Robust Inference on Macro Equations with Shock Proxies

Ryohei Oishi

University College London

March 2026

Structural macroeconomic equations and endogeneity

- Consider a linear(ized) structural macro equation represented as

$$y_t = w_t'\delta + u_t, \quad E[w_t u_t] \neq 0$$

- Take the New Keynesian Phillips curve as an example:

$$\pi_t = \lambda x_t + \gamma_f E_t(\pi_{t+1}) + \varepsilon_t^c$$

  Two sources of endogeneity
  - Measurement error in x_t and $E_t(\pi_{t+1})$
  - Simultaneous-equation bias due to ε_t^c

- Instrumental variables...? Predetermined variables may not be exogenous (c.f., DSGE) (Hansen and Singleton 1982, Galí and Gertler 1999)

A new approach in the literature: shock proxies as IV

- Barnichon and Mesters (2020 QJE): A sequence of **shock proxies**
  - Exogeneity: Proxies are more likely to satisfy exogeneity
  - Relevance: Weak instruments. **Anderson-Rubin (AR) test**
  - Many instruments: Almon parameterization

- Still, the econometric toolkit is far from complete
  1. The AR statistic has some disadvantages under overidentification (DF, rejection against overidentifying restrictions)
  2. Barnichon and Mesters (2020) tend to over-reject the null

My proposal: KLM and the restricted long-run variance

1. The **Almon-parametrized KLM** (Kleibergen) statistic: $DF = \#$ str params. Pivotal under the weak IV asymptotics
2. Apply the **restricted HAR long-run variance** Accurate size control for both the AR and KLM statistics

Two applications: the Phillips curve and fiscal rules

- Does the size-accuracy matter in practice? Yes, it does!
- With the proposed statistic, either a completely flat or a very steep Phillips curve is no longer rejected (cf. Barnichon and Mesters 2020)
- Finite-sample size distortion can result in an opposite conclusion
- I also study fiscal policy rules and multipliers, and find a statistically significant tax multiplier

Related Literature

- Weak-instrument-robust inference on macro equations:

  Barnichon Mesters (2020), Lewis Mertens (2024); Euler Eq: Yogo (2004), Ascari Magnusson Mavroeidis (2021); NKPC: Kleibergen Mavroeidis (2009), Magnusson Mavroeidis (2014), Mavroeidis Plagborg-Møller Stock (2014); Taylor rule: Mavroeidis (2010), Carlevaro Haque Magnusson (2025); among many many others

  Overidentification. Size accuracy. Widely-applicable method

- HAR inference and finite-sample size distortion:

  Lazarus Lewis Stock Watson (2018), Vogelsang (2018); Newey West (1986), Andrews (1991); Keifer Vogelsang (2000, 2002, 2005), Jansson (2004), Sun (2014)...

  Weak-IV robust inference. Restricted variance. Fixed-b

- Fiscal rule/multiplier estimation:

  Caldara Kamps (2017), Mertens Ravn (2013, 2014), Angelini Caggiano Castelnuovo Fanelli (2023), Keweloh Klein Prüser (2025); Montiel Olea Mertens (2018), Hebous and Zimmermann (2018)...

  Complements proxy-SVAR approaches. First statistically significant evidence on the tax multiplier with weak instruments

The model in one slide

- Consider the linear model

$$\boxed{y_t = w_t'\delta + u_t}$$

where $E[w_t u_t] \neq 0, \quad dim(\delta) = m \leq 3$

- Suppose there exists a sequence of shock proxies

$$\xi_{t:t-H} = (\xi_t, \cdots, \xi_{t-H})',$$

where $E[u_t \xi_s] = 0, \forall t, s$ (relevance condition discussed later)

- Apply the Almon parametrization

$$\boxed{z_t = \left(\sum_{h=0}^{H} \xi_{t-h}, \sum_{h=0}^{H} h\xi_{t-h}, \sum_{h=0}^{H} h^2 \xi_{t-h} \right)'}$$

and assume

$$\boxed{w_t = \Pi' z_t + v_t}$$

How to choose H? Why the Almon parametrization? "Regression in impulse response space"

