ESTIMATION IN AN INSTRUMENTAL VARIABLES MODEL WITH TREATMENT EFFECT HETEROGENEITY

JOB MARKET PAPER

MICHAL KOLESÁR*

DEPARTMENT OF ECONOMICS

HARVARD UNIVERSITY

VERSION 1.2.5, NOVEMBER 13, 2012†

Abstract

This paper analyzes estimators based on the classic linear instrumental variables model when the treatment effects are in fact heterogeneous, as in Imbens and Angrist (1994). I divide these estimators into two classes: two-step instrumental variables (TSIV) estimators that include the two-stage least squares (TSLS) estimator; and minimum distance estimators that include the limited information maximum likelihood (LIML) estimator. I show that if the local average treatment effects vary, estimators in the TSIV class typically all estimate the same convex combination of them. In contrast, estimands of minimum distance estimators may be outside of the convex hull of the local average treatment effects, and may therefore not correspond to a causal effect. This result questions the standard recommendation to use LIML when the number of instruments is large as a way of addressing the bias exhibited by TSLS in these settings. Instead, I propose a new TSIV estimator, a version of the jackknife instrumental variables estimator (UJIVE). Unlike TSLS or LIML, UJIVE is consistent for a convex combination of local average treatment effects under many instrument asymptotics that also allow for many covariates and heteroscedasticity. I therefore recommend that in settings with many instruments researchers use UJIVE, instead of TSLS or LIML.

Keywords: Instrumental Variables, Local Average Treatment Effects, Limited Information Maximum Likelihood, Jackknife.

*Electronic correspondence: mkolesar@fas.harvard.edu, http://www.people.fas.harvard.edu/~mkolesar/. I am deeply grateful to Gary Chamberlain and Guido Imbens for their constant guidance and support. I also thank Alberto Abadie, Nikhil Agarwal, Josh Angrist, Raj Chetty, Paul Goldsmith-Pinkham, Adam Guren, Pepe Montiel, Whitney Newey, Jim Stock, Fanyin Zheng, and participants at the Harvard Econometrics and Labor Workshops for helpful comments and suggestions.

†Version hash 3ec3c240, compiled November 27, 2012
1 Introduction

The classic linear instrumental variables model is commonly used to estimate treatment effects. When the individual treatment effect is independent of treatment status and covariates, estimators based on this model estimate the population average treatment effect (Heckman, 1997). However, since this assumption rules out selection into treatment based on anticipated gains from treatment, it is not very plausible in many empirical settings. It is therefore important to understand the properties of these estimators when the individual treatment effect is allowed to be correlated with treatment status.

The first contribution of this paper is to characterize the estimands of estimators based on the classic linear instrumental variables (iv) model when the treatment effects are unrestricted. I assume that the instruments satisfy the monotonicity condition of Imbens and Angrist (1994), so that for each pair of instrument values, we can identify a local average treatment effect (LATE). I show that the two-stage least squares (TSLS) estimator, under some mild assumptions about the first stage, estimates a convex combination of these local average treatment effects, weighted over different pairs of instrument values and covariates. On the other hand, unless all LATEs are the same, the estimand of the limited information maximum likelihood (LIML, Anderson and Rubin, 1949) depends on the covariance matrix of the reduced-form errors, and may lie outside the convex hull of the local average treatment effects. Therefore, the estimand may not correspond to a causal effect. Moreover, other estimators based on the classic linear IV model will, depending on how they are constructed, either estimate the same convex combination of LATEs as TSLS, or else behave similarly to LIML.

In particular, estimators that behave like TSLS can be thought of as two-step estimators. In the first step, they construct a single instrument, a predictor of the treatment status based on the first-stage regression. In the second step, an instrumental variables estimator that uses this constructed instrument as a single instrument is used to estimate the treatment effect. I refer to these estimators as two-step instrumental variables estimators. In the limit under standard asymptotics, the exact way of constructing the single instrument does not matter; all two-step IV estimators converge to the same probability limit as the infeasible instrumental variables estimator that uses a population linear predictor of the treatment status as a single instrument. In turn, the probability limit of this IV estimator corresponds to a weighted average of LATEs. The weights are non-negative if the single instrument itself satisfies monotonicity in that changing its value does not induce two-way flows in and out of treatment.

In contrast, estimators that behave like LIML are based on the property of the classic linear IV model that the coefficients on the instruments in the first-stage regression are proportional to the coefficients in the reduced-form outcome regression. These estimators, which I refer to as minimum distance estimators, minimize a minimum distance objective function that directly enforces this proportionality with respect to some weight matrix. The estimator of the treatment
effect is given by the estimator of the constant of proportionality. Goldberger and Olkin (1971) show that LI ML can be thought of in this way, with the weight matrix depending on the covariance matrix of the reduced-form errors.

This approach yields a different estimand under treatment effect heterogeneity because imposing proportionality of the reduced-form coefficients implies that the treatment and the outcome are treated symmetrically. In particular, it requires that the estimand of the reverse two-stage least squares estimator (RT SLS) be equal to the estimand of TS LS. The RTS LS estimator is obtained as the reciprocal of the TLS S estimator in the instrumental variables model that swaps the treatment and the outcome. This requirement makes sense if the instrumental variables model is supposed to solve an errors-in-variables problem (Zellner, 1970), or an omitted variable bias (Chamberlain, 2007). However, in the context of estimating treatment effects, the reduced-form coefficients are no longer proportional to each other unless all L ATE S are equal. Therefore, the TLS S and RTS LS estimands are in general different; the probability limit of the RTS LS estimator is the same as that of an instrumental variables estimator that uses a linear predictor of the outcome based on the reduced-form outcome regression as an instrument. This instrument induces a different weighting scheme for the L ATE S, and hence a different estimand, than using a linear predictor of the treatment status as an instrument. Unlike the TLS S weights, these weights are proportional to the effect size, with the bigger L ATE S receiving more weight.

There are two ways in which this difference between TLS S and RTS LS estimands can cause a minimum distance estimand to be outside the convex hull of L ATE S. First, if some L ATE S are negative, the RTS LS estimand gives them a negative weight, so that the estimand may end up being outside the convex hull of the L ATE S. Consequently, the minimum distance estimand, trying to equate RTS LS with TLS S, may end up being outside the convex hull. Second, even if the RTS LS estimand is inside the convex hull, if the weight matrix that is used to equate RTS LS with TLS S is non-diagonal, as is the case with LI ML, the minimum distance estimand is not guaranteed to lie between the RTS LS and TLS S estimands.

In settings with a few strong instruments, it is easy to avoid these problems by simply avoiding LI ML and using TLS S. However, when many instruments are used, TLS S may be severely biased even in large samples (Bound, Jaeger and Baker, 1995), and it is inconsistent under the many instrument asymptotic sequence of Kunitomo (1980), Morimune (1983), and Bekker (1994). Therefore, when the number of instruments is large, the standard recommendation has been to use LI ML, which is not only consistent under many instrument asymptotics, but also efficient among rotation invariant estimators and homoscedasticity (Chioda and Jansson, 2009; Anderson, Kunitomo and Matsushita, 2010). Recently, other estimators have been proposed that behave better than LI ML under heteroscedasticity. Hausman, Newey, Woutersen, Chao and Swanson (2012) propose a Fuller (1977) type modification to a jackknife version of LI ML (HL IM). Bekker and Crudu (2012) propose a similar estimator, which they call symmetric jackknife. However, all of these estimators are minimum distance estimators, and therefore not likely to work well under
treatment effect heterogeneity.

The second contribution of this paper is to propose a new estimator in the two-step IV class, the unbiased jackknife IV estimator (ujive), that remains consistent for a convex combination of lates even under many instrument asymptotics and heteroscedasticity. This estimator is similar to the jackknife instrumental variables estimator (jive, Phillips and Hale, 1977; Angrist, Imbens and Krueger, 1999) in that it also uses a “leave-one-out” jackknife-type predictor of the treatment in the first stage, but differs from jive in the way it deals with covariates. In particular, in constructing the single instrument in the two-step IV procedure, we need to partial out the effect of covariates on the treatment. Suppose, for example, that the instruments are classroom indicators, and the covariates are school indicators (school “fixed effects”). Then the jive estimate of the effect of covariates on the treatment status of individual i is given by an average treatment status of individuals in the same school as individual i. With a finite number of observations in each school, this estimate is noisy, and since it depends on the treatment status of individual i, the estimation error is correlated with the outcome. Therefore, the single constructed instrument is also correlated with the outcome, causing jive to be biased when the number of covariates is large (Ackerberg and Devereux, 2009). In contrast, the ujive estimate of the effect of the covariates is given by a sample average that excludes individual i, which guarantees that the prediction error will be uncorrelated with the outcome. As a result, unlike jive, ujive is consistent for a convex combination of lates even when we let the number of covariates, in addition to the number of instruments, increase in proportion to the sample size, as in Anatolyev (2011) and Kolesár, Chetty, Friedman, Glaeser and Imbens (2011).

The estimand of two-step IV estimators can be seen as one way of summarizing the effect of the treatment on outcome. For particular policy questions, however, we might be interested in a weighting scheme that is different than the one used by these estimators. For this purpose, a number of alternative approaches, not based on the classic IV model, have been proposed in the literature. For example, Frölich (2007) derives a non-parametric estimator for the largest subpopulation of compliers for which a treatment effect can be identified. When the instrument is binary, Abadie (2003) works out a semi-parametric approach to approximating a treatment response function, and Hirano, Imbens, Rubin and Zhou (2000) and Yau and Little (2001) use a parametric approach to estimate a lATE that does not condition on covariates. To keep the paper focused, I do not try to compare the classic IV estimators with these alternative approaches.

The rest of the paper is organized as follows. In Section 2, I set up the problem of estimating causal effects in a potential outcomes framework. In Section 3, I review assumptions underlying the classic linear IV model, and I introduce the classes of two-step IV and minimum distance estimators that are based on this model. In Section 4, I introduce the local average treatment effects framework of Imbens and Angrist (1994). In Section 5, I derive the first main result of the paper, the estimands of two-step IV and minimum distance estimators under the lATE assumptions. In Section 6, I derive the second main result of the paper that ujive is consistent for
a convex combination of late asymptotics. Section 7 concludes. Proofs are collected in the Appendix.

2 Potential outcomes framework

We want to learn about the causal effect of a treatment on some outcome of interest using a random sample of \( n \) individuals indexed by \( i = 1, \ldots, n \). For clarity of exposition, I focus on the case when the treatment is binary. Let \( T_i \) be an indicator for receiving treatment, so that \( T_i = 1 \) if individual \( i \) receives treatment and zero otherwise. Let \( Y_i(1) \) and \( Y_i(0) \) denote the potential outcomes in the treated and untreated states. The treatment effect for individual \( i \) is then given by \( \tau_i = Y_i(1) - Y_i(0) \).

The fundamental problem is that for each individual, we only observe the potential outcome corresponding to the observed treatment state, \( Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) \); the other potential outcome is not observed. Therefore, we cannot compute \( \tau_i \) directly for any individual. Moreover, there is a concern that anticipated potential outcomes affect selection into treatment, so that comparing the average outcome of the subsample of individuals who are treated in our sample with those who are not is likely to lead to a biased estimate of the population average treatment effect \( E[\tau_i] \).

We do, however, observe instruments \( Q_i \) with support \( Q \) that help to identify average treatment effects for at least some subpopulations. Following the notation in Imbens and Angrist (1994), for each possible realization \( q \in Q \), let \( T_i(q) \) denote the potential treatment variable that equals one if individual \( i \) would receive treatment if their instrument value was changed to \( Q_i = q \), and equals zero if they would not receive treatment. The observed treatment status is given by \( T_i = T_i(Q_i) \); the other potential treatments are not observed.

We also observe a vector of covariates \( X_i \) with support \( X \). I include these covariates explicitly for two related reasons. First, in many empirical applications the identification assumptions that underlie the instrumental variables framework may only be plausible conditional on \( X_i \). One simple approach in this case is to carry out the analysis separately for all values of the covariates. However, when the covariate set is detailed so that the support \( X \) is rich, this approach is unlikely to be satisfactory. Second, even when the identification assumptions are plausible unconditionally, inference without covariates might not be precise enough. A common solution to both of these problems in practice is to estimate a single model with covariates. It is therefore important to understand how the presence of covariates affects inference.

In summary, the observed data vector for each individual is given by \( (Y_i, T_i, Q_i, X_i) \). For simplicity I will assume that the support of the observed data vector is given by the Cartesian product \( \mathbb{R} \times \{0,1\} \times Q \times X \). This will ensure that we can freely manipulate \( T_i \) and \( Q_i \) while keeping \( X_i \) constant, so that the set of potential outcomes and treatments \( \{Y_i(t), T_i(q)\}_{t \in \{0,1\}, q \in Q} \) is well-defined.
Two important functions of the distribution of the observed data are given by the two regression functions

\[ \begin{align*}
    r(q, x) &= \mathbb{E}[Y_i \mid Q_i = q, X_i = x], \\
    p(q, x) &= \mathbb{E}[T_i \mid Q_i = q, X_i = x].
\end{align*} \] (1)

Since the treatment is binary, \( p(q, x) \) equals the conditional treatment probability, \( \mathbb{P}(T_i = 1 \mid Q_i = q, X_i = x) \), also known as the propensity score. When viewed as a random variable, I will denote it by \( p_i = p(Q_i, X_i) \). Similarly, \( r(q, x) \) denotes the conditional expectation of the outcome, and I denote it by \( R_i = r(Q_i, X_i) \) when viewed as a random variable. Without further assumptions, these regression function are not directly informative about the objects of interest—the treatment effects. They are therefore known as the reduced form equations.

In general, both reduced form equations will be non-linear. The linear IV estimators that I will consider are based on a linear approximation to the true non-linear reduced form

\[ \begin{align*}
    R_i^L &= \mathbb{E}^*[Y_i \mid Z_i, W_i] = Z_i^\prime \pi_1 + W_i^\prime \psi_1, \\
    P_i^L &= \mathbb{E}^*[T_i \mid Z_i, W_i] = Z_i^\prime \pi_2 + W_i^\prime \psi_2,
\end{align*} \] (2)

where \( \mathbb{E}^* \) denotes population (minimum mean-squared-error) linear projection\(^1\), and \( Z_i = z(Q_i, X_i) \) and \( W_i = w(X_i) \) are expansions of the original instruments and covariates, with \( \dim(Z_i) = K \) and \( \dim(W_i) = L \). I assume that \( W_i \) spans a column of ones. The estimators that I will consider will use these constructed instruments and covariates.

