TREATMENT EFFECT HETEROGENEITY AND WEAK INSTRUMENTS

Michal Kolesár*

April 2, 2024

We can use instrumental variables (IV) regression to solve a number of issues:

1. Errors-in-variables (see, for example, Zellner 1970);
2. Deal with omitted variable bias: we'd like to recover β in the projection $E[Y_i \mid D_i, A_i] = D_i\beta + A_i'\gamma$, but A_i is not observed (see, for example, Chamberlain 2007);
3. Estimate a simultaneous equations model, such as a demand-and-supply system (see, for example, Angrist, Graddy, and Imbens 2000); or
4. Estimate treatment effects when the unconfoundedness assumption fails.[1]

Which goal we have in mind often doesn't affect the estimation and inference procedures, but it does affect the interpretation of the results—for concreteness, we will focus here on the last goal. In this lecture we'll consider:

1. Implications of treatment effect heterogeneity for estimation and inference; and
2. Weak instrument issues.

In later lectures, we'll consider two particular IV designs that have been increasingly common: "judges" designs, and the associated many instrument issues, and shift-share designs.

1. SETUP AND REVIEW OF TEXTBOOK MODEL

We'll use the usual sampling framework, assuming $\mathcal{D}_i = (Y_i, D_i, Z_i, W_i)$ are drawn i.i.d. from a large population, where D_i is the treatment, Z_i is a k-vector of instruments, and W_i is an ℓ-vector of controls. Let $X_i = (Z_i', W_i')'$. Our parameters of interest will be defined with respect to this population

*Email: mkolesar@princeton.edu.

1. This is not quite the same thing as the second issue, just because the interpretation of β in the projection $E[Y_i \mid D_i, A_i]$ is not necessarily causal.

Research Question. Think about alternative populations of interest we talked about in the context of ordinary least squares (OLS) regressions in current setting. ⊠

Define the reduced form and the first stage:

$$Y_i = Z_i'\delta + W_i'\psi_Y + u_{Y,i}, \tag{1}$$

$$D_i = Z_i'\pi + W_i'\psi_D + u_{D,i}. \tag{2}$$

The coefficients π and ψ here are defined as (unconditional) best linear predictors. Under regularity conditions, arguments in the appendix show that the OLS estimators $\hat{\delta}$ and $\hat{\pi}$ satisfy[2]

$$\sqrt{n}\left(\begin{pmatrix}\hat{\delta}\\ \hat{\pi}\\ \ddot{Z}'D/n\end{pmatrix} - \begin{pmatrix}\delta\\ \pi\\ Q\pi\end{pmatrix}\right) \Rightarrow \mathcal{N}\left(0, \mathcal{V}_0\right), \tag{3}$$

Here $\ddot{Z} = Z - H_W Z$ denotes the residual from the sample projection of Z_i onto the covariates W_i, $H_W = W(W'W)^{-1}W'$ is the hat matrix, $\tilde{Z}_i = Z_i - E[Z_i W_i'] E[W_i W_i']^{-1} W_i$ is the residual from the population projection of Z_i onto W_i, $Q = E[\tilde{Z}_i \tilde{Z}_i']$, and

$$\mathcal{V}_0 = \text{var}\begin{pmatrix} u_i \otimes Q^{-1}\tilde{Z}_i \\ (\tilde{Z}_i\tilde{Z}_i' - Q)\pi + u_{2i}\tilde{Z}_i \end{pmatrix},$$

where $u_i = (u_{Y,i}, u_{D,i})'$.

The asymptotic variance of the $\mathcal{V}_0$ can be consistently estimated using the Eicker-Huber-White (EHW) variance estimator, with the sample residual $\ddot{Z}_i$ replacing $\tilde{Z}_i$, $\hat{u}_i$ replacing u_i, and sample averages replacing population expectations. Here $\hat{u}_i = (\hat{u}_{Y,i}, \hat{u}_{D,i})$, with $\hat{u}_Y = \ddot{Y} - \ddot{Z}\hat{\delta}$ and $\hat{u}_D = \ddot{D} - \ddot{Z}\hat{\pi}$.

We'd like to use these reduced-form estimates for estimation and inference on treatment effects. To that end, we'll need to make some substantive assumptions. We'll start with the textbook assumption that the treatment effects $Y_i(1) - Y_i(0)$ are constant and linear,

Assumption 1 (Constant treatment effects). $Y_i(d) = Y_i(0) + d\beta$.

In this case, letting $\gamma = E[W_i'W_i]^{-1}E[W_i Y_i(0)]$ (notice the analogy in definition of γ for causal inference on superpopulation in the previous set of notes), we can write

$$Y_i = D_i\beta + W_i'\gamma + \epsilon_i, \tag{4}$$

where $\epsilon_i = Y_i(0) - W_i'\gamma$ is called the *structural error*. The name comes from thinking about eq. (4) as a structural equation modeling the outcomes Y_i, in which case γ would have an economic meaning—here we relax that and simply define it as the best linear predictor of $Y_i(0)$.

The key substantive assumption that we need is:

2. We include the asymptotic distribution of $\ddot{Z}'D/n$ in this result since it'll be useful when we consider asymptotics under treatment effect heterogeneity.

Assumption 2 (Random assignment). $E[Z_i \mid Y_i(d), W_i] = \Delta' W_i$ for some $\ell \times k$ matrix Δ.

This assumption entails three conditions:

RANDOM ASSIGNMENT Z is mean-independent of the potential outcomes given W

EXCLUSION RESTRICTION the potential outcomes $Y_i(d, z)$ in fact only depend on d

LINEARITY $E[Z \mid W]$ is linear in W.

The linearity condition is an analog of assuming that the propensity score is linear in controls that we imposed when analyzing OLS in previous lecture. It implies that $\tilde{Z}_i = Z_i - \Delta' W_i$. Notice this is a "design-based" identification condition, in that it only restricts the assignment of the instrument, but it makes no restrictions on the potential outcomes. One could alternatively pursue "model-based" identification.

Research Question. Can we generalize the model-based identification arguments from OLS to the present context? ⊠

Assumptions 1 and 2 deliver the moment condition

$$E[X_i \epsilon_i] = 0$$

that underlies textbook IV theory. To derive the moment condition, note that $Z_i \epsilon_i = Z_i(Y_i(0) - W_i' \gamma)$, and taking conditional expectations then, by Assumption 2, yields $\Delta' W_i (Y_i(0) - W_i' \gamma)$. This is mean zero by definition of γ. Similarly, $W_i \epsilon_i = W_i(Y_i(0) - W_i' \gamma)$ is mean zero by definition of γ.

1.1. *Estimation and Inference*

The textbook model delivers a restriction on the reduced form: substituting the first stage (2) into eq. (4) implies that the reduced-form coefficients are proportional to each other:

$$\delta = \pi \beta. \tag{5}$$

Furthermore, $\psi_Y = \gamma + \psi_D \beta$, and $u_{Yi} = u_{Di} \beta + \epsilon_i$. Therefore, the variance of the structural error is linked to $\Omega(X_i) = E[u_i u_i' \mid X_i]$:

$$\sigma^2(X_i) := \operatorname{var}(\epsilon_i \mid X_i) = \operatorname{var}(u_{Yi} - u_{Di} \beta \mid X_i) = b' \Omega(X_i) b \qquad b = \begin{pmatrix} 1 \\ -\beta \end{pmatrix}.$$

The standard approach to estimation is based on the moment condition $E[X_i \epsilon_i] = 0$. If ϵ_i is homoskedastic, the optimal weight matrix is proportional to $E[X_i X_i']^{-1}$, and using the sample analog $(X'X)^{-1}$ yields the two-stage least squares (TSLS) estimator that minimizes $(Y - D\beta - W\gamma)' H_X (Y - D\beta - W\gamma)$, where $H_X = X(X'X)^{-1}X'$ is the hat

matrix. Minimizing this over (β, θ) yields:

$$\hat{\beta}_{\text{TSLS}} = \frac{D' H_{\ddot{Z}} Y}{D' H_{\ddot{Z}} D} = \frac{\hat{\pi} \ddot{Z}' \ddot{Z} \hat{\delta}}{\hat{\pi} \ddot{Z}' \ddot{Z} \hat{\pi}}, \qquad \hat{\gamma}_{\text{TSLS}} = (W'W)^{-1} W'(Y - D\hat{\beta}_{\text{TSLS}}).$$

If $k = 1$, then the weight matrix doesn't matter, and we simply have

$$\hat{\beta}_{\text{TSLS}} = \frac{\hat{\delta}}{\hat{\pi}}.$$

By standard generalized method of moments (GMM) arguments, or by applying the delta method to eq. (3) (see Appendix), we obtain

$$\sqrt{n}(\hat{\beta}_{\text{TSLS}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_1), \qquad \mathcal{V}_1 = \frac{E[\sigma^2(X_i)(\tilde{Z}_i' \pi)^2]}{(\pi' Q \pi)^2}. \tag{6}$$

This variance is usually estimated as

$$\hat{V}_1 = n \frac{\sum_i \hat{\epsilon}^2_{\text{TSLS},i} (\ddot{Z}_i' \hat{\pi})^2}{[\sum_i (\ddot{Z}_i' \hat{\pi})^2]^2}, \quad \hat{\epsilon}_{\text{TSLS}} = \ddot{Y} - \ddot{D}\hat{\beta}_{\text{TSLS}},$$

or equivalently $\hat{\epsilon}_{\text{TSLS},i} = Y_i - D_i \hat{\beta}_{\text{TSLS}} - W_i' \hat{\gamma}_{\text{TSLS}}$. Under homoskedastic errors, this simplifies to

$$\hat{V}_{1,ho} = \frac{\hat{\sigma}^2}{n^{-1} \sum_i (\ddot{Z}_i' \hat{\pi})^2} = \frac{\hat{\sigma}^2}{n^{-1} D' H_{\ddot{Z}} D},$$

where $\hat{\sigma}^2 = n^{-1} \sum_i \hat{\epsilon}^2_{\text{TSLS},i}$.

Remark 1. Note that we could alternatively estimate $\sigma^2(X_i)$ as $(\hat{u}_{Y,i} - \hat{u}_{D,i}\hat{\beta}_{\text{TSLS}})^2$. If $k = 1$, then these two approaches are equivalent, since, using $\hat{\pi}\hat{\beta}_{\text{TSLS}} = \hat{\delta}$, we get $\hat{u}_1 - \hat{u}_2 \hat{\beta}_{\text{TSLS}} = \ddot{Y} - \ddot{Z}\hat{\delta} - \ddot{D}\hat{\beta}_{\text{TSLS}} + \ddot{Z}\hat{\pi}\hat{\beta}_{\text{TSLS}} = \hat{\epsilon}$. However, if $k > 1$, this approach yields a different variance estimator, with different properties under weak instruments, many instruments, or heterogeneous treatment effects.

LIML An alternative approach to estimation that goes back to Anderson and Rubin (1949) is to assume that the structural and first-stage errors $(\epsilon_i, u_{D,i})$ are homoskedastic and jointly normal conditional on X_i, and estimate β by maximum likelihood. The resulting estimator is called limited information maximum likelihood (LIML) (viewing the problem as a simultaneous equation model, we don't use full information by not fully modeling the simultaneity). The estimator takes the form

$$\hat{\beta}_{\text{LIML}} = \underset{\beta}{\text{argmin}} \frac{(1, -\beta)\hat{\Pi}' \ddot{Z} \ddot{Z}' \hat{\Pi} (1, -\beta)}{(1, -\beta) S (1, -\beta)'} = \frac{D' H_{\ddot{Z}} Y - \kappa S_{12}}{D' H_{\ddot{Z}} D - \kappa S_{22}},$$
$$\kappa = \min \text{eig}(S^{-1} \hat{\Pi}' \ddot{Z} \ddot{Z}' \hat{\Pi}).$$

where $\hat{\Pi} = (\hat{\delta}, \hat{\pi})$, and $S = [(\ddot{Y}, \ddot{D}) - \ddot{Z}\hat{\Pi}]'[(\ddot{Y}, \ddot{D}) - \ddot{Z}\hat{\Pi}]/(n - k - \ell)$ is an estimator of $\text{var}(u_i)$ based on the reduced-form residuals. Note that TSLS can also be written in this

form, with $\kappa = 0$.

