}$ only through its part stemming from $\mathbf{x}$.

A3. Full rank: $E(\mathbf{x}'\mathbf{y})$, and $E(\mathbf{x}'\mathbf{x})$ are finite and identified from data. $E(\mathbf{x}'\mathbf{x})$ is non-singular.

A4. The structural error term: The structural error term $\mathbf{u}$ is independently distributed and can be homoscedastic or heteroscedastic. If $\mathbf{u}$ is heteroscedastic, then, we assume that $E(\mathbf{u}\mathbf{u}') = \mathbf{H}(\theta) \neq \sigma_0^2\mathbf{I}$, where



a. $cor(\mathbf{x}, \mathbf{u}) > 0$ b. $cor(\mathbf{x}, \mathbf{u}) < 0$

Figure 2: Orientation of a valid SIV $\mathbf{z_0} \in \mathscr{W}(\mathbf{x}, \mathbf{y})$ relative to $\mathbf{y}$, $\mathbf{x}$ given the error term $\mathbf{u}$. Panel (a) shows the case when $cor(\mathbf{x}, \mathbf{u}) > 0$, and panel (b) shows the case when $cor(\mathbf{x}, \mathbf{u}) < 0$.

σ_0^2 is constant, and $\mathbf{I}$ is an identity matrix. It is assumed that the covariance matrix of the error terms, $\mathbf{H}$, can be estimated by $E(\hat{\mathbf{u}}\hat{\mathbf{u}}') = \mathbf{H}(\hat{\sigma}_0^2 + h(\hat{\theta}))$, where $h(\cdot)$ is any strictly positive, twice differentiable function such that $h(0) = 0$, $\frac{\partial h(0)}{\partial \mathbf{z}} \neq 0$, $\hat{\sigma}_0$ is a positive constant, and $\hat{\mathbf{u}}$ is the estimate of $\mathbf{u}$.

3.2 The dual tendency (DT) condition for coplanar IVs

Assume that there is vector $\mathbf{z_0}$ such that $E(\mathbf{u}|\mathbf{z_0}) = 0$ holds. By employing coplanar vectors $\mathbf{x}$ and $\mathbf{r}$, we can represent this vector as $\mathbf{z_0} = \mathbf{x} + k\delta_0\mathbf{r}$. This condition implies that parameter γ_0 in Eq. (5), $\mathbf{x} = \gamma_0\mathbf{z}_0 + \mathbf{e_0}$, is predetermined by the given coplanar vectors $\mathbf{x}$ and $\mathbf{r}$ through their SIV relationship with $\mathbf{z_0}$. We state this conclusion as the following lemma.

Lemma 3.1 *Suppose that the assumptions A1, A2, A3 hold and vector* $\mathbf{z_0} \in \mathscr{W}(\mathbf{x}, \mathbf{y})$ *is such that* $E(\mathbf{u}|\mathbf{z_0}) = 0$ *holds. Then, the first-stage parameter estimate in* $\mathbf{x} = \gamma\mathbf{s} + \mathbf{e}$ *can be expressed as* $\gamma = \gamma_0 + g(\mathbf{x}, \mathbf{s}, \mathbf{z_0})$*, where* $g(\cdot)$ *stands for a twice-differentiable function that captures the effect of deviation of* $\mathbf{s}$ *from the true IV* $\mathbf{z_0}$*.*

Proof See Appendix A.3.

Next, we build upon this lemma and state the following theorem.

Theorem 3.2 (DT condition) *Suppose that the assumptions A1, A2, A3 hold and the structural error term satisfies* $E(\mathbf{u}\mathbf{u}' \mid \mathbf{z_0}) = 0$*. Then, for the model given by equations* (4) *and* (5)*, the conditions ("orthogonality")* $E(\mathbf{u}|\mathbf{z_0}) = \mathbf{0}$ *and ("first-stage homoscedasticity")* $E(\mathbf{e_0}\mathbf{e}_0'|\mathbf{z_0}) = \mathbf{0}$ *hold simultaneously.*

Proof See Appendix A.4.

We refer to the condition established in Theorem 3.2 as the "dual tendency" (DT) condition, as it shows that the valid IV should satisfy both $E(\mathbf{u}|\mathbf{z_0}) = \mathbf{0}$ and $E(\mathbf{e}\mathbf{e}'|\mathbf{z_0}) = \mathbf{0}$. Importantly, the DT condition holds only with respect to $\mathbf{z_0}$, the coplanar projection of a valid IV, $\mathbf{z}$. This is because the complementary projection $\mathbf{z_v}$ does not have any effect on $\mathbf{u}$ as $\mathbf{z_v}$ is orthogonal to plane $\mathscr{W}(\mathbf{x}, \mathbf{y})$.

3.3 Identification using the SIV method: homoscedastic case

Since we can construct any SIV $\mathbf{s} \in \mathscr{W}$ by selecting $\delta \in (0, \bar{\delta})$, therefore, we can determine an SIV that satisfies Theorem 3.2 and thereby, identify a valid SIV. This result is stated in the following lemma.

Lemma 3.3 (DT condition for SIV) *Suppose that the assumptions A1, A2, A3 hold and* $\mathbf{z_0}$ *is a valid unobserved IV such that* $E(\mathbf{u} \mid \mathbf{z_0}) = 0$*. The structural error term satisfies* $E(\mathbf{u}\mathbf{u}' \mid \mathbf{z_0}) = 0$*. Then, a valid SIV such that* $E(\mathbf{u} \mid \mathbf{s}^* = \mathbf{z_0}) = 0$ *is identified by* $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ *where* $\delta_0 = \arg_\delta[E(\mathbf{e}\mathbf{e}' \mid \mathbf{s}^*) = 0]$*, and* $\mathbf{e}$ *is the first-stage error term.*

Proof See Appendix A.5.

Based on the results of Lemma 3.3, we determine the SIV estimator.

Corollary 3.4 (Standard SIV Estimator) *Suppose that an SIV given by* $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ *satisfies Lemma 3.3. Then,* β*, the parameter in* (4)*, is identified by an IV estimator:* $\hat{\beta}_{IV} = (\mathbf{x}'\mathbf{s}^*)^{-\mathbf{1}}\mathbf{x}'\mathbf{y}$*.*

Proof See Appendix A.6.

3.3.1 Intuitive explanation of the DT condition and the SIV identification

The intuition behind Theorem 3.2 is more easily understood through the control function (CF) approach. For the model specified by equations (4) and (5), the control function approach implies the following expression:

$$E(\mathbf{y} \mid \mathbf{x}, \mathbf{e_0}) = f(\mathbf{x}) + E(\mathbf{u} \mid \mathbf{x}, \mathbf{e_0}) = f(\mathbf{x}) + h(\mathbf{e_0}),$$

where $h(\mathbf{e_0})$ is defined as:

$$E(\mathbf{u} \mid \mathbf{x}, \mathbf{e_0}) = E(\mathbf{u} \mid \mathbf{z_0}, \mathbf{e_0}) = E(\mathbf{u} \mid \mathbf{e_0}) = h(\mathbf{e_0}).$$

In the linear case, we have $f(\mathbf{x}) = \beta\mathbf{x}$ and $h(\mathbf{e_0}) = \rho\mathbf{e_0}$ (see Wooldridge, 2015). The function $h(\mathbf{e_0})$ serves as the control function that accounts for endogeneity (Arellano, 2003). In other words, under the CF approach, the expression $E(\mathbf{y}) = \beta\mathbf{x} + \rho\mathbf{e_0}$ implies that $E(\mathbf{u}) = \rho E(\mathbf{e_0})$. The latter relationship and the homoscedasticity of the structural error term, $E(\mathbf{u}\mathbf{u}'|\mathbf{z_0}) = 0$, imply that $E(\mathbf{e_0}\mathbf{e_0}'|\mathbf{z_0}) = 0$. Given that $E(\mathbf{u}'\mathbf{z_0}) = 0$, this is the basic description of the DT condition that holds when $\mathbf{u}$ is homoscedastic.

Moreover, ρ is determined as (see Wooldridge, 2015):

$$\rho = \frac{E(\mathbf{e_0}'\mathbf{u})}{E(\mathbf{e_0}\mathbf{e_0}')},$$

where $\mathbf{e_0}$ is the error term from the first-stage equation, and $E(\mathbf{e_0}\mathbf{e_0}')$ represents its variance. This relationship captures how the control function approach models the endogeneity inherent in the system.

Let's assume that $\mathbf{z_0}$ is a valid IV such that $E(\mathbf{u}|\mathbf{z_0}) = 0$. When an SIV $\mathbf{s}$ differs from the valid IV $\mathbf{z_0}$, the first-stage error term $\mathbf{e} = \mathbf{x} - \gamma\mathbf{s}$ can be expressed as:

$$\mathbf{e} = \mathbf{e_0} + \mathbf{e_s},$$

where $\mathbf{e_0} = \mathbf{x} - \gamma_0\mathbf{z_0}$, and $\mathbf{e_s}$ represents the difference between the true error $\mathbf{e_0}$ and the estimated error $\mathbf{e}$. As the residual term, we assume that $\mathbf{e_s}$ is independent of $\mathbf{e_0}$. Using this, we can express the estimate of the parameter ρ as:

$$\hat{\rho} = \frac{E(\mathbf{e}'\mathbf{u})}{E(\mathbf{e}\mathbf{e}')} = \frac{E(\mathbf{e_0}'\mathbf{u}) + E(\mathbf{e_s}'\mathbf{u})}{E(\mathbf{e_0}\mathbf{e_0}') + E(\mathbf{e_s}\mathbf{e_s}') + 2E(\mathbf{e_0})E(\mathbf{e_s})}.$$

By construction $E(\mathbf{e_0}) = E(\mathbf{e_s}) = 0$. Accounting for this, we state

$$\hat{\rho} = \frac{E(\mathbf{e_0}'\mathbf{u}) + E(\mathbf{e_s}'\mathbf{u})}{E(\mathbf{e_0}\mathbf{e_0}') + E(\mathbf{e_s}\mathbf{e_s}')}.$$

It follows that $\hat{\rho} \xrightarrow{p} \rho$ only when $\operatorname*{plim}_{n\to\infty} \mathbf{e_s} = \mathbf{0}$. By construction, if $cor(\mathbf{z_0}, \mathbf{s}) \neq 1$, then, $E(\mathbf{e_s}\mathbf{e_s}' \mid \mathbf{s}) \neq 0$. Therefore, when $\mathbf{e_s} \neq \mathbf{0}$,

$$E(\mathbf{e}\mathbf{e}' \mid \mathbf{s}) = E(\mathbf{e_0}\mathbf{e_0}' \mid \mathbf{s}) + E(\mathbf{e_s}\mathbf{e_s}' \mid \mathbf{s}) \neq 0.$$

Only when $\mathbf{s}^* = \mathbf{z_0}$, we have $\mathbf{e_s} = \mathbf{0}$, thus, $E(\mathbf{e_s}\mathbf{e_s}' \mid \mathbf{s}^*) = 0$. This implies that

$$E(\mathbf{e}\mathbf{e}' \mid \mathbf{s}^* = \mathbf{z_0}) = E(\mathbf{e_0}\mathbf{e_0}' \mid \mathbf{z_0}).$$

This result tells us that when the structural error term, $E(\mathbf{uu}'|\mathbf{z_0}) = 0$, and thus $E(\mathbf{e_0e_0'}|\mathbf{z_0}) = 0$, we have

$$E(\mathbf{ee}' \mid \mathbf{s}^* = \mathbf{z_0}) = 0.$$

Therefore, finding $\mathbf{s}^*$ such that $E(\mathbf{ee}' \mid \mathbf{s}^*) = 0$ implies that this vector satisfies $\mathbf{s}^* = \mathbf{z_0}$, and thus, $E(\mathbf{u} \mid \mathbf{s}^*) = 0$.

In Section 3.5, we will relax the assumption $E(\mathbf{e_0e_0'} \mid \mathbf{z_0}) = 0$ and consider the heteroscedastic case.

3.4 Identifying the sign of $cor(\mathbf{x},\mathbf{u})$

Lemma 2.3 implies that for the correct identification of parameters using the SIV method, we need to know the correct sign of $cor(\mathbf{x},\mathbf{u})$. We can identify the sign of $cor(\mathbf{x},\mathbf{u})$ using the outcome of Theorem 3.3. Specifically, only when we assume the correct sign for $cor(\mathbf{x},\mathbf{u})$, we can identify the SIV, $\mathbf{s}^*$, that satisfies condition $\mathbf{E}(\mathbf{ee}' \mid \mathbf{s}^*) = \mathbf{0}$ for $\delta_0 > 0$. If we assume a wrong sign, then such an SIV violates Theorem 3.3, and thus, it does not satisfy condition $\mathbf{E}(\mathbf{ee}' \mid \mathbf{s}^*) = \mathbf{0}$ for all $\delta > 0$. Based on such reasoning, we state the following condition for the sign restriction.

Corollary 3.5 *Suppose that the assumptions A1, A2, A3 hold,* $\mathbf{u}$ *is homoscedastic, and* $cor(\mathbf{x},\mathbf{u}) \neq 0$*. Then, the true sign of* $cor(\mathbf{x},\mathbf{u})$ *is determined by the assumed sign that yields* $\delta_0 > 0$*, such that* $E(\mathbf{ee}' \mid \mathbf{s}^*) = \mathbf{0}$ *holds for* $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ *and* $\mathbf{e} = \mathbf{x} - \gamma\mathbf{s}$*.*

Proof See Appendix A.7.

It is important to know additional properties of the locus of $E(\mathbf{ee}' \mid \mathbf{s})$ for $\delta \in (0, \bar{\delta})$ that can help identify the true sign of $cor(\mathbf{x},\mathbf{u})$. Firstly, when $\delta = 0$, $E(\mathbf{ee}' \mid \mathbf{s}) = 0$ as in this case $\mathbf{x} \equiv \mathbf{s}$. According to Theorem 3.3, when $\delta = \delta_0 > 0$, we have $E(\mathbf{u} \mid \mathbf{s}) = 0$ and $E(\mathbf{ee}' \mid \mathbf{s}) = 0$ given the sign is assumed correctly. This allows us to state the following remark.

Remark 1. The locus of $E(\mathbf{ee}' \mid \mathbf{s})$, constructed under the correct assumption for the sign of $cor(\mathbf{x},\mathbf{u})$, exhibits an non-decreasing trend in absolute values and sign change on $\delta \in (0, \bar{\delta})$.

On the other hand, if we search for $\mathbf{s}^*$ by assuming a wrong sign for $cor(\mathbf{x},\mathbf{u})$ or if there is no endogeneity so that $cor(\mathbf{x},\mathbf{u}) = 0$, then $E(\mathbf{ee}' \mid \mathbf{s})$ must not have any sign change as $\delta = \delta_0 > 0$ is not feasible where we have $E(\mathbf{u} \mid \mathbf{s}) = 0$ holds. The locus of $E(\mathbf{ee}' \mid \mathbf{s})$ may exhibit decreasing trend in absolute values on $\delta \in (0, \bar{\delta})$.

In this light, one can compute two versions of the locus of $E(\mathbf{ee}' \mid \mathbf{s})$ assuming $cor(\mathbf{x},\mathbf{u}) > 0$ and $cor(\mathbf{x},\mathbf{u}) <$, alternatively. By testing the series representing these loci applying the conditions specified in Remark 1, one can determine the true sign of $cor(\mathbf{x},\mathbf{u})$.