For example, in Angrist and Krueger (1991), the basic instruments \( Q_i \) were three quarter of birth indicators, and the constructed instruments \( Z_i \) were obtained by interacting \( Q_i \) with year of birth and state of birth indicators. A similar specification was used in Dobbie and Fryer (2011), who study the effect of Harlem Children Zone (hcz) charters on educational outcomes. In particular, Dobbie and Fryer (2011) construct \( Z_i \) by interacting an indicator for living within hcz, \( Q_i \), with cohort, so that \( Z_{i,\ell} = Q_i I_{X_i=\ell} \), where \( \ell \) indexes cohorts, \( \ell \in \{1, \ldots, L\} \). If we also set \( W_{i,\ell} = I_{X_i=\ell} \), and cohort is the only covariate that we observe, then the linear approximation is exact, and \( P_i = P_i^L \). With continuous instruments and covariates, we could use series expansions to construct \( Z_i \) and \( W_i \). Of course, we can also simply set \( z(Q_i, X_i) = Q_i \) and \( w(X_i) = X_i \). I make the distinction between the original instruments and covariates, \( (Q_i, X_i) \), and the constructed ones, \( (Z_i, W_i) \), because it will matter for the estimands of these estimators under treatment effect heterogeneity how exactly the instruments were constructed.

I use matrix notation to help keep the definitions and results compact. I denote the \( n \)-component vector with \( i \)-th element \( Y_i \) by \( Y \). Similarly, let \( T, W, Z, P, P^L, R \) and \( R^L \) denote vectors

\(^1\)In other words, the linear projection of \( A_i \) onto \( B_i \), \( \mathbb{E}^*[A_i \mid B_i] = B_i^\prime \gamma \), minimizes \( \min_{\gamma} \mathbb{E}[(A_i - B_i^\prime \gamma)^2] \). If the covariance matrix of \( B_i \) is non-singular so that \( \mathbb{E}[B_i B_i^\prime] \) is invertible, then the solution is uniquely given by \( \gamma = \mathbb{E}[B_i B_i^\prime]^{-1} \mathbb{E}[B_i A_i] \).
and matrices with rows $T_i$, $W_i$, $Z_i$, $P_i$, $P_i^L$, $R_i$ and $R_i^L$. For any full-rank $n \times m$ matrix $A$, let $H_A = A(A' A)^{-1} A'$ denote the associated $n \times n$ projection matrix (also known as the hat matrix), and let $D_A$ be an $n \times n$ diagonal matrix with $(H_A)_{ii}$ on the diagonal. Let $I_m$ denote the $m \times m$ identity matrix, and let $M_A = I_n - H_A$ denote the annihilator matrix. Let $A_{\perp} = M_W A$ denote the residual from the sample projection of $A$ onto $W$, and let $\bar{A}_i = A_i - E^*[A_i \mid W_i]$ denote the residual from the population projection of $A_i$ onto $W_i$. Also, let a.s. denote almost surely (i.e. with probability one).

3 Classic linear IV model and estimators

The classic linear iv model is usually defined in terms of a structural equation (see, for example Wooldridge, 2002, Chapter 5)

$$Y_i = W_i' \delta + T_i \beta + \epsilon_i,$$

where the covariates $W_i$ and the instruments $Z_i$ are assumed to be uncorrelated with the structural error $\epsilon_i$:

$$E[\epsilon_i W_i] = 0, \quad E[\epsilon_i Z_i] = 0. \quad (6)$$

The second assumption is that the instruments are relevant in the sense that the coefficient $\pi_2$ in Equation (4) is non-zero. The parameter of interest is $\beta$, and it represents the causal effect of $T_i$ on $Y_i$. Equations (5)–(6) can be compactly written as a moment condition

$$E^*[Y_i - W_i' \delta - T_i \beta \mid Z_i, W_i] = 0. \quad (7)$$

In this section, I use the potential outcomes framework to formulate assumptions that deliver the moment condition (7) and that give $\beta$ a direct causal interpretation as the population average treatment effect, $E[\tau_i]$. This allows me to more easily link the classic iv model to the LATE framework of Imbens and Angrist (1994). Second, I define two classes of estimators of $\beta$: the class of two-step iv estimators (that includes the two-stage least squares estimator) and the class of minimum-distance estimators (that includes LIML). This classification will be more useful when considering the behaviour of the classic iv estimators under the LATE framework than the traditional division into estimators that fit into the k-class (Nagar, 1959; Theil, 1961, 1971), and estimators that do not.

3.A Assumptions underlying the classic linear IV model

Interpreting $\beta$ in the moment condition (7) as the population average treatment effect requires three assumptions that correspond to Assumptions IV, CTE and L below. First, the instruments need to be valid in the sense that they only affect potential outcomes through their effect on the
treatment. Second, we need to restrict treatment effect heterogeneity. Third, we need to make some functional form assumptions.

In order to state formally what properties valid instruments should have, we need to include $Q$ in the definition of potential outcomes. Let $Y_i(t, q)$ be the potential outcome when individual $i$ receives treatment $t$ and instrument $q$, so that the observed outcome is given by $Y_i = Y_i(T_i, Q_i)$.

**Assumption IV**

(i) (Random assignment) \{\(Y_i(t, q), T_i(q)\)\}_{t \in \{0,1\}, q \in Q} \perp \perp X_i;

(ii) (Exclusion restriction) \(P(Y_i(t, q) = Y_i(t, q') | X_i) = 1\) for all \((t, q, q')\) a.s.; and

(iii) (Relevance) The distribution of $P^L_i$ conditional on $X_i$ is non-degenerate with positive probability.

Part (i) requires that conditional on covariates, the instruments are as good as randomly assigned in the sense that they are independent of potential outcomes and potential treatments. Part (ii) requires that the instruments only affect outcomes through their effect on the treatment. This assumption justifies writing the potential outcomes as functions of the treatment only, so that $Y_i(t) = Y_i(t, q)$.

Finally, Part (iii) is a rank condition—requires that the constructed instruments $Z_i$ have a non-zero effect on the treatment, at least for some values of covariates; it ensures that the coefficient $\pi_2$ in Equation (4) is non-zero. Since $Z_i = z(Q_i, X_i)$, a necessary condition is that the original instruments $Q_i$ have a non-zero effect on the treatment.

**Assumption CTE (Constant Average Treatment Effects)** For all \((t, q, x) \in \{0, 1\} \times Q \times X\),

\[
\mathbb{E}[\tau_i | Q_i = q, T_i = t, X_i = x] = \beta.
\]

Although Assumption CTE allows the individual treatment effects $\tau_i$ to vary, it requires that the source of heterogeneity in the individual treatment effects be unrelated to observables. In particular, it does not allow individuals’ treatment status to be correlated with gains from treatment, ruling out what Heckman, Urzua and Vytlacil (2006) call essential heterogeneity, or sorting on gains from treatment. This makes it implausible in many empirical applications—I will relax it in the next section when I introduce the late framework of Imbens and Angrist (1994).

By the Law of iterated expectations, the parameter $\beta$ corresponds to the average treatment effect $\mathbb{E}[\tau_i]$. Some textbook discussions of the classic linear IV model (Wooldridge, 2002; Angrist and Pischke, 2009) use a stronger version of this assumption by imposing $\tau_i = \beta$ for all $i$, ruling out any heterogeneity in the treatment effect, but such restrictive assumption is not needed.

Assumption CTE implies that

\[
0 = \mathbb{E}[Y_i(1) - Y_i(0) - \beta | Q_i = q, T_i = t, X_i = x]
= \mathbb{E}[T_i (Y_i(1) - Y_i(0) - \beta) | Q_i = q, T_i = t, X_i = x]
= \mathbb{E}[Y_i - Y_i(0) - T_i \beta | Q_i = q, X_i = x],
\]

(8)
where the last line follows from \( Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i) \) and the Law of iterated expectations.

To turn Equation (8) into the moment condition (7), we need that

\[
\mathbb{E}^*[Y_i(0) | Z_i, W_i] = \mathbb{E}^*[Y_i(0) | W_i].
\] (9)

If there are no covariates beyond the intercept, so that \( W_i = 1 \), then this equality holds automatically. However, since Assumption IV allows for cases in which the assignment of instrument is only random conditional on covariates, it only implies that \( \mathbb{E}[Y_i(0) | Z_i, X_i] = \mathbb{E}[Y_i(0) | X_i] \).

If there are no covariates beyond the intercept, so that \( W_i = 1 \), then this equality holds automatically. However, since Assumption IV allows for cases in which the assignment of instrument is only random conditional on covariates, it only implies that \( \mathbb{E}[Y_i(0) | Z_i, X_i] = \mathbb{E}[Y_i(0) | X_i] \). If the conditional expectation \( \mathbb{E}[Y_i(0) | X_i] \) is not linear in \( W_i \), then controlling for \( W_i \) in a linear way does not fully control for the effect of the covariates on \( Y_i(0) \). Consequently, \( \tilde{Z}_i = Z_i - \mathbb{E}^*[Z_i | W_i] \) (part of \( Z_i \) orthogonal to \( W_i \)) may be correlated with \( Y_i(0) - \mathbb{E}^*[Y_i(0) | W_i] \), and the coefficient on \( Z_i \) on the left-hand side of (9) may be non-zero. Therefore, some textbook discussions (Wooldridge, 2002; Angrist and Pischke, 2009) make the assumption that \( \mathbb{E}[Y_i(0) | X_i] = W_i'\delta \), so that controlling for \( W_i \) in a linear way controls fully for the effect of the covariates on \( Y_i(0) \). Unfortunately, this assumption has the undesirable implication that, in principle, sufficient variation in the covariates alone is enough to identify \( \beta \) since non-linear functions of \( W_i \), such as squares of \( W_i \), are valid instruments. Moreover, since it involves potential, rather than observed outcomes, it is not directly testable.

Here I focus on the other way we can ensure that \( \tilde{Z}_i \) is not correlated with \( Y_i(0) - \mathbb{E}^*[Y_i(0) | W_i] \)—by restricting the expectation of \( Z_i \) conditional on \( X_i \) to be linear in \( W_i \).

**Assumption L (Linearity)** \( \mathbb{E}[Z_i | X_i] = \mathbb{E}^*[Z_i | W_i] \).

Assumption L ensures that for controlling the effect of covariates on the instruments by a linear projection on \( W_i \) is as good as conditioning on \( X_i \). There are three important special cases in which Assumption L holds automatically. First, if there are no covariates. Second, if \( X_i \) is discrete and \( W_i \) is saturated, consisting of dummies for different values of \( X_i \). Third, if \( Z_i \) is a function of \( Q_i \) only, and \( Q_i \) is independent of \( X_i \), in which case \( \mathbb{E}[Z_i | X_i] = \mathbb{E}[Z_i] \). This happens, for example, when \( Q_i \) is some randomly assigned encouragement to take the treatment, and the covariates are added after the randomization to increase precision of inference.

Abadie (2003) shows that another consequence of Assumption L is that the parameter \( \delta \) in Equation (7) can now be interpreted as providing the best linear approximation to \( \mathbb{E}[Y_i(0) | X_i] \) in the sense of minimizing the mean-square error \( \mathbb{E}[(\mathbb{E}[Y_i(0) | X_i] - W_i'\delta)^2] \).

### 3.B Two-step IV estimators

An implication of the moment condition (7) is that \( \beta \) can be identified using a single instrument

\[
\hat{p}_i = \mathbb{E}^*[T_i | Z_i, W_i] - \mathbb{E}^*[T_i | W_i] = \tilde{Z}_i'\tau_2,
\]

the linear approximation to the propensity score (4) with

\[\text{By the residual regression formula (9) holds iff } \mathbb{E}[Y_i(0)Z_i] = 0. \text{ Assumption L implies that } \mathbb{E}[Z_i | X_i] = 0. \text{ Therefore, by the law of iterated expectations, we have } \mathbb{E}[Y_i(0)Z_i | X_i] = \mathbb{E}[\mathbb{E}[Y_i(0) | X_i, Z_i]Z_i | X_i] = \mathbb{E}[Y_i(0) | X_i] \mathbb{E}[Z_i | X_i] = 0 \text{ where the second equality follows from Assumption IV.}\]
the covariates partialled out. $\tilde{P}_i^L$ can be thought of as an approximation to $E[T_i \mid Q_i, X_i] - E[T_i \mid X_i] = P_i - E[P_i \mid X_i]$, which measures how strong the instrument assigned to individual $i$ is (in terms of how likely it is to induce an individual into taking the treatment), relative to other instruments they could have been assigned, holding the covariates fixed. Since $\tilde{P}_i^L$ is linear in $Z_i$ and $W_i$, the moment condition implies that

$$0 = E^*[Y_i - W_i^T \delta - T_i \beta \mid \tilde{P}_i^L] = E^*[Y_i - T_i \beta \mid \tilde{P}_i^L],$$

where the second equality follows from $E[W_i \tilde{P}_i^L] = 0$. Rearranging this expression, we obtain

$$\beta = \frac{E[\tilde{P}_i^L Y_i]}{E[\tilde{P}_i^L T_i]},$$

so that the IV estimator that uses $\tilde{P}_i^L$ as a single instrument, $\hat{\beta}_{IV} = \sum_i \tilde{P}_i^L Y_i / \sum_i \tilde{P}_i^L T_i$, is consistent for $\beta$. Moreover, if the error $\epsilon_i = Y_i - W_i^T \delta - T_i \beta$ is homoscedastic, so that $\text{var}(\epsilon_i^2 \mid X_i, Q_i) = \sigma^2$, then this estimator is asymptotically efficient.