- Regression in impulse response space (Barnichon and Mesters, 2020)

$$\mathcal{R}_h^\pi = \gamma_f \mathcal{R}_{h+1}^\pi + \lambda \mathcal{R}_h^x, \text{ for } h = 0, 1, 2, \ldots, H.$$

where $\mathcal{R}_h^y = E(y_t \xi_{t:t-H})$ is the vector of impulse responses of variable y to the shock ξ

- Empirical IRFs are typically different from zero for three to five years: $H \in [12, 20]$ for quarterly data provides information for identification but could be "too many"

- The Almon-param mitigates the many-instrument problem

- Smooth macro IRFs. A shape restriction: the IRF is quadratic
$u_t = \theta^{(0)}\xi_t + \theta^{(1)}\xi_{t-1} + \cdots \theta^{(H)}\xi_{t-H}, \; \theta^{(h)} = a + bh + c^2h$
$\Leftrightarrow u_t = \left(\sum_{h=0}^H \xi_{t-h}, \sum_{h=0}^H h\xi_{t-h}, \sum_{h=0}^H h^2\xi_{t-h}\right)\theta_a, \theta_a = (a,b,c)'$

Weak IV robust inference with Almon parametrization

- Goal: Construct confidence sets robust to weak instruments
- How: Test inversion
- Weak-instrument robust test: correct asymptotic coverage for $H_0 : \delta = \delta_0$ regardless of the instrument strength

1. Just-identification: the Almon-parameterized AR statistic
   Barnichon and Mesters (2020)
2. Over-identification: the Almon-parameterized KLM statistic, $DF = m < k = 3$

- Applying the restricted HAR long-run variance is important for size control
- Next: Test statistics, asymptotics, and finite-sample properties

AR test with the restricted long-run variance

- Intuition: Consider regression $u_t(\delta_0) := y_t - w_t'\delta_0 = z_t'\theta_a + \eta_t$. θ_a should be zero if the exogeneity condition holds

$$\boxed{AR_a^r(\delta_0) = \hat{\theta}_a' \widetilde{\Sigma}_{\theta_a}^{-1} \hat{\theta}_a = \frac{(y - W\delta_0)' P_Z (y - W\delta_0)}{\tilde{s}_u^2}},$$

where

$$\hat{\theta}_a = (Z'Z)^{-1} Z'(y - W\delta_0)$$

$$\widetilde{\Sigma}_{\theta_a} = (Z'Z)^{-1} \tilde{s}_u^2$$

$$\widetilde{s}_u^2 = \frac{1}{n-H} \sum_{t=H+1}^{n} \sum_{s=H+1}^{n} \widetilde{u}_t \widetilde{u}_s \kappa((t-s)/b_n), \quad \kappa(\cdot) \text{ is some kernel},$$

$$\widetilde{u}_t = \widetilde{\widetilde{u}}_t - \frac{1}{n-H} \sum_{t=H+1}^{n} \widetilde{\widetilde{u}}_t, \quad \widetilde{\widetilde{u}}_t = (y_t - w_t'\delta_0) - z_t' 0.$$

AR test with the unrestricted long-run variance

- Barnichon and Mesters (2020) use the unrestricted long-run variance, $\hat{s}_u^2$: AR_a
- Just replace $\tilde{\tilde{u}}_t$ with the following unrestricted version

$$\hat{u}_t = (y_t - w_t'\delta_0) - z_t'\hat{\theta}_a$$

The limiting distribution of the AR test statistics

Theorem (Barnichon and Mestes (2020))