A third approach to estimation would be to base estimation directly on the restriction (5). In particular, we could form a minimum distance estimator based on this restriction, yielding the objective function

$$\begin{pmatrix} \hat{\delta} - \pi\beta \\ \hat{\pi} - \pi \end{pmatrix}' W \begin{pmatrix} \hat{\delta} - \pi\beta \\ \hat{\pi} - \pi \end{pmatrix}.$$

Under homoskedasticity, the variance of $(\hat{\delta}', \hat{\pi}')'$ is given by $\Omega \otimes Q^{-1}$, so that the feasible weight matrix $W = S^{-1} \otimes \ddot{Z}'\ddot{Z}/n$ will be the optimal. In this case, Goldberger and Olkin (1971) show that the minimum distance estimator of β is numerically equivalent to LIML: so we can think of LIML as an estimator that exploits the proportionality restriction (5).

Proof. Letting $a = (\beta, 1)'$, we can write the objective function as

$$\begin{aligned} \operatorname{vec}(\hat{\Pi} - \pi a')(S^{-1} \otimes (\ddot{Z}'\ddot{Z}/n)) \operatorname{vec}(\hat{\Pi} - \pi a') &= \operatorname{vec}(\hat{\Pi} - \pi a') \operatorname{vec}((\ddot{Z}'\ddot{Z}/n)(\hat{\Pi} - \pi a')S^{-1}) \\ &= \operatorname{tr}((\hat{\Pi} - \pi a')'(\ddot{Z}'\ddot{Z}/n)(\hat{\Pi} - \pi a')S^{-1}) \\ &= \operatorname{tr}(\hat{\Pi}'(\ddot{Z}'\ddot{Z}/n)\hat{\Pi}) - 2a'S^{-1}\hat{\Pi}'(\ddot{Z}'\ddot{Z}/n)\pi + \pi'(\ddot{Z}'\ddot{Z}/n)\pi a'S^{-1}a \end{aligned}$$

The first-order condition with respect to π implies $\hat{\pi}_{\text{LIML}} = \hat{\Pi}S^{-1}a/(aS^{-1}a)$, so that the objective function with π concentrated out can be written as

$$\operatorname{tr}(T) - a'S^{-1}TS^{-1}a/(aS^{-1}a)$$

where $T = \hat{\Pi}'(\ddot{Z}'\ddot{Z}/n)\hat{\Pi}$. The result then follows from the identity $a'S^{-1}TS^{-1}a/(aS^{-1}a) = b'Tb/b'Sb + \operatorname{tr}(S^{-1}T)$, where $b = (1, -\beta)'$. □

Since the minimum distance objective function doesn't rely on normality or homoskedasticity of the errors, it is clear that LIML will remain asymptotically normal and consistent even without these assumptions—in fact, one can show that it is first-order asymptotically equivalent to TSLS. One can therefore use the estimator $\hat{V}_1$ above for standard errors, with $\hat{\epsilon}_i$ replaced with $\hat{\epsilon}_{i,\text{LIML}}$.[3]

1.2. *What can go wrong?*

So in the textbook model, one could use either TSLS or LIML for estimation, and standard errors based on $\hat{V}_1$ on inference; both approaches are asymptotically equivalent. There are three main reasons why this approach may fail in finite samples:

1. The model is wrong: the proportionality restriction (5), or, equivalently, the moment condition $E[X_i\epsilon_i] = E[X_i(Y_i - D_i\beta - W_i'\gamma)] = 0$ does not hold. This can happen if Assumption 2 doesn't hold (for example, the exclusion restriction fails),

3. One could also replace $\hat{\pi}$ in the standard error formula with the maximum likelihood or minimum distance estimator of π. One could also use the information matrix of the limited information likelihood to get standard errors, although one needs to use the sandwich formula to ensure validity under heteroskedasticity.

or if there is treatment effect heterogeneity (Assumption 1 fails). Note that as far as inference is concerned, this is only an issue if $k > 1$.

Question 1. Why?

2. The delta method fails. As shown in the appendix, since we can write $\beta = \delta' Q\pi / \pi' Q\pi$, it follows that $\hat{\beta}_{\text{TSLS}} = g(\hat{\delta}, \hat{\pi}, \hat{Q})$, where $\hat{Q} = \ddot{Z}'\ddot{Z}/n$. Therefore, we can apply the delta method to eq. (3) with this function g to obtain the asymptotic distribution of TSLS. For the delta method to work, we need g to be continuously differentiable at (δ, π, Q). This will fail if $\pi = 0$. By continuity, this implies that the delta method will work poorly if π is close to zero. This is a weak instrument problem.

3. Inference on the reduced form is unreliable, in that confidence intervals (CIs) for δ or π based on the normal approximation in eq. (3) and robust standard errors are unreliable. This may happen for the usual reasons that EHW standard errors fail, as we discussed in previous lecture. Such issues may be important in practice, as shown in Young (2022). The second reason why this may be an issue is that the number of instruments k is large relative to sample size: if the number of instruments is large relative to sample size, then the reduced-form estimators $\hat{\delta}$ and $\hat{\pi}$ may not be approximately normally distributed.

We'll now explore the first two issues in detail. We'll defer the treatment of the third issue (k is large relative to sample size) to next lecture.

2. TREATMENT EFFECT HETEROGENEITY

We can't do much about failure of Assumption 2 (apart from more flexibly controlling for the covariates if we're worried about linearity of $E[D_i \mid W_i]$), so we'll focus on the implications of the failure of Assumption 1. There are two main implications:

1. TSLS will estimate a weighed average of local average treatment effects (LATEs), but LIML will not in general estimate an object that has a causal interpretation. In the words of Heckman and Vytlacil (2005):

> The relevant question regarding the choice of instrumental variables in the general class of models studied in this paper is "What parameter is being identified by the instrument?" rather than the traditional question of "What is the efficient combination of instruments for a fixed parameter?"—the question that has traditionally occupied the attention of econometricians who study instrumental variables.

2. The standard error for $\hat{\beta}_{\text{TSLS}}$ based on $\hat{V}_1$ is no longer valid.

Imbens and Angrist (1994) show that if there are no covariates, D_i is binary and Assumption 2 holds, then TSLS estimates a weighted average of LATEs, provided an additional monotonicity condition holds. As was the case with OLS under treatment effect heterogeneity, depending on the policy in question, these weights may not be particularly policy relevant, but one can defend the focus on the TSLS estimand in much the same way as when we discussed the OLS estimand. A similar result obtains under multivalued D_i, as shown in Angrist and Imbens (1995). It is straightforward to extend these results to allow for covariates (under the assumption that $E[Z_i \mid W_i]$ is linear), and you can try doing so as an exercise.

To give the result with a binary treatment, suppose that the conditional distribution of Z_i given $W_i = w$ has J_w support points. Given $W_i = w$, order the support points $\{z_{jw}\}_{j=1}^{J_w}$ so that the propensity score $E[D_i \mid Z_i = z_{jw}, W_i = w] = p(z_{j,w}, W_i) = p_{j,w}$ is increasing in j. Let $\alpha(p_{j,w}, w)$ denote the LATE for people for individuals with $W_i = w$ who get treated when the instrument they receive corresponds to propensity score $p_{j+1,w}$ or higher, but not otherwise. Then

$$\beta := \frac{\delta' Q \pi}{\pi' Q \pi} = \int \sum_{j=1}^{J_w - 1} \frac{\lambda_j(w)}{\int \sum_{j=1}^{J_w-1} \lambda_j(w)\, dF_W(w)} \alpha(p_{j,w}, w)\, dF_W(w), \tag{7}$$

where $\lambda_j(w) = (p_{j+1,w} - p_{j,w}) P(p(Z_i, W_i) > p_{j,w} \mid W_i = w) E[\tilde{Z}_i' \pi \mid W_i, p(Z_i, W_i) > p_{j,w}]$. These weights will be positive if the last term is positive. Like in the case with OLS, these weights are not particularly pretty, though I suspect they may have an efficiency interpretation.

Like with OLS, including interactions of the instrument with the covariates would make them prettier, though that means we may run into many instrument issues.

In contrast, the estimand of LIML generally no longer has a causal interpretation in the sense that it estimates a convex combination of LATEs, unless all LATEs happen to be the same, as shown in Kolesár (2013). The reason is that under treatment effect heterogeneity, the proportionality restriction (5) no longer holds. Because LIML imposes this condition with a non-diagonal weight matrix, its behavior (in terms of its probability limit and its asymptotic variance) is quite sensitive to failures of this restriction. On the other hand, TSLS constructs a single instrument $\hat{Z}_i = \ddot{Z}_i' \hat{\pi}$ based on the first stage, and estimates β using with an IV estimator using single constructed instrument, $\hat{\beta}_{\text{TSLS}} = \hat{Z}_i' Y_i / \hat{Z}_i' D_i$: this makes it much more robust to treatment effect heterogeneity. Because of this result, one should not use LIML in practice if one has doubts about eq. (5) holding.

- Interact binary instrument with covariates.

2.2. *Inference*

Since the model in eq. (4) is misspecified, it will also generally matter if one wishes to do inference conditionally on X_i, conditionally on some other variable, or unconditionally, in analogy to our discussion in the previous lecture. Here we'll focus on unconditional inference; see Evdokimov and Kolesár (2018) for conditional results, if the estimand is defined as $\delta'\ddot{Z}'\ddot{Z}\pi/\pi'\ddot{Z}'\ddot{Z}\pi$ instead.

In particular, as shown in the appendix, an application of the delta method yields

$$\sqrt{n}(\hat{\beta}_{\text{TSLS}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_2), \qquad \mathcal{V}_2 = \frac{E[((\tilde{Z}_i'\pi_\Delta)u_{2i} + \epsilon_i(\tilde{Z}_i'\pi))^2]}{(\pi' Q \pi)^2}, \tag{8}$$

where

$$\epsilon_i = \tilde{Y}_i - \tilde{D}_i\beta = Y_i - D_i\beta - W_i'\gamma = \tilde{Z}_i\pi_\Delta + u_{Y,i} - u_{D,i}\beta, \tag{9}$$

$$\gamma = E[W_iW_i]^{-1}E[W_i'(Y_i - D_i\beta)] = \Delta\pi_\Delta + \psi_Y - \psi_D\beta, \tag{10}$$

and $\pi_\Delta = \delta - \pi\beta$. Note that if eq. (5) holds, then $\pi_\Delta = 0$, and $\mathcal{V}_2 = \mathcal{V}_1$. In general, however, the variances will be different, and we need to instead use the variance estimator

$$\hat{\mathcal{V}}_2 = n\frac{\sum_{i=1}^n[((\tilde{Z}_i'(\hat{\delta} - \hat{\pi}\hat{\beta}))\hat{u}_{2i} + \hat{\epsilon}_{\text{TSLS},i}(\tilde{Z}_i'\hat{\delta}))^2]}{(\sum_i \tilde{Z}_i\hat{\pi})^2},$$

Remark 2. The result that the usual standard errors are different under treatment effect heterogeneity can be found in the appendix in Imbens and Angrist (1994); it is also derived in Lee (2018) (if one simplifies the expression given in that paper, one obtains the expression above), or in Evdokimov and Kolesár (2018) (who in addition assume that $E[u_i \mid X_i] = 0$, which we don't assume here).

Remark 3 (Single instrument). If $k = 1$, then $\pi_\Delta = 0$, since δ and π are scalars, and $\beta = \delta/\pi$. Hence, $\hat{\mathcal{V}}_2 = \hat{\mathcal{V}}_1$, $\mathcal{V}_2 = \mathcal{V}_1$. Also, in this case, LIML and TSLS coincide.