Since $\tilde{P}_i^L$ is not directly observed, such an estimator is not feasible. Two step IV estimators implement a feasible version of $\hat{\beta}_{IV}$. In the first step, they construct an estimate $\hat{P}_i$ of $\tilde{P}_i^L$. In the second step, an IV estimator that uses this constructed instrument as a single instrument is used to estimate the treatment effect:

$$\hat{\beta}_P = \frac{\hat{P}_i^T Y_i}{\hat{P}_i^T T_i}. (10)$$

The class of two-step IV estimators is given by all estimators that admit this representation, where $\hat{P}$ is a function of $T$, $W$ and $Z$, including:

- The two-stage least squares (TSLS) estimator, which replaces $\pi_2$ and $\psi_2$ in (4) by their least-squares estimates, leading to $\hat{P} = H_{Z_i} T$;

- The bias-corrected two-stage least squares estimator (Nagar, 1959), which adjusts the TSLS propensity score estimator to $\hat{P} = ((1 - k)M_W + kH_{Z_i}) T$ to improve its finite-sample properties, where $k = 1/(1 - (K - 2)/n)$;

- The jackknife instrumental variables estimator (Phillips and Hale, 1977; Angrist et al., 1999), with $\hat{P} = M_W (I_n - (I_n - D_{(Z,W)})^{-1} M_{(Z,W)}) T$.

Under regularity conditions, the estimation error in the first step does not matter, and all of these estimators are consistent for $\beta$, and asymptotically efficient under homoscedasticity.

### 3.C Minimum distance estimators

Another implication of the conditional moment restriction (7) is that if we project $Y_i$ and $T_i$ onto $Z_i$ and $W_i$, the coefficients on $Z_i$ will be proportional to each other. To see this, by linearity of
linear projections, we obtain:

$$E^*[Y_i \mid Z_i, W_i] = W_i'\delta + E^*[T_i \mid Z_i, W_i]\beta.$$  \hfill (11)

Therefore, the coefficients in the linear projections (3)-(4) are related to the coefficients \((\beta, \delta)\) by

$$\delta = \psi_1 - \psi_2\beta,$$

and

$$\pi_1 = \pi_2\beta.$$  \hfill (12)

This proportionality restriction can be imposed directly in estimation of \(\beta\) by using a minimum distance objective function

$$(\text{vec}(\hat{\Pi}) - a \otimes \pi_2)'\Phi(\text{vec}(\hat{\Pi}) - a \otimes \pi_2), \quad a = \begin{pmatrix} \beta \\ 1 \end{pmatrix},$$  \hfill (13)

where \(\hat{\Pi} = (Z'_\perp Z'_\perp)^{-1}Z'_\perp(Y, T)\) is an unrestricted least-squares estimator of \(\Pi = (\pi_1, \pi_2)\), and \(\Phi\) is some weight matrix. Goldberger and Olkin (1971) show that the limited information maximum likelihood (LIML) estimator minimizes this objective function if the weight matrix is given by

$$\hat{\Phi} = \hat{\Omega}^{-1} \otimes Z'_\perp Z'_\perp/n,$$

where \(\hat{\Omega} = \hat{\Pi}'\hat{\Pi}\) is an estimator of the covariance matrix of the reduced-form errors \(V_{1i} = Y_i - E^*[Y_i \mid Z_i, W_i]\) and \(V_{2i} = T_i - E^*[T_i \mid Z_i, W_i]\) based on the unrestricted least-squares residuals. To understand the sensitivity of minimum distance estimators to departures from the assumption of constant treatment effects (Assumption CTE), it is helpful to work with a slightly different minimum distance objective function. Define a 2-by-2 matrix

$$\Xi = \Pi'E[\tilde{Z}_1\tilde{Z}_1']\Pi.$$  \hfill (14)

In Section 5, I will show that this matrix plays a key role in understanding the behaviour of classic linear IV estimators under the LATE framework. The proportionality restriction (12) implies a rank restriction on \(\Xi\), namely that \(\Xi = \Lambda a a'\), where \(\Lambda = \Xi_{22} = \pi_2'E[\tilde{Z}_1\tilde{Z}_1']\pi_2\). This restriction is essentially a restriction on the second moments of \(\hat{\Pi}\) if \(E[\tilde{Z}_1\tilde{Z}_1']\) is proportional to the identity matrix. If the weight matrix \(\hat{\Phi}\) has a Kronecker structure, \(\hat{\Phi} = \hat{S}^{-1} \otimes Z'_\perp Z_\perp/n\) for some positive definite matrix \(\hat{S} \in \mathbb{R}^{2 \times 2}\), minimizing the objective function (13) yields the same estimator of \(\beta\) as a minimum distance estimator based on the rank-restriction on \(\Xi\) given by$^3$

$$\hat{D}(\beta, \Lambda) = \text{vec}(\hat{\Xi} - \Lambda a a')(\hat{S}^{-1} \otimes \hat{S}^{-1})\text{vec}(\hat{\Xi} - \Lambda a a'),$$  \hfill (15)

$^3$See Kolesár (2012) for derivation.
where \( \hat{\xi} = (Y, T)'H_{Z_1}(Y, T)/n = \hat{\Gamma}'(Z_1'Z_1/n)\hat{\Gamma} \) is an unrestricted estimator of \( \xi \). The class of minimum distance estimators is given by all estimators that minimize the objective function (15) for some unrestricted estimator \( \hat{\xi} \) of \( \xi \) and some weight matrix \( \hat{S}^{-1} \otimes \hat{S}^{-1} \). These estimators can be written as
\[
\hat{\beta}_{\hat{\xi},\hat{S}} = \frac{\hat{\xi}_{12} - \hat{S}_{12} \min \text{eig}(\hat{S}^{-1} \hat{S})}{\hat{\xi}_{22} - \hat{S}_{22} \min \text{eig}(\hat{S}^{-1} \hat{S})},
\]
(16)

Apart from \( \text{liml} \), the class of minimum distance estimators includes:

- \( \Omega \)-class estimators of Keller (1975), which, like \( \text{liml} \), set \( \hat{\xi} = (Y, T)'H_{Z_1}(Y, T)/n \), but \( S \) is free to be any positive definite matrix. The choice \( \hat{S} = I_2 \) leads to the symmetrically normalized two-stage least squares estimator studied in Keller (1975), Hillier (1990) and Alonso-Borrego and Arellano (1999).

- The symmetric jackknife estimator of (Bekker and Crudu, 2012), which sets
\[
\hat{\xi} = \frac{1}{n} \left( Y \ T \right)' \left( H_{Z_1} - \frac{1}{2} \left( H_{Z_1} C + C' H_{Z_1} \right) - \frac{1}{4} C' H W C \right) \left( Y \ T \right),
\]
\[
\hat{S} = \frac{1}{n - K - L} \left( Y \ T \right)' M_{(Z,W)} D_{(Z,W)} \left( I_n - D_{(Z,W)} \right)^{-1} M_{(Z,W)} \left( Y \ T \right),
\]
where \( C = D_{(Z,W)} \left( I_n - D_{(Z,W)} \right)^{-1} M_{(Z,W)} \).

- If there are no covariates \( W \), then the \( \text{liml} \) estimator of Hausman et al. (2012) also admits this minimum distance representation, with
\[
\hat{\xi} = \frac{1}{n} \left( Y \ T \right)' \left( H_Z - D_Z \right) \left( Y \ T \right), \quad \hat{S} = \frac{1}{n - K} \left( Y \ T \right)' \left( M_Z + D_Z \right) \left( Y \ T \right).
\]

Under homoscedasticity, any weight matrix \( S \) produces an asymptotically efficient estimator (Alonso-Borrego and Arellano, 1999). Consequently all of these estimators are asymptotically efficient under these conditions, and first-order asymptotically equivalent to the optimal two-step iv estimator. The reason for this is that under standard asymptotics, the estimator \( \hat{\xi}_{12}/\hat{\xi}_{22} \) of \( \beta \) that does not use the information about \( \beta \) contained in \( \xi_{11} \) is asymptotically equivalent to a minimum distance estimator that uses the optimal weight matrix. This is easy to check since, in fact, \( \hat{\xi}_{12}/\hat{\xi}_{22} \) is the two-stage least squares estimator.

4 Local average treatment effects approach

Instead of restricting treatment effect heterogeneity, the local average treatment effects framework of Imbens and Angrist (1994) replaces Assumption CTE by a monotonicity assumption that restricts how a treatment response to changing the value of the instrument may vary across people:
Assumption M (Monotonicity) For all \( q, q' \in Q \) either \( \mathbb{P}(T_i(q) \geq T_i(q') \mid X_i) = 1 \) or \( \mathbb{P}(T_i(q) \leq T_i(q') \mid X_i) = 1 \) a.s.

This assumption maintains that changing the instruments from \( q \) to \( q' \) affects all individuals with the same value of \( X_i \) in the same direction—it rules out situations in which, in response to a change in \( Q_i \), some people drop out of treatment and others select into it. If \( Q_i \) is an encouragement to take the treatment, for example, then monotonicity requires that encouraging people to take the treatment makes everyone more likely to take it. Vytlacil (2002) shows that Assumption M is equivalent to assuming a latent index model as first proposed by Heckman (1976), in which selection into the treatment is modeled by a latent index crossing a threshold.\(^4\)

For each value \( x \in \mathcal{X} \) and for each pair \((q, q')\), define a local average treatment effect (LATE):

\[
\tau(q, q'; x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i(q) \neq T_i(q'), X_i = x]. \tag{17}
\]

This is the treatment effect averaged over individuals with \( X_i = x \) who change their treatment status if we change their instrument from \( q \) to \( q' \). Angrist, Imbens and Rubin (1996) refer to this set of individuals as compliers. Imbens and Angrist (1994) show that under Assumptions IV, and M, so long as \( \mathbb{P}(T_i(q) \neq T_i(q') \mid X_i = x) > 0 \), these local average treatment effects can be identified from the reduced form regressions:

\[
\tau(q, q'; x) = \frac{r(q, x) - r(q', x)}{p(q, x) - p(q', x)}. \tag{18}
\]

If \( \mathbb{P}(T_i(q) \neq T_i(q') \mid X_i = x) = 0 \), then the set of compliers that the local average treatment effect (17) conditions on is empty, \( p(q, x) = p(q', x) \), and \( \tau(q, q'; x) \) is not identified. Since Assumption IV (iii) implies that the distribution of \( P_i \) conditional on \( X_i \) is non-degenerate with positive probability, it ensures that at least some local average treatment effects are identified. On the other hand, the population average treatment effect \( \mathbb{E}[\tau_i] \) is no longer identified once Assumption CTE is dropped unless the instrument \( Q_i \) is sufficiently strong to change everyone’s treatment status (known as “identification at infinity”). The reason is that without restricting treatment effect heterogeneity, we have no way of computing the treatment effect for individuals who don’t change their treatment status in response to a change in \( Q_i \).

To facilitate expressing estimands or estimators based on the linear IV model in terms of local average treatment effects, it will be useful to write \( \tau(q, q'; x) \) in terms of functions of the propensity score. Because of the equivalence between monotonicity and single index models, the instruments \( Q_i \) enter the model only through the propensity score (Heckman and Vytlacil, 1999; Heckman et al., 2006). Therefore, \( r(q, x) = \mathbb{E}[Y_i \mid P_i = p(q, x), X_i = x] \). Let \( \mathcal{P}_x \) be the support of \( P_i \) conditional

\[\text{In the Heckman (1976) model, the index is given by } T_i^* = p(Q_i, X_i) - U_i, \text{ where } U_i \text{ is an unobserved random variable, distributed independently of } (Q_i, X_i). T_i^* \text{ is interpreted as the expected net utility of selecting into treatment, so that } T_i = 1 \text{ if } T_i^* \geq 0.\]
on \( X_i = x \). Suppose that \( Q_i \) is discrete, so that \( \mathcal{P}_x \) has finitely many support points. Let \( J_x \) be the number of support points, with \( \mathcal{P}_x = \{ p_{1,x} < \ldots < p_{J_x,x} \} \). Define a marginal local average treatment effect:

\[
\alpha(p_{j,x}; x) = \frac{\mathbb{E}[Y_i \mid P_i = p_{j+1,x}, X_i = x] - \mathbb{E}[Y_i \mid P_i = p_{j,x}, X_i = x]}{p_{j+1,x} - p_{j,x}}, \quad j = 1, \ldots, J_x - 1.
\]

\( \alpha(p_{j,x}; x) \) is the local average treatment effect for individuals who get treated when the instrument they receive corresponds to propensity score with rank higher than \( j \) but not otherwise. We can express every local average treatment effect (17) for which the set of compliers is non-empty in terms of these marginal LATEs. In particular, let \( p(q, x) = p_{j,x} \) and that \( p(q', x) = p_{j',x} \), and suppose that \( j > j' \). Then we obtain:

\[
\tau(q, q'; x) = \frac{\sum_{m=j}^{j-1} (\mathbb{E}[Y_i \mid P_i = p_{m+1,x}, X_i = x] - \mathbb{E}[Y_i \mid P_i = p_{m,x}, X_i = x])}{p_{j,x} - p_{j',x}} = \sum_{m=j}^{j-1} \frac{p_{m+1,x} - p_{m,x}}{p_{j,x} - p_{j',x}} \alpha(p_m; x).
\]

If \( j' = j - 1 \), then \( \tau(q, q'; x) = \alpha(p_j; x) \).

If the support of \( \mathcal{P}_x \) is continuous, with \( \mathcal{P}_x = [p_x, \bar{p}_x] \), a similar result obtains if we replace the marginal LATEs \( \alpha(p_{j,x}; x) \) by their limit as \( q \to q' \), the marginal treatment effect (Heckman, 1997):

\[
\tau(q, q'; x) = \frac{1}{p(q, x) - p(q', x)} \int_{p(q', x)}^{p(q, x)} \text{mte}(p; x) \, dp, \quad \text{mte}(p; x) = \frac{\partial}{\partial p} \mathbb{E}[Y_i \mid P_i = p, X_i = x],
\]

where the equality follows from Equation (18) and the fundamental theorem of calculus. To keep the exposition simple, I will focus on the case with discrete instruments and finite support \( \mathcal{P}_x \). The results in this paper generalize easily to the continuous case by replacing \( \alpha(p_m; x) \) with the marginal treatment effect, \( (p_{m+1,x} - p_{m,x}) \) with \( dp \), and replacing sums with integrals.

## 5 Estimands under the LATE framework

This section presents the first main result of the paper: the estimands of two-step IV estimators and minimum distance estimators when we do not restrict treatment effect heterogeneity.

I derive this result in two steps. First, in Lemma 1 below, express their probability limits in terms of the reduced-form parameter \( \Xi \), defined in Equation (14)—this result does not require any modelling assumptions. Second, I assume the local average treatment effects framework, and I express these reduced-form limits in terms of local average treatment effects.