Let Assumption 1 hold. Under H_0: $\delta = \delta_0$ for $\frac{H}{n} \to c \in (0,1)$ as $n \to \infty$,

$$AR_a(\delta_0) \Rightarrow \chi^2(3).$$

No restrictions on Π in Assumption 1: Weak-IV robust

Corollary

Under the same conditions above,

$$AR_a^r(\delta_0) \Rightarrow \chi^2(3).$$

KLM statistic: overidentified models

- The Almon-parametrized KLM statistic:

$$\boxed{K_a^{r}(\delta_0) = \frac{(y - W\delta_0)' P_{Z\widetilde{\Pi}(\delta_0)}(y - W\delta_0)}{\tilde{s}_u^2}}$$

where

$$\widetilde{\Pi}(\delta_0) = (Z'Z)^{-1}Z'\left[W - (y - W\delta_0)\frac{\tilde{s}_{uV}}{\tilde{s}_u^2}\right]$$

- Intuition: Projection on $Z\widetilde{\Pi}$, instead of Z; Decorrelation trick. Quadratic form of the score of the concentrated log-likelihood
- $K_a(\delta_0)$ is defined similarly by using the unrestricted variance $\hat{s}_u^2$

The limiting distribution of the KLM statistics

- Three types of instrument validity:

  (i) Instruments are relevant: Π is fixed and full rank

  (ii) Instruments are weakly relevant: $\Pi = \Pi_n = K_n^{-1} C$ with C fixed and full rank, and $K_n = diag(n, n^2, n^3)$

  (iii) Instruments are irrelevant: $\Pi = 0$

Theorem

Let Assumption 1 hold. Under H_0: $\delta = \delta_0$ for $\frac{H}{n} \to c \in (0,1)$ as $n \to \infty$,

$$K_a^r(\delta_0), K_a(\delta_0) \Rightarrow \chi^2(m)$$

for all (i),(ii),(iii) cases.

Case (ii) is equivalent to the Staiger-Stock framework

- Case (ii) Instruments are weakly relevant: $\Pi = \Pi_n = K_n^{-1} C$ with C fixed and full rank, and $K_n = diag(n, n^2, n^3)$
- This is not "super" weak, but the standard local to zero weak-instrument asymptotics with Almon parameterization
- The signal component of the concentration parameter does not diverge: $\mu := \Pi'(Z'Z)\Pi = C'(K_n^{-1} Z'Z K_n^{-1}) C \Rightarrow C'\Psi C$
- The signal of additional information is the same

$$\pi_n \xi_t = c n^{-1/2} \underbrace{\xi_t}_{O_p(1)} = O_p(n^{-1/2})$$

$$\Pi_n' z_t = C' \begin{bmatrix} n^{-1/2} & 0 & 0 \\ 0 & n^{-1/2} & 0 \\ 0 & 0 & n^{-1/2} \end{bmatrix} \underbrace{\begin{bmatrix} n^{-1/2} \sum_{h=0}^{H} \xi_{t-h} \\ n^{-3/2} \sum_{h=0}^{H} h \xi_{t-h} \\ n^{-5/2} \sum_{h=0}^{H} h^2 \xi_{t-h} \end{bmatrix}}_{O_p(1) \text{ by FCLT}} = O_p(n^{-1/2})$$

Finite-sample properties

- $\{AR_a^r, AR_a\}$ and $\{K_a^r, K_a\}$ are asymptotically equivalent, but have different finite-sample properties
- The HAR inference literature: standard inference—using the **unrestricted** long-run variance with χ^2 **critical values**—induces over-rejection in finite samples (Lazarus, Lewis, Stock, Watson 2018). This applies to AR_a of Barnichon and Mesters (2020)
- Using the **restricted** long-run variance helps to reduce finite-sample size distortion (LLSW 2018; Vogelsang 2018). This applies to AR_a^r and K_a^r
- Combining AR_a and K_a with the fixed-b asymptotics may also help to control size (a series of Kiefer and Vogelsang; Sun 2014)

Simulation setup

- DGP based on a hybrid New Keynesian Phillips curve:

$$\pi_t = \gamma_b \pi_{t-1} + \gamma_f E_t(\pi_{t+1}) + \lambda x_t + e_t$$

$$x_t = \rho_1 x_{t-1} + \rho_2 x_{t-2} + \varepsilon_t + \nu e_t$$

  where $e_t = \rho e_{t-1} + \sqrt{1-\rho^2}\zeta_t$ with $\zeta_t \sim N(0,1)$ and $\varepsilon_t \sim N(0, \sigma_i^2)$