3.5 Identification using the robust SIV method

Now, let us assume that $E(\mathbf{uu}' \mid \mathbf{z_0}) \neq \mathbf{0}$ so that we have a heteroscedastic structural error term. Based on the CF approach, since $E(\mathbf{u}) = \rho E(\mathbf{e})$, it follows that $E(\mathbf{ee}' \mid \mathbf{z_0}) \neq 0$. Here, $\mathbf{e}$ is the true first-stage error term. Then, the condition $E(\mathbf{ee}' \mid \mathbf{s}) = 0$ might be biased and therefore unable to identify the valid SIV. Thus, we need to have a robust method to identify the valid SIV in the case with the residual heteroscedasticity implied by $E(\mathbf{ee}' \mid \mathbf{z_0}) \neq 0$.

In the presence of heteroscedasticity, we have $E(\mathbf{ee}') = \mathbf{H} \neq \sigma_0^2\mathbf{I}$, where σ_0^2 is constant, and $\mathbf{I}$ is an identity matrix. According to the Generalized Least Squares (GLS) estimator method, the spherical properties of the disturbances can be restored by transforming the original variables as follows:

$$\mathbf{P'x} = \mathbf{P's}\gamma + \mathbf{P'e}, \tag{7}$$

where $\mathbf{P'P} = \mathbf{H^{-1}}$. Based on (7), it implies that

$$\mathbf{e_g} = \mathbf{P'e}. \tag{8}$$

If we were able to determine the true spherical first-stage error term, $\mathbf{e_g}$, then just could have used $E(\mathbf{e_g e_g'} \mid \mathbf{z_0}) = \mathbf{0}$ as the DT condition. Unfortunately, the full content of $\mathbf{H}$ will be known only in very few applications. In practice, we may use an estimate $\hat{\mathbf{H}}$ instead of the true $\mathbf{H}$. Thus, we can have only the estimates of the first-stage error terms obtained using the OLS and the generalised least squares (GLS) denoted as $\hat{\mathbf{e}}$ and $\hat{\mathbf{e}}_\mathbf{g}$ respectively (see ch.8, Hill et al., 2010, more on GLS). Since it possible that $\hat{\mathbf{e}}_\mathbf{g} \neq \mathbf{e_g}$, we cannot claim that the spherical properties of the disturbances are fully restored by the above transformation, and therefore, we cannot use $E(\hat{\mathbf{e}}_\mathbf{g} \hat{\mathbf{e}}_\mathbf{g}' \mid \mathbf{z_0}) = \mathbf{0}$ as the DT condition.

Importantly, the degree of the conditional heteroscedasticity in $\mathbf{e}$ and $\mathbf{e_g}$ is expected to differ as the latter is obtained using the transformed variables. Note that $\mathbf{e_g}$ is not orthogonal to $\mathbf{s}$, instead, $\mathbf{e_g}$ is orthogonal to the span $(\mathbf{H^{-1}s})$. It follows that (p.80, Kuan, 2004)

$$\mathbf{ee'} \leq \mathbf{e_g e_g'}. \tag{9}$$

Since both error terms are obtained using the same SIV, $\mathbf{s}$, we can write the difference in the degree of the heteroscedasticity of these two estimates of the error term as follows:

$$\Delta = \mathbf{E}(\mathbf{e_g e_g'}|\mathbf{s}) - \mathbf{E}(\mathbf{ee'}|\mathbf{s}) = \mathrm{tr}[(\mathbf{H}-\mathbf{I})\mathbf{H})], \tag{10}$$

where $\mathbf{I}$ is and identity matrix of corresponding size, $\mathbf{H} = \mathbf{E}(\mathbf{ee'}|\mathbf{s})$, and $\mathrm{tr}(\cdot)$ is the trace of the matrix, which is the sum of its diagonal elements.

To estimate Δ, we need to estimate $\mathbf{H}$. More often, we have a general idea about the structure of $\mathbf{H}$ and assume that it is a function of just a few additional parameters, i.e. $\mathbf{H} = \mathbf{H}(\theta)$. In particular, using the feasible GLS (FGLS) method, we derive a consistent estimator for $\mathbf{H}$, i.e. $\hat{\mathbf{H}} = \mathbf{H}(\hat{\theta}, \mathbf{s})$ and use this specification instead of $\mathbf{H}$.

Based on our assumption A4, we assume that the conditional variance function $\mathbf{H}(\theta)$ follows the linear form below:

$$\sigma_i^2 = E(e_i^2|s_i, z_{0i}) = \alpha + \zeta\psi_i + \alpha_1 z_{0i},$$

where $(\psi = \mathbf{s} - \mathbf{z_0})$ is a deviation of an SIV from the valid IV and $\mathbf{z_0}$ is the valid IV contributing to the heteroscedasticity of the error term. The equation for the variance can be estimated as the following regression (see Hill et al., 2010, p.386 and Davidson and MacKinnon, 2009, p.265-67).

$$\hat{e}_i^2 = b + \phi s_i + v_i, \tag{11}$$

which implies the underlying relationship $\phi\mathbf{s} \equiv \zeta\psi + \alpha_1\mathbf{z_0}$, and $\hat{e}_i^2$ is the element of $\hat{\mathbf{e}}\hat{\mathbf{e}}'$ stemming from the OLS.

Using the regression for $\hat{\mathbf{e}}\hat{\mathbf{e}}'$, we state the following estimate of the conditional variances, where the hat symbol implies that the predicted conditional variances are obtained using the first-stage predicted error terms:

$$\mathbf{E}[\hat{\mathbf{e}}\hat{\mathbf{e}}' \mid \theta, \mathbf{s}] \equiv \mathbf{H}(\hat{\theta}, \mathbf{s}) = \hat{b} + \hat{\zeta}\psi + \hat{a}_1\mathbf{z_0}. \tag{12}$$

Since the FGLS approach uses the estimated conditional variance as in (12), $\hat{\mathbf{H}} \equiv \mathbf{H}(\hat{\theta}, \mathbf{s})$, it implies that $\hat{\mathbf{P}} = \hat{\mathbf{H}}^{-\mathbf{1/2}}$. Substituting this into (10), we write:

$$\hat{\Delta} = \mathrm{tr}[(\hat{\mathbf{H}} - \mathbf{I})\hat{\mathbf{H}}]. \tag{13}$$

Suppose that $\hat{\theta}$ is a consistent estimator of θ for given $\mathbf{s}$. Then, $\hat{\mathbf{H}} = \mathbf{H}(\hat{\theta}, \mathbf{s})$ implies that $\hat{\mathbf{H}}$ is asymptotically equivalent to using the true $\mathbf{H}$. Therefore, $\hat{\Delta}$ is asymptotically equivalent to Δ (see p.267, Green, 2003):

$$\operatorname*{plim}_{n\to\infty} \hat{\Delta} = \Delta \equiv \mathrm{tr}[(\mathbf{H} - \mathbf{I})\mathbf{H}]. \tag{14}$$

We denote the locus of values of $\hat{\Delta}(\delta)$ for all $\delta \in (0, \bar{\delta})$ as the following function:

$$\mathbf{D}_\Delta : \delta \in \mathbb{R}_+ \to \Delta \in \mathbb{R}. \tag{15}$$

Next, we can state a modified property of a valid SIV in the presence of heteroscedasticity.

Lemma 3.6 (Robust DT Condition) *Suppose that the assumptions A1, A2, A3 hold and vector* $\mathbf{z_0} \in \mathscr{W}(\mathbf{x}, \mathbf{y})$ *is such that* $E(\mathbf{u}|\mathbf{z_0}) = 0$ *and* $E(\mathbf{e}\mathbf{e}' \mid \mathbf{z_0}) \neq 0$ *holds, where* $\mathbf{e}$ *is the OLS estimate of the first-stage error term. The difference in the degree of the heteroscedasticity,* Δ*, is estimated by* (13) *and their locus over* $\delta \in (0, \bar{\delta})$ *is given by function* $\mathbf{D}_\Delta$ *as in* (15)*. Then, a valid SIV such that* $E(\mathbf{u} \mid \mathbf{s}^* = \mathbf{z_0}) = 0$ *is identified by* $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ *where* $\delta_0 = \arg\min_\delta(\mathbf{D}_\Delta)$.

Proof See Appendix A.8.

Since Lemma 3.6 specifies the robust condition for identification of a valid SIV, we can state the following result.

Corollary 3.7 (Robust SIV Estimator) *Suppose that an SIV given by* $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ *satisfies Lemma 3.6. Then,* β*, the parameter in* (4)*, is identified by an IV estimator:* $\hat{\beta}_{IV} = (\mathbf{x}'\mathbf{s}^*)^{-\mathbf{1}}\mathbf{x}'\mathbf{y}$.

Proof See Appendix A.9.

3.5.1 An empirical implementation of the robust DT condition

Lemma 3.6 implies that $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ is such that $\mathbf{s}^* = \mathbf{z_0}$, and thus, $E(\mathbf{u}'\mathbf{s}^*) = 0$ holds. The main point here is that although we do not observe the valid IV $\mathbf{z_0}$, we can determine the point where $\delta_0 = \arg\min_\delta(\mathbf{D}_\Delta)$ holds, which reveals us the valid SIV, $\mathbf{s}^* = \mathbf{z_0}$. To find the value of $\delta_0 = \arg\min_\delta(\mathbf{D}_\Delta)$, we need to find a way to measure the difference in the variance of the error terms on the SIVs for the OLS and FGLS estimates given by (13).

In the context of our problem, we have two random variables determined on $\delta \in (0, \hat{\delta})$: $Y = \mathbf{e_g}\mathbf{e_g}'$ and $X = \mathbf{e}\mathbf{e}'$. Note that $\Delta(\delta) = E(\mathbf{e_g}\mathbf{e_g}'|\mathbf{s}) - E(\mathbf{e}\mathbf{e}'|\mathbf{s})$ represents the true but unknown difference between the population variances of $\mathbf{e}$ and $\mathbf{e_g}$ for the given value of δ. However, we can use only the observed realizations of these variances based on the random variables $E(\hat{\mathbf{e}}_\mathbf{g}\hat{\mathbf{e}}_\mathbf{g}'|\mathbf{s})$ and $E(\hat{\mathbf{e}}\hat{\mathbf{e}}'|\mathbf{s})$. Since we need to compare the expected values of two random variables based on their sample data, analyzing the difference between the expected values of random variables naturally leads to a consideration of the difference between their distributions. This is because the behavior of random variables is fully characterized by their probability distributions, as explained below. The assertion that two random variables can be compared using their cumulative distribution functions (CDFs) is supported by established results in probability theory, which are briefly outlined below.

Assume that $X_1, X_2, \ldots$ are independent and identically distributed random variables in $\mathbb{R}$ with common cumulative distribution function $F(x)$. A sequence of random variables $X_1, X_2, \ldots$ converges in distribution to a random variable X, shown by $X_n \xrightarrow{d} X$, if $\lim_{n\to\infty} F_n(x) = F(x)$, for all x at which $F(x)$ is continuous. A continuous function is such a function that maps convergent sequences into convergent sequences: if $x_n \to x$, then $g(x_n) \to g(x)$. This statement is proved using the continuous mapping theorem (see Theorem 2.3, van der Vaart, 1998). According to the Glivenko-Cantelli theorem, the empirical distribution function, $F_n(x)$, converges uniformly to the true distribution function almost surely. That is, $\|F_n - F\|_\infty = \sup_{x\in\mathbb{R}} \left|F_n(x) - F(x)\right| \longrightarrow 0$ almost surely (p. 265, van der Vaart, 1998).

In our case, $\mathbf{E}(\mathbf{e_g e_g'}|\mathbf{s}) - \mathbf{E}(\mathbf{ee'}|\mathbf{s})$ is the true (asymptotic) difference in heteroscedasticity based on FGLS and OLS. However, we need to argue about the sample based difference asymptotic:

$$\left(E(\mathbf{\hat{e}_g \hat{e}_g'}|\mathbf{s}) - E(\mathbf{\hat{e}\hat{e}'}|\mathbf{s})\right) \xrightarrow{d} \left(\mathbf{E}(\mathbf{e_g e_g'}|\mathbf{s}) - \mathbf{E}(\mathbf{ee'}|\mathbf{s})\right).$$

According to the Glivenko-Cantelli theorem, the distributional convergence, $E(\mathbf{\hat{e}_g \hat{e}_g'}|\mathbf{S}) \xrightarrow{d} E(\mathbf{e_g e_g'}|\mathbf{s})$ holds if $\lim_{n\to\infty} G_n(E(\mathbf{\hat{e}_g \hat{e}_g'}|\mathbf{s})) = G(\mathbf{e_g e_g'}|\mathbf{s})$, for all $\mathbf{\hat{e}_g \hat{e}_g'}$. Analogously, $E(\mathbf{\hat{e}\hat{e}'}|\mathbf{s}) \xrightarrow{d} E(\mathbf{ee'}|\mathbf{s})$ holds if $\lim_{n\to\infty} F_n(E(\mathbf{\hat{e}\hat{e}'}|\mathbf{s})) = F(E(\mathbf{ee'}|\mathbf{s}))$, for all $\mathbf{\hat{e}\hat{e}'}$. Therefore, to compare the expected values of two random variables based on their samples, we need to evaluate differences between their distributions, which can arise from samples of $\mathbf{\hat{e}_g \hat{e}_g'}$ and $\mathbf{\hat{e}\hat{e}'}$. Notably, we assume that these variables may have differing distributions. Therefore, we cannot simply consider the distribution of the difference of these variables.

The statistical tool-set suggests that we can compare these two random variables using the stochastic dominance method (e.g., McFadden, 1989). Stochastic dominance (SD) is the partial order of the CDFs of random variables. Specifically, if $G(x) < F(x)$ for $x \in \mathbb{R}$, where G and F represent the distribution functions of Y and X, respectively, then a random variable X is first-order stochastically dominated by a random variable Y. It is important to note that the direction of stochastic dominance is opposite to the direction of dominance of the cumulative distribution functions (CDFs). This suggests the following relationships:

$$\begin{aligned} G(x) < F(x), \text{iff}\, P(Y \le x) < P(X \le x); \\ \text{alternatively, iff}\, P(Y > x) > P(X > x). \end{aligned} \tag{16}$$

Specifically, (16) shows that the random variable X produces observations above any fixed threshold $x \in \mathbb{R}$ more frequently than the random variable Y. Since $Y = E(\mathbf{\hat{e}_g \hat{e}_g'}|\mathbf{s})$ and $X = E(\mathbf{\hat{e}\hat{e}'}|\mathbf{s})$, cumulative distribution functions (CDFs), $G(E(\mathbf{\hat{e}_g \hat{e}_g'}|\mathbf{s}))$ and $F(E(\mathbf{\hat{e}\hat{e}'}|\mathbf{s}))$, are constructed from the sample data to approximate the true cumulative distribution functions. Based on equation (9), we anticipate that the conditional variance of FGLS residuals will be larger than that of OLS residuals. Therefore, we expect $G(E(\mathbf{\hat{e}_g \hat{e}_g'}|\mathbf{s})) < F(E(\mathbf{\hat{e}\hat{e}'}|\mathbf{s}))$.

To find the minimum difference point, we need to have a distance metric that quantifies how different two random variables (or their distributions) are. The most widespread metric for first-order stochastic dominance seems to be the the Kolmogorov-Smirnov-type statistic that may be used to test whether two underlying one-dimensional probability distributions differ (McFadden, 1989; Conover, 1998; Daniel, 1990).

The Kolmogorov-Smirnov distance is given by

$$D_{n,m} = \sup_x |G_n(x) - F_m(x)|, \tag{17}$$

where G_n and F_m represent the distribution functions of the first and second samples, with sizes n and m respectively. The term sup refers to the supremum function. We use these distances to draw conclusions about the difference between X and Y in the following manner: We find all the differences between$G(E(\hat{\mathbf{e}}_{\mathbf{g}}\hat{\mathbf{e}}'_{\mathbf{g}}|\mathbf{s}))$ and $F(E(\hat{\mathbf{e}}\hat{\mathbf{e}}'|\mathbf{s}))$ for all $\delta \in (0, \bar{\delta})$, and construct the locus: $\mathbf{D}_E = \{D(\delta = 0_+)..., D(\delta = \bar{\delta}_-)\}$, where $0_+ = \inf(\delta \in (0, \bar{\delta}))$ and $\bar{\delta}_- = \sup(\delta \in (0, \bar{\delta}))$.