**Lemma 1** Suppose that the data \( \{ Y_i, T_i, Q_i, Z_i, X_i, W_i \}_{i=1}^n \) are i.i.d with finite second moments.
(i) Consider a two-step IV estimator \( \hat{\beta}_P \) that satisfies \( \hat{P}'Y/n \xrightarrow{p} E[\hat{P}'Y] = E[P_Y] \neq 0 \), where \( \hat{P}_i^L = E^*[T_i | Z_i, W_i] - E^*[T_i | W_i] \). Then:

\[
\hat{\beta}_P \xrightarrow{p} \frac{E[\hat{P}'_i Y_i]}{E[\hat{P}'_i T_i]} = \frac{\Xi_{12}}{\Xi_{22}}.
\]

(ii) Consider the reverse two-stage least squares estimator given by \( \hat{\beta}_{rtsls} = \frac{Y' H_Z Z \hat{Y} Y}{Y' H_Z Z} \). Suppose that \( \Xi_{12} \neq 0 \) and that \( E[(Z_i, W_i)(Z_i, W_i)'] \) is full rank. Then:

\[
\hat{\beta}_{rtsls} \xrightarrow{p} \frac{E[\hat{R}'_i Y_i]}{E[\hat{R}'_i T_i]} = \frac{\Xi_{11}}{\Xi_{12}},
\]

where \( \hat{R}_i^L = E^*[Y_i | Z_i, W_i] - E^*[Y_i | W_i] \).

(iii) Consider a minimum distance estimator \( \hat{\beta}_{\hat{\Xi},S} \) that satisfies \( \hat{S} \xrightarrow{p} S \) for some positive definite matrix \( S \), and \( \hat{\Xi} \xrightarrow{p} \Xi \). Suppose that \( \Xi_{22} \neq \min \text{eig}(S^{-1}\Xi)S_{22} \). Then \( \hat{\beta}_{\hat{\Xi},S} \xrightarrow{p} \beta_S \), where \( \beta_S \) minimizes the objective function

\[
D_S(\beta, \Lambda) = \text{vec}(\Xi - aa'\Lambda)'(S^{-1} \otimes S^{-1}) \text{vec}(\Xi - aa'\Lambda),
\]

and it is given by

\[
\beta_S = \frac{\Xi_{12} - S_{12} \min \text{eig}(S^{-1}\Xi)}{\Xi_{22} - S_{22} \min \text{eig}(S^{-1}\Xi)}.
\]

Lemma 1 shows that understanding how the reduced-form parameter \( \Xi \) relates to local average treatment effects is the key to understanding the properties of estimators based on the classic linear IV model.

In particular, Part (i) shows that the probability limit of TLS and other two-step IV estimators is simply given by \( \Xi_{12}/\Xi_{22} \), the estimand of an IV estimator that uses the linear predictor of the treatment (with the effect of the covariates partialled out), \( \hat{P}_i^L \), as a single instrument. It makes a high-level assumption that the first-step estimator \( \hat{P}_i \) converges to its population target, \( \hat{P}_i^L \). The primitive conditions for this depend on the estimator, but for TLS, a sufficient condition is that \( E[(Z_i, W_i)(Z_i, W_i)'] \) is full rank.

Part (ii) introduces a new estimator, the reverse two-stage least squares estimator (RTSLS). It is obtained as the reciprocal of the TLS estimator in the instrumental variables model that swaps the treatment and the outcome:

\[
\hat{\beta}_{rtsls} = \left( \frac{\hat{R}'_{rtsls} T}{\hat{R}'_{rtsls} Y} \right)^{-1} = \frac{Y' H_Z Z \hat{Y} Y}{Y' H_Z Z} \frac{Y' H_Z Z'}{Y' H_Z Z'},
\]

where \( R_{rtsls} = H_Z Y \) is the first-step estimator of \( \hat{R}_i^L = R_i - E^*[R_i | W_i] \) based on a least-squares estimation of Equation (3). Lemma 1 shows that the probability limit of this estimator is given by
Ξ₁₁/Ξ₁₂, the estimand of an IV estimator that uses the linear predictor of the outcome (again with the covariates partialled out), ˆRᵢ, as a single instrument. The reason for introducing this estimator is that one way of thinking about what a minimum distance estimand tries to do is to try to think of it as trying to be close to both Ξ₁₂/Ξ₂₂ and Ξ₁₁/Ξ₁₂, using the weight matrix S as a distance metric.

Part (iii) formalizes this notion. The regularity condition Ξ₂₂ ≠ \text{min eig}(S^{-1}Ξ)S₂₂ ensures that the limiting objective function has a well-defined minimum. Again, the primitive conditions for ˆΞ → Ξ depend on the estimator, but for \text{LIML}, a sufficient condition is that \( \mathbb{E}[(Z_i, W_i)(Z_i, W_i)'] \) is full rank.

The rationale for trying to equate the two-step IV estimand Ξ₁₂/Ξ₂₂ with the rtsls estimand Ξ₁₁/Ξ₁₂ is that the classic linear IV model is symmetric in Y and T; instead of instrumenting for T in Equation (7) like tsls does, we can multiply it by 1/β, instrument for Y, and take the reciprocal of the resulting estimator, obtaining rtsls. Both tsls and rtsls converge to the same probability limit, equal to the population average treatment effect \( \beta \). By pooling the information about \( \beta \) contained in tsls with the information contained in rtsls, minimum distance estimators have more attractive finite sample properties in classic IV model than two-step IV estimators, which don’t use information about \( \beta \) contained in rtsls (Phillips, 1983; Hillier, 1990). They are also more efficient under many instrument asymptotics (Hausman et al., 2012).

The key question is how the interpretation of two-step IV, rtsls, and minimum distance estimands changes under the late framework when Assumption CTE in the classic IV model is replaced by the Assumption M. I first answer this question for two-step IV and rtsls estimands in Theorem 1 and Corollary 1 below by expressing the two ratios Ξ₁₁/Ξ₁₂ and Ξ₁₂/Ξ₂₂ in terms of the marginal local average treatment effects \( \alpha(\cdot) \) defined in Equation (19).

**Theorem 1** Suppose that Assumptions IV, M and L hold. Let \( F^X \) denote the distribution of \( X_i \). Then

\[
\frac{Ξ₁₂}{Ξ₂₂} = \int \frac{\theta_j(x)}{\sum_{j=1}^{J-1} \theta_j(x)} \alpha(p_j, x) \, dF^X(x),
\]

and, if Ξ₁₂ ≠ 0

\[
\frac{Ξ₁₁}{Ξ₁₂} = \int \frac{\zeta_j(x)}{\sum_{j=1}^{J-1} \zeta_j(x)} \alpha(p_j, x) \, dF^X(x),
\]

16
where
\[
\theta_j(x) = (p_{j+1,x} - p_{j,x}) \mathbb{P}(P_i > p_{j,x} \mid X_i = x) \mathbb{E}[\hat{\theta}_i^L \mid X_i = x, P_i > p_{j,x}],
\]
\[
\zeta_j(x) = (p_{j+1,x} - p_{j,x}) \mathbb{P}(P_i > p_{j,x} \mid X_i = x) \mathbb{E}[\hat{\zeta}_i^L \mid X_i = x, P_i > p_{j,x}].
\]

Theorem 1 shows that both \( \Xi_{12}/\Xi_{22} \) and \( \Xi_{11}/\Xi_{12} \) can be expressed as an affine combination of (marginal) local average treatment effects (the weights integrate to one, but are not necessarily positive).

For the two-step IV weights \( \theta_j(x)/\int \sum_{j=1}^{L} \theta_j(x) \, dF(x) \) to be positive, we need that the single instrument \( \hat{\theta}_i^L \) is monotone in the propensity score \( P_i \). This ensures that the last term \( \mathbb{E}[\hat{\theta}_i^L \mid X_i = x, P_i > p_{j,x}] \) is always positive. In other words, we need the linear approximation \( P_i \) to the true propensity score \( P_i \) to be good enough in the sense that changing the value of \( \hat{\theta}_i^L \) does not induce two-way flows in and out of treatment (see Heckman and Vytlacil (2005) and Heckman et al. (2006) for discussion of this issue). If the linear approximation to the propensity score is exact, so that \( P_i = P_i^L \), then the weights are guaranteed to be positive. A leading case in which this condition holds automatically is when \( Q_i \) and \( X_i \) are both finite, and we estimate a saturated model in which the instruments \( Z_i \) are generated by interacting indicators for different values of \( Q_i \) with indicators for different values of \( X_i \). In this case, Angrist and Imbens (1995) obtain a similar expression for the weights \( \theta_j(x)/\int \sum_{j=1}^{L} \theta_j(x) \, dF(x) \).

Similarly, the \( \text{rtsls} \) weights \( \zeta_j(x)/\int \sum_{j=1}^{L} \zeta_j(x) \, dF(x) \) are positive if the single instrument \( \hat{\zeta}_i^L = \mathbb{E}^*[Y_i \mid W_i, Z_i] - \mathbb{E}^*[Y_i \mid W_i] \) used by \( \text{rtsls} \) is monotone in the propensity score \( P_i \). The next corollary gives a necessary and sufficient condition for this condition to hold if the linear approximations (3)–(4) are exact.

**Corollary 1** Suppose that the linear approximations (3)–(4) are exact, so that \( \mathbb{E}[Y_i \mid Q_i, X_i] = \mathbb{E}^*[Y_i \mid Z_i, W_i] \) and \( \mathbb{E}[T_i \mid Q_i, X_i] = \mathbb{E}^*[T_i \mid Z_i, W_i] \), and that Assumptions IV, M, and L hold. Then the weights \( \theta_j(x)/\int \sum_{j=1}^{L} \theta_j(x) \, dF(x) \) are positive, and the weights \( \zeta_j(x)/\int \sum_{j=1}^{L} \zeta_j(x) \, dF(x) \) are positive if all marginal \( \text{lates} \) \( \{a(p_j(x); x)\} \) have the same sign. In the special case that \( J_x = 2 \) for all \( x \),
\[
\theta_1(x) = \text{var}(P_i \mid X_i = x), \quad \zeta_1(x) = \text{var}(P_i \mid X_i = x) a(p_{1,x}; x).
\]

The proof relies on the fact that if the linear approximations (3)–(4) are exact, then the conditional expectation of \( R_i = R_i^L \) can be decomposed as
\[
\mathbb{E}[R_i^L \mid P_i = p_{j,x}, X_i = x] = \mathbb{E}[R_i^L \mid P_i = p_{1,x}, X_i = x] + \sum_{j=1}^{L} a(p_{j,x}; x)(p_{j+1,x} - p_{j,x}),
\]
so that \( R_i^L \) is only monotone in the propensity score if the marginal \( \text{lates} \) \( a(\cdot) \) all have the same sign. The other implication of this decomposition is that it demonstrates that the conditional expectation of the instrument \( R_i^L \), and hence the term \( \mathbb{E}[\hat{R}_i^L \mid X_i = x, P_i > p_{j,x}] \) depend on the size...
of the marginal treatment effects $\alpha(\cdot)$. In the special case that $Q_i$ is binary, so that $f_x = 2$, and the instruments $Z_i$ are generated by interacting $Q_i$ with the covariates, this results in the weights $\zeta$ to be exactly equal to the product of the marginal treatment effect with the two-step IV weights $\theta$. Therefore, larger local average treatment effects receive more weight, and negative local average treatment effects receive a negative weight in this case.

Taken together, Lemma 1, Theorem 1 and Corollary 1 show that under Assumptions IV, M and L, two-step IV estimators estimate a convex combination of local average treatment effects, so long as the linear approximation $P^L_i$ to the true propensity score $P_i$ is monotone in $P_i$. In the special case with a binary $Q_i$, Corollary 1 shows that these weights are given by the variance of the propensity score, so that better identified lates receive more weight. If in fact all lates are equal, then this weighting scheme ensures that under homoscedasticity, asymptotic variance of two-step IV estimators is minimized. On the other hand, the weighting used by the RTSLS estimand is different, depends on the size of the local average treatment effects, and may result in an estimand outside of the convex hull of lates if some lates are positive and some are negative.

Because it gives more weight to larger lates, the RTSLS estimand will always be larger than the two-step IV estimand. This result holds in general by the Cauchy-Schwarz inequality since $\Xi$ is a covariance matrix of $(\tilde{R}^L_i, \tilde{P}^L_i)$,

$$\Xi = E[(\tilde{R}^L_i, \tilde{P}^L_i)'(\tilde{R}^L_i, \tilde{P}^L_i)],$$

so that $\Xi_{11} \geq \Xi_{22}$. Hence, $\Xi_{11}/\Xi_{12} \geq \Xi_{12}/\Xi_{22}$, with equality only if $\tilde{P}^L_i$ is perfectly correlated with $\tilde{R}^L_i$, in which case the RTSLS weights are proportional to the two-step IV weights. There are only two ways how this can happen: either all local average treatment effects are equal, or else the dimension of $Z_i$ is one, so that the IV model (7) is exactly identified. In general with more than one instrument, it will be the case that $\Xi_{11}/\Xi_{12} > \Xi_{12}/\Xi_{22}$.

The result that the RTSLS estimand is in general different from the two-step IV estimand has important implications for minimum distance estimators. On the one hand, combining RTSLS with TLSLS leads to more attractive properties of minimum distance estimators in the classic IV model under which $\Xi_{11}/\Xi_{12} = \Xi_{12}/\Xi_{22}$. On the other hand, trying to equate RTSLS and TLSLS when their estimands are in fact different makes minimum distance estimators unattractive under treatment effect heterogeneity; as I discuss next, it may cause the minimum distance estimands to no longer correspond to a causal effect.

If the local average treatment effects are not all equal, then $\Xi_{11}/\Xi_{12} \neq \Xi_{12}/\Xi_{22}$, and the probability limit of a minimum distance estimator depends on the weight matrix $S$. If the weight matrix is diagonal, then the minimum distance estimand lies between TLSLS and RTSLS estimands—this was first shown in Zellner (1970) in an errors-in-variables context. Therefore, the symmetrically normalized two-stage least squares estimator (see page 12 for definition), for example, which uses the identity matrix as a weight matrix will always lie between two-step IV...
and RTSLS estimands. The relative weight given to the RTSLS and TLS estimands depends on the ratio \( S_{11}/S_{22} \). In particular if the ratio \( S_{11}/S_{22} \) is small, then the penalty from being far away from \( \Xi_{11}/\Xi_{12} \) is large, so the minimum distance estimand will be close to the RTSLS estimand. On the other hand, if \( S_{11}/S_{22} \) is large, then the minimum distance estimand will be close to the two-step iv estimand (see Zellner (1970) and Keller (1975) for a detailed discussion). Heuristically, if we concentrate \( \Lambda \) out of the objective function \( D_S(\beta, \Lambda) \) given in (21), we obtain that

\[
\beta_S = \arg \min_{\beta} \frac{\Xi_{22} \beta^2 - 2\Xi_{12} \beta + \Xi_{11}}{S_{11} + S_{22} \beta^2}.
\]

Now, if we set \( S_{22} = 0 \), then \( \beta_S = \Xi_{12}/\Xi_{22} \), and if we set \( S_{11} = 0 \), then we obtain \( \beta_S = \Xi_{11}/\Xi_{12} \).