- $[\gamma_b, \gamma_f] = [0.6, 0.4]$, implying $\gamma_b + \gamma_f = 1$, a common restriction in empirical studies (also $[\gamma_b, \gamma_f] = [0.6, 0.3]$ in the paper)

- Transform the model by imposing the restriction: $\delta = (\gamma_f, \lambda)$

- Run $10,000$ simulations and compute the rejection frequency with different ρ and σ_i

Simulation results with $\gamma_b + \gamma_f = 1$ ($\alpha = 5\%, T = 100$)

σ_i	ρ	AR_a^r	AR_a	K_a^r	K_a	K^r	K
0.10	0.0	5.0	10.4	3.5	8.8	86	91
0.25	0.0	4.9	10.4	3.9	8.7	78	82
0.50	0.0	5.4	10.1	4.8	8.8	63	62
1.00	0.0	6.6	9.5	5.7	8.2	46	34
0.10	0.5	5.5	12.4	3.5	10.0	8	21
0.25	0.5	5.1	12.1	3.3	9.5	8	19
0.50	0.5	5.3	12.1	3.5	9.8	10	20
1.00	0.5	6.3	11.6	4.4	9.6	13	21

- AR_a^r and K_a^r have accurate size, while AR_a and K_a over-reject
- Both K^r and K overject: The Almon-parametrization is necessary
- Fixed-b improves AR_a, K_a, but too conservative for AR_a^r, K_a^r (in the paper)
- Robust when $\alpha = 0.1, 0.32$ and without $\gamma_b + \gamma_f = 1$ for AR_a, AR_a^r

Phillips curve

- Analyze the hybrid New Keynesian Phillips curve

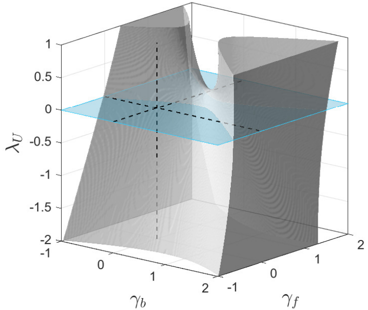
$$\pi_t = \gamma_b \pi_{t-1} + \gamma_f E_t(\pi_{t+1}) + \lambda x_t + \varepsilon_t^s$$

  with and without the common restriction $\gamma_b + \gamma_f = 1$
- Instruments are monetary policy shock proxies

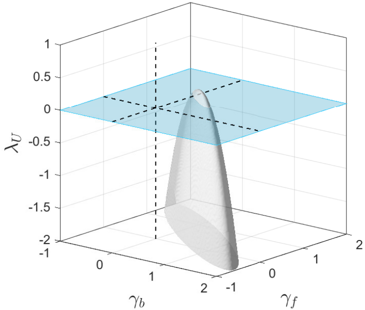
Data

- I use the same dataset as Barnichon and Mesters (2020)
- Variables:
  - Inflation: PCE excluding food and energy prices
  - Unemployment gap: HP-detrended unemployment rate
  - Output gap: HP-detrended GDP
- Instruments:
  - 1969–2007 Romer–Romer narrative series
  - 1990–2017 High-frequency identified (HFI) series: Δ_{FOMC} 3-month-ahead FF futures + Δ_{FOMC} 10-year yield
- The weak-IV test of Lewis and Mertens (2025): these proxies are weak Results