Based on the above arguments, we state the following result.

Corollary 3.8 *Suppose that the assumptions A1, A2, A3 hold. Then,* $\delta_0 = \arg\min_\delta(\mathbf{D}_E) \xrightarrow{p} \arg\min_\delta(\mathbf{D}_\Delta)$.

Proof See Appendix A.10.

Thus, using $\delta_0 = \arg\min_\delta(\mathbf{D}_E)$, one can find the SIV, $\hat{\mathbf{s}}^* = \mathbf{x} + k\hat{\delta}_0\mathbf{r}$ such that $\hat{\mathbf{s}}^* \xrightarrow{\mathbf{p}} \mathbf{s}^* = \mathbf{z_0}$, where $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$, that satisfies Lemma 3.6; and thus, satisfies the validity condition: $E(\mathbf{u}'\mathbf{s}^*) = 0$. Therefore, $\operatorname*{plim}_{n\to\infty} E(\mathbf{u} \mid \hat{\mathbf{s}}^*) = 0$.

A parametric approach

As a measure of the dependency of the variance of the residuals on the SIVs, we can use the explained sum of squares of the regressions in the form as in (11) denoted by SSR. We find it as a squared sum of n observations of the residuals

$$\text{SSR} = \sum_{i=0}^{n} (\hat{e}_i - \bar{\hat{e}})^2.$$

Analogously, we find a similar measure for the estimate of the residuals stemming from FGLS:

$$\text{SSR}_g = \sum_{i=0}^{n} (\hat{e}_{gi} - \bar{\hat{e}}_g)^2.$$

To compare the OLS and FGLS cases, we scale the above measures of the degree of heteroscedasticity using the sum of squared first-stage errors, denoted as SSE. For the OLS case, we have

$$\text{SSE} = \sum_{i=0}^{n} \hat{e}_i^2,$$

and for the FGLS case, we have

$$\text{SSE}_g = \sum_{i=0}^{n} \hat{e}_{gi}^2.$$

Using these values, we can compute statistics based on the sum of squares. For the OLS case,

$$X^2 = \frac{\text{SSR}/2}{(\text{SSE}/n)^2}, \tag{18}$$

and for the FGLS case,

$$X_g^2 = \frac{\text{SSR}_g/2}{(\text{SSE}_g/n)^2}. \tag{19}$$

Since the residuals $\hat{\mathbf{e}}$ and $\hat{\mathbf{e}}_{\mathbf{g}}$ are independently distributed, by Cochran's theorem (Cochran, 1952) one can assume that $X^2 \sim \chi^2(1)$ and $X_g^2 \sim \chi^2(1)$ random variables with 1 degree of freedom (Soch et al., 2024).

Later, we will relax this assumption and also use nonparametric methods. Given that $\hat{\mathbf{e}}$ and $\hat{\mathbf{e}}_\mathbf{g}$ are both determined by the value of δ, we can compute X^2 and X_g^2 for each $\delta \in (0, \bar{\delta})$.

In our context, we compare CDFs of X^2 and X_g^2. Given that the underlying distributions of X^2 and X_g^2 follow a χ^2 distribution, we consider data series of equal sizes, where $n = m$. The cumulative distribution functions (CDFs) are defined as $G_n \equiv P[\chi^2(1) < X^2(\delta)]$ and $F_n \equiv P[\chi^2(1) < X_g^2(\delta)]$. Therefore, we compare $P[\chi^2(1) < X^2(\delta)]$ and $P[\chi^2(1) < X_g^2(\delta)]$ and determine

$$D(\delta) = P[\chi^2(1) < X^2(\delta)] - P[\chi^2(1) < X_g^2(\delta)],$$

for all $\delta \in (0, \bar{\delta})$, and construct the locus: $\mathbf{D}_E = \{D(\delta = 0_+)..., D(\delta = \bar{\delta}_-)\}$, where $0_+ = \inf(\delta \in (0, \bar{\delta}))$ and $\bar{\delta}_- = \sup(\delta \in (0, \bar{\delta}))$. Since $X \sim \chi^2_{d_1}$ and $Y \sim \chi^2_{d_2}$ are independent, then $\frac{X/d_1}{Y/d_2} \sim \mathrm{F}(d_1, d_2)$. Thus, alternatively, one can use the F-distribution and determine

$$D(\delta) = P\left[F(1,1) < \frac{X^2(\delta)}{X_g^2(\delta)}\right].$$

Then, using $\delta_0 = \arg\min_\delta(\mathbf{D}_E)$, one can find the SIV, $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$, that satisfies Lemma 3.6; and thus, satisfies the validity condition: $E(\mathbf{u}'\mathbf{s}^*) = 0$.

A nonparametric approach

We also use the empirical distribution functions instead of the χ^2 distribution to ascertain that our results are robust to deviations from normal distribution. Empirical evidence suggests that the Anderson-Darling test is often more powerful than the Kolmogorov-Smirnov test (Anderson and Darling, 1952, 1954; Stephens, 1974). So, for the empirical CDF approach, we use the Anderson-Darling test distance. The Anderson-Darling test is based on the distance proposed by Pettitt (1976) uses a weighted quadratic distance between F_n and G_m:

$$A^2_{n,m} = \frac{nm}{(n+m)^2} \sum_{k=1}^{n+m} \frac{(F_n(x_k) - G_m(x_k))^2}{H_{n+m}(x_k)(1 - H_{n+m}(x_k))}. \tag{20}$$

where H_{n+m} is the empirical CDF of $x_1, \ldots, x_{n+m}$. In the context of our case, we compute empirical CDFs for $G_n(E(\widehat{\hat{\mathbf{e}}\hat{\mathbf{e}}'|\mathbf{s}}))$ and $F_n(E(\widehat{\hat{\mathbf{e}}_\mathbf{g}\hat{\mathbf{e}}'_\mathbf{g}|\mathbf{s}}))$ by accounting that $m = n$. To calculate H_{2n}, first we concatenate $E(\widehat{\hat{\mathbf{e}}\hat{\mathbf{e}}'|\mathbf{s}})$ and $E(\widehat{\hat{\mathbf{e}}_\mathbf{g}\hat{\mathbf{e}}'_\mathbf{g}|\mathbf{s}})$ as $\mathbf{Z} = [\mathbf{E}(\widehat{\hat{\mathbf{e}}\hat{\mathbf{e}}'|\mathbf{s}})|\mathbf{E}(\widehat{\hat{\mathbf{e}}_\mathbf{g}\hat{\mathbf{e}}'_\mathbf{g}|\mathbf{s}})]$. Then, using this $\mathbf{Z}$, we find H_{2n} as the empirical CDF of the elements of $\mathbf{Z}$.

By determining the Anderson-Darling distance for all combinations of F_n and G_n calculated over the range of $\delta \in (0, \bar{\delta})$, we construct a locus of values $\mathbf{D}_E = \{D(\delta = 0_+)..., D(\delta = \bar{\delta}_-)\}$, where $D(\delta) \equiv A^2_{n,n}(\delta)$, and $0_+ = \inf(\delta \in (0, \bar{\delta}))$ and $\bar{\delta}_- = \sup(\delta \in (0, \bar{\delta}))$. Then, analogously with the parametric CDFs, we find the value of $\delta_0 = \arg\min_\delta(\mathbf{D}_E)$, which determines the SIV $\hat{\mathbf{s}}^* = \mathbf{x} + k\hat{\delta}_0\mathbf{r}$ such that $\hat{\mathbf{s}}^* \xrightarrow{\mathbf{p}} \mathbf{s}^* = \mathbf{z_0}$, and therefore, $\operatorname{plim}_{n\to\infty} E(\mathbf{u} \mid \hat{\mathbf{s}}^*) = 0$.

We can summarise the argument above with the following lemma, which defines the practical implementation of the robust SIV method.

Lemma 3.9 *Suppose that the assumptions A1, A2, A3, Corollary 3.8 hold. There is an unobservable valid IV $\mathbf{z_0}$ such that $E(\mathbf{u} \mid \mathbf{z_0}) = 0$. The locus of values of $D(\delta)$, for all $\delta \in (0, \bar{\delta})$, is given by*

$$\mathbf{D}_E = \{D(\delta = 0_+)..., D(\delta = \bar{\delta}_-)\},$$

where $0_+ = \inf(\delta \in (0,\bar{\delta}))$ *and* $\bar{\delta}_- = \sup(\delta \in (0,\bar{\delta}))$. *Then,* $\delta_0 = \arg\min_\delta(\mathbf{D}_E)$ *identifies the SIV* $\hat{\mathbf{s}}^* = \mathbf{x} + k\hat{\delta}_0\mathbf{r}$ *such that* $\hat{\mathbf{s}}^* \xrightarrow{\mathbf{p}} \mathbf{z_0}$*, and therefore,* $\operatorname*{plim}_{n\to\infty} E(\mathbf{u} \mid \hat{\mathbf{s}}^*) = 0$.

Proof See Appendix A.11.

3.6 The consistency of the SIV estimator

It is known that asymptotic identification is a necessary and sufficient condition for consistency. The parameter vector $\mathbf{b_{SIV}}$ is asymptotically identified if two asymptotic identification conditions are satisfied. The first condition is that, with parameter vector β_0 of the true DGP as a special case of the model (1),

$$\alpha(\beta_0) = \text{plim}\frac{1}{n}\mathbf{V}'(\mathbf{y} - \beta_\mathbf{0}\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} V_i' u_i = 0,$$

should hold. Here, $\mathbf{V}$ is the matrix of exogenous variables, which has the same dimension as vector $\mathbf{x}$ in the exact identification case. The second condition requires that $\alpha(\beta) \neq 0$ for all $\beta \neq \beta_\mathbf{0}$.

By showing that both of these conditions hold, we state the following lemma.

Lemma 3.10 *The SIV estimator is a consistent estimator.*

Proof See Appendix A.12

The asymptotic distribution of the SIV estimator can be straightforwardly determined, since after determining the matrix $\mathbf{V}$, the rest of our approach is just the usual IV method; thus, we can use the existing results for the IV method.

4 Application

The proposed SIV method offers a robust and flexible approach to address endogeneity in a wide range of empirical settings. In this section, we demonstrate the practical application of the SIV method through various examples, showcasing its versatility across different data types, model specifications, and research contexts. To gain statistical power, we applied re-sampling to the artificial data and we applied the basic bootstrapping sampling method to obtain the mean values of the estimates based on sample draws from the empirical dataset. This way, we can also evaluate the variability of the endogenous effect we are estimating. The outline of the SIV-based estimation procedure is given in Appendix B.

4.1 Demonstration of the SIV method using an artificial dataset

We begin by illustrating the SIV method using an artificial dataset, where the true data-generating process is known, and we ensure that the error term is correlated to the explanatory variable. This simulation exercise allows us to validate the performance of the SIV estimator and compare it with traditional methods, such as the OLS.

To generate the artificial dataset, we use the approach developed by Jones and Pewsey (2009). They suggest a transformation function $H(\tilde{\mathbf{x}};\varepsilon,\kappa) = \sinh[\kappa \sinh^{-1}(\tilde{\mathbf{x}}) - \varepsilon]$, where $\varepsilon \in \mathbb{R}$ and $\kappa \in \mathbb{R}_+$. When this transformation is applied to the normal cumulative distribution function (CDF) $S(\tilde{\mathbf{x}};\varepsilon,\kappa) = \Phi[H(\tilde{\mathbf{x}};\varepsilon,\kappa)]$, it produces a unimodal distribution whose parameters (ε,κ) control skewness and kurtosis, respectively. Thus, if $\varepsilon = 0$ and $\kappa = 1$, we obtain the original normal distribution.

We consider an example of data based on an artificial DGP. A detailed description of the procedure to construct the DGP to generate the artificial dataset is given in Appendix C. First, we generate a sequence v of N numbers. Then we use this sequence to generate the series of the endogenous variable using the transformation function by Jones and Pewsey (2009): $\tilde{\mathbf{x}} = 5 \cdot H(v, 0, 0.9) + 1$. Next, we generate the disturbance term $\mathbf{u}$ that is correlated with $\tilde{\mathbf{x}}$. We also generate an additional exogenous vector as $\mathbf{w} = rnorm(N, 0, 10)$. Then, we compute the outcome variable as follows:

$$\tilde{\mathbf{y}} = \mathbf{1} + 2\tilde{\mathbf{x}} + \mathbf{w} + \mathbf{u}. \tag{21}$$

Next, we apply the SIV method to the model based on (21) assuming that $\mathbf{u}$ is unobservable.

Table 1: Comparison of OLS, SIV, Robust SIV-parametric (RSIV-p) and Robust SIV-nonparametric (RSIV-n) estimates using simulated data.

	Dependent variable: $\tilde{\mathbf{y}}$			
Regressors	*OLS*	*SIV*	$RSIV_p$	$RSIV_n$
$\tilde{\mathbf{x}}$	2.99***	2.36***	1.94***	2.02
	(0.099)	(0.099)	(0.101)	(0.101)
$\mathbf{w}$	0.99***	0.99***	0.99***	0.99***
	(0.004)	(0.004)	(0.004)	(0.004)
Constant	-0.17	0.57***	1.06***	0.96***
	(0.148)	(0.149)	(0.150)	(0.150)

Note: *** indicates p<0.01. Standard errors are in parentheses.
$\tilde{\mathbf{x}} = 5 \cdot H(v, 0, 0.9) + 1$, where $H(v, \varepsilon, \delta) = \sinh[\delta \sinh^{-1}(v) - \varepsilon]$ and v is a sequence of 10^5 numbers in $(-15, 15)$ with a step equal to $30/10^5$. For reproducibility of the DGP, use `set.seed(1001)` in R 4.0.3; $\mathbf{w} = rnorm(N, 20, 10)$. The DGP is $\tilde{\mathbf{y}} = \mathbf{1} + \mathbf{2}\tilde{\mathbf{x}} + \mathbf{w} + \mathbf{u}$, where $\mathbf{u} = \tilde{\mathbf{x}} + \mathbf{rnorm}(\mathbf{N}, \mathbf{0}, \mathbf{sd}(\tilde{\mathbf{x}}) - \mathbf{mean}(\tilde{\mathbf{x}}) + \mathbf{7} \cdot \varepsilon)$.
The true disturbance term has $cov(\mathbf{x}, \mathbf{u}) > 0$.
For SIVs, mean values of the estimates are shown based on 100 sample draws of 5000 observations selected with replacement from the DGP data.

First, we linearly project out the exogenous variables ($\mathbf{w}$) including the constant term and use $\mathbf{y}$ and $\mathbf{x}$ determined as the residual of $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ regressed on the exogenous variables. By design of equation (21), the true value of β_1 relating $\mathbf{x}$ to $\mathbf{y}$ is equal to 2.