If \( S \) is non-diagonal, however, then the minimum distance estimand may lie outside the interval formed by the two-step iv and RTSLS estimands. This is typically the case for LIML, for which \( S \) equals the covariance matrix of the reduced-form errors, which is typically non-diagonal. To see how this may happen, consider a simple model in which we observe draws of a vector \( A_i \), distributed according to the bivariate Normal distribution with mean \((\mu_1, \mu_2)\)' and covariance matrix \( \Omega \). If \( \mu_1 = \mu_2 \), and \( \Omega \) is known, then the optimal estimator is given by:

\[
\hat{\mu} = \frac{i' \Omega^{-1} \bar{A}}{i' \Omega^{-1} i}, \quad i = \begin{pmatrix} 1 \\ 1 \end{pmatrix},
\]

where \( \bar{A} = n^{-1} \sum_{i=1}^n A_i \). The probability limit of this estimator is given by

\[
\mu_\Omega = \frac{(\Omega_{22} - \Omega_{12}) \mu_1 + (\Omega_{11} - \Omega_{12}) \mu_2}{\Omega_{22} + \Omega_{11} - 2\Omega_{12}}.
\]

If \( \Omega_{12} = 0 \), then \( \mu_\Omega \) lies between \( \mu_1 \) and \( \mu_2 \). If, however, \( \Omega \) is non-diagonal, then this may no longer be the case—if, for example, \( \mu_2 = 0 \) and \( \mu_1 \) is positive, then \( \mu_\Omega \) will be negative if \( \Omega_{22} < \Omega_{12} \).

There are two ways, therefore, in which a minimum distance estimand may end up being outside of the convex hull of LTS. First, if some LTS are positive and some are negative, and the RTSLS estimand is outside of the convex hull, then so long as the weight matrix \( S \) gives sufficient weight to RTSLS, the minimum distance estimand will also be outside of the convex hull. Second, even if the RTSLS estimand lies inside the convex hull, if the weight matrix \( S \) is non-diagonal, the minimum distance estimand may end up being outside of the convex hull. These possibilities make LIML and other minimum distance estimators an unattractive estimator choice in settings with possible treatment effect heterogeneity.
6 Estimation with many instruments

In this section, I derive the second main result of the paper that a version of the jackknife instrumental variables estimator, the unbiased jackknife instrumental variables estimator (ujive), is consistent for a convex combination of lates under a many instrument asymptotic sequence in which both the number of instruments and the number of covariates is allowed to increase in proportion with the sample size. In settings with many instruments and treatment effect heterogeneity, ujive is therefore a more attractive estimator than tsls, which is inconsistent under many instrument asymptotics. It is also more attractive than liml, the standard alternative to tsls when many instruments are used, since, as shown in Section 5, liml may converge to a quantity outside of the convex hull of local average treatment effects even under standard asymptotics.

To illustrate the issues that arise with a large number of instruments, as well as to motivate ujive, I first discuss a simple example in which the instruments $Z_i$ are indicators for group membership. I then give the general consistency theorem.

6.A A simple example with groups as instruments

Consider the problem of estimating the effect of incarceration on post-release criminality, as in Aizer and Doyle, Jr. (2011) and Nagin and Snodgrass (2011). The identification strategy in these papers relies on the fact that cases are randomly assigned to judges who vary in their sentencing severity. This suggests using judge indicators as instruments for incarceration.\(^5\) Hence, with $K+1$ judges, $\dim(Z_i) = K$, and $Z_{ik} = \mathbb{I}_{Q_i = k}$, where $Q_i$ denotes the judge assigned to individual $i$ (I omit the indicator for the last judge so that we can include the intercept). In this context, the monotonicity assumption requires that the judges can be ordered in terms of how strict they are. The local average treatment effects are defined for each pair of judges and correspond to the average treatment effect for individuals who would get incarcerated if assigned to the stricter judge of the two, but would not get incarcerated if assigned to the more lenient judge. If the effect of incarceration for more serious offenders (who get incarcerated unless assigned to the most lenient judges) is different from the effect for individuals who committed less serious crimes (who only get incarcerated if assigned to the strictest judges), then these lates will differ.

In the absence of covariates (beyond the intercept), the propensity score for individual $i$ is given by $P_i = \mathbb{E}[T_i \mid Q_i]$, and it corresponds simply to the incarceration propensity of the judge assigned to individual $i$. Because the first stage is saturated, the linear approximation (4) is exact, and $\tilde{P}_i = \hat{P}_i = P_i - \mathbb{E}[P_i]$. Intuitively, $\hat{P}_i$ measures how strict the judge assigned to individual $i$ is compared to other judges.

Let $J_k$ denote the number of cases assigned to judge $k$. The two-stage least squares estimator

\(^5\) A similar strategy is also used in Dobbie and Song (2012), who study the effect of being granted bankruptcy protection on subsequent earnings, using judge indicators as instruments.
We have (Bekker, 1994) that let the dimension of $Z_i$ where the last line follows from the law of large numbers applied to all four expressions in parentheses, and the fact that $E[Y_i] = 1$, $E[T_i] = 0$, and $E[\hat{P}_i] = P_i$. The resulting tsls estimator is given by

$$\hat{P}_{\text{tsls}} = \frac{n^{-1} \sum_i \hat{P}_{i,\text{tsls}} Y_i}{n^{-1} \sum_i \hat{P}_{i,\text{tsls}} T_i}. \quad (22)$$

There are two basic ways of doing asymptotics in this setting. The first option is to let the number of cases per judge grow to infinity while keeping the number of judges fixed. This corresponds to the standard asymptotics. As the number of cases per judge $Q_i$ increases, $\hat{P}_{i,\text{tsls}} \xrightarrow{p} P_i = p_{Q_i}$, and the numerator and the denominator in (22) converge to $E[\hat{P}_i Y_i]$ and $E[\hat{P}_i T_i]$, respectively. By Lemma 1 and Theorem 1, $\hat{P}_{\text{tsls}}$ therefore converges to a weighted average of local average treatment effects. However, with a large number of judges and small number of cases per judge, these asymptotics do not capture the finite-sample properties of the estimator very well.

The other possibility is to keep the number of cases per judge, $J_k$, fixed, and let the number of judges $K \to \infty$. This corresponds to the many instrument asymptotics (Kunitomo, 1980; Morimune, 1983; Bekker, 1994) that let the dimension of $Z_i$ increase in proportion with the sample size. Under these asymptotics, $P_i$ can no longer be consistently estimated, and so the exact way in which it is estimated will matter. The problem with the tsls estimator $\hat{P}_{i,\text{tsls}}$ is that since it includes own observation $T_i$, its estimation error is correlated with $Y_i$ and $T_i$. As a result, the numerator and the denominator in (22) no longer converge to $E[\hat{P}_i Y_i]$ and $E[\hat{P}_i T_i]$. To see this, let $V_{1,i} = Y_i - R_i$ and $V_{2,i} = T_i - \hat{P}_i$ denote errors in the reduced form (1)–(2), and let $K/n \to \kappa > 0$, so that the average number of cases per judge converges to $1/\kappa < \infty$. Then we can write $\hat{P}_{i,\text{tsls}} = P_i + J_{Q_i}^{-1} \sum_j Q_i = Q_i V_{2,j}$. We have

$$\frac{1}{n} \sum_i \hat{P}_{i,\text{tsls}} Y_i = \frac{1}{n} \sum_i \hat{T}_{i,\text{tsls}} Y_i - \frac{1}{n^2} \sum_i \sum_j T_j Y_i$$

$$= \frac{K}{n} \left( \frac{1}{K} \sum_{jk} \frac{1}{J_k} \sum_{Q_i = k} V_{2,j} \sum_{Q_i = k} Y_i \right) + \left( \frac{1}{n} \sum_i P_i Y_i \right) - \left( \frac{1}{n} \sum_j T_j \right) \left( \frac{1}{n} \sum_i Y_i \right) \quad (23)$$

where the last line follows from the law of large numbers applied to all four expressions in parentheses, and the fact that $E[\sum_{Q_i = k} V_{2,j} \sum_{Q_i = k} Y_i] = E[V_{2,j} Y_i] = E[V_{2,i} V_{1,i}]$. Similarly, for the denominator, $n^{-1} \sum_i \hat{P}_{i,\text{tsls}} Y_i \xrightarrow{p} \kappa \text{var}(V_{2,i}) + E[T_i \hat{P}_i]$. Therefore, tsls is inconsistent for its
target, \( \mathbb{E}[\tilde{P}_i Y_i] / \mathbb{E}[\tilde{P}_i T_i] \).

There are two basic ways of adjusting the \textsc{tsls} estimator to make it work under many instruments. First is to estimate the unconditional covariance matrix of \( V_i = (V_{1i}, V_{2i}) \) and subtract an estimate of the bias. This is exactly the idea behind the bias-corrected two-stage least squares estimator of Nagar (1959) and Donald and Newey (2001). Unfortunately, the estimator of the bias is only consistent under homoscedasticity (Bekker and van der Ploeg, 2005; Ackerberg and Devereux, 2009), and it is unclear how to estimate \( \text{var}(V_i) \) consistently when \( \text{var}(V_i \mid Q_i, X_i) \) is heteroscedastic. We use in Chetty, Friedman, Hilger, Saez, Schanzenbach and Yagan (2011).

The estimation error \( P_i - \hat{P}_i \) is given by

\[
\hat{P}_{i,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - n^{-1} \sum_{j=1}^{n} \hat{T}_{j,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - n^{-1} \sum_{j=1}^{n} T_j.
\]

The estimation error \( P_i - \hat{P}_i, \text{JIVE} = \sum_j Q_j = Q_i, j \neq i V_j \) is no longer correlated with \( T_i \) or \( Y_i \), and the \text{JIVE} estimator is consistent for a convex combination of \textsc{lates} under both types of asymptotics.

So far, the discussion has abstracted from the presence of covariates. However, in practice judges are only randomly assigned at the county level. Therefore, with data from several counties, we need to include county indicators (sometimes called “fixed effects”) as covariates. Hence, with \( L \) counties, \( \text{dim}(W_i) = L \), and \( W_{i\ell} = \mathbb{1}_{X_i = \ell} \), where \( X_i \) denotes the county of individual \( i \). The propensity score \( P_i \) is no longer correlated with the incarceration propensity of judge \( Q_i \). However, we now have \( \hat{P}_{i} = \hat{P}_i - \mathbb{E}[P_i \mid X_i] \), so that \( \hat{P}_i \) measures how strict judge \( Q_i \) is compared to other judges that individual \( i \) could have been assigned in the county. The \text{JIVE} estimator of \( P_i \) now becomes

\[
\hat{P}_{i,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - C_{X_i}^{-1} \sum_{j: X_j = X_i} \hat{T}_{j,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - C_{X_i}^{-1} \sum_{j: X_j = X_i} T_j,
\]

where \( C_\ell \) is the number of cases in county \( \ell \). With a large number of counties, a natural way of thinking about the sampling is to let the number of counties \( L \to \infty \), while keeping the number of judges per county and the number of cases per judge fixed. This is similar to the many instrument asymptotics in that the number of judges increases in proportion to the sample size, \( K/n \to \kappa > 0 \), except that instead of keeping the number of counties fixed, we also let them to grow in proportion with sample size, so that \( L/n \to \lambda \). This modification of the many instrument asymptotic sequence was proposed by Anatolyev (2011) and Kolesár \textit{et al.} (2011), and it is also used in Chetty, Friedman, Hilger, Saez, Schanzenbach and Yagan (2011).
Under these asymptotics, the jive estimator is biased. The problem is not its estimate of the propensity score—we still have that $n^{-1} \sum_i \hat{T}_{i,jive} Y_i \overset{p}{\rightarrow} E[P_i Y_i]$, and $n^{-1} \sum_i \hat{T}_{i,jive} T_i \overset{p}{\rightarrow} E[P_i T_i]$. Instead, the source of bias comes from its estimate of the average strictness of judges in county $X_i$, $C^{-1}_{X_i} \sum_j: x_j=x_i T_j$. By the same logic as in the case of $\text{tsls}$ with many instruments, the problem is that this estimate includes own observation $T_i$, so that the estimation error $E[P_i | X_i] - C^{-1}_{X_i} \sum_j: x_j=x_i T_j$ is correlated with $Y_i$ and $T_i$. By arguments similar to those used to derive Equation (23), we have

$$\hat{P}_{i,jive} = \frac{n^{-1} \sum_i \hat{T}_{i,jive} Y_i}{n^{-1} \sum_i \hat{T}_{i,jive} T_i} \overset{p}{\rightarrow} \frac{E[\hat{P}_i Y_i] - \lambda \text{cov}(V_{1,i}, V_{2,i})}{E[\hat{P}_i T_i] - \lambda \text{var}(V_{2,i})}. \quad (25)$$

This probability limit may differ substantially from the target $E[\hat{P}_i Y_i] / E[\hat{P}_i T_i]$, especially in settings in which the Rothenberg (1984) concentration parameter $E[\hat{P}_i T_i] / \text{var}(V_{2,i}) = \pi_2 E[Z_i \hat{Z}_i'] \pi_2 / \text{var}(V_{2,i})$ is small. As a result, jive to be severely biased in finite samples as demonstrated by Ackerberg and Devereux (2009).