1969–2007: The Romer-Romer shock



(a) AR_a^r with $\chi^2(3)$

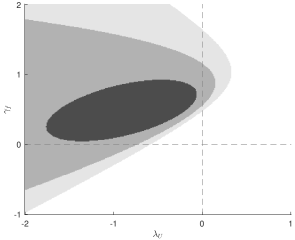


(b) AR_a with $\chi^2(3)$

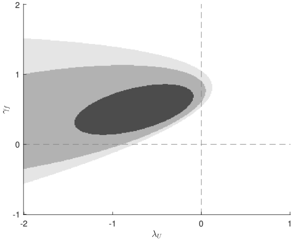
Figure: 95% confidence sets

- Much wider confidence sets with AR_a than AR_a^r
- May reach an opposite conclusion between AR_a^r and AR_a regarding the significance of λ_U and the area of γ_f and γ_b

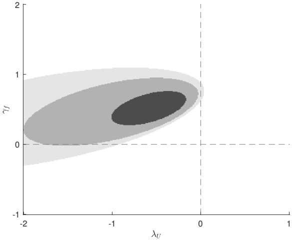
Similar results with the common restriction $\gamma_b + \gamma_f = 1$



(a) AR_a^r with $\chi^2(3)$



(b) AR_a with $\chi^2(3)$



(c) Barnichon and Mesters

- Non-significant slope is consistent with Lewis and Mertens (2022)
- γ_f could be larger than one and negative
- The figure in Barnichon and Mesters (2020) shows much tighter confidence sets, such as significant λ_U, because they use $\chi^2(2)$

Robustness

Similar results when...

- the forcing variable is the output gap
- the sample period is 1990–2017 with the HFI proxy
- AR_a with the fixed-b critical values is used
- K_a^r and K_a are used under $\gamma_b + \gamma_f = 1$

Fiscal policy rules (tax or spending) and the multipliers

- The simple fiscal rule

$$p_t = \psi^p_{gdp} y_t + e^p_t$$

  where $p \in \{tr, g\}$

- There exists a mapping between ψ^p_{gdp} and the impact multiplier given a reduced form VAR model (Caldara and Kamps 2017, CK)

- The general fiscal rule (in the paper)

$$p_t = \psi^p_{gdp} y_t + \psi^p_{\pi} \pi_t + \psi^p_r r_t + e^p_t$$

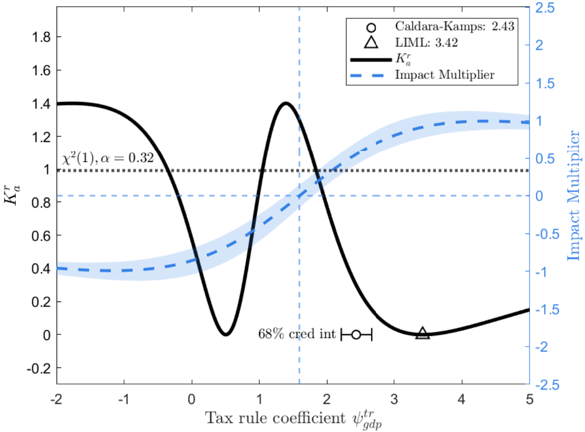
- Pros: No assumption on the specification of a proxy-SVAR and identification strength

- Cons: Dynamic multiplier, different sample period between IV and data

Data

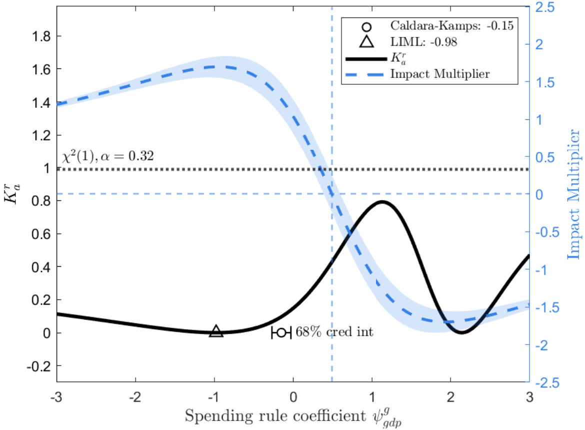
- I use the same dataset as CK
- The reduced-form VAR has 5 variables: federal tax revenue, government spending (consumption + investment), GDP, CPI, and 3-month T-Bill rate. The first three are real per capita values
- Instruments: Fernald (2012) TFP shocks + Oil shock of Hamilton (2003) for the general rule
- Sample period: 1950-2006 using quarterly data
- The weak-IV test of Lewis and Mertens (2025): these proxies are weak Results