In Table 1, we present the mean values of the estimates based on 100 sample draws from the DGP given by (21). For the reproducibility of the samples, we used $set.seed(3 * (i))$, where i is the order number of the draw. We then compute δ_0 based on the condition of Lemma 3.3 for the simple SIV case, and Lemma 3.9 for the robust SIV cases. The robust SIV cases are estimated (i) parametrically, assuming a χ^2 distribution for the conditional variances of the first-stage disturbances, and (ii) nonparametrically, using empirical CDFs. We synthesize SIVs using $\mathbf{s}^* = \mathbf{x} - \delta_0 \mathbf{r}$, where $\mathbf{r} = (\mathbf{I} - \mathbf{P_x})\mathbf{y}$, where $\mathbf{P_x} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{\mathbf{-1}}\mathbf{x}$ is projection onto $\mathbf{x}$. In other words, we find the SIV that satisfies the DT condition for each sample and then estimate β_1 using that SIV. Based on 100 sample draws, we have obtained the following 95% confidence intervals: for $mean(\hat{\beta}_1^{SIV})$: 2.36 ± 0.19, $mean(\hat{\beta}_1^{RSIV_p})$: 1.94 ± 0.19, $mean(\hat{\beta}_1^{RSIV_n})$: 2.02 ± 0.19, and for $mean(\hat{\beta}_1^{OLS})$: 2.99 ± 0.19. The Wald test confirms that the average values of $\hat{\beta}_1^{SIV}$ and $\hat{\beta}_1^{RSIV}$ are statistically not different from the true value of $\beta_1 = 2$. On the other hand, the sample mean of the OLS estimates of β_1 is significantly different than the true value of $\beta_1 = 2.0$ confirmed by the Wald test. Overall, the evidence indicates that the SIV method yields accurate estimates for the linear regression parameter considered in the example.

4.2 Application of the SIV method to empirical data

We consider the exercise that uses the Mroz (1987) data to estimate the hours of work by married women. This study used the data on 428 working, married women to estimate the labour supply equation:

$$\mathbf{hours} = b_0 + b_1\mathbf{lwage} + b_2\mathbf{educ} + b_3\mathbf{age} + b_4\mathbf{kidslt6} + b_5\mathbf{kidsge6} + b_6\mathbf{nwifeinc} + \mathbf{u}.$$

In this labor supply model for married women, we anticipate a positive effect of the wage rate (**lwage**) on labor supplied, meaning higher wages increase labor supply. However, because the supply and demand for labor impact both the quantity supplied and the wage rate, using OLS for estimation is inappropriate due to endogeneity. Additionally, the demand for labor negatively affects the wage rate, suggesting that feedback from the demand side on labor supply is expected to be negative. This also implies that $cov(\mathbf{u}, \mathbf{lwage}) < 0$.

The DT condition, according to Corollary 3.5, confirms that $cov(\mathbf{u}, \mathbf{lwage}) < 0$ is the correct relationship between the endogenous regressor and the unobservables. That is, only when we assume $cov(\mathbf{u}, \mathbf{lwage}) < 0$, we find $\delta_0 > 0$ that satisfies $cov(\mathbf{e_0^2}, \mathbf{s_0}) = 0$, where $\mathbf{e_0}$ is the first-stage error term obtained by using the SIV $\mathbf{s_0} = x + \delta_0\mathbf{r}$. The original estimation uses the level of experience(**exper**) and its squared value (**expersq**) as the instruments for the wage rate. The results of the estimations are given in Table 2.

Table 2: Estimates of the effect of wages on work hours using Mroz (1987) data. Comparison of OLS, IV, SIV, and Robust SIV methods.

Endogenous regressor	*Dependent variable:* work hours				
	OLS	*IV*	*SIV*	$RSIV_p$	$RSIV_n$
lwage	−17.40	1,544.81***	1,183.99***	1,172.86***	1,172.86***
	(54.215)	(480.73)	(117.45)	(116.28)	(116.28)
95% CI: $\hat{\beta}_{SIV}$			1183.9±230.8	1172.8±228.6	1172.8± 228.6
Weak instruments		0	0	0	0
Wu-Hausman		0	0	0	0
Sargan		0.35			
Observations	428	428	428	428	428
Adjusted R^2	0.05	−1.81	−1.05	−1.03	−1.03
Residual Std. Error (df = 421)	755.1	1,301.9	1,105.9	1,105.9	1,105.9
F Statistic	5.04*** (df = 6; 421)				

Note: *** indicates p<0.01. Standard errors are in parentheses.
For SIVs, mean values of the estimates and CI are shown based on 50 bootstrap sample draws.
Endogenous variable: **lwage** is the log of wages; $cov(\mathbf{u}, \mathbf{lwage}) < 0$.
Conventional IVs in model (2): years of experience, **exper** and $\mathbf{exper^2}$.
SIV: The mean DT condition parameter: $\bar{\delta}_0 = 1.07$, for *RSIV-p*: $\bar{\delta}_{Rp} = 1.07$ *and RSIV-n*: $\bar{\delta}_{Rn} = 1.07$.
$\mathbf{s} = \mathbf{x} + \delta_0\mathbf{r}$, $\mathbf{r} = (\mathbf{I} - \mathbf{P_x})\mathbf{y}$, where $\mathbf{P_x} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}$ is the orthogonal projection matrix onto $\mathbf{x}$.
$\mathbf{x}$ and $\mathbf{y}$ are residuals of the regressions of **educ** and **lwage** on all the exogenous variables.
Exogenous variables: **educ**, **age**, **kidslt6**, **kidsge6**, **nwifeinc** from Mroz (1987).

To gain statistical power, we applied basic bootstrapping sampling method to obtain the mean values of the estimates based on 50 sample draws from the original dataset. This way, we can also evaluate the variability of the endogenous effect we are estimating using the SIV method. The SIV method indicates that the hours of work increase with the wage rate, albeit less than in the original IV case. The 95% CI of the effect of *lwage* is relatively narrow, indicating strong reliability of the estimates. Overall, the SIV results are in line with the traditional IV results. In addition, the robust SIV estimations are not statistically different from the simple SIV estimate, indicating that there is no problem caused by heteroscedasticity and the overall robustness of the SIV method.

4.2.1 Application example 2: The effect of Protestantism on literacy

Next, we consider Becker and Woessmann (2009) who investigate the impact of Protestantism on literacy rates across regions, addressing the potential endogeneity of religious affiliation, based on the following specification

$$\textbf{Literacy rate} = constant + \beta(\textbf{Protestant share}) + \mathbf{V}'\gamma + \mathbf{u},$$

$$\textbf{Protestant share} = constant + \pi(\textbf{Distance to Wittenberg}) + \mathbf{V}'\delta + \mathbf{v},$$

where **V** is a vector of demographic and regional controls.

Table 3: Estimates of the effect of Protestantism on literacy rates using Becker and Woessmann (2009) data. Comparison of OLS, IV, SIV, Robust SIV-parametric ($RSIV_p$) and Robust SIV-nonparametric methods ($RSIV_n$).

Endogenous regressor	*Dependent variable: Literacy rate*				
	OLS	IV	SIV	$RSIV_p$	$RSIV_n$
f_prot	0.100^{***}	0.187^{***}	0.586^{***}	0.582^{***}	0.584^{***}
	(0.010)	(0.028)	(0.064)	(0.063)	(0.064)
95% CI: $\hat{\beta}_{SIV}$			0.586±0.12	0.582±0.12	0.584± 0.12
Weak instruments		0	0	0	0
Wu-Hausman		0	0	0	0
Observations	452	452	452	452	452
Adjusted R^2	0.73	0.68	−0.74	−0.71	−0.72
Residual Std. Error (df = 437)	6.58	7.13	16.7	16.58	16.64
F Statistic	88.1^{***}				
	(df = 14; 437)				

Note: *** indicates p<0.01. Standard errors are in parentheses.
For SIVs, mean values of the estimates and CI are shown based on 50 bootstrap sample draws.
Endogenous variable: **f_prot** is the share of Protestant population in the locality; $cov(\mathbf{u}, \mathbf{f\_prot}) < 0$.
Conventional IV: The distance to Wittenberg, **kmwittenberg**.
SIV: The mean DT condition parameter: $\bar{\delta}_0 = 2.23$, for *RSIV-p*: $\bar{\delta}_{Rp} = 2.2$ *and RSIV-n*: $\bar{\delta}_{Rn} = 2.32$.
x and **y** are residuals of the regressions of **f_prot** and **f_rw** on all the exogenous variables.
$\mathbf{s} = \mathbf{x} + \delta_0\mathbf{r}$, where $\mathbf{P_x} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}$ is the orthogonal projection matrix onto **x**.
Exogenous variables: **f_jew**, **f_young**, **f_fem**, **f_ortsgeb**, **f_pruss**, **hhsize**, **lnpop**, **gpop**, **f_miss**, **f_blind**, **f_deaf**, **f_dumb**, **uni1517** from Becker and Woessmann (2009).

Becker and Woessmann (2009) exploited the distance from individual counties to Wittenberg, the city where Martin Luther initiated the Protestant Reformation, as an instrument for the Protestant share of the population during the 1870s. Their analysis suggested a negative correlation between Protestantism and unobservable factors hindering literacy. Our analysis confirms the sign assumption by Becker and Woessmann (2009) for $cor(\mathbf{x}, \mathbf{u})$. Following Corollary 3.5 and the subsequent discussion, the correct sign for $cor(\mathbf{x}, \mathbf{u})$ determined by the case where the magnitude of $E(\mathbf{e^2}, \mathbf{s})$ increases on the range $\delta \in (0, \delta_0)$, where $\delta_0 = arg_\delta(E(\mathbf{e^2}, \mathbf{s}) = 0)$.

The SIV results reported are obtained as the mean values of the estimates based on 50 sample draws from the original dataset using basic bootstrapping. The SIV method (simple and robust versions) finds that the effect of Protestantism was almost three times greater than the estimate stemming from the original IV estimate (see Table 3). The lack of statistical difference between the simple SIV and robust SIV results indicates that this example has no issue of random heteroscedasticity.

4.2.2 Application example 3: The effect of 401(k) program participation on having IRA savings account

Finally, we demonstrate the versatility of the SIV method by applying it to a binary-data model, where the outcome variable is dichotomous. Specifically, we apply the SIV method to the model on the effects of 401(k) retirement programs on the probability of having savings accounts. The data comes from Abadie (2003), and the model is presented in Wooldridge (2013, p. 539). In this case, the dependent and endogenous explanatory variables are binary. The question of this study is to find out whether participating in a 401(k) retirement program results in a decrease or an increase in the probability of having an individual retirement account (IRA). The problem here is that individuals who participate in 401(k) retirement programs may have stronger preferences for savings, so even if there were no such programs, they would have saved more than those who do not participate. Thus, using the OLS methods would result in upward-biased estimates of the effects of tax-deferred retirement programs. That is, such behaviour implies that $cov(\mathbf{p401k}, \mathbf{u}) > 0$. This assumption is confirmed by the DT condition, according to Corollary 3.5 by yielding $\delta_0 > 0$ that satisfies $cov(\mathbf{s_0}, \mathbf{e_0^2}) = 0$. The estimations indicate that when we use the conventional IV proposed by Abadie (2003), the estimate of the effect of participation in a 401(k) program on the probability of having an IRA becomes not statistically significant.

The SIV results are obtained as the mean values of the estimates based on 50 sample draws from the original dataset using basic bootstrapping. For the SIV method (see Table 4), we obtain a negative and statistically significant effect of participation in a 401(k) program on the probability of having an IRA. That is, the SIV results indicate that participating in 401(k) retirement programs largely crowds out individual retirement account (IRA) participation. In other words, the estimates of the effect of participation in a 401(k) program on the probability of having an IRA, obtained through the OLS and the TSLS using traditional instruments, are both misleading. In addition, the results of the robust SIV estimation are not significantly different from the simple SIV results, thus, we can conclude that heteroscedasticity is not a big problem in this case.

5 Conclusion

The SIV method presented in this paper addresses the challenges associated with traditional instrumental variable (IV) approaches. It eliminates the need to find valid external instruments, reduces the weak instrument problem, and avoids the complexities related to over-identification. Our innovative approach utilizes vector representation in the regression space to show that any valid instrument can be expressed as a linear combination of the vectors formed by the outcome and endogenous variables. This offers a new perspective on instrumental variables and provides a robust, data-driven alternative to conventional IV techniques.

The SIV method is built upon two key innovations. Our first step involves introducing the "dual tendency" (DT) moments condition, a necessary requirement for identifying valid synthetic instrumental variables. The condition specifies that given that the structural error term is homoscedastic, a valid SIV defined within the plane formed by the outcome and endogenous variables, must concurrently meet the criteria of orthogonality, $E(\mathbf{u} \mid \mathbf{s}) = 0$, and first-stage error term homoscedasticty, $E(\mathbf{ee'} \mid \mathbf{s}) = 0$. As a more general method, a robust DT condition can deal with cases when the structural error and first-stage terms are

Table 4: Estimates of the effect of 401(k) program participation on the probability of having an IRA savings account using Abadie (2003) data. Comparison of OLS, IV, SIV, Robust SIV-parametric (RSIV-p) and Robust SIV-nonparametric methods (RSIV-n).

Endogenous regressor	*Dependent variable:* Probability of having IRA				
	OLS	*IV*	*SIV*	*RSIV-p*	*RSIV-n*
p401k	0.051^{***}	0.017	-0.913^{***}	-0.904^{***}	-0.904^{***}
	(0.010)	(0.013)	(0.020)	(0.020)	(0.020)
95% CI: $\hat{\beta}_{SIV}$			-0.913 ± 0.04	-0.904 ± 0.04	-0.904 ± 0.04
Weak instruments		0	0	0	0
Wu-Hausman		0	0	0	0
Observations	9,275	9,275	9,275	9,275	9,275
Adjusted R^2	0.18	0.18	−0.71	−0.70	−0.70
Residual Std. Error (df = 9267)	0.39	0.39	0.56	0.57	0.56
F Statistic	302.05^{***} (df = 7; 9267)				

Note: *** indicates p<0.01. Standard errors are in parentheses.
For SIVs, mean values of the estimates and CI are shown based on 50 bootstrap sample draws.
Endogenous variable: **p401k** is the probability of participating in 401(k) retirement programs.
The dependent and endogenous variables are binary; $cov(\mathbf{u}, \mathbf{p401k}) > 0$.
Conventional IVs: Eligibility for 401(k), **e401k**.
SIV: The mean DT condition parameter: $\bar{\delta}_0 = 1.05$, for *RSIV-p*: $\bar{\delta}_{Rp} = 1.04$ *and RSIV-n*: $\bar{\delta}_{Rn} = 1.04$.
$\mathbf{s} = \mathbf{x} + \delta_0 \mathbf{r}$, $\mathbf{r} = (\mathbf{I} - \mathbf{P_x})\mathbf{y}$, where $\mathbf{P_x} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}$ is the orthogonal projection matrix onto $\mathbf{x}$.
Exogenous variables: **inc**, **incsq**, **age**, **agesq**, **marr**, **fsize** from Abadie (2003).

heteroscedastic. In this case, a valid SIV satisfies the criteria of orthogonality only when the difference between the conditional variances of first-stage error term stemming from OLS and FGLS estimations using the SIV is minimized. Overall, the DT condition provides a powerful, data-driven criterion for confirming the validity of synthetic instruments, bridging the gap between unobservable orthogonality conditions and observable data characteristics. This approach not only ensures the validity of our synthetic instruments but also provides a practical means of verification, addressing a critical challenge in instrumental variable estimation.

Furthermore, our methodology allows for the accurate identification of the correct sign of $cov(\mathbf{x}, \mathbf{u})$ from the data, which is essential for preventing flawed models and inaccurate conclusions in practical research. By exploiting geometric properties and the characteristics of the DT condition, the SIV method offers vital insights into the essence of endogeneity in models.

The robustness and efficacy of the SIV method have been rigorously demonstrated through both simulated and real-world empirical applications. In our simulation studies using artificial data with known parameters, the SIV method consistently recovered the true parameter values with high accuracy. This performance in controlled settings provides strong evidence for the method's ability to correctly identify causal effects in the presence of endogeneity. More compellingly, when applied to diverse empirical datasets, the SIV method yielded results that closely aligned with existing findings in the literature or with theoretically expected values. For instance, in our application to labor supply data, the SIV estimates of wage elasticity were consistent with previous studies while offering improved precision. Similarly, in our analysis of the effects of Protestantism on literacy rates and the impact of 401(k) programs on savings behavior, the SIV method produced estimates that were not only statistically significant but also economically meaningful and in line with theoretical predictions. These empirical validations across various domains underscore the versatility and reliability of the SIV method in real-world research contexts, suggesting its potential to become a valuable tool in the econometrician's toolkit for addressing endogeneity in a wide range of applications.