The unbiased jackknife instrumental variables estimator (ujive) that I propose solves the bias problem of jive by also leaving out own observation when estimating $E[P_i | X_i]$:

$$\hat{P}_{i,ujive} = \hat{T}_{i,jive} - \frac{1}{C_{X_i} - 1} \sum_{j: x_j=x_i, j \neq i} T_j.$$  

Intuitively, $\hat{P}_i$ is a sample measure of how strict judge $Q_i$ is relative to other judges in country $X_i$ in a sample that excludes individual $i$. This estimator of $\hat{P}_i$ was first used in Chetty et al. (2011) in a setting with the same formal structure as the current example. In particular, Chetty et al. (2011) used classroom indicators as instruments for test score, conditioning on schools. The next subsection gives a general formula for ujive, and proves that it is consistent under many instrument asymptotics that also allow for many covariates.

6.B Consistency of UJIVE under many instruments

Consider now the general case. Let $\phi$ denote the coefficient on $W_i$ in the linear projection $E^*[T_i | W_i]$. To define ujive, decompose $\hat{P}_i^L$, the linear approximation to the propensity score with the effect of covariates partialled out, as

$$\hat{P}_i^L = E^*[T_i | Z_i, W_i] - E^*[T_i | W_i] = Z_i' \pi_2 + W_i' \psi_2 - W_i' \phi.$$

Let $\hat{\pi}_2,i$ and $\hat{\psi}_2,i$ be the least-squares estimates of $\pi_2$ and $\psi_2$ based on a sample with observation $i$ removed. Similarly, let $\hat{\phi}_i$ be the least-squares estimate of $\phi$ based on a sample with observation $i$ removed. The ujive estimator is a two-step iv estimator with the first-step estimator of $\hat{P}_i^L$ given
where \( \hat{\beta}_{\text{ijve}} = Z_i' \hat{\gamma}_2(i) + W_i' \hat{\gamma}_1(i) - W_i' \hat{\phi}_i \).

In matrix notation
\[
\hat{P}_{\text{ijve}} = \hat{T}_{\text{ijve}} - (I_n - D_W)^{-1}(H_W - D_W)T,
\]
where \( \hat{T}_{\text{ijve}} = (I_n - D_{(Z,W)})^{-1}(H_{(Z,W)} - D_{(Z,W)})T \). Using \( \hat{P}_{\text{ijve}} \) as a single instrument in an IV estimator then yields
\[
\hat{\beta}_{\text{ijve}} = \frac{\hat{P}_{\text{ijve}}Y}{\hat{P}_{\text{ijve}}T}.
\]

In contrast, while the jive estimator of \( \mathbb{E}^*[Y_i \mid Z_i, W_i] \) is identical to \( \hat{T}_{\text{ijve}} \), its estimator of \( \mathbb{E}^*[Y_i \mid W_i] \) is given by a sample projection of \( \hat{T}_{\text{ijve}} \) onto \( W \), so that \( \hat{P}_{\text{jive}} = \hat{T}_{\text{ijve}} - H_W \hat{T}_{\text{ijve}} \) (see the jive formula on page 10).

To formally define the many instrument asymptotic framework, I need to allow the distribution of random variables to change with the sample size. To reflect this, let the random variables be indexed by \( n \), so that, for instance, \( Y_n = (Y_{n,1}, \ldots, Y_{n,n})' \) denotes the vector of observed outcomes when the sample size is \( n \). In addition, let \( P^X_{n,i} = \mathbb{E}[T_{n,i} \mid X_{n,i}] \) and \( R^X_{n,i} = \mathbb{E}[Y_{n,i} \mid X_{n,i}] \) denote the expectations of \( T_{n,i} \) and \( Y_{n,i} \) conditional on \( X_{n,i} \) only, so that \( P^X_{n,i} = \mathbb{E}[P_{n,i} \mid X_{n,i}] \) and \( R^X_{n,i} = \mathbb{E}[R_{n,i} \mid X_{n,i}] \). The many instrument asymptotic framework I consider is summarized by the following assumptions:

**Assumption R (Regularity conditions)**

(i) \( \{ (Y_{n,i}, T_{n,i}, X_{n,i}, Q_{n,i}) : i = 1, \ldots, n \} \) is a triangular array of i.i.d. random variables, the \( n \)th row having distribution \( F_{n}^{Y,T,X,Q} \). \( F_{n}^{Y,T,X,Q} \) converges in distribution to \( F^{Y,T,X,Q} \);

(ii) There is a positive constant \( C_1 \), such that \( \sup_n \sup_{i \leq n} \text{var}(Y_{n,i} \mid Q_{n,i}, X_{n,i}) \leq C_1 \), and \( \sup_n \sup_{i \leq n} \text{var}(Y_{n,i} \mid X_{n,i}) \leq C_1 \) a.s. Also, as \( n \to \infty \),

\[
\mathbb{E}[(R^2_{n,i}, P^2_{n,i})] \to \mathbb{E}[(R^2, P^2, |RP|) < \infty,
\]

\[
\mathbb{E}[(R^X_{n,i}, P^X_{n,i})] \to \mathbb{E}[(R^X, P^X, |R^X P^X|) < \infty,
\]

where \( (R, P, R^X, P^X) \) is distributed according to the limiting distribution \( F^{R,P,R^X,P^X} \); and

(iii) \( \text{rank}(Z_n, W_n) = K + L \) and \( (H_{(Z_n, W_n)})_{ii} < C_2 \) for some \( C_2 < 1 \) a.s., where \( Z_{n,i} = z(Q_{n,i}, X_{n,i}) \) and \( W_{n,i} = w(X_{n,i}) \), with \( \dim(Z_{n,i}) = K \) and \( \dim(W_{n,i}) = L \). The functions \( z \) and \( w \) may depend on \( n \).

**Assumption MI (Many instruments)** As \( n \to \infty \):

(i) \( K/n \to \kappa \) and \( L/n \to \lambda \) for some \( \kappa, \lambda \geq 0 \);

(ii) \( \sum_i [\mathbb{E}[T_{n,i} \mid X_{n,i}] - \mathbb{E}^*[T_{n,i} \mid W_{n,i}]^2]/n \to 0 \) a.s.; and

(iii) \( \sum_i [\mathbb{E}[T_{n,i} \mid Q_{n,i}, X_{n,i}] - \mathbb{E}^*[T_{n,i} \mid Z_{n,i}, W_{n,i}]^2]/n \to 0 \) a.s.
Assumption $R$ (i) allows the distribution of the data to change with the sample size, converging to some limiting distribution $F_{Y,T,X,Q}$. Part (ii) requires that the second moments of conditional expectations of $Y_{n,i}$ and $T_{n,i}$ exist and are well-behaved in the limit. It is necessary for sample averages such as $n^{-1} \sum_{i=1}^{n} R_{n,i}^2$, to have a well-specified probability limit. The restriction $\text{rank}(Z,W) = K + L$ in Part (iii) is a normalization. The assumption that $(H_{i(Z,W)})_{ii} < C_2$ requires that no single observation has too much leverage. It implies that $(K + L)/n < C_2$ since $n^{-1} \sum_i (H_{i(Z,W)})_{ii} = (K + L)/n$.

Assumption $MI$ (i) generalizes the many instrument asymptotic sequence by also allowing the number of covariates to increase with the sample size. In terms of the incarceration example, the original Bekker (1994) many instruments sequence keeps the number of counties as well as the number of cases per judge fixed, and lets the number of judges per county increase to infinity. Under Assumption $MI$, we can think of generating the data by sampling $L$ counties form some large population of counties. In Angrist and Krueger (1991), where $Z_i$ is generated by interacting quarter of birth with $L$ state of birth and year of birth indicators, Assumption $MI$ (i) lets the number of states and years $L \to \infty$, while keeping the number of individuals observed in each state and year fixed. Finally, Assumption $MI$ also accommodates models in which $z$ and $w$ are some approximating functions, such as splines or polynomials in the basic instruments and covariates $Q_i$ and $X_i$. This corresponds to fixing the distribution of the data, so that $F_{n,Y,T,X,Q} = F_{Y,T,X,Q}$, and letting the number of terms in the approximating functions $w$ and $z$ increase with the sample size. Parts (ii)–(iii) then require that these approximating functions get to their population targets in the limit, and allow me to relax the requirement imposed by Assumption $L$ that expectation of $Z_i$ conditional on $X_i$ is exactly linear in $W_i$ in the sample. These conditions are similar to the assumptions in Bekker (1994) and Hansen, Hausman and Newey (2008).

Note that I do not make any assumptions about the coefficients on $Z_{n,i}$ and $W_{n,i}$ in the projections $\mathbb{E}^*[T_{n,i} | W_{n,i}, Z_{n,i}]$ and $\mathbb{E}^*[T_{n,i} | W_{n,i}]$. Under additional assumptions, such as sparsity (only few coefficients in these linear projections matter), approximations to $\hat{P}_{n,i}$ other than $\hat{P}_{n,i,UIJVE}$ will work (see, for example, Belloni, Chen, Chernozhukov and Hansen, 2012).

**Theorem 2** Suppose that Assumptions $R$ and $MI$ hold, and that the limiting distribution $F_{Y,T,X,Q}$ of the data satisfies Assumptions $IV$, and $M$. Then:

$$
\hat{\theta}_{UIJVE} \to_{P} \frac{\mathbb{E}[Y(P - \mathbb{E}[P | X])]}{\mathbb{E}[T(P - \mathbb{E}[P | X])]} = \int \sum_{j=1}^{L} \frac{\theta_j(x)}{\sum_{j=1}^{L} \theta_j(x)} \mathbb{P}(P_{j|x} | X) \ dF^X(x),
$$

where $(Y, T, P, X)$ are distributed according to the limiting distribution $F_{Y,T,P,X}$, and

$$
\theta_j(x) = (P_{j+1|x} - P_{j,x}) \mathbb{P}(P > P_{j,x} | X = x) \left( \mathbb{E}[P | X = x, P > P_{j,x}] - \mathbb{E}[P | X = x] \right).
$$
Thus, ujive estimates a convex combination of local average treatment effects. This conclusion is robust to many instruments, many covariates, and heteroscedasticity.

7 Conclusion

In this paper, I derived estimands of estimators based on a classic linear IV model under treatment effect heterogeneity. I assumed that the instruments satisfy the monotonicity condition of Angrist and Imbens (1995), so that for each pair of instrument values, we can identify a local average treatment effect (LATE). If the LATES for all possible instrument pairs are all equal to each other, then all classic estimators estimate this common local average treatment effect. If the LATES vary, then, under mild assumptions, estimators in the class of two-step IV estimators estimate the same convex combination of them. This class includes the two-stage least squares estimator (TSLS). The estimand of LIML, however, is different, depends on the reduced-form covariance matrix, and may be outside of the convex hull of the local average treatment effects. This possibility makes LIML unattractive in settings with treatment effect heterogeneity.

Unfortunately, the TSLS estimator is inconsistent under many instrument asymptotics, making it a poor choice of estimator in settings with a large number of instruments. I showed that a different two-step IV estimator, the unbiased jackknife IV estimator (UJIVE), on the other hand, remains consistent for a convex combination of LATEs under a many instrument asymptotic sequence that allows for heteroscedasticity, and lets the number of instruments and covariates increase in proportion with the sample size. I therefore recommend that in settings with many instruments, empirical researchers use UJIVE instead of LIML or TSLS.
Appendix A Auxiliary Lemmata

First I define some notation and collect some basic results that I use throughout Appendices A and B. I use the notation \(|\cdot|_F\) to denote the Frobenius norm, so that for any matrix \(A\), 
\[|A|_F = \sqrt{\text{tr}(AA^T)}.\]
By the Cauchy-Schwarz inequality, we then have \(\text{tr}(A'B) \leq |A||B|_F\). The Frobenius norm is sub-multiplicative, so that \(|AB|_F \leq |A||B|_F\) for any matrices \(A\) and \(B\). Let \(\mathcal{Z}_n = \{Q_i, X_i\}_{i=1}^n\) denote the collection of covariates and instruments. Also, let

\[
G_n = (I_n - D_{(z_n,w_n)})^{-1}(H_{(z_n,w_n)} - D_{(z_n,w_n)}) - (I_n - D_{w_n})^{-1}(H_{w_n} - D_{w_n}).
\]

Then \(\hat{P}^{\text{IVE}} = G_n T_n\).

**Lemma 2** Let \(\hat{P}_i = \mathbb{E}^*[T_i \mid Z_i, W_i] - \mathbb{E}^*[T_i \mid W_i]\), and let \(\tilde{R}_i = \mathbb{E}*[Y_i \mid Z_i, W_i] - \mathbb{E}*[Y_i \mid W_i]\). Then:

(i) \(\Xi_{12} = \mathbb{E}[Y_i \hat{P}_i]\) (ii) \(\Xi_{12} = \mathbb{E}[T_i \tilde{R}_i]\) (iii) \(\Xi_{11} = \mathbb{E}[Y_i \tilde{R}_i]\) (iv) \(\Xi_{22} = \mathbb{E}[T_i \hat{P}_i]\)

**Proof.** Consider Part (i). Observe that 
\[
\mathbb{E}[Y_i \hat{P}_i] = \mathbb{E}[(Z_i' \pi_1 + W_i' \phi_1)' \hat{P}_i] \\
= \mathbb{E}[(Z_i' \pi_1 + W_i' \phi_1)'Z_i' \pi_2] \\
= \mathbb{E}[\pi_1 Z_i Z_i \pi_2] = \mathbb{E}[\pi_1 Z_i Z_i \pi_2] \\
= \mathbb{E}[T_i \tilde{R}_i]
\]

where the first line follows from Equation (3) and the fact that \(\hat{P}_i\) is linear in \(Z_i\) and \(W_i\), the second line follows from \(\hat{P}_i = Z_i' \pi_1\), the third line follows from \(\mathbb{E}[W_i Z_i] = 0\), and the last line follows by definition of \(\Xi_{12}\). Parts (ii)-(iv) follow by similar arguments, using the substitutions \(\tilde{R}_i = Z_i' \pi_1\) and \(\hat{P}_i = Z_i' \pi_2\). \(\square\)

**Lemma 3** Let \(A_i = a(Q_i, X_i)\) be some function of the instruments and covariates such that \(\mathbb{E}[A_i \mid X_i] = 0\). Then, under Assumptions IV and M

\[
\mathbb{E}[Y_i A_i] = \int \sum_{j=1}^{j-1} \mathbb{E}[A_i \mid X_i = x, P_i > p_{j,x}] P(P_i > p_{j,x} \mid X_i = x) \, dF_X(x),
\]

\[
\mathbb{E}[T_i A_i] = \int \sum_{j=1}^{j-1} (p_{j,x} - p_{j-1,x}) \mathbb{E}[A_i \mid X_i = x, P_i > p_{j,x}] P(P_i > p_{j,x} \mid X_i = x) \, dF_X(x).
\]