Remark 4 (Known propensity score). We can interpret TSLS as an IV estimator that uses a single constructed instrument $\hat{Z}_i = Z_i'\hat{\pi}$, and controls W_i. You can think of $\hat{Z}_i$ as an estimate of $Z_i'\pi$, which under some conditions (which ones are they?) is the best single instrument. If D_i is binary, and there are no covariates, we can interpret $Z_i'\pi$ as the propensity score. In other words, β in eq. (7), and γ in eq. (10) can be thought of as solutions to the exactly identified moment condition

$$E\left[\begin{pmatrix}\pi' Z_i \\ W_i\end{pmatrix}(Y_i - D_i\beta - W_i'\gamma)\right] = 0,$$

and TSLS can also be thought of as a GMM estimator based on a sample analog of this moment condition, with $\hat{\pi} = (\ddot{Z}'\ddot{Z})^{-1}\ddot{Z}'D$ replacing the unknown nuisance parameter π. Suppose you're an oracle who knows the propensity score. Then you can use the above

moment condition, but without having to use the estimate $\hat{\pi}$. This yields the estimates $\hat{\beta}_{\text{oracle}} = (\pi'\ddot{Z}'D)^{-1}\pi'\ddot{Z}'Y$ and $\hat{\gamma}_{\text{oracle}} = (W'W)^{-1}W'(Y - D\hat{\beta}_{\text{oracle}})$. Since the model based on the above moment condition is exactly identified, by standard GMM arguments, regardless of the presence of treatment effect heterogeneity, $\sqrt{n}(\hat{\beta}_{\text{oracle}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_1)$, with

$$\mathcal{V}_1 = \frac{E[\epsilon_i^2(\tilde{Z}_i'\pi)^2]}{E[(\tilde{Z}_i'\pi)^2]^2},$$

and ϵ_i defined in eq. (9). So under constant treatment effects, TSLS (and LIML) achieve oracle efficiency: the fact that the first-stage (i.e. the propensity score) is estimated doesn't affect the asymptotic variance (we'll explore the reason for this in later lectures). But under heterogeneous treatment effects, we do not get to this oracle: the difference between $\mathcal{V}_1$ and $\mathcal{V}_2$ exactly accounts for the fact that π is estimated.

Remark 5 (Known first stage). There is an alternative oracle estimator of β, based on the moment condition

$$E\left[\begin{pmatrix}\pi'Z_i\\ W_i\end{pmatrix}(Y_i - (Z_i'\pi)\beta - W_i'\delta)\right] = 0, \qquad \delta = \gamma + \psi_2\beta.$$

with $\hat{\beta}_{\text{oracleo}} = (\pi\ddot{Z}'\ddot{Z}\pi)^{-1}\pi'\ddot{Z}'Y$. This estimator is an analog of the TSLS estimator with a known first stage: if we know π, we don't need to run the first stage of the two-stage procedure, only the second stage. The asymptotic variance of this oracle is given by

$$\mathcal{V}_3 = \frac{E[(\epsilon_i + u_{2i}\beta)^2(\tilde{Z}_i'\pi)^2]}{E[(\tilde{Z}_i'\pi)^2]^2}.$$

In contrast to the previous case, knowing the first stage *does* have an effect on the asymptotic variance of this estimator: $\mathcal{V}_3 \neq \mathcal{V}_2$ even under homogeneous treatment effects. You may recall from your undergraduate days that it is incorrect to report standard errors from the regression of Y_i onto $\hat{Z}_i = Z_i'\hat{\pi}$ and onto W_i as the TSLS standard errors (as mentioned, e.g., on page 97 in Wooldridge 2010)—these standard errors estimate $\mathcal{V}_3$. So when person A says that (under constant treatment effects), not knowing the first stage does affect standard errors, and person B says that it does not, both are right, they just have a different oracle (a different set of moment conditions) in mind.

3. WEAK INSTRUMENTS

In many applications, the first-stage coefficients may be close to zero. In such cases, the delta method may not work well, so that CIs based on $\hat{V}_1$ or $\hat{V}_2$ may undercover, and the TSLS estimator may be biased (in the sense that its distribution will not be centered at β, even in large samples).

Example 1. As an example, consider the problem of estimating the elasticity of intertemporal substitution (EIS). One approach involves estimating the log-linearized Euler equa-

tion based on a portfolio choice problem of an agent with Epstein-Zin preferences, which is given by (see Campbell (2003) and Yogo (2004) for derivation)

$$E_t[\Delta c_{t+1} - \mu_j - \psi r_{j,t+1}] = 0$$

where $r_{j,t+1}$ is return on asset j, Δc_{t+1} is consumption growth, μ_j is a constant, and ψ is the EIS, and E_t denotes expectation conditional on the agent's information set at time t.

We could try to estimate ψ by running the regression

$$\Delta c_{t+1} = \mu_j + \psi r_{j,t+1} + e_t.$$

However, the error term in that regression, $e_t = \Delta c_{t+1} - E_t[\Delta c_{t+1}] - \psi(r_{j,t+1} - E_t[r_{j,t+1}])$ is going to be correlated with $r_{j,t+1}$. On the other hand, e_t will be by definition uncorrelated with any variables in the information set at time t, so we can use those as instruments. Alternatively, we could instrument for Δc_{t+1} using the same instruments, and estimate $1/\psi$. Instrumenting for Δc_{t+1}, computing $\hat{\psi}_{inv} = 1/\psi$ using TSLS, and then reporting $\hat{\psi}_{inv}^{-1}$ as an estimate of ψ is known as reverse TSLS.

With one instrument, the two approaches are numerically equivalent. When there is more than one instrument, both IV regressions are asymptotically equivalent under standard asymptotics, so we would expect the estimates to be approximately similar in finite samples.

The problem is that empirical estimates of both ψ and of $1/\psi$ are small. For instance, Campbell (2003, Table 9) reports a 95% CI $[-0.14, 0.28]$ for ψ, and CI $[-0.73, 2.14]$ for $1/\psi$, using quarterly U.S. data (1947–1998) on non-durable consumption and T-bill returns.

The reason is that the equivalence breaks down when the instruments are weak, and weak instruments are likely the culprit here because both consumption growth and asset returns are notoriously difficult to predict. ⊠

Remark 6 (Aside). The fact that the two approaches, estimating ψ and $1/\psi$ no longer give compatible results when instruments are weak led Hahn and Hausman (2002) to propose a test for weak instruments based on the difference between $\hat{\psi}$ and $\hat{\psi}_{inv}^{-1}$. Unfortunately, the power of the test against weak or irrelevant instruments is low and the tests are not consistent against irrelevant instruments (Hausman, Stock, and Yogo 2005).

In a recent review, Andrews, Stock, and Sun (2019) document that a substantial fraction of IV regressions in papers recently published in the *American Economic Review* have first-stage F statistics under 20, which means weak instruments are frequently encountered in practice.

Before we dive into the issues and solutions, here are the key takeaways:

- Most of the theoretical literature focused on the simplest case with constant treatment effects and homoskedastic errors for tractability. However, not all the recommendations from this literature carry over to the general case with heteroskedasticity, clustering, or serial dependence, and to heterogeneous treatment effects.

- For detecting weak instruments, it is popular to use the $F > 10$ rule of thumb. This rule relies heavily on homoskedasticity, and the rule looks quite different under heteroskedasticity. Moreover, it is unclear how one should use it in practice.

- If $k = 1$, the usual CIs work well unless the endogeneity problem is very severe. Furthermore, one can report the Anderson and Rubin (1949) CIs as a robustness check: they are robust to weak instruments, and efficient under strong instruments. The tF procedure of Lee et al. (2022) is another alternative. Since the critical values it uses are determined under extreme endogeneity, it quantifies the precision gains of the Wald test due to (implicitly or explicitly) ruling out extreme endogeneity—these precision gains can be substantial if $F \leq 20$. If one knows the sign of the first stage, the (median) bias of TSLS is minimal if one conditions on the estimated first-stage being right-signed.

- If $k > 1$, things are more complicated, and very hard if there is more than one endogenous variable. None of the existing testing procedures are robust to treatment effect heterogeneity

  Research Question. Can we develop such a procedure? ⊠

  Similarly, the finding that estimators such as LIML are more robust to the weak instrument problem is sensitive to the failure of constant treatment effects assumption.

- Of course, weak instrument and weak identification issues are present in more complicated, non-linear models. Diagnostics and solutions in these models are an active area of research.

3.1. *The weak instrument problem with a single instrument*

To see what can go wrong when the instruments are weak, let's consider the extreme case in which the instruments are irrelevant, so that $\pi = 0$. For simplicity, suppose that we have a single instrument. Then, by the central limit theorem (CLT) since $\tilde{Z}'D = \tilde{Z}'u_2$,

$$\hat{\beta}_{\text{TSLS}} = \frac{\tilde{Z}'Y}{\tilde{Z}'D} = \frac{\tilde{Z}'D\beta + \tilde{Z}'\epsilon}{\tilde{Z}'u_2} = \beta + \frac{n^{-1/2}\tilde{Z}'\epsilon}{n^{-1/2}\tilde{Z}'u_2} \Rightarrow \beta + \frac{\Sigma_{11}^{1/2}\mathcal{Z}_\epsilon}{\Sigma_{22}^{1/2}\mathcal{Z}_2},$$

where $\mathcal{Z}_\epsilon$ and $\mathcal{Z}_2$ are standard normal with covariance $\rho = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$, and $\Sigma = \text{var}(\tilde{Z}_i(\epsilon_i, u_{2i})')$. We can think of ρ as measuring the endogeneity problem: under homoskedasticity, ρ measures the correlation between the structural error ϵ_i in Equation (4) and the first-stage error. If $\rho = 0$, then OLS is consistent. We therefore refer to ρ as (the degree of) endogeneity.

To simplify this a little, let's decompose $\mathcal{Z}_\epsilon$ into a part that's perfectly correlated with

$\mathcal{Z}_2$ and part that's independent letting

$$\mathcal{Z}_\perp = (1-\rho^2)^{-1/2}\mathcal{Z}_\epsilon - \rho\mathcal{Z}_2,$$

so that $\mathcal{Z}_\perp$ and $\mathcal{Z}_2$ are independent standard normal, and $\mathcal{Z}_\epsilon = \rho\mathcal{Z}_2 + \sqrt{1-\rho^2}\mathcal{Z}_\perp$. Plugging this in:

$$\hat{\beta} \Rightarrow \beta + \frac{\Sigma_{12}}{\Sigma_{22}} + \sqrt{\frac{(1-\rho^2)\Sigma_{11}}{\Sigma_{22}}}C, \tag{11}$$

where $C = \mathcal{Z}_\perp/\mathcal{Z}_2$ has a Cauchy distribution.