The SIV method has advantages but also limitations. The method's computational intensity in finding the optimal δ_0 may pose challenges for large datasets. Additionally, it is primarily designed for models with a single endogenous variable in cross-sectional settings, and extending it to scenarios with multiple endogenous variables, nonlinear models, or panel data structures presents theoretical and practical challenges. Recognizing these limitations highlights significant directions for further study and refinement of the SIV approach.

In conclusion, the SIV method has the potential to significantly improve causal inference across various disciplines, such as epidemiology, social sciences, and policy evaluation. It offers a robust tool for assessing causal impacts, constructing valid instruments from existing data, and determining the true sign of endogeneity. This could lead to more accurate assessments of causal relationships, better-informed decision-making, and a deeper understanding of various outcomes in social, economic, and health contexts. Furthermore, the method is applicable in newly ererging domains where causal inference is becoming more and more important, such as data science and machine learning. Its ability to take endogeneity in large-scale observational data into account could improve the interpretability and reliability of predictive models.

Appendices

A Proofs

A.1 Lemma 2.1

Proof of Lemma2.1 Let us consider a vector $\mathbf{u}$ such that $\mathbf{u} \in \mathscr{W}(\mathbf{x},\mathbf{y})$. Let $\mathscr{S}(\mathbf{z})$ be a plane that is orthogonal to vector $\mathbf{u}$. Now, take some vector $\mathbf{z} \in \mathscr{S}(\mathbf{z})$. Since any vector on $\mathscr{S}(\mathbf{z})$ is orthogonal to $\mathbf{u}$, it follows that $\mathbf{z} \perp \mathbf{u}$ and thus, $E(\mathbf{u} \mid \mathbf{z}) = \mathbf{0}$, as $\mathbf{z} \in \mathscr{S}(\mathbf{z})$. Let vector $\mathbf{z_0}$ be the orthogonal projection of $\mathbf{z}$ onto the plane spanned by $\mathbf{x}$ and $\mathbf{y}$. By definition of the orthogonal projection, we have $\mathbf{z_0} \in \mathscr{W}(\mathbf{x},\mathbf{y})$. Since the original vector $\mathbf{z} \in \mathscr{S}(\mathbf{z})$ is represented as the vector sum of $\mathbf{z_0}$ and the orthogonal component vector $\mathbf{z_v}$, we also have $\mathbf{z_0} \in \mathscr{S}(\mathbf{z})$. Thus, the vector $\mathbf{z_0}$ is defined at the intersection of $\mathscr{W}(\mathbf{x},\mathbf{y})$ and $\mathscr{S}(\mathbf{z})$. Since plane $\mathscr{S}(\mathbf{z})$ is orthogonal to $\mathbf{u}$, vector $\mathbf{z_0}$ should also satisfy the orthogonality condition; thus, $E(\mathbf{u} \mid \mathbf{z_0}) = \mathbf{0} : \mathbf{z} \in \mathscr{W} \cap \mathbf{S}$. However, if we had $E(\mathbf{u} \mid \mathbf{z_0}) \neq \mathbf{0}$, it would imply that $\mathscr{S}(\mathbf{z}) \not\perp \mathbf{u}$. This outcome would contradict the assumption that $\mathscr{S}(\mathbf{z}) \perp \mathbf{u}$. Therefore, for $\mathscr{S}(\mathbf{z}) \perp \mathbf{u}$ to hold, it must be that $E(\mathbf{u} \mid \mathbf{z_0}) = \mathbf{0}$. As $\mathbf{z}', \mathbf{z} \in \mathscr{S}(\mathbf{z}) \perp \mathbf{u}$, $E(\mathbf{u} \mid \mathbf{z_0}) = \mathbf{0}$ implies that $E(\mathbf{u} \mid \mathbf{z}) = \mathbf{0}$. ■

A.2 Proof of Lemma 2.3

Proof According to Lemma 2.2, a vector $\mathbf{z}$ located in the segment spanned by $\mathbf{x}$ and $\mathbf{r}$ can be written as $\mathbf{z} = \zeta\mathbf{x} + \omega\mathbf{r}$ where $\zeta, \omega \in \mathbb{R}$. We can determine a vector $\mathbf{z_0} \parallel \mathbf{z}$ multiplying $\mathbf{z}$ by a scalar $\frac{1}{\zeta}$, so that, $\mathbf{z_0} = \mathbf{x} + \frac{\omega}{\zeta}\mathbf{r}$. Since we consider only the vectors $\mathbf{z_0}$ that satisfy $corr(\mathbf{x},\mathbf{z_0}) > 0$, ζ is a non-zero positive parameter, whereas $\omega > 0$ if $corr(\mathbf{r},\mathbf{z_0}) > 0$ and $\omega < 0$ if $corr(\mathbf{r},\mathbf{z_0}) < 0$. Therefore, $\frac{\omega}{\zeta}$ is determined. Given that $corr(\mathbf{x},\mathbf{r}) = 0$, $corr(\mathbf{x},\mathbf{u}) > 0$ and $corr(\mathbf{y},\mathbf{r}) > 0$ imply that $corr(\mathbf{u},\mathbf{r}) > 0$. Then for $corr(\mathbf{z_0},\mathbf{u}) = 0$ to hold, $corr(\mathbf{r},\mathbf{z_0}) < 0$ must hold. By symmetry, when $corr(\mathbf{x},\mathbf{u}) < 0$, for $corr(\mathbf{z_0},\mathbf{u}) = 0$ to hold, $corr(\mathbf{r},\mathbf{z_0}) > 0$ must hold. Then, denoting $\delta \equiv |\omega|$ and $k = -sign(corr(\mathbf{x},\mathbf{u}))$, we can write: $\mathbf{z} = \mathbf{x} + k\delta\mathbf{r}$. ■

A.3 Proof of Lemma 3.1

Assume that there is vector $\mathbf{z_0}$ such that $E(\mathbf{u} \mid \mathbf{z_0}) = 0$ holds. By employing coplanar vectors $\mathbf{x}$ and $\mathbf{r}$, we can represent this vector as $\mathbf{z_0} = \mathbf{x} + k\delta_0\mathbf{r}$. This condition implies that parameter γ_0, in $\mathbf{x} = \gamma_0 \mathbf{z}_0 + \mathbf{e_0}$, is predetermined by the given coplanar vectors $\mathbf{x}$ and $\mathbf{r}$ through their SIV relationship with $\mathbf{z_0}$. For any SIV $\mathbf{s} \mid \mathbf{s} \neq \mathbf{z_0}$ determined in the same quadrant of $\mathscr{W}(\mathbf{x},\mathbf{y})$, as the true SIV $\mathbf{z_0} \in \mathscr{W}(\mathbf{x},\mathbf{y})$, and given by $\mathbf{s} = \mathbf{x} + k\delta\mathbf{r} \mid_{\delta \neq \delta_0}$, we have $\gamma \neq \gamma_0$ for the first-stage estimate. Since $\mathbf{s}$ and $\mathbf{z_0}$ are coplanar vectors, we can write an SIV $\mathbf{s}$, in the form $\mathbf{s} = \mathbf{z_0} + (\mathbf{s} - \mathbf{z_0})$. We recall that the OLS estimate is calculated by $\gamma = \frac{cov(\mathbf{x},\mathbf{s})}{var(\mathbf{s})}$. Now, we re-write $\gamma = \frac{cov(\mathbf{x},(\mathbf{z_0}+(\mathbf{s}-\mathbf{z_0})))}{var(\mathbf{z_0}+(\mathbf{s}-\mathbf{z_0}))}$. One can re-formulate this expressions as

$$\gamma = \frac{cov(\mathbf{x},\mathbf{z_0})}{var(\mathbf{z_0})} + \frac{cov(\mathbf{x},(\mathbf{z_0}+(\mathbf{s}-\mathbf{z_0})))}{var(\mathbf{z_0}+(\mathbf{s}-\mathbf{z_0}))} - \frac{cov(\mathbf{x},\mathbf{z_0})}{var(\mathbf{z_0})}. \tag{A1}$$

Given that $\gamma_0 = \frac{cov(\mathbf{x},\mathbf{z_0})}{var(\mathbf{z_0})}$, we can write (A1) as

$$\gamma = \gamma_0 + g(\mathbf{x},\mathbf{s},\mathbf{z_0}), \tag{A2}$$

where $g(\cdot) = \frac{cov(\mathbf{x},(\mathbf{z_0}+(\mathbf{s}-\mathbf{z_0})))}{var(\mathbf{z_0}+(\mathbf{s}-\mathbf{z_0}))} - \frac{cov(\mathbf{x},\mathbf{z_0})}{var(\mathbf{z_0})}$ stands for a function that captures the bias in γ caused by the deviation of $\mathbf{s}$ from the true IV $\mathbf{z_0}$. Since all these vectors are determine in a separable Hilbert space, this function is contious and twice differentiable. ■

A.4 Proof of Theorem 3.2

Assume that $\mathbf{z_0}$ is such that $E(\mathbf{u}|\mathbf{z_0}) = 0$ holds. Let us use an SIV $\mathbf{s} = \mathbf{x} + k\delta\mathbf{r}$ to instrument $\mathbf{x}$. Recall that $\mathbf{e} = \mathbf{x} - \gamma\mathbf{s}$. Then, $\mathbf{ee}' = ||\mathbf{x}||^2 + \gamma^2||\mathbf{s}||^2 - 2\gamma\langle\mathbf{x},\mathbf{s}\rangle$, where $||\cdot||$ denotes the vector norm, and $\langle\cdot,\cdot\rangle$ denotes the dot product. Let us consider the derivative given by $\frac{\partial \mathbf{ee}'}{\partial \mathbf{s}}$.

$$\frac{\partial \mathbf{ee}'}{\partial \mathbf{s}} = \frac{d}{d\mathbf{s}}\left(||\mathbf{x}||^2 + \gamma^2||\mathbf{s}||^2 - 2\gamma\langle\mathbf{x},\mathbf{s}\rangle\right) = 2\gamma\frac{\partial\gamma}{\partial\mathbf{s}}||\mathbf{s}||^2 + 2\gamma^2\mathbf{s} - 2\frac{\partial\gamma}{\partial\mathbf{s}}\langle\mathbf{x},\mathbf{s}\rangle - 2\gamma\mathbf{x} \tag{A3}$$

Here, we take into account that according to (A2) in Lemma 3.1, $\gamma = \gamma_0 + g(\mathbf{x},\mathbf{s},\mathbf{z_0})$. This implies that

$$\frac{\partial\gamma}{\partial\mathbf{s}} = \frac{\partial(\gamma_0 + g(\mathbf{x},\mathbf{s},\mathbf{z_0}))}{\partial\mathbf{s}} = \frac{\partial\gamma_0}{\partial\mathbf{s}} + \frac{\partial(g(\mathbf{x},\mathbf{s},\mathbf{z_0}))}{\partial\mathbf{s}}. \tag{A4}$$

By definition $\frac{\partial\gamma_0}{\partial\mathbf{s}} = 0$ and $g(\mathbf{x},\mathbf{s}^*,\mathbf{z_0}) = 0$, where $\mathbf{s}^* = \mathbf{z_0}$. The latter implies that $\frac{\partial(g(\mathbf{x},\mathbf{s}^*,\mathbf{z_0}))}{\partial\mathbf{s}} = 0$. Thus, when $\mathbf{s}^* = \mathbf{z_0}$ and accounting for (A2), from (A3), we have the following:

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \mathbf{s}} \mid_{\mathbf{s}^*=\mathbf{z_0}} = 2\gamma_0^2\mathbf{z_0} - 2\gamma_0\mathbf{x} = 2\gamma_0(\gamma_0\mathbf{z_0} - \mathbf{x}). \tag{A5}$$

Now, recall that by the construction of the first-stage residuals, we have $E(\gamma_0\mathbf{z_0} - \mathbf{x}) = 0$. Therefore, it follows that $E\left(\frac{\partial\mathbf{ee}'}{\partial\mathbf{s}} \mid_{\mathbf{s}^*=\mathbf{z_0}}\right) = 0$, which implies that $E(\mathbf{ee}'|\mathbf{s}^*) = 0$ holds. Since by definition $E(\mathbf{u}|\mathbf{z_0}) = 0$, this implies that $E(\mathbf{e}'\mathbf{e}|\mathbf{s}^*) = 0$ holds simultaneously with $E(\mathbf{u}|\mathbf{s}^*) = 0$.■

A.5 Proof of Lemma 3.3

Proof Given that $\mathbf{s} = \mathbf{x} + k\delta\mathbf{r}$ with $k = (-1)sign[cov(\mathbf{x},\mathbf{u})]$, we can write:

$$\mathbf{P_s} = \mathbf{s}(\mathbf{s}'\mathbf{s})^{-\mathbf{1}}\mathbf{s} = (\mathbf{x} + k\delta\mathbf{r})\left((\mathbf{x} + k\delta\mathbf{r})'(\mathbf{x} + k\delta\mathbf{r})\right)^{-1}(\mathbf{x} + k\delta\mathbf{r}).$$

Thus, we have the first-stage error term as:

$$\mathbf{e} = (\mathbf{I} - \mathbf{P_s})\mathbf{x} = [\mathbf{I} - (\mathbf{x} + k\delta\mathbf{r})\left((\mathbf{x} + k\delta\mathbf{r})'(\mathbf{x} + k\delta\mathbf{r})\right)^{-1}(\mathbf{x} + k\delta\mathbf{r})]\mathbf{x}. \tag{A6}$$

Theorem 3.2 implies that, given $\mathbf{E}(\mathbf{uu}' \mid \mathbf{z_0}) = \mathbf{0}$, the condition $\mathbf{E}(\mathbf{ee}' \mid \mathbf{s} = \mathbf{z_0}) = \mathbf{0}$ holds for a valid SIV. In the equation given by $\mathbf{E}(\mathbf{ee}' \mid \mathbf{s}^*) = \mathbf{0}$, the only unknown is parameter δ. Therefore, by solving the optimization problem $\delta_0 = \arg_\delta \mathrm{E}(\mathbf{ee}'|\mathbf{s}^*) = 0$ subject to $0 < \delta < \bar{\delta}$ to find the optimal δ_0, we determine a valid SIV $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ such that $E(\mathbf{u}'\mathbf{s}^*) = \mathbf{0}$.■

A.6 Proof of Corollary 3.4

Proof According to Lemma 3.3, the DT condition, $\delta_0 = \arg_\delta \mathrm{E}(\mathbf{ee}'|\mathbf{s}^*) = 0$, determines a valid SIV, $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ with $k = (-1)sign[cov(\mathbf{x}, \mathbf{u})]$ such that $E(\mathbf{u}'\mathbf{s}^*) = 0$. Then, β, a parameter of the model in in (4), is identified by an IV estimator: $\hat{\beta}_{IV} = (\mathbf{x}'\mathbf{s}^*)^{-\mathbf{1}}\mathbf{x}'\mathbf{y}$.