**Proof.** First consider \(\mathbb{E}[Y_i A_i]\). By the Law of iterated expectations,

\[
\mathbb{E}[Y_i A_i] = \int \sum_a \mathbb{E}[Y_i \mid X_i = x, A_i = a, P_i = p_{j,x}] P(P_i = p_{j,x}, A_i = a \mid X_i = x) \, dF_X(x)
\]

\[
= \int \sum a \mathbb{E}[Y_i \mid X_i = x, P_i = p_{j,x}] P(P_i = p_{j,x}, A_i = a \mid X_i = x) \, dF_X(x)
\]

where the second line follows from the fact that under Assumptions IV and M, the conditional expectation of \(Y_i\)
depends only on $P_i$ and $X_i$. Using the substitution

$$E[Y_i \mid X_i = x, P_i = p_{j,x}] = E[Y_i \mid X_i = x, P_i = p_{1,x}] + \sum_{j=1}^{j-1} a(p_{j,x}; x)(p_{j+1,x} - p_{j,x}),$$

we can expand the expression (27) as

$$E[Y_i A_i] = \int \sum_j \sum_a \sum_{j=1}^{j-1} a(p_{j,x}; x)(p_{j+1,x} - p_{j,x}) \mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x) dF_X(x) + \int E[A_i \mid X_i = x] E[Y_i \mid X_i = x, P_i = p_{1,x}] dF_X(x)$$

$$= \int \sum_j \sum_a \sum_{j=1}^{j-1} a(p_{j,x}; x)(p_{j+1,x} - p_{j,x}) \mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x) dF_X(x)$$

$$= \int \sum_{j=1}^{j-1} a(p_{j,x}; x)(p_{j+1,x} - p_{j,x}) \sum_a \mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x) dF_X(x),$$

where the second line follows since $E[A_i \mid X_i] = 0$ by assumption, and the last line follows from changing the order of summation. Therefore, by definition of conditional expectation:

$$E[Y_i A_i] = \int \sum_{j=1}^{j-1} a(p_{j,x}; x)(p_{j+1,x} - p_{j,x}) \sum_{j=1}^{j-1} \mathbb{E}[A_i \mid P_i = p_{j,x}, X_i = x] \mathbb{P}(P_i = p_{j,x} \mid X_i = x) dF_X(x)$$

$$= \int \sum_{j=1}^{j-1} a(p_{j,x}; x)(p_{j+1,x} - p_{j,x}) \mathbb{E}[A_i \mid X_i = x, P_i > p_{j,x}] \mathbb{P}(P_i > p_{j,x} \mid X_i = x) dF_X(x).$$

The expression for $E[T_i A_i]$ can be derived using the same arguments, except that we substitute

$$E[T_i \mid X_i = x, P_i = p_{j,x}] = p_{1,x} + \sum_{j=1}^{j-1} (p_{j+1,x} - p_{j,x}).$$

I use the following results from Chao, Swanson, Hausman, Newey and Woutersen (2012) and Politis, Romano and Wolf (1999) to prove Lemma 6 and Lemma 7 below:

**Lemma 4 (Chao et al., 2012, Lemma A.1)** Suppose that, conditional on some set of random variables $\mathcal{F}, \{\{A_i, B_i\}_{i=1}^{n}\}$ is independent a.s., where $A_i$ and $B_i$ are some scalars random variables. Let $H$ be a symmetric idempotent matrix with rank $K$. Let $\mathbb{E}[A_i \mid \mathcal{F}] = \tilde{a}_i$, $\mathbb{E}[B_i \mid \mathcal{F}] = \tilde{b}_i$, and $\sigma_A^2 = \max_{i \leq n} \text{var}(A_i \mid \mathcal{F})$, $\sigma_B^2 = \max_{i \leq n} \text{var}(B_i \mid \mathcal{F})$. Then there exists a positive constant $C$ such that

$$\mathbb{E}\left[\left(\sum_{i,j \neq i} (A_i H_{ij} B_j - \tilde{a}_i H_{ij} \tilde{b}_j)\right)^2 \mid \mathcal{F}\right] \leq C(K \sigma_A^2 \sigma_B^2 + \sigma_A^2 \tilde{b}^T \tilde{b} + \sigma_B^2 \tilde{a}^T \tilde{a}).$$

**Lemma 5 (Lemma 1.3.2., Politis et al., 1999)** Suppose $(A_{n,1}, \ldots, A_{n,n})$ is a triangular array of i.i.d. random variables, the $n$th row having distribution $F_n^A$, Assume $F_n^A$ converges in distribution to $F^A$, and $\mathbb{E}[|A_{n,1}|] \rightarrow \mathbb{E}[|A|] < \infty$, as $n \rightarrow \infty$, where $A$ is distributed according to $F^A$. Then $n^{-1} \sum_{i=1}^n A_{n,i} \rightarrow \mathbb{E}[A]$ as $n \rightarrow \infty$.

**Lemma 6** Suppose that Assumptions R and MI hold. Then
(i) \( T_n'G_nT_n/n = P_n'G_nP_n/n + o_p(1) \); and
(ii) \( Y_n'G_nT_n/n = R_n'G_nP_n/n + o_p(1) \),
where \( G_n \) is defined in Equation (26).

**Proof.** I will prove Part (i), Part (ii) follows by similar arguments. Let \( A_{n,i}^W = (1 - (H_{n,i})_{ii})^{-1}T_n \), and let \( A_{n,i}^{Z,W} = (1 - (H_{Z,n,W})_{ii})^{-1}T_n \), and denote by \( P_{n,i} = (1 - (H_{n,i})_{ii})^{-1}P_n \) and \( P_{n,i}^{Z,W} = (1 - (H_{Z,n,W})_{ii})^{-1}P_n \) their expectations conditional on \( Z_n \). Note that since 0 \( \leq \) \( P_{n,i} \leq 1 \), it follows that \( \text{var}(P_{n,i} | Z_n) \leq 1 \). Then we can write:

\[
T_n'G_nT_n - P_n'G_nP_n = \sum_{i} \sum_{j \neq i} \left( A_{n,i}^W (H_{n,i,W})_{ji} T_j - a_{n,i}^W (H_{n,i,W})_{ji} P_i \right) + \sum_{i} \sum_{j \neq i} \left( A_{n,i}^{W} (H_{n,i,W})_{ji} T_j - a_{n,i}^{W} (H_{n,i,W})_{ji} P_i \right).
\]

Therefore, obtain that

\[
\mathbb{E}[(T_n'G_nT_n - P_n'G_nP_n)^2/n^2 | Z_n] \leq \mathbb{E} \left[ \left( \sum_{i} \sum_{j \neq i} \left( A_{n,i}^{Z,W} (H_{Z,n,W})_{ji} T_j - a_{n,i}^{Z,W} (H_{Z,n,W})_{ji} P_i \right) / n \right) \right]^2 | Z_n]
\]

\[
+ \mathbb{E} \left[ \left( \sum_{i} \sum_{j \neq i} \left( A_{n,i}^{W} (H_{n,i,W})_{ji} T_j - a_{n,i}^{W} (H_{n,i,W})_{ji} P_i \right) / n \right) \right]^2 | Z_n]
\]

\[
\leq C \frac{n^2}{n^2} \left[ \frac{(K + L)}{(1 - C_2)^2} + \frac{2}{(1 - C_2)^2} \right] + \frac{C}{n^2} \left[ \frac{L}{(1 - C_2)^2} + \frac{2}{(1 - C_2)^2} \right] P_n^2 P_n,
\]

where the first line follows from triangle inequality, and the second line follows from applying Lemma 4 with \( F = Z_n \), and the implication of Assumption R that

\[
\sup_{i \leq n} \sup_{n} \text{var} \left( A_{n,i}^{Z,W} | Z_n \right) \leq \frac{1}{(1 - C_2)^2}, \quad \sup_{i \leq n} \text{var} \left( A_{n,i}^{W} | Z_n \right) \leq \frac{1}{(1 - C_2)^2}.
\]

Next, by Assumption R, we can apply the law of large numbers given in Lemma 5 to \( n^{-1} \sum_{i=1}^n p_{n,i}^2 \), to get \( n^{-1} \sum_{i=1}^n p_{n,i}^2 = \mathbb{E}[P^2] + o_p(1) = O_p(1) \). Also, \( (K + L)/n^2 = o(1) \) by Assumption MI, so that

\[
\mathbb{E}[(T_n'G_nT_n - P_n'G_nP_n)^2/n^2 | Z_n] \leq o_p(1).
\]

Therefore, by Markov inequality and the dominated convergence theorem,

\[
T_n'G_nT_n/n = P_n'G_nP_n/n + o_p(1),
\]

which proves assertion (i). \( \Box \)

**Lemma 7** Suppose Assumptions R and MI hold. Let \( (Y, T, P, R) \) be distributed according to the limiting distribution \( F_{Y,T,R,P} \). Then

(i) \( P_n'G_nP_n/n = \mathbb{E}[T(P - \mathbb{E}[P | X])] + o_p(1) \); and

(ii) \( R_n'G_nP_n/n = \mathbb{E}[Y(P - \mathbb{E}[P | X])] + o_p(1) \),
where \( G_n \) is defined in Equation (26).

29
Proof. Again, I will only prove Part (i), Part (ii) follows by similar arguments. To this end, write \( P_n' G_n P_n / n \) as
\[
P_n'(I_n - (I_n - D(Z_n, W_n))^{-1} M(Z_n, W_n) - P_n'(I_n - D W_n)^{-1} (H W_n - D W_n) P_n / n.
\]
I will prove the assertion in two steps. First, I will prove that
\[
P_n'(I_n - (I_n - D(Z_n, W_n))^{-1} M(Z_n, W_n) P_n / n = E[TP] + o_p(1).
\]
(29)
Second, I will prove that
\[
P_n'(I_n - D W_n)^{-1} (H W_n - D W_n) P_n / n = E[TE[P | X]] + o_p(1).
\]
(30)
Combining (29) with (30) then yields the result.

To prove (29), note that
\[
\sum_{i} (I_n - (I_n - D(Z_n, W_n))^{-1} M(Z_n, W_n) P_n / n)
\]
and
\[
\sum_{i} (I_n - D W_n)^{-1} (H W_n - D W_n) P_n / n)
\]
where the first equality follows from \( M(Z_n, W_n) P_n = 0 \), the second equality follows from the definition of Frobenius norm, and the last line follows from the fact that \( H(Z_n, W_n) \) is positive semi-definite so that \( tr(H(Z_n, W_n) (P_n - P_n') (P_n - P_n')') / n \geq 0 \) and Assumption MI. Therefore, we obtain
\[
\frac{n}{2} \left( \sum_{i} (I_n - (I_n - D(Z_n, W_n))^{-1} M(Z_n, W_n) P_n / n) \right)^2 
\]
and
\[
\frac{n}{2} \left( \sum_{i} (I_n - D W_n)^{-1} (H W_n - D W_n) P_n / n) \right)^2
\]
where the first line follows since the Frobenius norm is sub-multiplicative, the second line follows from the definition of Frobenius norm, the third line follows from the result (31) and Assumption R, and the last line follows from applying the law of large numbers given in Lemma 5 to \( n^{-1} \sum_{i=1}^{n} p_{n,i}^2 \). Therefore, we obtain
\[
P_n'(I_n - (I_n - D(Z_n, W_n))^{-1} M(Z_n, W_n) P_n / n = P_n' P_n / n + o_p(1).
\]
Since by Assumption R and Lemma 5, \( n^{-1} \sum_{i=1}^{n} p_{n,i}^2 \rightarrow E[P^2] = E[TP] \), assertion (29) follows.

Now I prove assertion (30). Let \( A_{n,i} = (1 - (H W_n))^{-1} P_{n,i} \), and denote by \( a_{n,i} = (1 - (H W_n))^{-1} P_{n,i} \), its expectation conditional on \( \{X_{n,i}\}_{i=1}^{n} \), where \( P_{n,i} = E[P_{n,i} | X_{n,i}] \). Note that since \( 0 \leq P_{n,i} \leq 1 \), it follows that \( \text{var}(P_{n,i} | X_{n,i}) \leq 1 \). Therefore, applying Lemma 4 with \( F = \{X_{n,i}\}_{i=1}^{n} \), we obtain
\[
E \left[ \left( \sum_{i \neq j} A_{n,i} (H W_n)_{ij} P_{n,j} - a_{n,j} (H W_n)_{ij} E[P_{n,j} | W_{n,i}] \right) / n \right]^2 \right\}_{i=1}^{n} \leq \frac{C}{n^2} \left( L \frac{1}{(1 - C)^2} + \frac{2}{(1 - C)^2} \sum_{i} (E[P_{n,i} | X_{n,i}])^2 \right)
\]
\[
\leq \frac{C}{n^2} \left( L \frac{1}{(1 - C)^2} + \frac{2n}{(1 - C)^2} \right) = o_p(1),
\]
30
where the first line follows from the implication of Assumption $R$, $(1 - (H_{ij})^{-1}) \leq 1/(1 - C_2)$, the second line follows from $|P_{ij}| \leq 1$, and the last line follows from $L \leq n$. It therefore follows by Markov inequality and the dominated convergence theorem that
\[
P'_n(I_n - D_{W_i})^{-1}(H_{W_i} - D_{W_i})P_n/n = (P_n^X)'(I_n - D_{W_i})^{-1}(H_{W_i} - D_{W_i})P_n^X/n + o_p(1)
\]
where $P_n^X$ is an $n$-vector with $i$th element given by $P_{n,i}^X$. Let $P^X_{n,i} = E^*[P_{n,i} | W_{n,i}]$. Now, by arguments as in Equations (31) and (32) with $P_n$ replaced by $P_n^X$ and $P_n'$ replaced by $P^X_n$, we have that:
\[
\| (P_n^X)'(I_n - D_{W_i})^{-1}M_{W_i}P_n^X/n \|_F = o_p(1).
\]
Since $(I_n - D_{W_i})^{-1}(H_{W_i} - D_{W_i}) = I_n - (I_n - D_{W_i})^{-1}M_{W_i}$, it follows that
\[
P'_n(I_n - D_{W_i})^{-1}(H_{W_i} - D_{W_i})P_n/n = n^{-1} \sum_i E[P_{n,i} | X_{n,i}]E[P_{n,i} | X_{n,i}] + o_p(1).
\]
By Assumption $R$, we can apply Lemma 5 to obtain
\[
P'_n(I_n - D_{W_i})^{-1}(H_{W_i} - D_{W_i})P_n/n = E \left[ E[P | X]^2 \right] + o_p(1)
\]
which prove assertion (30).

\[\square\]

Appendix B    Proofs

**Proof of Lemma 1.** First consider part (i). Since $E[\hat{P}^L_{1,n}] \neq 0$, it follows by the continuous mapping theorem that $\hat{P} = E[\hat{P}^L_{1,n}] / E[H^L_{1,n}]$. Part (i) then follows by Lemma 2.