- If $\tilde{Z}_i$ is mean-independent of (ϵ_i, u_{2i}), then $\beta + \Sigma_{12}/\Sigma_{22} = \beta + \rho\sqrt{\Sigma_{11}/\Sigma_{22}}$ is the probability limit of OLS. So asymptotically, TSLS is median-biased, and centered around the OLS limit. Because of the Cauchy distribution, it has thick tails. So it's like taking the OLS estimator, and adding heavy-tailed noise to it.
- TSLS is inconsistent, and doesn't converge to anything even asymptotically: its distribution doesn't get less spread out even asymptotically. So there is a sharp discontinuity in the asymptotic distribution, with the rate of convergence changing, depending on whether $\pi = 0$. As a result, the bootstrap, the m-out-of-n bootstrap, and subsampling will not work either (see, for instance Andrews and Guggenberger 2010)
- The Wald statistic (i.e. the usual t-ratio) is given by

$$W = \frac{\hat{\beta}-\beta}{\sqrt{\frac{\sum_i \hat{\epsilon}^2_{\text{TSLS},i}(\tilde{Z}_i\hat{\pi})^2}{[\sum_i(\tilde{Z}_i\hat{\pi})^2]^2}}} = \frac{\hat{\beta}-\beta}{\sqrt{\frac{n^{-1}\sum_i \hat{\epsilon}^2_{\text{TSLS},i}\tilde{Z}_i^2}{[n^{-1/2}\sum_i \tilde{Z}_i u_{2i}]^2}}}$$

  Now, by CLT and the law of large numbers (LLN), $n^{-1}\sum_i \hat{\epsilon}^2_{\text{TSLS},i}\tilde{Z}_i^2 = n^{-1}\sum_i(\tilde{\epsilon}_i + \tilde{u}_i(\beta - \hat{\beta}_{\text{TSLS}}))^2\tilde{Z}_i^2 \Rightarrow \Sigma_{11}(1 - 2\rho\mathcal{Z}_\epsilon/\mathcal{Z}_2 + \mathcal{Z}_\epsilon^2/\mathcal{Z}_2^2)$, and $n^{-1/2}\sum_i \tilde{Z}_i u_{2i} \Rightarrow \Sigma_{22}^{1/2}\mathcal{Z}_2$, so by the continuous mapping theorem,

$$W \Rightarrow \frac{\mathcal{Z}_\epsilon/\mathcal{Z}_2}{\sqrt{\mathcal{Z}_2^{-2}(1 - 2\rho\mathcal{Z}_\epsilon/\mathcal{Z}_2 + \mathcal{Z}_\epsilon^2/\mathcal{Z}_2^2)}} = \frac{\mathcal{Z}_\perp/\mathcal{Z}_2 + \rho/\sqrt{1-\rho^2}}{\sqrt{1/\mathcal{Z}_2^2 + Z_\perp^2/\mathcal{Z}_2^4}}.$$

  Depending on the value of the structural correlation ρ implied by the null, the null rejection probability can be a lot lower, or a lot higher than 5%. See Figure 1.

Remark 7. Early evidence of problems with weak instruments goes back to Nelson and Startz (1990b, 1990a) who show that, in the exactly identified case, if the instruments are weak, the density of the IV estimator in finite samples may be very far away from its "asymptotic" distribution. However, their striking results appear to stem from the fact that they study the model

$$Y = D\beta + \epsilon, \qquad D = Z\pi + \underbrace{\epsilon_\perp\lambda + \epsilon}_{u_2},$$

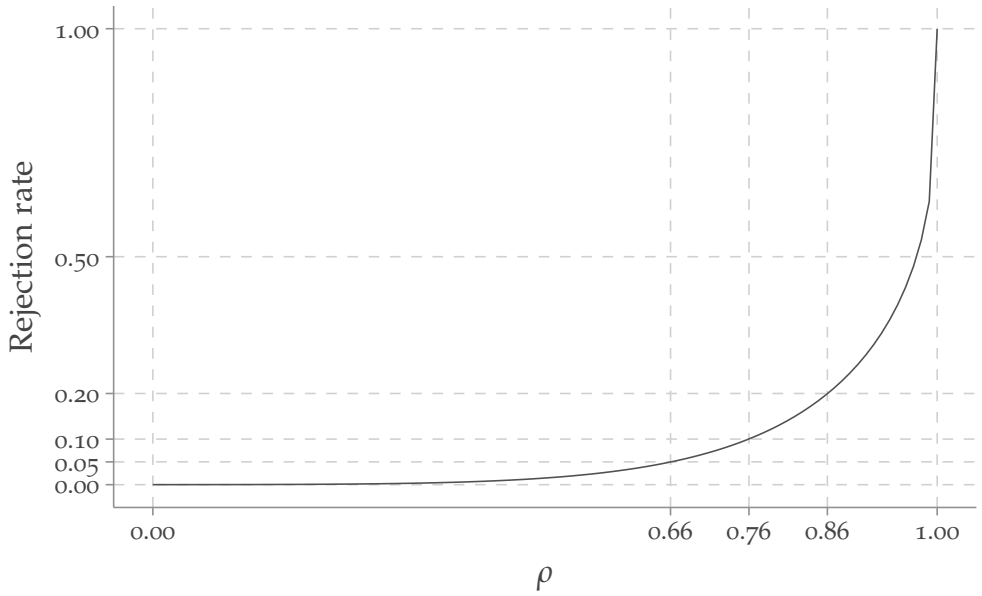

Figure 1: Rejection rate of the Wald test with nominal level 0.05 when $k = 1$, with irrelevant instruments.

and treat both Z *and* $\epsilon_\perp$ as fixed, meaning that part of the first-stage error u_2 is fixed. Bound, Jaeger, and Baker (1995) give convincing evidence of weak instrument issues based on the Angrist and Krueger (1991) study: in their Table 3, they use randomly generated information instead of the quarter of birth. The results look strikingly "reasonable".

To capture the weak instrument problem, we can think of the issue as doing inference on β given a single observation $(\hat{\delta}, \hat{\pi})$ that is distributed

$$\operatorname{vec}(\hat{\Pi}) \sim \mathcal{N}\left(\begin{pmatrix}\pi\beta\\ \pi\end{pmatrix}, \mathcal{V}\right), \qquad \mathcal{V} = E[\Omega(X_i) \otimes Q^{-1}\tilde{Z}_i\tilde{Z}_i'Q^{-1}]/n, \tag{12}$$

with $\mathcal{V}$ and Q known.

This suppresses any complications arising from issue 3, that is non-normality of the reduced form estimates, or difficulties with estimating $\mathcal{V}_0$ and focuses attention solely on the weak instruments problem. This can be formally justified in two ways:

1. Assume that $\hat{\mathcal{V}}_0 \xrightarrow{p} \mathcal{V}_0$, and that eq. (3) holds. Then (12) is the right limit experiment if we make no further assumptions, as argued in Müller (2011).

2. Assume that $\pi = C/\sqrt{n}$ for some fixed constant C. This is the idea of Staiger and Stock (1997). This is in similar spirit as the local-to-unity asymptotics in time series autoregressions: here π is local to 0. As long as we restrict ourselves to tests that are functions of $\hat{\Pi}$, then under this sequence, the problem is asymptotically equivalent the problem of inference given a single observation in the model (12).

These asymptotics are known as weak instrument asymptotics.

In the limit experiment (12), given the known reduced-form error $\mathcal{V}$, the distribution of $\hat{\beta}$, the Wald statistic, or any other object is governed by the unknown parameters (β, π). Oftentimes, it turns out to be more convenient to reparametrize the problem. If $k = 1$, in particular, it is convenient to instead use the parametrization $(E[F], \rho)$, where $E[F] = \pi_2^2/\mathcal{V}_{22} + 1$ is the expectation of the first-stage F statistic, $F = \hat{\pi}^2/\mathcal{V}_{22}$, and

$$\rho(\mathcal{V}, \beta) = \frac{\mathcal{V}_{12}/\mathcal{V}_{22} - \beta}{\sqrt{\mathcal{V}_{11}/\mathcal{V}_{22} - 2\mathcal{V}_{12}\beta/\mathcal{V}_{22} + \beta^2}} \tag{13}$$

is the degree of endogeneity (as discussed above). Then, letting $\mathcal{Z}_\epsilon$ and $\mathcal{Z}_2$ denote standard normal random variables with correlation ρ, we have

$$\hat{\beta} - \beta \sim (\mathcal{V}_{11}/\mathcal{V}_{22} + \beta^2 - 2\beta\mathcal{V}_{12}/\mathcal{V}_{22})^{1/2} \frac{\mathcal{Z}_\epsilon}{\sqrt{E[F] - 1} + \mathcal{Z}_2},$$

and the Wald statistic has the distribution

$$W \sim \frac{\text{sign}(\sqrt{E[F] - 1} + \mathcal{Z}_2)\mathcal{Z}_\epsilon}{\sqrt{Z_\epsilon^2/(\sqrt{E[F] - 1} + \mathcal{Z}_2)^2 - 2\rho\mathcal{Z}_\epsilon/(\sqrt{E[F] - 1} + \mathcal{Z}_2) + 1}}. \tag{14}$$

See appendix for derivation.

3.2. *Detecting weak instruments*

Let's consider the general case with potentially several instruments. To test for weak instruments, we first need to define what we mean by the term. Stock and Yogo (2005) start with eq. (12), under homoskedastic errors, so that

$$\mathcal{V} = (\Omega/n) \otimes Q^{-1}. \tag{15}$$

The parameter space is given by (β, π). They give two definitions, each of which gives a set of values $\Pi_W \subseteq \mathbb{R}^k$ for π for which the instruments are weak:

1. The bias of TSLS relative to OLS is bigger than 0.1 for some β.
2. The Wald test based on TSLS has size over 10% for some β.

We can actually read off when the instruments are weak from Table 1 in Richardson (1968), who showed that (we give a derivation based on Sawa (1972) in the appendix)

$$b_{\text{TSLS}} := \frac{E[\hat{\beta}_{\text{TSLS}} - \beta]}{\beta_{WOLS} - \beta} = 1 - \frac{\mu^2}{2} e^{-\mu^2/2} \int_0^1 x^{k/2-1} e^{\frac{\mu^2}{2}x} dx = \\ = 1 - \frac{k(E[F] - 1)}{2} \int_0^1 x^{k/2-1} e^{(x-1)k(E[F]-1)/2} dt, \tag{16}$$

where $\beta_{WOLS} = \Omega_{12}/\Omega_{22}$ is the limit of OLS under weak-instrument asymptotics, and $E[F] = n\pi' Q\pi / k\Omega_{22} + 1$ is the expectation of the first-stage F under homoskedasticity (henceforth, we ignore the difference between Q and $\ddot{Z}'\ddot{Z}/n$ for simplicity). Sometimes, instead of $E[F]$, people use the non-centrality parameter of the F-statistic, $k(E[F]-1) = n\pi' Q\pi / k\Omega_{22}$, called the *concentration parameter*.

Some notable take-aways:

- The relative TSLS bias (relative to OLS) depends only on $E[F]$, and not the value of β or degree of endogeneity. This suggests that (at least under homoskedasticity), $E[F]$ is the right quantity to measure instrument strength
- The expression is only well-defined if $k \geq 2$. More generally, $E[\hat{\beta}^p_{\text{TSLS}}]$ exists only for $p \leq k-1$ (Richardson 1968).

  Question 2. What is the implication of this for Monte Carlos?

- For k even, we can use repeated application of integration by parts to evaluate the integral, so that, for instance,

$$b_{\text{TSLS}} = \begin{cases} e^{1-E[F]} & \text{if } k = 2, \\ \frac{1-e^{2-2E[F]}}{2E[F]-2} & \text{if } k = 4. \end{cases}$$

  For k odd, we can evaluate the integral numerically.

- By evaluating the relative bias and using the fact that $k \cdot F$ has a non-central χ^2_k distribution, we can derive critical values for testing the hypothesis $H_0 \colon b_{\text{TSLS}} \leq 0.1$. For example, for $k = 2$ we need $E[F] = 1 - \log(0.1) \approx 3.30$ for TSLS bias to be 10% of OLS bias, which leads to the critical value 7.85. For $k = 3, 4, 5$, we obtain critical values $9.18, 10.23, 10.78$, and so on. These critical values fluctuate between 10 and 11.5 for larger values of F. These roughly match Table 5.1 in Stock and Yogo (2005) for $k \geq 3$. I am not sure why they don't report the critical value for $k = 2$. For $k = 1$, there is no bias, and hence no critical value.

  You can see for yourself by running 6 lines of R code:

```
b <- function(mu, k)
    1-integrate(function(x) mu/2*x^(k/2-1)*exp((x-1)*mu/2),
                                lower=0, upper=1)$value
crit <- function(k)
    qchisq(p=0.95, df=k, ncp=uniroot(function(mu)
              b(mu, k)-0.1, lower=0.1, upper=10*k)$root)/k
```

- The $F > 10$ rule of thumb was suggested in Staiger and Stock (1997). Since the critical value is "close" to 10 regardless of the number of instruments k, this calculation "justifies" it.

For the second definition, the critical value increases in k (see Tables 5.1 and 5.2 in Stock and Yogo 2005), starting with 16.38 when $k = 1$. The fact that they increase in k is actually quite important, why?