The simple tax rule and the multiplier



- The tax multiplier is significantly positive and approximately unity
- CK's credible interval is consistent with the figure, but narrower
- Short-run restriction of Blanchard and Perotti (2002) $\psi_{gdp}^{tr} = 2.08$ is not rejected (marginally), but the uncertainty is large

The simple spending rule and the multiplier



- ψ_{gdp}^g is not statistically significant, while the multiplier is, loosely speaking, likely to be above unity
- CK's credible interval is consistent with the figure, but narrower
- The restriction of Blanchard and Perotti (2002) $\psi_{gdp}^g = 0$ is not rejected, but negative ψ_{gdp}^g is equaly consistent with the data

Conclusion

Theoretically:

1. A weak-IV robust method for overidentified macro equations
2. HAR size control in finite samples

The recipe:

- Just-identified: Almon-parametrized AR statistic
- Overidentified: Almon-parametrized KLM statistic
- For both, use the restricted long-run variance

Empirically:

1. The previous findings of a significant slope of the NKPC are mostly driven by size distortion
2. A significant tax multiplier using a weak-IV robust method; however, the uncertainty remains large

- Assumption
- Supplementary results

Assumptions I

ASSUMPTION 1. The observations $\{y_t, w_t, z_t\}$ are generated by the linear IV model

$$y_t = w_t'\delta + u_t \Leftrightarrow y = W\delta + u$$

$$w_t = z_t'\Pi + v_t \Leftrightarrow W = Z\Pi + V$$

and

$$z_t = \left(\sum_{h=0}^{H} \xi_{t-h}, \sum_{h=0}^{H} h\xi_{t-h}, \sum_{h=0}^{H} h^2\xi_{t-h} \right)'$$

for $t = H+1, \ldots, n$, where δ is $m \times 1$, Π is $k \times m$ with $m \leq k = 3$, and $\eta_t = (\xi_t, u_t, v_t')$.

Theorem

Assumptions II

Further, assume that

(i) for all t and s, we have (a) $\mathrm{E}[\eta_t] = 0$, (b) $\mathrm{E}[u_t \xi_s] = 0$, and (c) $\mathrm{E}[v_t \xi_s] = 0$,

(ii) for some $r > 2$ and finite constant Δ, we have $\sup_t \|\eta_t\|_{2r} \leq \Delta$,

(iii) η_t is L_2-NED (near-epoch dependence) of size $-\frac{r-1}{r-2}$ with $d_t = 1$ on V_t, where $\{V_t\}$ is an α-mixing process of size $-\frac{r}{r-2}$, Theorem

Assumptions III

(iv) for integers $p, q \geq 0$ we have uniformly in n and H, with $H < n$, that

$$\omega^2_{\xi,p,n,H} = \mathrm{Var}\left[\sum_{t=H+1}^{n} t^p \xi_t\right] = \omega^2_{\xi,p}(n-H)^{2p+1} + o((n-H)^{2p+1})$$

$$\Omega_{uv,q,n,H} = \mathrm{Var}\left[\sum_{t=H+1}^{n} t^q (u_t, v_t')'\right] = \Omega_{uv,q}(n-H)^{2q+1} + o((n-H)^{2q+1}),$$

where $\Omega_{uv,q} = \begin{bmatrix} \omega^2_{u,q} & \omega_{uv,q} \\ \omega_{vu,q} & \Omega_{vv,q} \end{bmatrix}$ with finite $\omega^2_{\xi,p} > 0$ and positive definite $\Omega_{uv,q}$, and $\omega^2_{u,0}$ denotes the long-run variance of u_t and $\omega_{uv,0}$ denotes the long-run covariance of u_t and v_t. Theorem