Next, consider Part (ii). The reverse two-stage least squares estimator can be written as:
\[
\hat{\beta}_{rtsls} = \frac{\tilde{\beta}'_1 (Z'_1 Z'_1/n) \hat{\beta}_1}{\tilde{\beta}'_2 (Z'_1 Z'_1/n) \hat{\beta}_2},
\]
where $\hat{\beta}_1$ and $\hat{\beta}_2$ are least-squares estimators of $\pi_1$ and $\pi_2$ based on Equations (3)–(4). Since $E[(Z_i, W_i')(Z_i, W_i)]'$ is full rank, and the data is iid with finite second moments, these least-squares estimators are consistent for $\pi_1$ and $\pi_2$. Also, by the law of large numbers, the fact that $E[W_i W_i']$ is full rank, and the continuous mapping theorem:
\[
Z'_1 Z'_1/n = Z'_1 Z'_1/n - (Z'_1 W'_1/n) (W'_1 W'_1/n)^{-1} (W'_1 Z'_1/n)
\]
\[
\xrightarrow{p} E[Z'_1 Z'_1/n] - E[Z'_1 W'_1/n] (E[W_i W_i'])^{-1} E[W_i Z'_1/n] = E[\tilde{Z}'_1 \tilde{Z}'_1/n].
\]
Hence
\[
\hat{\beta}_{rtsls} \xrightarrow{p} \frac{\pi_1 E[\tilde{Z}'_1 \tilde{Z}'_1/n]}{\pi_2 E[\tilde{Z}'_1 \tilde{Z}'_1/n]}. \frac{\pi_1 E[\tilde{Z}'_1 \tilde{Z}'_1/n]}{\pi_2 E[\tilde{Z}'_1 \tilde{Z}'_1/n]},
\]
which, combined with Lemma 2, proves the assertion.

Finally, too prove Part (iii), it suffices to show that
\[
\min \text{eig}(\hat{S}^{-1} \hat{\Sigma}) \xrightarrow{p} \min \text{eig}(S^{-1} \Sigma), \quad (33)
\]
since then $\hat{\beta}_{rtsls} \xrightarrow{p} \beta_{rtsls}$ by the continuous mapping theorem. To show (33), note that $\min \text{eig}(\hat{S}^{-1} \hat{\Sigma})$ is the minimum of
the function
\[ \hat{D}_S(\omega) = \frac{\omega' \hat{\omega}}{\omega' \hat{S} \omega}, \quad \omega \in S^1, \]
where \( S^1 \) denotes the unit circle in \( \mathbb{R}^2 \), a compact space. Therefore, if \( \hat{D}_S(\omega) \) converges uniformly to the limiting function \( D_S(\omega) = \omega' \Xi \omega / (\omega' \hat{S} \omega) \), then \( \min_* \hat{D}_S(\omega) \xrightarrow{p} \min_* D_S(\omega) \) by standard arguments (see, for example Newey and McFadden, 1994). To prove uniform convergence, I will use the arguments in Chao and Swanson (2005). Fix some \( \omega \in S^1 \), and note that:

\[
|D_S(\omega) - D_S(\omega)| = \left| \frac{\omega' \hat{\omega}}{\omega' \hat{S} \omega} - \frac{\omega' \hat{\omega}}{\omega' \hat{S} \omega} \right| = \frac{1}{|\omega' \hat{S} \omega|} |\omega' \hat{\omega} - \omega' \hat{S} \omega|
\]

\[
= \frac{1}{|\omega' \hat{S} \omega|} |\omega' (\hat{T} - S) \omega - D_S(\omega) \omega' (\hat{T} - S) \omega|
\]

\[
\leq \frac{1}{|\omega' \hat{S} \omega|} (|\omega' (\hat{T} - S) \omega| + D_S(\omega) |\omega' (\hat{T} - S) \omega|),
\]
where the first line follows from the definition of \( \hat{D}_S \), the second line follows from the definition of \( D_S \), and the third line follows by triangle inequality. I now bound all three terms in the last expression uniformly in \( \omega \). Since the trace operator is an inner product under Frobenius norm \( \| A \|_F = \sqrt{\text{tr}(A^T A)} \), by Cauchy-Schwarz inequality:

\[
|\omega' (\hat{T} - S) \omega| = |\text{tr} (\omega' (\hat{T} - S) \omega)| \leq \sqrt{\text{tr}(\omega' \omega)} \| \hat{T} - S \|_F
\]

\[
= \| \hat{T} - S \|_F = o_p(1),
\]
where the second line follows from \( \omega' \omega = 1 \) since \( \omega \in S^1 \) and \( \hat{T} \xrightarrow{p} S \) so that \( \| \hat{T} - S \|_F = o_p(1) \). By an identical argument, we also have \( |\omega' (\hat{T} - S) \omega| = o_p(1) \). Finally, to bound \( 1/|\omega' \hat{S} \omega| \), note that since \( \hat{S} \xrightarrow{p} S > 0 \), \( \omega' \hat{S} \omega > 0 \) with probability approaching 1, so that \( 1/|\omega' \hat{S} \omega| < C \) for some \( C < \infty \) with probability approaching 1. Hence:

\[
|D_S(\omega) - D_S(\omega)| \leq o_p(1) + o_p(1)D_S(\omega),
\]
where \( D_S(\omega) \) is bounded by max eig(\( S^{-1} \Xi) \), it follows that \( \sup_\omega |\hat{D}_S(\omega) - D_S(\omega)| = o_p(1) \) as required. 

**Proof of Theorem 1.** By Lemma 1, \( \Xi_{12} = \mathbb{E}[Y_i \hat{P}_i^L] \) and \( \Xi_{22} = \mathbb{E}[T_i \hat{P}_i^L] \). Since by Assumption L, \( \mathbb{E}[\hat{P}_i^L | X_i] = 0 \), we can apply Lemma 3 with \( A_i = \hat{P}_i^L \) to get

\[
\Xi_{12} = \mathbb{E} \left[ \sum_{j=1}^{j-1} a(p_{j,x} | x) (p_{j,x} - p_{j-1,x}) \mathbb{E}[\hat{P}_i | X_i = x, P_{j,x} > p_{j,x} | P_{j,x} > p_{j,x} | X_i = x] \right] dF_X(x)
\]

\[
\Xi_{22} = \mathbb{E} \left[ \sum_{j=1}^{j-1} (p_{j,x} - p_{j-1,x}) \mathbb{E}[\hat{P}_i | X_i = x, P_{j,x} > p_{j,x} | P_{j,x} > p_{j,x} | X_i = x] \right] dF_X(x),
\]
which yields the result for \( \Xi_{12} / \Xi_{22} \).

Second, by Lemma 1, \( \Xi_{12} = \mathbb{E}[T_i \hat{R}_i^L] \) and \( \Xi_{11} = \mathbb{E}[Y_i \hat{R}_i^L] \). Since by Assumption L, \( \mathbb{E}[\hat{R}_i^L | X_i] = 0 \), applying Lemma 3 with \( A_i = \hat{R}_i^L \) yields the result for \( \Xi_{11} / \Xi_{12} \).

**Proof of Corollary 1.** Let \( P_i = p_{j,x} | X_i = x = s_{j,x} \). If the linear approximations (3)–(4) are exact, then \( \hat{P}_i = P_i^L \) and \( \hat{R}_i = R_i^L \). We can therefore write:

\[
\mathbb{E}[R_i^L | P_i = p_{j,x}, X_i = x] = \mathbb{E}[R_i^L | P_i = p_{j,x}, X_i = x] + \sum_{f=1}^{j-1} a(p_{f,x} | x) (p_{f+1,x} - p_{f,x}),
\]

32
so that
\[
\mathbb{E}[R_i^+ \mid X_i = x] = \mathbb{E}[R_i^+ \mid P_i = p_{1,x}, X_i = x] + \sum_{j=1}^{I-1} \sum_{m=j+1}^{I} \alpha(p_{j,x}; x)(p_{f,1,x} - p_{f,x}) s_{m,x},
\]
and
\[
\mathbb{E}[R_i^+ \mid P_i > p_{j,x}, X_i = x] = \mathbb{E}[R_i^+ \mid P_i = p_{1,x}, X_i = x] + \sum_{j=1}^{I-1} \sum_{m=j+1}^{I} \alpha(p_{j,x}; x)(p_{f,1,x} - p_{f,x}) + \sum_{m=j+1}^{I} \sum_{m'=m+1}^{I} \frac{1}{s_{m,x}} s_{m,x}.
\]

Therefore,
\[
\mathbb{E}[\tilde{R}_i^+ \mid P_i > p_{j,x}, X_i = x] \mathbb{P}(P_i > p_{j,x} \mid X_i = x) = \sum_{j=1}^{I-1} \sum_{m=j+1}^{I} \alpha(p_{j,x}; x)(p_{f,1,x} - p_{f,x}) \sum_{m'=m+1}^{I} \frac{1}{s_{m',x}} \sum_{m=1}^{I} s_{m',x} \sum_{m=1}^{I} s_{m,x}
\]
\[
= \sum_{j=1}^{I-1} \sum_{m=j+1}^{I} \alpha(p_{j,x}; x)(p_{f,1,x} - p_{f,x}) \sum_{m'=m+1}^{I} \frac{1}{s_{m',x}} \sum_{m=1}^{I} s_{m',x} \sum_{m=1}^{I} s_{m,x}.
\]

By Theorem 1, we therefore have:
\[
\zeta_j(x) = (p_{j+1,x} - p_{j,x}) \sum_{m=j+1}^{I} \alpha(p_{j,x}; x)(p_{f,1,x} - p_{f,x}) \sum_{m'=m+1}^{I} \frac{1}{s_{m',x}} \sum_{m=1}^{I} s_{m',x} \sum_{m=1}^{I} s_{m,x}
\]
\[
+ (p_{j+1,x} - p_{j,x}) \sum_{m=1}^{I} \frac{1}{s_{m,x}} s_{m,x}.
\]

If \( \alpha(p_{j,x}; x) \geq 0 \) for all \( j \) and \( x \), then all the terms in this expression are non-negative, so that \( \zeta_j(x) \) is non-negative.

To obtain the expressions for \( \theta_1(x) \) and \( \theta_2(x) \) in the special case that \( J_x = 2 \), note that since \( \mathbb{E}[P_i \mid X_i = x] = s_{1,x}p_{1,x} + s_{2,x}p_{2,x} \), we have
\[
\theta_1(x) = (p_{2,x} - p_{1,x})[p_{2,x} - \mathbb{E}[P_i \mid X_i = x]] s_{2,x}
\]
\[
= (p_{2,x} - p_{1,x})[p_{2,x}(1 - s_{2,x}) - p_{1,x}s_{1,x}] s_{2,x} = (p_{2,x} - p_{1,x})^2 s_{1,x} s_{2,x}.
\]

On the other hand,
\[
\text{var}(P_i \mid X_i = x) = (p_{2,x} - \mathbb{E}[P_i \mid X_i = x])^2 s_{2,x} + (p_{1,x} - \mathbb{E}[P_i \mid X_i = x])^2 s_{1,x} = (p_{2,x} - p_{1,x})^2 s_{1,x} s_{2,x}.
\]

Secondly, since \( R_i^+ = R_j \),
\[
\mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}] = \mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}]
\]
\[
\mathbb{E}[Y_i \mid X_i = x, P_i = p_{2,x}] = \mathbb{E}[Y_i \mid X_i = x, P_i = p_{2,x}] + \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x}),
\]
so that
\[
\mathbb{E}[\tilde{R}_i^+ \mid X_i = x, P_i = p_{1,x}] = \mathbb{E}[\tilde{R}_i^+ \mid X_i = x, P_i = p_{2,x}]
\]
\[
= \mathbb{E}[Y_i \mid X_i = x, P_i = p_{0,x}] + \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x}) - \mathbb{E}[Y_i \mid X_i = x]
\]
\[
= \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x})(1 - s_{2,x})
\]
\[
= \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x}) s_{1,x}.
\]
Therefore, it follows that:

\[ \zeta_1(x) = (p_{2,x} - p_{1,x})^2s_{1,x}s_{2,x}\alpha(p_{1,x}; x), \]

which completes the proof. 

\( \square \)

**Proof of Theorem 2.** Using the matrix notation from Equation (26),

\[ \hat{\beta}_{UIVE} = \frac{Y_n'G_n'Y_n}{T_nG_n'T_n} \]

By Lemma 6 and Lemma 7,

\[ \frac{Y_n'G_n'Y_n}{T_nG_n'T_n} = \mathbb{E}[Y(P - \mathbb{E}[P | X])] + o_p(1), \]
\[ \frac{T_nG_n'T_n}{T_nG_n'T_n} = \mathbb{E}[T(P - \mathbb{E}[P | X])] + o_p(1). \]

Next, since the limiting distribution of the data satisfies Assumption IV (iii), \( \mathbb{E}[T(P - \mathbb{E}[P | X])] > 0 \), so that by the continuous mapping theorem,

\[ \frac{T_n'G_nY_n}{T_n'G_nT_n} = \frac{\mathbb{E}[Y(P - \mathbb{E}[P | X])] }{\mathbb{E}[T(P - \mathbb{E}[P | X])]} + o_p(1). \]

The assertion of the Theorem then follows by applying Lemma 3 with \( A_i = P - \mathbb{E}[P | X] \). 

\( \square \)
References


— and CRUDU, F. (2012). Symmetric Jackknife Instrumental Variable Estimation, working paper, University of Groningen. 3, 12


MORIMUNE, K. (1983). Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size. *Econometrica, 51*(3), 821–841. 3, 21


37