Remark 8. A alternative second definition would be to specify a set of parameters (β, π) for which the Wald test overrejects. This would lead to a very different test, especially if we're willing to a priori restrict ρ (or equivalently β): since the least favorable value of ρ is 1, one could in such cases tolerate much smaller values of $E[F]$.

In particular, suppose that $k = 1$. Then, using Figure 2, we can compute the rejection rates of the Wald test for different values of $E[F]$ and ρ (if there is heteroskedasticity, we use the heteroskedasticity-robust version of F, we drop the subscript here). Figure 2 gives the plot, which now also appears in Angrist and Kolesár (2024).

Without restricting the value of ρ, if we want to keep the rejection rate below 10%, we need to ensure that $E[F] \geq 6.88$ according to the plot. The first-stage F statistic (the square of the t-statistic) is distributed non-central χ^2_1 with non-centrality parameter $E[F] - 1$. This distribution is stochastically increasing in the non-centrality parameter, so we can use the first-stage F statistic along with the critical value based on the 95% quantile of non-central χ^2_1 with non-centrality parameter 5.88, which yields the Stock and Yogo (2005) critical value: `qchisq(p=0.95, df = 1, ncp = 5.88)=16.56`. An $F > 10$ cutoff would apply if we want to keep the rejection rate below 13.4%.

Note, however, that the set of parameters where overrejection occurs is rather restricted: in particular, if we're willing to put a priori bounds on the value of β that would lead to $|\rho| \leq 0.76$ (for a given covariance matrix of the reduced form coefficients $\mathcal{V}$), then we never need to worry about the weak instrument problem in the sense that the Wald test will not overreject by more than 5%. Angrist and Kolesár (2024) argue that in many applications in labor economics, such large values of endogeneity are unlikely.

If $k > 1$ and the errors are not homoskedastic, $\mathcal{V}$ doesn't have the Kronecker structure in eq. (15) above, and the $F > 10$ rule of thumb should not be used, whether calculated using the robust F statistic

$$F_{\mathrm{r}} = \frac{1}{k}\hat{\pi}'\mathcal{V}_{22}^{-1}\hat{\pi} = \frac{n}{k}\hat{\pi}'QE[\Omega_{22}(X_i)\tilde{Z}_i\tilde{Z}_i']Q\hat{\pi}$$

or the homoskedastic version, F_{h}

As an alternative, Montiel Olea and Pflueger (2013) instead suggest a pre-test based on what they call an effective first-stage F,

$$F_{\mathrm{eff}} = \frac{\hat{\pi}'Q\hat{\pi}}{\mathrm{tr}(\mathcal{V}_{22}Q)} = \frac{k/n \cdot \Omega_{22}F_{\mathrm{h}}}{\mathrm{tr}(Q^{-1}E[\Omega_{22}(X_i)\tilde{Z}_i\tilde{Z}_i']/n)}$$

Note if $k = 1$, then $F_{\mathrm{eff}} = F_{\mathrm{r}}$.

The reason for suggesting this statistic is that the bad behavior of the TSLS estimator is determined whether its denominator, given by $\hat{\pi}'Q\hat{\pi}$ in the limit experiment (12), is close to zero. F_{h} measures this object, while F_{r} measures the wrong object: its non-

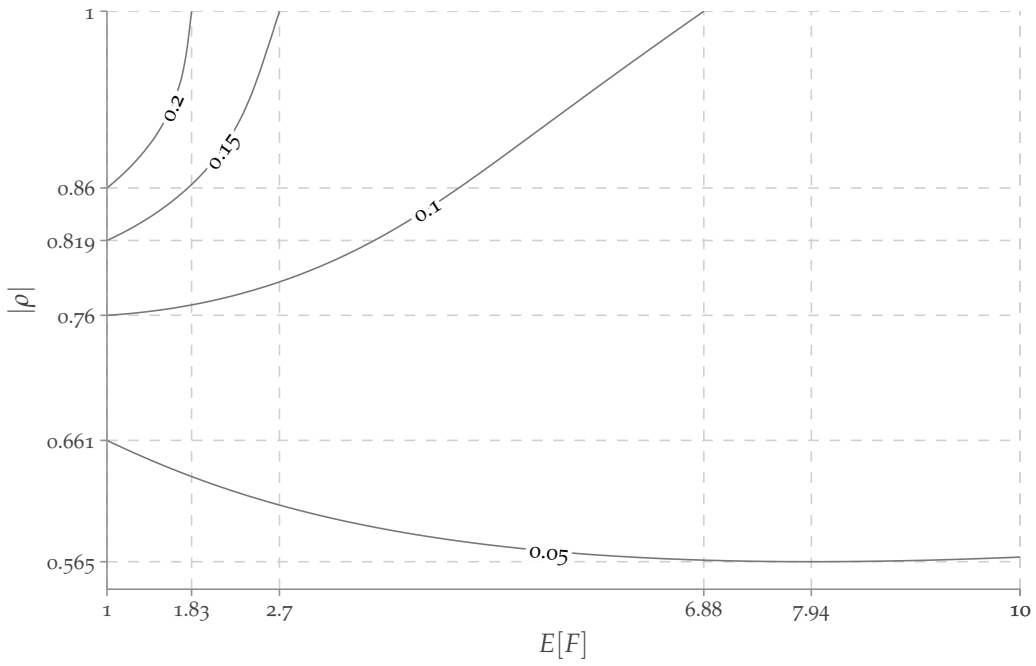

Figure 2: Contour plot of rejection rates of the Wald test with nominal level 0.05 when $k = 1$ as a function of $E[F]$ and ρ. Figure 1 gives a cross-section of this plot at $E[F] = 1$.

centrality parameter is not proportional to $\pi' Q \pi$. F_{h} gets the standard errors (scaling) wrong, however; F_{eff} gets them right "on average". Montiel Olea and Pflueger (2013) compute the critical values needed to ensure that an approximation to the bias of TSLS relative to the OLS bias doesn't exceed 10%.

To summarize:

- If $k = 1$, $F_{\text{eff}} = F_{\text{r}}$, and one can use the Stock and Yogo (2005) cutoff for ensuring that the size of the Wald test doesn't exceed 10%, that is $F_{\text{r}} \geq 16.6$
- If $k > 1$, use the Montiel Olea and Pflueger (2013) test, that ensures small TSLS bias

Note the discrepancy in what the rule is based upon: it's not possible to state the rule in terms of TSLS bias if $k = 1$, and, at the same time, a rule using the size of the Wald test based on TSLS when $k > 1$ would be very conservative.

However, screening on the first-stage F-statistic appears to compound, rather than reduce, inferential problems arising from weak instruments (see Andrews, Stock, and Sun (2019) for discussion and suggestive evidence), so one should use F-statistic as a diagnostic, not a decision rule.

When $k = 1$, then eq. (12) simplifies to a ratio of means problem,

$$\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} \sim \mathcal{N}_2 \left(\pi \begin{pmatrix} \beta \\ 1 \end{pmatrix}, \mathcal{V} \right), \qquad \mathcal{V} = \frac{E[\Omega(X_i)\tilde{Z}_i^2]}{nQ^2}$$

Doing inference about β is therefore equivalent to the problem of doing inference about ratio of means of a bivariate Gaussian variable. This connection has been pointed out by Zellner (1978), and Mariano and McDonald (1979).

The ratio of normal means is an old problem in statistics, dating back to at least Fieller (1932). Fieller (1940, 1954) proposed the following solution: under the null $H_0 \colon \beta = \beta_0$, the distribution of the statistic $\hat{\delta} - \hat{\pi}\beta_0$ does not depend on the nuisance parameter π: it's pivotal, so that

$$S(\beta_0) = \frac{(\hat{\delta} - \beta_0\hat{\pi})^2}{b_0'\mathcal{V}b_0} \sim_{H_0} \chi_1^2, \qquad b_0 = \begin{pmatrix} 1 \\ -\beta_0 \end{pmatrix}.$$

Therefore, the CI

$$\begin{aligned} &\left\{\beta_0 \in \mathbb{R} \colon S(\beta_0) \leq z_{1-\alpha/2}^2\right\} = \left\{\beta_0 \colon \frac{(\hat{\beta} - \beta_0)^2 F}{\mathcal{V}_{11}/\mathcal{V}_{22} - 2\beta_0\mathcal{V}_{12}/\mathcal{V}_{22} + \beta_0^2} \leq z_{1-\alpha/2}^2\right\} \\ &= \left\{\beta_0 \colon (F - z_{1-\alpha/2}^2)\beta_0^2 + 2(z_{1-\alpha/2}^2\mathcal{V}_{12}/\mathcal{V}_{22} - \hat{\beta}F)\beta_0 + \hat{\beta}^2 F - z_{1-\alpha/2}^2\mathcal{V}_{11}/\mathcal{V}_{22} \leq 0\right\}. \end{aligned}$$

will have coverage $1 - \alpha$ independently of the parameter π. Here $F = \hat{\pi}^2/\mathcal{V}_{22}$ denotes the first-stage F statistic.

Let $\bar{F} = F\mathcal{V}_{22}\hat{b}'\mathcal{V}\hat{b}/|\mathcal{V}|$ denote the F-statistic for testing $(\delta, \pi) = 0$, where $\hat{b} = (1, -\hat{\beta})'$. Note that $\bar{F} - F = F(\hat{\beta}\mathcal{V}_{22} - \mathcal{V}_{12})^2/|\mathcal{V}| \geq 0$.

Since the above display is a quadratic equation, there are three forms the CI can take.

1. If $F \geq z_{1-\alpha/2}^2$ (that is, we reject the null that $\pi = 0$), then the CI takes the form of an interval $[C_1, C_2]$, with endpoints given by

$$C_j = \hat{\beta} + z_{1-\alpha/2}^2 \frac{\hat{\beta} - \mathcal{V}_{12}/\mathcal{V}_{22}}{F - z_{1-\alpha/2}^2} \pm z_{1-\alpha/2} \frac{\sqrt{(\bar{F} - z_{1-\alpha/2}^2)|\mathcal{V}|}}{(F - z_{1-\alpha/2}^2)\mathcal{V}_{22}}$$

2. If $\bar{F} \geq z_{1-\alpha/2}^2 \geq F$, then it takes the form $(-\infty, C_2] \cup [C_1, \infty)$, with C_j given in the previous display

3. If $\bar{F} \leq z_{1-\alpha/2}^2$ (we don't reject the null that $\Pi = 0$, with critical value based on χ_1^2 rather than χ_2^2), then it is given by $\mathbb{R}$.

This CI is a special case of the CI proposed by Anderson and Rubin (1949) (discussed below), so it is often referred to as the Anderson-Rubin (AR) CI.

In contrast, the usual Wald CI is given by

$$\hat{\beta} \pm z_{1-\alpha/2} \frac{\sqrt{\hat{b}' \mathcal{V} \hat{b}}}{|\hat{\pi}|} = \hat{\beta} \pm z_{1-\alpha/2} \frac{\sqrt{\bar{F} |\mathcal{V}|}}{F \mathcal{V}_{22}}.$$

Comparing the two intervals, it follows that

- The AR CI is always longer than the usual Wald CI. It's obviously longer if $F \leq z^2_{1-\alpha/2}$. Otherwise, if $F \geq z^2_{1-\alpha/2}$, then it is longer so long as $(\bar{F} - z^2_{1-\alpha/2})/(F - z^2_{1-\alpha/2})^2 \geq \bar{F}/F^2 \iff 0 \geq F(F - \bar{F}) + \bar{F}(z^2_{1-\alpha/2} - F)$. But the right-hand side is always negative since $z^2_{1-\alpha/2} \leq F \leq \bar{F}$.
- Under standard asymptotics, $F = O_p(n)$, and $\bar{F} = O_p(n)$, so that

  $$C_j = \hat{\beta} + O_p(1/n) + z_{1-\alpha/2} \frac{\sqrt{\hat{b}' \mathcal{V} \hat{b}}}{|\hat{\pi}|} \sqrt{1 + O_p(1/n)},$$

  so that the AR and Wald CIs are asymptotically equivalent.
- This suggests that one should always use the AR CI: it works under weak instruments, and it is as efficient as the Wald CI under strong instruments. One problem with the CI, however, is that it is not bet-proof, as discussed in Müller and Norets (2016): because its coverage is exactly 95%, and because we know that if the CI is given by $\mathbb{R}$, its conditional coverage is 100%, this means that its conditional coverage (conditional on knowing the shape of the CI) when it's not the whole real line is lower than 95%: therefore one can make money by betting against it. The formal argument is given by Theorem 1 in Müller and Norets (2016). To make it bet-proof, we'd need to enlarge it when it doesn't consist of the whole real line.[4]
- As a test, the AR test has an appealing optimality property: it is uniformly most powerful (UMP) among all unbiased tests. This was shown in Moreira (2009); I give a self-contained proof in Section B.
- The AR test is used little in practice. Motivated by this, Lee et al. (2022) propose an alternative procedure called tF that is based on the observation that the worst-case rejection of the Wald test occurs at $\rho = 1$. When $\rho = 1$, the t-statistic depends on the data only through the first-stage F, and they use this observation to derive critical values $c_\alpha(F)$ that depend on it. These critical values tend to infinity as $F \to z^2_{1-\alpha/2}$.