A.7 Proof of Corollary 3.5

Proof Recall that when $cor(\mathbf{x}, \mathbf{u}) \neq 0$, the first-stage homoscedasticity DT condition holds for an SIV $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ only under the condition given by $E(\mathbf{e}'\mathbf{e} \mid \mathbf{s}^* = \mathbf{z_0}) = \mathbf{0}$, where $E(\mathbf{u}|\mathbf{z_0}) = 0$. Then, if for the assumed sign of $cor(\mathbf{x}, \mathbf{u})$, the condition $E(\mathbf{e_1}\mathbf{e_1}'|\mathbf{s_1^*}) = 0$ holds, where $\mathbf{e_1} = \mathbf{x} - \gamma_1\mathbf{s_1}$, $\mathbf{s_1^*} = \mathbf{x} + k\delta_{0,1}\mathbf{r}$, and $\delta_{0,1} > 0$, thus $\mathbf{s_1^*} = \mathbf{z_0}$ and this is a valid IV. On the other hand, if under the assumed sign of $cov(\mathbf{x}, \mathbf{u})$, we obtain $E(\mathbf{e_2}'\mathbf{e_2}|\mathbf{s_2}) \neq \mathbf{0}$ for all $\delta_2 > 0$, where $\mathbf{e_2} = \mathbf{x} - \gamma_2\mathbf{s_2}$, $\mathbf{s_2} = \mathbf{x} + k\delta_2\mathbf{r}$. Then, according to Lemma 3.3, we conclude that such SIVs are not valid on the whole feasible range for δ. In this case, the assumed sign of $cov(\mathbf{x}, \mathbf{u})$ does not align with the true sign of $cov(\mathbf{x}, \mathbf{u})$, if the assuming the opposite sign yields a valid SIV. Thus, the true sign of $cov(\mathbf{x}, \mathbf{u})$ is determined by the assumed sign that yields $\delta_0 > 0$, $\mathbf{s} = \mathbf{x} + k\delta_0\mathbf{r}$ and $\mathbf{e} = \mathbf{x} - \gamma\mathbf{s}$, such that $E(\mathbf{e}'\mathbf{e}|\mathbf{s}) = \mathbf{0}$ holds. If for both types of sign assumptions yield $E(\mathbf{e_2}'\mathbf{e_2}|\mathbf{s_2}) \neq \mathbf{0}$ for all $\delta_2 > 0$, then we conclude an absence of endogeneity in the model. ■

A.8 Proof of Lemma 3.6

Proof Recall from (15) that $\mathbf{D}_\Delta : \delta \in \mathbb{R} \to \Delta \in \mathbb{R}$, which implies that $\mathbf{D}_\Delta$ maps δ to corresponding difference between the conditional variances, $\Delta = E(\mathbf{e_g}\mathbf{e_g}'|\mathbf{s}) - E(\mathbf{ee}'|\mathbf{s}) = \mathrm{tr}((\mathbf{H} - \mathbf{I})\mathbf{H}) = \mathrm{tr}(\mathbf{H^2}) - \mathrm{tr}(\mathbf{H})$,

Let us find the first-order condition of the difference function $\mathbf{D}_\Delta$ with respect to ζ and using linearity:

$$\frac{\partial \mathbf{D}_\Delta}{\partial \zeta} = \frac{d}{d\zeta}\left(\mathrm{tr}(\mathbf{H^2}) - \mathrm{tr}(\mathbf{H})\right) = 2\mathrm{tr}\left(\frac{\partial \mathbf{H}}{\partial \zeta}\mathbf{H}\right) - \mathrm{tr}\left(\frac{\partial \mathbf{H}}{\partial \zeta}\right) = \mathrm{tr}\left((2\mathbf{H} - \mathbf{I})\frac{\partial \mathbf{H}}{\partial \zeta}\right) = 0. \tag{A7}$$

Since $\mathbf{H}(\theta) = b + \zeta\psi + a_1\mathbf{z_0}$, we have $\frac{\partial \mathbf{H}}{\partial \zeta} = \psi(\delta)$, then, from (A7), it follows:

$$\mathrm{tr}\left((2\mathbf{H} - \mathbf{I})\psi\right) = 0.$$

Since, $\mathrm{tr}(2\mathbf{H} - \mathbf{I}) \neq 0$, for the latter condition hold, we must have $\psi^* = \mathbf{0}$. This implies that, at this point, we have $\mathbf{s}^* = \mathbf{z_0}$. Next, we determine the sign of the second-order condition: $\frac{\partial^2 \mathbf{D}_\Delta}{\partial \zeta_1^2}$.

$$\frac{\partial^2 \mathbf{D}_\Delta}{\partial \zeta_1^2} = \mathrm{tr}(2\mathbf{H} - \mathbf{I}) - \mathrm{tr}(2\mathbf{H}\psi).$$

Then, for $\psi^* = \mathbf{0}$, we have, $\frac{\partial^2 \mathbf{D}_\Delta}{\partial \zeta_1^2} = \mathrm{tr}(2\mathbf{H} - \mathbf{I})$. Accounting for $\mathbf{ee}' \leq \mathbf{e_g e_g'}$ implied by (9), we establish that $(\mathbf{H} - \mathbf{I}) \geq 0$, which then implies that $\mathrm{tr}(2\mathbf{H} - \mathbf{I}) > 0$. Thus, the Hessian of this optimization problem is positive definite; therefore, at the point where $\mathbf{s}^* = \mathbf{z_0}$ holds, the function of the difference in the degrees of conditional heteroscedasticity reaches its minimum. That is, $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r} = \mathbf{z_0}$ holds, where $\delta_0 = \arg\min_\delta(\mathbf{D}_\Delta)$. Since by definition $E(\mathbf{u}|\mathbf{z_0}) = 0$, the latter result implies that $\delta_0 = \arg\min_\delta(\mathbf{D}_\Delta)$ holds simultaneously with $E(\mathbf{u}|\mathbf{s}^* = \mathbf{z_0}) = 0.\blacksquare$

A.9 Proof of Corollary 3.7

Proof According to Lemma 3.6, the robust DT condition, $\delta_0 = \arg\min_\delta(\mathbf{D}_\Delta)$, determines a valid SIV, $\mathbf{s}^* = \mathbf{x} + k\delta_0\mathbf{r}$ with $k = (-1)sign[cov(\mathbf{x}, \mathbf{u})]$ such that $E(\mathbf{u}'\mathbf{s}^*) = 0$. Then, β, a parameter of the model in in (4), is identified by an IV estimator: $\hat{\beta}_{IV} = (\mathbf{x}'\mathbf{s}^*)^{-\mathbf{1}}\mathbf{x}'\mathbf{y}$.

A.10 Proof of Corollary 3.8

Proof Recall that $\mathbf{\hat{e}_g \hat{e}_g'} \xrightarrow{d} \mathbf{e_g e_g'}$ holds if $\sup_\delta \left| G_n(\mathbf{\hat{e}_g \hat{e}_g'}) - G(\mathbf{\hat{e}_g \hat{e}_g'}) \right| \xrightarrow{a.s.} 0$, for all $\mathbf{\hat{e}_g \hat{e}_g'}$ determined on $\delta \in (0, \hat{\delta})$ by Glivenko–Cantelli theorem. Analogously, $\mathbf{\hat{e}\hat{e}'} \xrightarrow{d} \mathbf{ee}'$ holds if $\sup_\delta \left| F_n(\mathbf{\hat{e}\hat{e}'}) - F(\mathbf{\hat{e}\hat{e}'}) \right| \xrightarrow{a.s.} 0$, for all $\mathbf{\hat{e}\hat{e}'}$ determined on $\delta \in (0, \hat{\delta})$. Thus, $(\mathbf{\hat{e}_g \hat{e}_g'} - \mathbf{\hat{e}\hat{e}'}) \xrightarrow{d} (\mathbf{e_g e_g'} - \mathbf{ee}')$, if $\sup_\delta \left| G(\mathbf{\hat{e}_g \hat{e}_g'}) - F(\mathbf{\hat{e}\hat{e}'}) \right| \xrightarrow{p} \sup_\delta \left| G(\mathbf{e_g e_g'}) - F(\mathbf{ee}') \right|$. This implies that $D(\delta) \xrightarrow{p} \Delta(\delta)$. Given that $\mathbf{D}_E = [D(\delta), \forall\delta \in (0, \bar{\delta})]$ and $\mathbf{D}_\Delta = [\Delta(\delta), \forall\delta \in (0, \bar{\delta})]$, we conclude that $\arg\min_\delta(\mathbf{D}_E) \xrightarrow{p} \arg\min_\delta(\mathbf{D}_\Delta)$. $\blacksquare$

A.11 Proof of Lemma 3.9

According to Lemma 3.6, for the difference between the conditional variance of the first-stage error terms, $\Delta = \mathrm{tr}((\mathbf{P}'\mathbf{P} - \mathbf{I})E(\mathbf{ee}'|\mathbf{s}))$ determined for all $\delta \in (0, \bar{\delta})$, at point $\delta_0 = \arg\min_\delta(\mathbf{D}_\Delta)$, $E(\mathbf{u} \mid \mathbf{s}^* = \mathbf{z_0}) = 0$ holds. According to Corollary 3.8, $\arg\min_\delta(\mathbf{D}_E) \xrightarrow{p} \arg\min_\delta(\mathbf{D}_\Delta)$. Therefore, $\delta_0 = \arg\min_\delta(\mathbf{D}_E)$ identifies the SIV $\mathbf{\hat{s}}^* = \mathbf{x} + k\hat{\delta}_0\mathbf{r}$ such that $\mathbf{\hat{s}}^* \xrightarrow{\mathbf{p}} \mathbf{z_0}$ holds. Since by construction $E(\mathbf{u} \mid \mathbf{z_0}) = 0$, $\mathbf{\hat{s}}^* \xrightarrow{p} \mathbf{z_0}$ implies that $\operatorname*{plim}_{n\to\infty} E(\mathbf{u} \mid \mathbf{\hat{s}}^*) = 0.\blacksquare$

A.12 Proof of Lemma 3.10

Proof Let us denote by $\mathbf{V_0} = \mathbf{V}|\mathbf{s}^*$ the matrix of exogenous variables that includes the true IV $\mathbf{s}^*$ determined as an SIV that satisfies the DT condition, and denote by $\mathbf{X} = \mathbf{V}|\mathbf{x}$ the matrix of the regressors. A standard assumption for the IV estimator to be consistent is $\operatorname*{plim}_{n\to\infty} n^{-1}\mathbf{V_0'u} = \mathbf{0}$. That is, the error terms are asymptotically uncorrelated with the instruments. We can express the SIV estimator as

$$\begin{aligned} \mathbf{b_{SIV}} &= (\mathbf{V_0'x})^{-\mathbf{1}}\mathbf{V_0'X}\beta_\mathbf{0} + (\mathbf{V_0'X})^{-\mathbf{1}}\mathbf{V_0'u} \\ &= \beta_0 + (n^{-1}\mathbf{V_0'X})^{-\mathbf{1}}n^{-1}\mathbf{V_0'u}. \end{aligned} \tag{A8}$$

Since the $\operatorname*{plim}_{n\to\infty} n^{-1}(\mathbf{V_0'X})^{-\mathbf{1}}$ is deterministic and nonsingular by assumption, then $\mathbf{b_{SIV}}$ satisfies the first asymptotic identification condition because $\operatorname*{plim}_{n\to\infty} n^{-1}\mathbf{V_0'u} = \mathbf{0}$ holds due to Lemma 3.9.

Next, let us consider the second condition for $\beta \neq \beta_0$. For the true DGP, we have

$$\mathbf{y} = \mathbf{X}\beta_\mathbf{0} + \mathbf{u}. \tag{A9}$$

We transform (A9) to

$$\mathbf{y}-\beta\mathbf{x}=\mathbf{X}\beta_{\mathbf{0}}+\mathbf{u}-\mathbf{X}\beta=\mathbf{u}+\mathbf{X}(\beta_{\mathbf{0}}-\beta). \tag{A10}$$

Using(A10), we write

$$\begin{aligned}\alpha(\beta)&=\operatorname*{plim}_{n\to\infty} n^{-1}\mathbf{V}_{\mathbf{0}}'(\mathbf{y}-\beta\mathbf{x})\\&=\operatorname*{plim}_{n\to\infty} n^{-1}\mathbf{V}_{\mathbf{0}}'(\mathbf{u}+\mathbf{X}(\beta_{\mathbf{0}}-\beta)).\end{aligned}$$

Since $\operatorname*{plim}_{n\to\infty} n^{-1}\mathbf{V}_{\mathbf{0}}'\mathbf{u}=\mathbf{0}$, we have

$$\alpha(\beta)=\operatorname*{plim}_{n\to\infty} n^{-1}\mathbf{V}_{\mathbf{0}}'\mathbf{X}(\beta_{\mathbf{0}}-\beta).$$

It is known that $\operatorname*{plim}_{n\to\infty} n^{-1}(\mathbf{V}_{\mathbf{0}}'\mathbf{X})^{-\mathbf{1}}$can be assumed as deterministic and nonsingular, thus, the probability limit $\alpha(\beta)\neq 0$ as soon as $\beta\neq\beta_{\mathbf{0}}$.The second asymptotic identification condition is satisfied; therefore, the SIV estimator is consistent. ■

B Outline of the practical implementation of the SIV method

The following procedures are applied to either a dataset or a bootstrap sample. In the case of bootstrapping, it is essential to save the values of $\hat{\beta}$ and δ_0 obtained for each sample. The means of the sample $\hat{\beta}$ can be reported directly as the SIV estimates. Additionally, one can use the mean of the sample δ_0 to determine the SIV using this estimate. Subsequently, the IV estimation method can be applied using the obtained SIV as the instrumental variable.

B.1 Step-by-step guide to applying the SIV method

1. Define **y** and **x** as residuals from a projection onto the space spanned by $\tilde{\mathbf{V}}$, that is:

$$\mathbf{y}=(\mathbf{I}-\mathbf{P}\tilde{\mathbf{V}})\tilde{\mathbf{y}} \text{ and } \mathbf{x}=(\mathbf{I}-\mathbf{P}\tilde{\mathbf{V}})\tilde{\mathbf{x}},$$

where $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ denote the original vectors for the outcome variable and the endogenous regressor, **I** is the identity matrix, and $\mathbf{P}_{\tilde{\mathbf{V}}}$ is the projection matrix onto the vector space spanned by the matrix of exogenous regressors $\tilde{\mathbf{V}}$.

2. Choose a vector $\mathbf{r}\perp\mathbf{x}$ in plane $\mathscr{W}$ spanned by **x** and **y**. In practice, one can use the residual of the regression $\mathbf{x}=\beta\mathbf{y}+\varepsilon$

3. Assume the sign of $cov(\mathbf{x},\mathbf{u})$.

4. Assume the starting value for the scalar parameter δ to be a small positive number, e.g., 0.001.

5. Construct $\mathbf{s}=\mathbf{x}+k\delta\mathbf{r}$, where $k=(-1)\cdot\text{sign}(cov(\mathbf{x},\mathbf{u}))$.

6. Compute the residuals of the first-stage regression: $\mathbf{e}=\mathbf{x}-\gamma\mathbf{s}$.

7. Iterate steps (5) and (6) for $\delta\in(0<\delta<\bar{\delta})$, where $\bar{\delta}$ is determined by the correlation between **s** and **x**. The rule of thumb is that when $cor(\mathbf{x},\mathbf{s})$ is too small then **s** becomes a weak IV, so we should not be considering those SIVs.

8. Determine $\delta_0=\arg_\delta\{\text{E}[\mathbf{s}'(\mathbf{e}\mathbf{e}')]=0\}$ subject to $0<\delta<\bar{\delta}$ to find the optimal δ_0.

9. Compute $\mathbf{s}_0 = \mathbf{x} + k\delta_0\mathbf{r}$.

10. Check the alternative path by assuming a sign for $\text{cov}(\mathbf{x},\mathbf{u})$ opposite to the originally assumed sign. Repeat steps (5) to (9) with the new sign assumption.

11. Follow Corollary 3.5 and determine the true sign of $\text{cov}(\mathbf{x},\mathbf{u})$ based on the results obtained from both sign assumptions.

12. Use $\mathbf{s}_0$ found for the true sign as the instrumental variable in the two-stage least squares (2SLS) estimation procedure to obtain consistent estimates of the causal effect of **x** on **y**.