ESTIMATION AND SIGN-SCREENING No estimator can be fully immune to bias since it is impossible to construct a consistent or at least median unbiased estimator when the instruments are irrelevant.

4. The bet-proofness concept is related to the Cox (1958) problem: the alternative, shorter CI we considered there is not bet-proof.

However, we can construct an unbiased estimator if the sign of the first stage is known, as pointed out in Andrews and Armstrong (2017). In particular, they point out that we can construct a unique unbiased estimator of $1/\pi$ if $\pi \geq 0$ as

$$\frac{1}{\mathcal{V}_{22}^{1/2}} m(t_1),$$

where $m(x) = (1 - \Phi(x))/\phi(x)$ is the Mills' ratio, and $t_1 = \hat{\pi}/\mathcal{V}_{22}^{1/2}$ is the first-stage t-statistic (so $F = t_1^2$). Then $\hat{\pi}_\perp = \hat{\delta} - \mathcal{V}_{12}/\mathcal{V}_{22} \cdot \hat{\pi}$ (the part of $\hat{\delta}$ that's independent of $\hat{\pi}$) has expectation $\pi\beta - \mathcal{V}_{12}/\mathcal{V}_{22}\pi$, an unbiased estimator of β can be constructed as

$$\hat{\beta}_U = t_1 m(t_1)\hat{\beta}_{\text{TSLS}} + (1 - t_1 m(t_1))\beta_{WOLS},$$

where $\beta_{WOLS} = \mathcal{V}_{12}/\mathcal{V}_{22}$ is the weak-instrument limit of OLS. The curious thing is that if $t_1 \geq 0$—that is, the first-stage is right-signed—then the estimator shrinks TSLS towards OLS. The shrinkage interpretation of $\hat{\beta}_U$ seems surprising: since $\hat{\beta}_{\text{TSLS}}$ is biased towards OLS, shrinkage towards OLS increases bias. This counterintuitive fact arises because the estimator $\hat{\beta}_U$ is unbiased by virtue of averaging a conditional positive bias when $t_1 > 0$ and with conditional negative bias when $t_1 < 0$.

It is hard to imagine an analyst who is prepared to sign the population first stage while ignoring the sign of the estimated first stage. Such conditioning, however, strips $\hat{\beta}_U$ of its appeal. What about $\hat{\beta}_{\text{TSLS}}$? Angrist and Kolesár (2024) show that sign-screening actually halves the median bias of TSLS, without reducing coverage: in contrast with procedures that screen on the *magnitude* of the first-stage F statistic, screening on the sign of the corresponding t-statistic has to have little effect on rejection rates for a conventional Wald test.

Question 3. What is the practical takeaway?

3.4. *Overidentified case*

This case is much more complicated:

- We can generalize Fieller's idea: Since

  $$\hat{\delta} - \beta\hat{\pi} \sim \mathcal{N}(0, Q^{-1}E[b'\Omega(X_i)b\tilde{Z}_i\tilde{Z}_i']Q^{-1}/n),$$

  the test that rejects the null $H_0\colon \beta = \beta_0$ whenever

  $$AR = n(\hat{\delta} - \beta_0\hat{\beta}_2)' \left(QE[b'\Omega(X_i)b\tilde{Z}_i\tilde{Z}_i']^{-1}Q\right)(\hat{\delta} - \beta_0\hat{\beta}_2)$$

  is greater than the 95th quantile of χ_k^2 will have size 5%. This test is called the Anderson and Rubin (1949) test. The idea generalizes naturally to GMM models.

  However, the bet-proofness problem becomes even more severe, since the resulting

CI will be empty with positive probability. Furthermore, the CI is no longer efficient under standard asymptotics: it is longer than the Wald CI.

The reason that the CI may be empty is that it tests the joint null the TSLS estimand being equal to β_0, and δ being proportional to π.

- The conditional likelihood ratio (CLR) test suggested by Moreira (2003) is more powerful than AR, and enjoys some optimality properties under homoskedastic errors (Andrews, Moreira, and Stock 2006) (that do not, however, carry over to the heteroskedastic case). It is available in Stata.
- In the just identified case, we need values of $|\rho|$ that are unreasonably high in cross-section applications (very close to 1) to generate severe overrejection of the Wald test: it is hard to come up with empirical examples where Wald CIs are substantively misleading. This changes when $k > 1$, as we've seen from the intertemporal elasticitity of substitution example.
- There is no procedure when $k > 1$ that allows for treatment effect heterogeneity.

APPENDICES

A. DERIVATIONS

Proof of Equation (3). Let $\Delta = E[W_i W_i']^{-1} E[W_i Z_i']$. We can think of $\hat{\delta}, \hat{\pi}$ and $\ddot{Z}'D/n$ as method of moments estimators based on the moment condition $E[g(\mathcal{D}_i, \theta)]$, where $\theta = \text{vec}(\delta, \pi, Q\pi, \text{vec}(\Delta'))$, and

$$g(\mathcal{D}_i, \theta) = \begin{pmatrix} \text{vec}((Z_i - \Delta' W_i)(Y_i - Z_i'\delta, D_i - Z_i'\pi)) \\ (Z_i - \Delta' W_i) D_i - Q\pi \\ \text{vec}((Z_i - \Delta' W_i) W_i') \end{pmatrix} = \begin{pmatrix} \tilde{u}_i \otimes \tilde{Z}_i \\ \tilde{Z}_i D_i - Q\pi \\ W_i \otimes \tilde{Z}_i \end{pmatrix}, \tag{17}$$

where $\tilde{u}_i = u_i + \Psi' W_i$. The derivative of this moment condition is

$$\Gamma = -\begin{pmatrix} I_2 \otimes Q & 0 & \Psi' E[W_i W_i'] \otimes I_K \\ 0 & I_K & E[D_i W_i'] \otimes I_K \\ 0 & 0 & E[W_i W_i'] \otimes I_K \end{pmatrix}, \quad \Gamma^{-1} = \begin{pmatrix} I_2 \otimes Q^{-1} & 0 & -\Psi' \otimes Q^{-1} \\ 0 & I_K & -(\pi'\Delta + \psi') \otimes I_K \\ 0 & 0 & E[W_i W_i']^{-1} \otimes I_K \end{pmatrix},$$

while its variance is

$$\Sigma = E\begin{pmatrix} \tilde{u}_i \tilde{u}_i' \otimes \tilde{Z}_i \tilde{Z}_i' & \tilde{u}_i \otimes (\tilde{Z}_i \tilde{Z}_i' D_i - \tilde{Z}_i \pi' Q) & \tilde{u}_i W_i' \otimes \tilde{Z}_i \tilde{Z}_i' \\ \tilde{u}_i' \otimes (\tilde{Z}_i \tilde{Z}_i' D_i - Q\pi \tilde{Z}_i') & (\tilde{Z}_i D_i - Q\pi)(\tilde{Z}_i D_i - Q\pi)' & W_i' \otimes (\tilde{Z}_i \tilde{Z}_i' D_i - Q\pi \tilde{Z}_i') \\ W_i \tilde{u}_i' \otimes \tilde{Z}_i \tilde{Z}_i' & W_i \otimes (\tilde{Z}_i \tilde{Z}_i' D_i - \tilde{Z}_i \pi' Q) & W_i W_i' \otimes \tilde{Z}_i \tilde{Z}_i' \end{pmatrix}.$$

The upper block of $\Gamma^{-1}\Sigma\Gamma^{-1'}$ is given by

$$\mathcal{V}_0 = E\begin{pmatrix} \Omega(X_i) \otimes Q^{-1} \tilde{Z}_i \tilde{Z}_i' Q^{-1} & u_i \otimes (Q^{-1} \tilde{Z}_i (\tilde{Z}_i' \overline{D}_i - \pi' Q)) \\ u_i' \otimes ((\overline{D}_i \tilde{Z}_i - Q\pi) \tilde{Z}_i' Q^{-1}) & (\tilde{Z}_i \overline{D}_i - Q\pi)(\tilde{Z}_i \overline{D}_i - Q\pi)' \end{pmatrix} = \text{var}\begin{pmatrix} u_i \otimes Q^{-1} \tilde{Z}_i \\ \tilde{Z}_i \overline{D}_i - Q\pi \end{pmatrix},$$

where $\overline{D}_i = D_i - E[D_i W_i'] E[W_i W_i']^{-1} W_i = \tilde{Z}_i' \pi + u_{2i}$, and $\Omega(X_i) = E[u_i u_i' \mid X_i]$ is the conditional variance of the reduced-form errors. Since $\tilde{Z}_i \overline{D}_i - Q\pi = (\tilde{Z}_i \tilde{Z}_i' - Q)\pi + u_{2i} \tilde{Z}_i$, this yields the

result. □

Proof of Equation (6) and Equation (8). TSLS can be thought of as an estimator of

$$\beta = g(\theta) = \delta' Q\pi / \pi' Q\pi,$$

with θ defined as in eq. (17). The derivative of this function is given by

$$G = \frac{1}{\pi' Q\pi} \begin{pmatrix} b \otimes Q\pi \\ \pi_\Delta, \end{pmatrix}$$

where $\pi_\Delta = \delta - \beta\pi$. Hence, using the fact that $\pi' Q\pi_\Delta = 0$ by definition of β,

$$G'\mathcal{V}_0 G = E\left[u_{\Delta,i}^2(\pi'\tilde{Z}_i)^2 + 2u_{\Delta,i}\pi'\tilde{Z}_i(\overline{D}_i\tilde{Z}_i'\pi_\Delta) + (\overline{D}_i\tilde{Z}_i'\pi_\Delta)^2\right] = E\left[\left(\epsilon_i \cdot \pi'\tilde{Z}_i + u_{2i} \cdot \tilde{Z}_i'\pi_\Delta\right)^2\right]$$

where $b'u_i = u_{\Delta,i}$ and $\epsilon_i = \tilde{Z}_i\pi_\Delta + u_{\Delta,i} = (Y_i - D_i\beta) - W_i' E[W_i W_i]^{-1} E[W_i(Y_i - D_i\beta)]$. This yields eq. (8). If $\pi_\Delta = 0$, then one obtains eq. (6). □

Proof of Equation (16). In a model with normal homoskedastic reduced form errors

$$\begin{pmatrix} (\ddot{Z}'\ddot{Z})^{1/2}(\hat{\delta} - \hat{\pi}\beta)/\sqrt{b'\Omega b} \\ (\ddot{Z}'\ddot{Z})^{1/2}\hat{\pi}/\Omega_{22}^{1/2} \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes \lambda \sim \begin{pmatrix} \sqrt{1-\rho^2} & \rho \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} \mathcal{Z}_1 \\ \mathcal{Z}_2 \end{pmatrix},$$

where $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are standard normal k-vectors, and $\lambda = (\ddot{Z}'\ddot{Z})^{1/2}\pi/\Omega_{22}^{1/2}$. Recall also that $\beta_{WOLS} = \Omega_{12}/\Omega_{22} = \rho\sqrt{b'\Omega b/\Omega_{22}} + \beta$. Therefore,

$$\begin{aligned} \hat{\beta}_{\text{TSLS}} - \beta &= \frac{\hat{\pi}'\ddot{Z}'\ddot{Z}'(\hat{\delta} - \hat{\pi}\beta)}{\hat{\pi}'\ddot{Z}'\ddot{Z}'\hat{\pi}} \\ &= \sqrt{\frac{b'\Omega b}{\Omega_{22}}(1-\rho^2)}\frac{\mathcal{Z}_1'(\mathcal{Z}_2 + \lambda)}{(\mathcal{Z}_2+\lambda)'(\mathcal{Z}_2+\lambda)} + (\beta_{WOLS} - \beta)\frac{(\mathcal{Z}_2+\lambda)'\mathcal{Z}_2}{(\mathcal{Z}_2+\lambda)'(\mathcal{Z}_2+\lambda)}. \end{aligned}$$

By Proposition 1(2) in Bao and Kan (2013), the expectation of both terms exists. The expectation of the first term is zero by iterated expectations. As noted in the proof of Lemma 4 in Magnus (1986), for $x > 0$, it follows by setting $z = tx$ in the gamma function identity $(s-1)! = \Gamma(s) = \int_0^\infty z^{s-1}e^{-z}dz$ that $1/x^s = \frac{1}{(s-1)!}\int_0^\infty t^{s-1}e^{-tx}dt$. With $s = 1$, we therefore get, by Fubini's theorem, that for any X_1, X_2 with $X_2 > 0$,

$$E[X_1/X_2] = \int_0^\infty E[X_1 e^{-tX_2}]dt$$

Let $X_1 = (\mathcal{Z}_2+\lambda)'\lambda$, $X_2 = (\mathcal{Z}_2+\lambda)'(\mathcal{Z}_2+\lambda)$, and $\mu^2 = \lambda'\lambda$. We have

$$\begin{aligned} E[X_1^s e^{-tX_2}] &= \int \lambda'(z+\lambda)e^{-t(z+\lambda)'(z+\lambda)}\frac{1}{(2\pi)^{k/2}}e^{-z'z/2}dz \\ &= e^{-\frac{2t}{2t+1}\frac{\mu^2}{2}}\frac{1}{(2t+1)^{k/2}}E_{Z\sim\mathcal{N}_k(-\frac{2t}{2t+1}\lambda, I_k/(2t+1))}[\lambda'(Z+\lambda)] = e^{-\mu^2/2}e^{\frac{1}{2t+1}\frac{\mu^2}{2}}\frac{1}{(2t+1)^{k/2+1}}\mu^2, \end{aligned}$$

where the second line follows by "completing the square". Hence,

$$\frac{E[\hat{\beta}_{\text{TSLS}} - \beta]}{\beta_{WOLS} - \beta} = 1 - E[X_1/X_2] = 1 - \frac{\mu^2}{2} e^{-\mu^2/2} \int_0^\infty \frac{2}{(2t+1)^{k/2+1}} e^{\frac{1}{2t+1}\frac{\mu^2}{2}} dt$$
$$= 1 - \frac{\mu^2}{2} e^{-\mu^2/2} \int_0^1 x^{k/2-1} e^{x\mu^2/2} dx$$

which yields the result. Recall that

$$M(a,b,z) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 e^{zu} u^{a-1} (1-u)^{b-a-1} \, du \qquad \square$$

is Kummer's confluent hypergeometric function (sometimes denoted ${}_1F_1(a;b;z)$). We can therefore equivalently write the integral as $\frac{2}{k} M(k/2, k/2+1, \mu^2/2)$.

Proof of Equation (14). Write the limit experiment as

$$\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} \sim \begin{pmatrix} \beta\pi \\ \pi \end{pmatrix} + \begin{pmatrix} \sqrt{\mathcal{V}_{11} - \mathcal{V}_{12}^2/\mathcal{V}_{22}} & \mathcal{V}_{12}/\mathcal{V}_{22}^{1/2} \\ 0 & \mathcal{V}_{22}^{1/2} \end{pmatrix} \begin{pmatrix} \mathcal{Z}_\perp \\ \mathcal{Z}_2 \end{pmatrix},$$

where $\mathcal{Z}_\perp$ and $\mathcal{Z}_2$ are independent standard normal. The asymptotic variance of $\hat{\beta}$ is given by $E[\epsilon_i^2(\tilde{Z}_i'\pi)^2]/E[(\tilde{Z}_i'\pi)^2]^2 = E[\epsilon_i^2\tilde{Z}_i^2]/(E[\tilde{Z}_i^2]^2\pi^2) = (\mathcal{V}_{11} - 2\mathcal{V}_{12}\beta + \mathcal{V}_{22}\beta^2)/\pi^2$, and the asymptotic variance estimator is given by $(\mathcal{V}_{11} - 2\mathcal{V}_{12}\hat{\beta} + \mathcal{V}_{22}\hat{\beta}^2)/\hat{\pi}^2$. Thus, the Wald statistic can then be written as

$$W = \frac{\hat{\beta} - \beta}{\sqrt{(\hat{\pi}/\mathcal{V}_{22}^{1/2})^{-2}(\hat{\beta}^2 - 2\hat{\beta}\mathcal{V}_{12}/\mathcal{V}_{22} + \mathcal{V}_{11}/\mathcal{V}_{22})}}.$$

Let $\mathcal{Z}_\epsilon = (\mathcal{V}_{11}/\mathcal{V}_{22} + \beta^2 - 2\beta\mathcal{V}_{12}/\mathcal{V}_{22})^{-1/2}(\sqrt{\mathcal{V}_{11}/\mathcal{V}_{22} - \mathcal{V}_{12}^2/\mathcal{V}_{22}^2}\mathcal{Z}_\perp + (\mathcal{V}_{12}/\mathcal{V}_{22} - \beta)\mathcal{Z}_2)$. Note that $\mathcal{Z}_\epsilon$ is standard normal, with $E[\mathcal{Z}_\epsilon\mathcal{Z}_2] = \rho$. Then

$$\hat{\beta} - \beta = \frac{(\mathcal{V}_{11}/\mathcal{V}_{22} + \beta^2 - 2\beta\mathcal{V}_{12}/\mathcal{V}_{22})^{1/2}\mathcal{Z}_\epsilon}{\lambda + \mathcal{Z}_2}, \qquad \hat{\pi}/\mathcal{V}_{22}^{1/2} = \lambda + \mathcal{Z}_2,$$

Hence,

$$\hat{\beta}^2 - 2\hat{\beta}\mathcal{V}_{12}/\mathcal{V}_{22} + \mathcal{V}_{11}/\mathcal{V}_{22} = \left(\mathcal{V}_{11}/\mathcal{V}_{22} + \beta^2 - 2\beta\mathcal{V}_{12}/\mathcal{V}_{22}\right)\left(\frac{Z_\epsilon^2}{(\lambda+\mathcal{Z}_2)^2} - 2\rho\frac{\mathcal{Z}_\epsilon}{\lambda+\mathcal{Z}_2} + 1\right)$$

and the Wald statistic has the distribution

$$W = \frac{\mathcal{Z}_\epsilon/(\lambda+\mathcal{Z}_2)}{\sqrt{\frac{1}{(\lambda+\mathcal{Z}_2)^2}\left(\frac{Z_\epsilon^2}{(\lambda+\mathcal{Z}_2)^2} - 2\rho\frac{\mathcal{Z}_\epsilon}{\lambda+\mathcal{Z}_2} + 1\right)}} = \frac{\text{sign}(\lambda+\mathcal{Z}_2)\mathcal{Z}_\epsilon}{\sqrt{Z_\epsilon^2/(\lambda+\mathcal{Z}_2)^2 - 2\rho\mathcal{Z}_\epsilon/(\lambda+\mathcal{Z}_2) + 1}}.$$

For numerical results, observe that the rejection region of the Wald statistic is quadratic in $\mathcal{Z}_\epsilon$:

$$[(\lambda+\mathcal{Z}_2)^2 - z_{1-\alpha/2}^2]\mathcal{Z}_\epsilon^2 + 2\rho z_{1-\alpha/2}^2(\lambda+\mathcal{Z}_2)\mathcal{Z}_\epsilon - z_{1-\alpha/2}^2(\lambda+\mathcal{Z}_2)^2 \geq 0$$

If $(\lambda+\mathcal{Z}_2)^2 \geq z_{1-\alpha/2}^2$, we reject for large values of $\mathcal{Z}_\epsilon$. If $(\lambda+\mathcal{Z}_2)^2 \leq (1-\rho^2)z_{1-\alpha/2}^2$, then determinant, $4(\lambda+\mathcal{Z}_2)^2 z_{1-\alpha/2}^2\left[(\lambda+\mathcal{Z}_2)^2 - (1-\rho^2)z_{1-\alpha/2}^2\right]$, is negative, and we never reject. Otherwise, we reject for $\mathcal{Z}_\epsilon$ in an interval. To compute the rejection probabilities, condition on $\mathcal{Z}_2$, and use the fact that $P(\mathcal{Z}_\epsilon < x \mid \mathcal{Z}_2) = \Phi((x - \rho\mathcal{Z}_2)/\sqrt{1-\rho^2})$. $\square$

B. OPTIMALITY OF ANDERSON-RUBIN

We use the following result:

Theorem 9 (Lehmann and Romano 2005, Theorem 4.4.1). Suppose the statistic (U, T) has distribution $C(\theta, \vartheta)e^{\theta u + \vartheta' t}d\nu(u, t)$ with respect to some measure ν. Then a test that rejects if $U \notin [C_1(T), C_2(T)]$, where the cutoffs are chosen to ensure that the test is unbiased and has size α conditional on T is UMP unbiased unconditionally.

Proof. The proof uses the fact that the conditional distribution $U \mid T$ is $C_t(\theta)e^{\theta u}d\nu_t(u)$, which is an exponential family with no nuisance parameters, and T is sufficient for ϑ if we fix θ. Therefore, there exists a UMP unbiased test, that rejects if U is outside an interval, where the cutoffs are chosen to satisfy a size and an unbiasedness restriction. We then need to argue these conditional results imply UMP unbiasedness unconditionally. □

In our case, it follows from eq. (12), that the data are given by $Y := (\hat{\delta}, \hat{\pi})' \sim \mathcal{N}(\pi a, \Omega)$, where we define $\pi := \pi$, $a := (\beta, 1)'$, and $\Omega := \mathcal{V}$. Hence, the density is proportional to

$$e^{\pi a'\Omega^{-1}Y} = e^{\pi a_0'\Omega^{-1}Y + \pi(a-a_0)'\Omega^{-1}Y} = e^{\pi a_0'\Omega^{-1}Y + \theta e_1'\Omega^{-1}Y},$$

where we let $a_0 = (\beta_0, 1)$, $\theta = (\beta - \beta_0)\pi$, and π is the nuisance parameter ϑ. To get the density of $U = e_1'\Omega^{-1}Y$ conditional on $T = a_0'\Omega^{-1}Y$, observe that since $\left(b, e_2\right)^{-1} = \left(e_1, a\right)'$, where $b = (1, -\beta)'$, we have

$$\begin{pmatrix} e_1 & a \end{pmatrix}' \Omega^{-1} \begin{pmatrix} e_1 & a \end{pmatrix} = \left(\begin{pmatrix} b & e_2 \end{pmatrix}' \Omega \begin{pmatrix} b & e_2 \end{pmatrix} \right)^{-1} = \frac{1}{|\Omega|} \begin{pmatrix} \Omega_{22} & -e_2'\Omega b \\ -e_2'\Omega b & b'\Omega b \end{pmatrix}. \tag{18}$$