B.2 Step-by-step guide to the heteroscedasticity robust (parametric) SIV method

We assume that the true sign of $cor(\mathbf{x},\mathbf{u})$ is determined using the simple approach above.

1. Steps 1-4 of the above procedure

2. Construct $\mathbf{s} = \mathbf{x} + k\delta\mathbf{r}$, where $k = (-1)\cdot\text{sign}(\text{cov}(\mathbf{x},\mathbf{u}))$.

3. Compute the residuals of the first-stage regression: $\mathbf{e} = \mathbf{x} - \gamma_{OLS}\mathbf{s}$ and $\mathbf{e}_g = \mathbf{x} - \gamma_{FGLS}\mathbf{s}$.

4. Estimate predicted values of regressions $\mathbf{e^2} = \mathbf{s} + \varepsilon$ and $\mathbf{e_g^2} = \mathbf{s} + \varepsilon_\mathbf{g}$.

5. Compute $X^2 = \dfrac{\text{SSR}/2}{(\text{SSE}/n)^2}$ for OLS and $X_g^2 = \dfrac{\text{SSR}_g/2}{(\text{SSE}_g/n)^2}$, for FGLS case, where $\text{SSR} = \sum_{i=0}^{n}(\hat{\hat{e}}_i - \bar{\hat{e}})^2$ and $\text{SSR}_g = \sum_{i=0}^{n}(\hat{\hat{e}}_{gi} - \bar{\hat{e}}_g)^2$.

6. Determine $D(\delta) = P(\chi^2(1) < X^2(\delta)) - P(\chi^2(1) < X_g^2(\delta))$, for all $\delta \in (0,\bar{\delta})$, and construct the locus: $\mathbf{D} = \{D(\delta = 0)...,D(\delta = \bar{\delta})\}$.

7. Determine $\delta_0 = \arg\min_\delta(\mathbf{D})$.

8. Compute $\mathbf{s}_0 = \mathbf{x} + k\delta_0\mathbf{r}$.

9. Use $\mathbf{s}_0$ as the instrumental variable in the two-stage least squares (2SLS) estimation procedure to obtain consistent estimates of the causal effect of **x** on **y**.

C The artificial data generation procedure

First, we generate sequence v of N numbers in $(-15,15)$ with a step equal to $30/N$. Then we use this sequence to generate the series of the endogenous variable $\tilde{\mathbf{x}} = 5\cdot H(v,0,0.9) + 1$. Next, we generate the disturbance term **u** that is correlated with $\tilde{\mathbf{x}}$. For that purpose, generate an initial distribution by summing a uniform distribution and a positively skewed normal distribution:

$$\mathbf{u_1} = runif(N, min = -1, max = 1) + rsnorm(N, mean = 1, sd = 5, \xi = 1.2),$$

for $set.seed(5001)$ using R. Here, $runif(\cdot)$ generates a uniformly distributed random variable and $rsnorm(\mu,\sigma,\xi)$ generates a random variable with a skewed normal distribution, where $\xi = 1.2$ is the parameter of the skewness. We also generate an additional exogenous vector as $\mathbf{w} = rnorm(N,0,10)$. The disturbance term has two parts given as $\mathbf{u} = \mathbf{e} + \mathbf{v}$, where **e** is correlated with $\tilde{\mathbf{x}}$ and **v** is not correlated with $\tilde{\mathbf{x}}$. Thus, we generate a part of the disturbance term independent of $\tilde{\mathbf{x}}$ and $\mathbf{w}$ by regressing $\mathbf{u_1}$ against $\tilde{\mathbf{x}}$ and $\mathbf{w}$ and finding the residual $\mathbf{v} = \mathbf{u_1} - \zeta_0 - \zeta_1\tilde{\mathbf{x}} - \zeta_2\mathbf{w}$, and then, normalize it as $\mathbf{v} = \mathbf{v} \cdot mean(\tilde{\mathbf{x}})/2$. We generate the correlated part of the disturbance term as $\mathbf{e} = \tilde{\mathbf{x}} - mean(\tilde{\mathbf{x}}) + rnorm(N, 0, sd(\tilde{\mathbf{x}}))$. The final disturbance term is generated as

$$\mathbf{u} = q[\tilde{\mathbf{x}} - mean(\tilde{\mathbf{x}}) + rnorm(N, 0, sd(\tilde{\mathbf{x}}))] + \mathbf{v}, \tag{C1}$$

where $q \equiv sign[cor(\tilde{\mathbf{x}}, \mathbf{u})] > 0$.

D Sources of the data sets

1. Mroz. dta: T.A. Mroz (1987), The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, *Econometrica* 55, 765–799.

`http://www.cengage.com/aise/economics/wooldridge_3e_datasets/`

2. ipehd_qje2009_master.dta: Becker and Woessmann (2009), Was Weber Wrong? A Human Capital Theory of Protestant Economic History, *Quarterly Journal of Economics* 124 (2): 531–596.

`https://www.ifo.de/sites/default/files/ipehd_qje2009_data_tables.zip`

3. X401ksubs.dta: Introductory Econometrics: A Modern Approach, Fifth Edition, Jeffrey M. Wooldridge. Source: Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models, *Journal of Econometrics* 113(2), 231–263.

`http://www.cengage.com/aise/economics/wooldridge_3e_datasets/`

E A sample code to implement the SIV method

```
rm(list=ls()) #Removes all items in Environment!
##Libraries
library(stargazer)
library(dplyr)
library(lmtest)
library(haven)
library(zoo)
library(sandwich)
library(rMR)
library(ivreg)
##Functions used
# Two-sample Anderson-Darling statistic
ad2_stat <- function(x, y) {
  # Sample sizes
  n <- length(x)
  m <- length(y)

  # Pooled sample and pooled ecdf
  z <- c(x, y)
  z <- z[-which.max(z)] # Exclude the largest point
  H <- rank(z) / (n + m)

  # Statistic computation via ecdf()
  (n * m / (n + m)^2) * sum((ecdf(x)(z) - ecdf(y)(z))^2 / ((1 - H) * H))

}
### The function to test absolute non-decreasing trend and sign change for locus E(ee'siv)
check_sign_change <- function(x) {
  # Step 2: Check if there is a sign change
  signs <- sign(x)  # Get sign (-1, 0, or 1)
  sign_changes <- any(diff(signs) != 0, na.rm = TRUE)  # Detect if sign changes

  # Step 3: Return 1 if both conditions are met, otherwise 0
  return(as.integer(sign_changes))
}

check_initial_abs_increase <- function(x) {
  # Step 1: Check if the absolute values start by increasing (non-decreasing trend)
  abs_x <- abs(x)
  initial_increasing <- all(diff(abs_x[1:min(20, length(abs_x))]) >= 0)  # Check first 3 points or as many as available

  # Step 3: Return 1 if both conditions are met, otherwise 0
  return(as.integer(initial_increasing))
}
find_first_sign_change <- function(x) {
  sign_changes <- which(diff(sign(x)) != 0)  # Find indices where sign changes
  if (length(sign_changes) > 0) {
    return(sign_changes[1] + 1)  # Return the first occurrence (adjust for diff)
  } else {
    return(NA)  # Return NA if no sign change
  }
}

####### DATA
library(wooldridge)
#
mydata<-wooldridge::mroz
mydata<-mydata[complete.cases(mydata), ]
attach(mydata)

############Setting the regression variables. You need to input your variables here
H0 <- data.frame(hours, lwage,educ, age,kidslt6, kidsge6, nwifeinc)
# Vector of variables used in the regression. Firts var is outcome, the second var is endogenous regressor
IV <- data.frame(exper,expersq) ####IV variables
Y <- as.character(colnames(H0))[1] ###OUTCOME variable
X <-as.character(colnames(H0))[2] ###ENDOEGENOUS variable
H<- as.character(colnames(H0))[-(1:2)] ##### EXOGENOUS variables
#print(H)
formula_str <- paste(paste0(Y," ~ ", X,"+"), paste(H, collapse = " + "))## Construct formula as a string
formula <- as.formula(formula_str)# Convert to formula object
#print(formula)
```

```
iv_str <- as.character(colnames(IV))# Construct the external IVs as a string
ivs <- paste(paste(iv_str, collapse = " + "))# convert into part of the instruments object
instruments_str <- paste0(" ~ ",ivs, " + ", paste(H, collapse = " + "))# Construct instruments as a string
instruments <- as.formula(instruments_str)# Convert to instrument object
# determine the number of rows
N<-nrow(mydata)
#####Traditional methods
### OLS estimation
ols1 <- lm(formula, data = mydata)
summ.ols1 <- summary(ols1, vcov. = function(x) vcovHC(x, type="HC1"),
                      diagnostics=T)
##IV regression
iv1<-ivreg(formula, instruments, data=mydata)
summ.iv1 <- summary(iv1,      diagnostics=T)

#####SIV Method

###Basic vectors for  SIV calculation###
y1<-hours ### the outcome variabe
## Factoring out the effects of other exogenous variables
formula_str <- paste(paste0(Y," ~ "), paste(H, collapse = " + "))## Construct formula as a string
formula <- as.formula(formula_str)# Convert to formula object
fity<-lm(formula, data=mydata)
y<-resid(fity)
x1<-lwage### the endogenous variable
## Factoring out the effects of other exogenous variables
formula_str <- paste(paste0(X," ~ "), paste(H, collapse = " + "))## Construct formula as a string
formula <- as.formula(formula_str)# Convert to formula object
fitx<-lm(formula, data=mydata)
x<-resid(fitx)
#saving the transformed x and y
mydata$x<-(x-mean(x))
mydata$y<-(y-mean(y))
y0<-y
x0<-x
V=0
### Generating a vector orthogonal to x
fity<-lm(y0~(x0), data=mydata)
V<-resid(fity)
V<-(V-mean(V))/sd(V)
V<-V*sd(x0)
mydata$R<-V
############ Determining the sign of cor(x,u)
k=0
j=1
signc=matrix(ncol = 5, nrow = 2)
signc[1,1] <-1
signc[2,1] <--1
# ininc=matrix(ncol = 2, nrow = 2)
# ininc[1,1] <-1
# ininc[2,1] <--1

for (j in 1:2) {
  if(j<2){k=1}else{k=-1}#the assumed sign for cor(x,u)

  # IV regression
  data <- mydata
  dd<-3#end value for delta
  d<-0.01# starting value for delta
  delt<-0.01# step to change delta
  i<-1 ### starting value for a counter
  ## placeholders for variables
  m1=0
  m2=0
  while (d<dd){# we compute m1 until siv is close to become perpendicular to x
    data$siv<-(data$x-k*d*data$R) ### SIV
    ####OLS estimates
    rls<-(lm(data$x~data$siv, data=data))
    s.rls<-summary(rls,vcov. = function(x) vcovHC(x, type="HC1"), diagnostics=T)
    data$ev21<-resid(s.rls)
    m1[i]<-(cov(data$ev21^2,data$siv))
    m2[i] <- (cor(data$ev21^2,data$siv))
    d<-d+delt
    i<-i+1
  }
  par(mfrow = c(1, 2))
  plot(m1[0:300])
  plot(m2[0:300])

  m=m1
  signc[j,3] <- check_initial_abs_increase(m)#result
  signc[j,5] <- check_initial_abs_increase(m2)
  if(signc[j,3]!=1){
    index <- find_first_sign_change(m1)
    m=m1[index:i]
  }else{m=m1}
  signc[j,2] <-check_sign_change(m)
  signc[j,3] <- check_initial_abs_increase(m)#result
  signc[j,4] <- check_sign_change(m2)
}
ch=0
for(j in 1:2){cat("the assumed sign for cor(x,u):", signc[j,1], if(signc[j,3]*signc[j,4]==0)
{"is FALSE"
}else{"is TRUE"}, "\n")
  ch[j] <- signc[j,2]+signc[j,3]+signc[j,4]+signc[j,5]
}

k <- signc[which.max(ch),1]# the assumed sign for cor(x,u)
if(k!=0){### this loop works if there is endogeneity so that k is not zero
############ Initial settings for SIV
vvar=0
```

```
d0i=0
d0ri=0
d0rni=0
b2=0
b2r=0
b2rp=0
b2t=0
N<-nrow(mydata)
reps=2### need to change for larger bootsrap
S=round(N*.999)
l=1
fitc <- matrix(ncol = 1, nrow = reps)
sumb2=matrix(ncol = 1, nrow = reps)
fitcr <- matrix(ncol = 1, nrow = reps)
sumb2r=matrix(ncol = 1, nrow = reps)
fitcrn <- matrix(ncol = 1, nrow = reps)
sumb2rn=matrix(ncol = 1, nrow = reps)
fitct <- matrix(ncol = 1, nrow = reps)
sumb2t=matrix(ncol = 1, nrow = reps)
lowbp=0
upbp=0

##### Bootstrap sampling loop. You may use data <- mydata instead of data <- mydata[sample(1:N, S), TRUE]
#if you just want see how it works for the original sample data.
while (l<reps){
  # IV regression
  set.seed(3*(l))   # a different seed for each sub-sample
  data <- mydata[sample(1:N, S), TRUE]
  ####Computation of ml

  dd<-3#end value for delta
  d<-0.01# starting value for delta
  delt<-0.01# step to change delta
  i<-1 ### starting value for a counter
  ## placeholders for variables
  ml=0
  st=0
  ev22=0
  dv=0
  dv2=0
  x4=0
  l1=0
  l2=0
  while (d<dd){# we compute ml until siv is close to become perpendicular to x
    data$siv<-(data$x-k*d*data$R) ### SIV
    ####OLS estimates
    rls<-(lm(x~siv, data=data))
    s.rls<-summary(rls,vcov. = function(x) vcovHC(x, type="HC1"), diagnostics=T)
    data$ev21<-resid(s.rls)
    ### FGLS estimates
    ehatsq <- resid(rls)^2
    sighatsq.ols  <- lm(log(ehatsq)~siv, data=data)
    data$vari <- sqrt(exp(fitted(sighatsq.ols)))
    vvar[i] <- var(data$vari)
    fgls <-lm(x~siv, weights=1/vari, data=data)
    data$ev22<-resid(fgls)
    ##  ## Homoscedastic estimate for a simple SIV case
    m1[i]<-(cor(data$ev21^2,data$siv))
    ### Parametric computations for heteroscedastic case
    n=length(data$ev21)
    l1 <- summary(lm((ev21^2)~siv,data=data))
    l2 <-summary( lm((ev22^2)~siv,data=data))
    ssr1 <- sumsq(predict(lm((ev21^2)~siv,data=data))-mean(data$ev21^2))
    sse1=sumsq(data$ev21)
    x1= (ssr1/2)/(sse1/n^2)^2
    ssr2 <- sumsq(predict(lm((ev22^2)~siv,data=data))-mean(data$ev22^2))
    sse2=sumsq(data$ev22)
    x2= (ssr2/2)/(sse2/n^2)^2
    dv[i] <-pchisq(x2, df =1,lower.tail=FALSE)-pchisq(x1, df =1,lower.tail=FALSE)#
    x3 <- x1/x2#sumsq(predict(lm((ev22^2)~siv,data=data))/predict(lm((ev21^2)~siv,data=data))/predict(lm((ev21^2)~siv,data=data)))
    dv2[i] <- pf(x3, df = 1, df2 = 1, lower.tail = TRUE)
    # Non-parametric CDF computations for heterscedastic case x3 <- x1/x2#sumsq(predict(lm((ev22^2)~siv,data=data))/predict(lm((ev21^2)~siv,data=data))/predict(lm((ev21^2)~siv,data=data)))
      dv2[i] <- pf(x3, df = 1, df2 = 1, lower.tail = TRUE)
    samp1 <- (predict(lm((ev21^2)~siv,data=data)))^2
    samp2 <- (predict(lm((ev22^2)~siv,data=data)))^2
    xx0 <- samp1
    yy0 <- samp2
    ad0 <- ad2_stat(x = xx0, y = yy0)
    x4[i] <- 1-ad0
    st[i]<-d
    d<-d+delt
    i<-i+1
  }
  ### updating the formula for regressions
  formula_str <- paste(paste0(Y," ~ ", X,"+"), paste(H, collapse = " + "))## Construct formula as a string
  formula <- as.formula(formula_str)# Convert to formula object
  ### update with your own instruments: instruments<-~ siv+all exogenous variables
  iv_str <- paste(paste0(" ~ ", "siv","+"), paste(H, collapse = " + "))## Construct formula as a string
  instruments <- as.formula(iv_str)# Convert to formula object
  #instruments<-~siv+educ+ age+kidslt6+ kidsge6+ nwifeinc
  # formula <-hours~lwage+educ+ age+kidslt6+ kidsge6+ nwifeinc
  #### DT condition of homoscedatic case
  d0 <- (which.min(abs(m1)))*delt
  d0i[1] <- d0
  data$siv<-(data$x-k*d0*data$R)
  iv2<-ivreg(formula, instruments, data=data)
  summ.iv2 <- summary(iv2, diagnostics=T)#, vcov. = function(x) vcovHC(x, type="HC1"), diagnostics=T)
  ## saving the estimation parameters for each sample
  fitc[1] <- iv2$coefficients[2]
  sumb2[l] <-  summ.iv2$coefficients[2,2]

  ### DT point for heteroscedastic case- parametric approach
  d0r <-  which.min(dv2)*delt
  d0ri[l] <- d0r
  data$siv<-(data$x-k*d0r*data$R)
  iv3<-ivreg(formula, instruments, data=data)
  summ.iv3 <- summary(iv3, diagnostics=T)#,  vcov. = function(x) vcovHC(x, type="HC1"), diagnostics=T)
  ## saving the estimation paramters for each sample
  fitcr[l] <- iv3$coefficients[2]
  sumb2r[l] <-  summ.iv3$coefficients[2,2]

  #### DT point for heteroscedastic case- non-parametric approach
  d0rn <- which.min(x4)*delt
  d0rni[l] <- d0rn
  data$siv<-(data$x-k*d0rn*data$R)
  iv4<-ivreg(formula, instruments, data=data)
  summ.iv4 <- summary(iv4, diagnostics=T)#,
  ## saving the estimation paramters for each sample
  fitcrn[l] <- iv4$coefficients[2]
  sumb2rn[l] <-  summ.iv4$coefficients[2,2]
  l <- l+1
}

## Distribution of sample paramters
# the simple homogenous case
fitc <- fitc[complete.cases(fitc)]
sumb2<- sumb2[complete.cases(sumb2)]
alpha <- 0.05 # chosen significance level
b2[1] <- mean(fitc)# the endogenous parameter beta
df <- df.residual(iv2)# degrees of freedom for t-stat
seb2 <-mean(sumb2)# the average standard error of the parameter beta
tc <- qt(1-alpha/2, df) ## t-statistic
lowbp[1] <- b2[1]-tc*seb2  # lower bound for beta
upbp[1] <- b2[1]+tc*seb2   # upper bound for beta

#### The parametric heterogenous case
fitcr <- fitcr[complete.cases(fitcr)]
sumb2r<- sumb2r[complete.cases(sumb2r)]
alpha <- 0.05 # chosen significance level
b2[2] <- mean(fitcr)
df <- df.residual(iv3)
seb2r <-mean(sumb2r)
tc <- qt(1-alpha/2, df)
lowbp[2] <- b2[2]-tc*seb2r  # lower bound
upbp[2] <- b2[2]+tc*seb2r   # upper bound

#################The non-parametric heterogenous case
fitcrn <- fitcrn[complete.cases(fitcrn)]
sumb2rn<- sumb2rn[complete.cases(sumb2rn)]
alpha <- 0.05 # chosen significance level
b2[3] <- mean(fitcrn)
```

```
df <- df.residual(iv4)
seb2rn <-mean(sumb2rn)
tc <- qt(1-alpha/2, df)
lowbp[3] <- b2[3]-tc*seb2rn  # lower bound
upbp[3] <- b2[3]+tc*seb2rn   # upper bound

### Table for CI of beta
mv<-data.frame(lowbp,b2,upbp)
colnames(mv)<-c("low beta","mean b2", "high beta")
rownames(mv)<- c("SIV","SIVRr","SIVRn")#, "nearc4")
###
(mv)

### final satge estimations
###### Simple homogenous assumption case
d0i <-  d0i[complete.cases(d0i)]
d0m <- mean(d0i)
mydata$siv<-(mydata$x-k*d0m*mydata$R)
iv2<-ivreg(formula, instruments, data=mydata)
summ.iv2 <- summary(iv2, diagnostics=T)#
############Paramteric heterogenous case
d0ri <-  d0ri[complete.cases(d0ri)]
d0rm <- mean(d0ri)
mydata$siv<-(mydata$x-k*d0rm*mydata$R)
iv3<-ivreg(formula, instruments, data=mydata)
summ.iv3 <- summary(iv2, diagnostics=T)#

#### Non-parameteric heterogenous case
d0rni <-  d0rni[complete.cases(d0rni)]
d0rnm <- mean(d0rni)
mydata$siv<-(mydata$x-k*d0rnm*mydata$R)
iv4<-ivreg(formula, instruments, data=mydata)
summ.iv4 <- summary(iv2, diagnostics=T)#
}else{
  print("NO endogeneity problem. All SIV estimates are the same as the OLS")
  d0m=0.001
mydata$siv<-(mydata$x-k*d0m*mydata$R)
iv2<-ivreg(formula, instruments, data=mydata)
summ.iv2 <- summary(iv2, diagnostics=T)#
  d0rm=0.001
  mydata$siv<-(mydata$x-k*d0rm*mydata$R)
  iv3<-ivreg(formula, instruments, data=mydata)
  summ.iv3 <- summary(iv2, diagnostics=T)#

  d0rnm=0.001
  mydata$siv<-(mydata$x-k*d0rm*mydata$R)
  iv4<-ivreg(formula, instruments, data=mydata)
  summ.iv4 <- summary(iv4, diagnostics=T)#
  }
# The estimation output
stargazer(ols1, iv1, iv2, iv3, iv4,  # Include iv4 in the list of models
          type = "text",
          omit = "reg",
          dep.var.caption = "Work hours",
          dep.var.labels.include = FALSE,
          model.numbers = FALSE,
          model.names = FALSE,
          column.labels = c("OLS", "IV", "SIV", "RSIV-p", "RSIV-n"),  # Add a label for iv4
          no.space = TRUE,
          add.lines = list(
            c("Weak instruments", "",
              round(summ.iv1$diagnostics[1, "p-value"], 2),
              round(summ.iv2$diagnostics[1, "p-value"], 2),
              round(summ.iv3$diagnostics[1, "p-value"], 2),
              round(summ.iv4$diagnostics[1, "p-value"], 2)),  # Add p-value for iv4
            c("Wu-Hausman", "",
              round(summ.iv1$diagnostics[2, "p-value"], 2),
              round(summ.iv2$diagnostics[2, "p-value"], 2),
              round(summ.iv3$diagnostics[2, "p-value"], 2),
              round(summ.iv4$diagnostics[2, "p-value"], 2)),  # Add p-value for iv4
            c("Sargan", "",
              round(summ.iv1$diagnostics[3, "p-value"], 2),
              round(summ.iv2$diagnostics[3, "p-value"], 2),
              round(summ.iv3$diagnostics[3, "p-value"], 3),
              round(summ.iv4$diagnostics[3, "p-value"], 3))  # Add p-value for iv4
          ),
          multicolumn = FALSE)
###End of the main code
```