Post-multiplying both sides by $\left(b, e_2\right)' = \left(\begin{smallmatrix} 1 & -\beta \\ 0 & 1 \end{smallmatrix}\right)$ then yields

$$\begin{pmatrix} e_1 & a \end{pmatrix}' \Omega^{-1} = \frac{1}{|\Omega|} \begin{pmatrix} \Omega_{22} & -e_2'\Omega b - \Omega_{22}\beta \\ -e_2'\Omega b & b'\Omega b + e_2'\Omega b\beta \end{pmatrix}$$

Pre-multiplying both sides by $(b'\Omega b, e_2'\Omega b)$ yields

$$b'\Omega b e_1'\Omega^{-1} + e_2'\Omega b a'\Omega^{-1} = b, \tag{19}$$

where we use the fact that

$$\left| \begin{pmatrix} b & e_2 \end{pmatrix}' \Omega \begin{pmatrix} b & e_2 \end{pmatrix} \right| = |\Omega|.$$

Now,

$$\begin{pmatrix} e_1'\Omega^{-1}Y \\ a_0'\Omega^{-1}Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} e_1'\Omega^{-1}a\pi \\ a_0'\Omega^{-1}a\pi \end{pmatrix}, \begin{pmatrix} e_1 & a_0 \end{pmatrix}' \Omega^{-1} \begin{pmatrix} e_1 & a_0 \end{pmatrix} \right).$$

Thus, by the formula for the conditional normal distribution,

$$e_1'\Omega^{-1}Y \mid a_0'\Omega^{-1}Y \sim \mathcal{N}\left(m, \frac{1}{|\Omega| \cdot a_0'\Omega^{-1}a_0}\right),$$

where

$$\begin{aligned} m &= e_1'\Omega^{-1}a\pi + \frac{e_1'\Omega^{-1}a_0}{a_0'\Omega^{-1}a_0}(a_0'\Omega^{-1}Y - a_0'\Omega^{-1}a\pi) \\ &= e_1'\Omega^{-1}a\pi - \frac{e_2'\Omega b_0}{b_0'\Omega b_0}(a_0'\Omega^{-1}Y - a_0'\Omega^{-1}a\pi) = \frac{1}{b_0'\Omega b_0}b_0'a\pi - \frac{e_2'\Omega b_0}{b_0'\Omega b_0}a_0'\Omega^{-1}Y \\ &= \frac{\theta}{b_0'\Omega b_0} - \frac{e_2'\Omega b_0}{b_0'\Omega b_0}a_0'\Omega^{-1}Y. \end{aligned}$$

Here the second equality uses eq. (18), and the third equality uses eq. (19). Since $\theta = 0$ under the null, for both a one- and a two-sided test, we reject for large values of the z-statistic

$$\frac{\left[e_1'\Omega^{-1} + \frac{e_2'\Omega b_0}{b_0'\Omega b_0}a_0'\Omega^{-1}\right]Y}{1/\sqrt{|\Omega|a_0'\Omega^{-1}a_0}} = \frac{\frac{1}{b_0'\Omega b_0}b_0'Y}{1/\sqrt{|\Omega|a_0'\Omega^{-1}a_0}} = \frac{b_0'Y}{\sqrt{b_0'\Omega b_0}},$$

where the first equality uses eq. (19), and the second one uses eq. (18).

REFERENCES

Anderson, Theodore W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *The Annals of Mathematical Statistics* 20, no. 1 (March): 46–63. https://doi.org/10.1214/aoms/1177730090.

Andrews, Donald W. K., and Patrik Guggenberger. 2010. "Asymptotic Size and a Problem with Subsampling and with the m Out Of n Bootstrap." *Econometric Theory* 26, no. 02 (April): 426. https://doi.org/10.1017/S0266466609100051.

Andrews, Donald W. K., Marcelo J. Moreira, and James H. Stock. 2006. "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression." *Econometrica* 74, no. 3 (May): 715–752. https://doi.org/10.1111/j.1468-0262.2006.00680.x.

Andrews, Isaiah, and Timothy B. Armstrong. 2017. "Unbiased Instrumental Variables Estimation under Known First-Stage Sign." *Quantitative Economics* 8, no. 2 (July): 479–503. https://doi.org/10.3982/QE700.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11, no. 1 (August): 727–753. https://doi.org/10.1146/annurev-economics-080218-025643.

Angrist, Joshua, and Michal Kolesár. 2024. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." *Journal of Econometrics* 240, no. 2 (March): 105398. https://doi.org/10.1016/j.jeconom.2022.12.012.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens. 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies* 67, no. 3 (July): 499–527. https://doi.org/10.1111/1467-937X.00141.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity." *Journal of the American Statistical Association* 90, no. 430 (June): 431–442. https://doi.org/10.1080/01621459.1995.10476535.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, no. 4 (November): 979–1014. https://doi.org/10.2307/2937954.

Bao, Yong, and Raymond Kan. 2013. "On the Moments of Ratios of Quadratic Forms in Normal Random Variables." *Journal of Multivariate Analysis* 117 (May): 229–245. https://doi.org/10.1016/j.jmva.2013.03.002.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90, no. 430 (June): 443–450. https://doi.org/10.1080/01621459.1995.10476536.

Campbell, John Y. 2003. "Consumption-Based Asset Pricing." Chap. 13 in *Handbook of the Economics of Finance*, edited by G. M. Constantinides, M. Harris, and R. Stulz, vol. 1B, 803–887. Amsterdam: Elsevier. https://doi.org/10.1016/S1574-0102(03)01022-7.

Chamberlain, Gary. 2007. "Decision Theory Applied to an Instrumental Variables Model." *Econometrica* 75, no. 3 (May): 609–652. https://doi.org/10.1111/j.1468-0262.2007.00764.x.

Cox, David Roxbee. 1958. "Some Problems Connected with Statistical Inference." *The Annals of Mathematical Statistics* 29, no. 2 (June): 357–372. https://doi.org/10.1214/aoms/1177706618.

Evdokimov, Kirill, and Michal Kolesár. 2018. "Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects." January. https://www.princeton.edu/~mkolesar/papers/het_iv.pdf.

Fieller, Edgar C. 1932. "The Distribution of the Index in a Normal Bivariate Population." *Biometrika* 24, nos. 3/4 (November): 428–440. https://doi.org/10.1093/biomet/24.3-4.428.

Fieller, Edgar C. 1940. "The Biological Standardization of Insulin." *Supplement to the Journal of the Royal Statistical Society* 7 (1): 1–64. https://doi.org/10.2307/2983630.

———. 1954. "Some Problems in Interval Estimation." *Journal of the Royal Statistical Society. Series B (Methodological)* 16, no. 2 (July): 175–185. https://doi.org/10.1111/j.2517-6161.1954.tb00159.x.

Goldberger, Arthur S., and Ingram Olkin. 1971. "A Minimum-Distance Interpretation of Limited-Information Estimation." *Econometrica* 39, no. 3 (May): 635–639. https://doi.org/10.2307/1913273.

Hahn, Jinyong, and Jerry A. Hausman. 2002. "A New Specification Test for the Validity of Instrumental Variables." *Econometrica* 70, no. 1 (January): 163–189. https://doi.org/10.1111/1468-0262.00272.

Hausman, Jerry A., James H. Stock, and Motohiro Yogo. 2005. "Asymptotic Properties of the Hahn-Hausman Test for Weak-Instruments." *Economics Letters* 89, no. 3 (December): 333–342. https://doi.org/10.1016/j.econlet.2005.06.007.

Heckman, James J., and Edward J. Vytlacil. 2005. "Structural Equations, Treatment Effects and Econometric Policy Evaluation." *Econometrica* 73, no. 3 (May): 669–738. https://doi.org/10.1111/j.1468-0262.2005.00594.x.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62, no. 2 (March): 467–475. https://doi.org/10.2307/2951620.

Kolesár, Michal. 2013. "Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity." Working paper, Princeton University, November. https://www.princeton.edu/~mkolesar/papers/late_estimation.pdf.

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter. 2022. "Valid t-Ratio Inference for IV." *American Economic Review* 112, no. 10 (October): 3260–3290. https://doi.org/10.1257/aer.20211063.

Lee, Seojeong. 2018. "A Consistent Variance Estimator for 2SLS When Instruments Identify Different LATEs." *Journal of Business & Economic Statistics* 36, no. 3 (July): 400–410. https://doi.org/10.1080/07350015.2016.1186555.

Lehmann, Erich L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses*. 3rd ed. New York, NY: Springer. https://doi.org/10.1007/0-387-27605-X.

Magnus, Jan R. 1986. "The Exact Moments of a Ratio of Quadratic Forms in Normal Variables." *Annales d'Économie et de Statistique*, no. 4, 95–109. https://doi.org/10.2307/20075629.

Mariano, Roberto S., and James B. McDonald. 1979. "A Note on the Distribution Functions of LIML and 2SLS Structural Coefficient in the Exactly Identified Case." *Journal of the American Statistical Association* 74, no. 368 (December): 847–848. https://doi.org/10.1080/01621459.1979.10481040.

Montiel Olea, José Luis, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics* 31, no. 3 (July): 358–369. https://doi.org/10.1080/00401706.2013.806694.

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71, no. 4 (July): 1027–1048. https://doi.org/10.1111/1468-0262.00438.

———. 2009. "Tests with Correct Size When Instruments Can Be Arbitrarily Weak." *Journal of Econometrics* 152, no. 2 (October): 131–140. https://doi.org/10.1016/j.jeconom.2009.01.012.

Müller, Ulrich K. 2011. "Efficient Tests under a Weak Convergence Assumption." *Econometrica* 79, no. 2 (March): 395–435. https://doi.org/10.3982/ECTA7793.

Müller, Ulrich K., and Andriy Norets. 2016. "Credibility of Confidence Sets in Nonstandard Econometric Problems." *Econometrica* 84, no. 6 (November): 2183–2213. https://doi.org/10.3982/ECTA14023.

Nelson, Charles R., and Richard Startz. 1990a. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58, no. 4 (July): 967–976. https://doi.org/10.2307/2938359.

———. 1990b. "The Distribution of the Instrumental Variables Estimator and Its *t*-Ratio When the Instrument Is a Poor More." *The Journal of Business,* A Conference in Honor of Merton H. Miller's Contributions to Finance and Economics, 63, no. 1 (January): S125–S140. https://doi.org/10.1086/296497.

Richardson, David H. 1968. "The Exact Distribution of a Structural Coefficient Estimator." *Journal of the American Statistical Association* 63, no. 324 (December): 1214–1226. https://doi.org/10.1080/01621459.1968.10480921.

Sawa, Takamitsu. 1972. "Finite-Sample Properties of the k-Class Estimators." *Econometrica* 40, no. 4 (July): 653–680. https://doi.org/10.2307/1912960.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65, no. 3 (May): 557–586. https://doi.org/10.2307/2171753.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." Chap. 5 in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by Donald W. K. Andrews and James H. Stock, 80–108. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511614491.006.

Wooldridge, Jeffrey Marc. 2010. *Econometric Analysis of Cross Section and Panel Data.* 2nd ed. Cambridge, MA: MIT Press. ISBN: 0-262-23258-8.

Yogo, Motohiro. 2004. "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak." *Review of Economics and Statistics* 86, no. 3 (August): 797–810. https://doi.org/10.1162/0034653041811770.

Young, Alwyn. 2022. "Consistency without Inference: Instrumental Variables in Practical Application." *European Economic Review* 147 (August): 104112. https://doi.org/10.1016/j.euroecorev.2022.104112.

Zellner, Arnold. 1970. "Estimation of Regression Relationships Containing Unobservable Independent Variables." *International Economic Review* 11, no. 3 (October): 441–454. https://doi.org/10.2307/2525323.

———. 1978. "Estimation of Functions of Population Mean and Regression Coefficients Including Structural Coefficients: A Minimum Expected Loss (MELO) Approach." *Journal of Econometrics* 8, no. 2 (October): 127–158. https://doi.org/10.1016/0304-4076(78)90024-6.