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*(2), 231 – 263.

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature 59*(2), 391–425.

Adão, R., M. Kolesár, and E. Morales (2019, 08). Shift-share designs: Theory and inference. *The Quarterly Journal of Economics 134*(4), 1949–2010.

Anderson, T. W. and D. A. Darling (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics 23*(2), 193 – 212.

Anderson, T. W. and D. A. Darling (1954). A test of goodness of fit. *Journal of the American Statistical Association 49*(268), 765–769.

Angrist, J. D. and A. B. Krueger (2001, December). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives 15*(4), 69–85.

Arellano, M. (2003). Endogeneity and instruments in nonparametric models. Comments to papers by Darolles, Florens & Renault; and Blundell & Powell. In L. H. M. Dewatripont and S. Turnovsky (Eds.), *Advances in Economics and Econometrics, Theory and Applications*, Volume 2. Cambridge University Press, Cambridge.

Autor, D. H., D. Dorn, and G. H. Hanson (2013). The china syndrome: Local labor market impacts of import competition in the united states. *American Economic Review 103*(6), 2121–2168.

Bartik, T. J. (1991). *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

Becker, S. O. and L. Woessmann (2009). Was Weber Wrong? A Human Capital Theory of Protestant Economic History. *The Quarterly Journal of Economics 124*(2), 531–596.

Blanchard, O. J. and L. F. Katz (1992). Regional evolutions. *Brookings Papers on Economic Activity 1*, 1–75.

Borusyak, K. and P. Hull (2023). Nonrandom exposure to exogenous shocks. *Econometrica 91*(6), 2155–2185.

Borusyak, K., P. Hull, and X. Jaravel (2021, 06). Quasi-experimental shift-share research designs. *The Review of Economic Studies 89*(1), 181–213.

Borusyak, K., P. Hull, and X. Jaravel (2024, 01). Design-based identification with formula instruments: a review. *The Econometrics Journal 28*(1), 83–108.

Bound, J., D. A. Jaeger, and R. M. Baker (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*(430), 443–450.

Butler, R. J. (2016). The simple geometry of correlated regressors and iv corrections. *International Journal of Statistics in Medical Research 5*, 182–188.

Card, D. (2009). Immigration and inequality. *American Economic Review: Papers & Proceedings 99*(2), 1–21.

Chernozhukov, V. and C. Hansen (2008). The reduced form: A simple approach to inference with weak instruments. *Economics Letters 100*(1), 68 – 71.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics 23*(3), 315–345.

Conover, W. J. (1998). *Practical Nonparametric Statistics* (3rd ed.). Hoboken, NJ: Wiley.

Daniel, W. (1990). *Applied Nonparametric Statistics*. Duxbury advanced series in statistics and decision sciences. PWS-KENT Pub.

Davidson, R. and J. G. MacKinnon (2009). *Econometric Theory and Methods* (Second ed.). Oxford University Press, New York, USA.

Diamond, R. (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980–2000. *American Economic Review 106*(3), 479–524.

DiTraglia, F. J. and C. García-Jimeno (2021). A framework for eliciting, incorporating, and disciplining identification beliefs in linear models. *Journal of Business & Economic Statistics 39*(4), 1038–1053.

Ebbes, P., M. Wedel, U. Böckenholt, et al. (2005). Solving and testing for regressor-error (in)dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics 3*, 365–392.

Erickson, T. and T. M. Whited (2002). Two-step GMM estimation of the errors-in-variables model using high-order moments. *Econometric Theory 18*, 776–799.

Gallo, J. L. and A. Páez (2013). Using synthetic variables in instrumental variable estimation of spatial series models. *Environment and Planning A: Economy and Space 45*(9), 2227–2242.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020, August). Bartik instruments: What, when, why, and how. *American Economic Review 110*(8), 2586–2624.

Green, W. H. (2003). *Econometric Analysis*. Pearson Education International.

Greenstone, M., A. Mas, and H.-L. Nguyen (2020). Do credit market shocks affect the real economy? quasiexperimental evidence from the great recession and 'normal' economic times. *American Economic Journal: Economic Policy 12*(1), 200–225.

Haschka, R. E. (2022). Handling endogenous regressors using copulas: A generalization to linear panel models with fixed effects and correlated regressors. *Journal of Marketing Research 59*(4), 860–881.

Haschka, R. E. (2024). Endogeneity in stochastic frontier models with 'wrong' skewness: Copula approach without external instruments. *Statistical Methods and Applications 33*, 807–826.

Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *The Journal of Economic Perspectives 15*(4), 57–67.

Heckman, J. and R. Pinto (2024). Econometric causality: The central role of thought experiments. *Journal of Econometrics 243*(1), 105719.

Hill, R. C., W. E. Griffiths, and G. C. Lim (2010). *Principles of Econometrics* (3 ed.). Wiley.

Hummels, D., R. Jorgensen, J. Munch, and C. Xiang (2014). The wage effects of offshoring: Evidence from danish matched worker–firm data. *American Economic Review 104*(6), 1597–1629.

Imbens, G. W. (2024). Causal inference in the social sciences. *Annual Review of Statistics and Its Application 11*(Volume 11, 2024), 123–152.

Jaravel, X. (2019). The unequal gains from product innovations: Evidence from the us retail sector. *Quarterly Journal of Economics 134*(2), 715–783.

Jones, M. C. and A. Pewsey (2009, 10). Sinh-arcsinh distributions. *Biometrika 96*(4), 761–780.

Klein, R. and F. Vella (2010). Estimating a class of triangular simultaneous equations models without exclusion restrictions. *Journal of Econometrics 154*(2), 154 – 164.

Kuan, C.-M. (2004). *Statistics: Concepts and Methods* (2nd ed.). Taipei: Huatai Publisher.

Lewbel, A. (1997). Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D. *Econometrica 65*(5), 1201–1213.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics 30*(1), 67–80.

McFadden, D. (1989). Testing for stochastic dominance. In T. B. Fomby and T. K. Seo (Eds.), *Studies in the Economics of Uncertainty: In Honor of Josef Hadar*. New York, Berlin, London, and Tokyo: Springer.

Moon, H. R. and F. Schorfheide (2009). Estimation with overidentifying inequality moment conditions. *Journal of Econometrics 153*(2), 136–154.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica 55*(4), 765–799.

Nakamura, E. and J. Steinsson (2014). Fiscal stimulus in a monetary union: Evidence from us regions. *American Economic Review 104*(3), 753–792.

Oberfield, E. and D. Raval (2021). Micro data and macro technology. *Econometrica 89*(2), 703–732.

Park, S. and S. Gupta (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science 31*(4), 567–586.

Peri, G., K. Shih, and C. Sparber (2016). Stem workers, h-1b visas, and productivity in us cities. *Journal of Labor Economics 49*(3), 277–307.

Pettitt, A. N. (1976). A two-sample anderson-darling rank statistic. *Biometrika 63*(1), 161–168.

Rigobon, R. (2003). Identification through heteroskedasticity. *The Review of Economics and Statistics 85*(4), 777–792.

Saiz, A. (2010). The geographic determinants of housing supply. *Quarterly Journal of Economics 125*(3), 1253–1296.

Soch, J., T. B. of Statistical Proofs, Maja, P. Monticone, T. J. Faulkenberry, A. Kipnis, K. Petrykowski, C. Allefeld, H. Atze, A. Knapp, C. D. McInerney, Lo4ding00, and amvosk (2024, jan). Statproofbook/statproofbook.github.io: Statproofbook 2023.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association 69*, 730–737.

Stock, J. H., J. H. Wright, and M. Yogo (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics 20*(4), 518–529.

Tang, D., D. Kong, and L. Wang (2024). The synthetic instrument: From sparse association to sparse causation.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

Vives-i Bastida, J. and A. Gulek (2023, May 10). Synthetic IV estimation in panels. Available at SSRN:https://ssrn.com/abstract=4716511.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5 ed.). South-Western Cengage Learning.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources 50*(2), 420–445.

Xu, C. (2022). Reshaping global trade: The immediate and long-run effects of bank failures. *Quarterly Journal of Economics 137*(4), 2107–2161.