Assumptions IV

(v) $b_n = o(n)$ and $\kappa(\cdot) \in \mathcal{K}$ where

$$\mathcal{K} = \left\{ \begin{array}{l} \kappa(\cdot) : \mathbb{R} \to [-1, 1], \kappa(0) = 1, \\ \kappa(x) = \kappa(-x) \; \forall x \in \mathbb{R}, \int_{-\infty}^{\infty} |\kappa(x)| dx < \infty, \\ \int_{-\infty}^{\infty} \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa(x) e^{ivx} dx \right| dv < \infty, \\ \kappa(\cdot) \text{ is continuous at 0 and at all except for} \\ \text{a finite number of points.} \end{array} \right\}$$

Theorem

Assumptions V

Definition (Near epoch dependent (NED) stochastic processes)

A sequence of integrable random vectors $\{X_t\}$ is L_2-NED on a stochastic sequence $\{V_t\}$ on probability space $(\Omega, \mathcal{F}, P)$ if for $m \geq 0$

$$\|X_t - \mathrm{E}(X_t | \mathcal{F}_{t-m}^{t+m})\|_2 < d_t v_m,$$

where $\mathcal{F}_s^t = \sigma(V_s, \ldots, V_t) \subset \mathcal{F}$, $t \geq s$, d_t is a sequence of non-negative constants, and $v_m \to 0$ as $m \to \infty$. We will say that the sequence is L_2-NED of size $-s$ when $v_m = O(m^{-s-\varepsilon})$ for some $\varepsilon > 0$.

Theorem

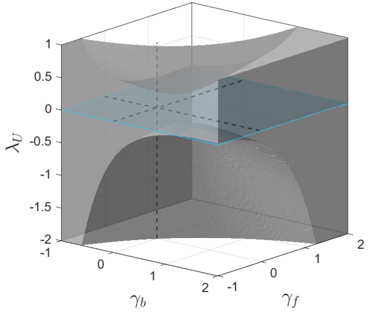
Weak-IV test: RR and HFI are weak instruments

	$\gamma_b, \lambda_U, \gamma_f$		$\gamma_b + \gamma_f = 1, \lambda_U$	
Unemployment gap	g	cv	g	cv
RR	9.3	26.0	10.9	26.0
HFI	4.6	22.9	4.8	22.9

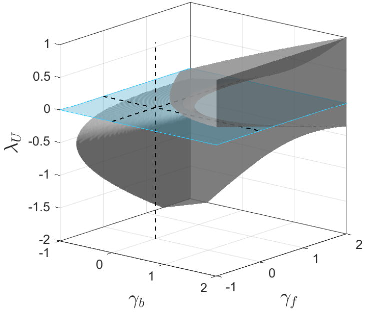
	$\gamma_b, \lambda_Y, \gamma_f$		$\gamma_b + \gamma_f = 1, \lambda_Y$	
Output gap	g	cv	g	cv
RR	5.4	25.0	5.8	25.0
HFI	8.3	21.7	8.8	21.7

Note: An instrument strength test by Lewis and Mertens (2025), testing the null hypothesis of weak instrument bias less than or equal to 10% of the worst-case benchmark. g is the test statistic, and cv is the critical value at the 5% significance level. Back

1990–2017: The HFI proxy



(a) AR_a with $\chi^2(3)$

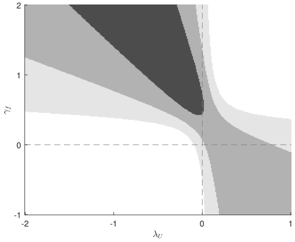


(b) AR_a^r with $\chi^2(3)$

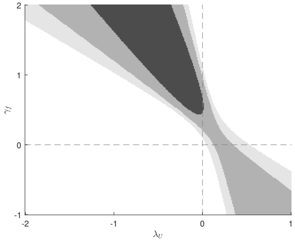
Figure: 95% confidence sets

- Similar results with the Romer-Romer proxy
- A wider region of positive λ_U. Negative γ_f, γ_b are not rejected

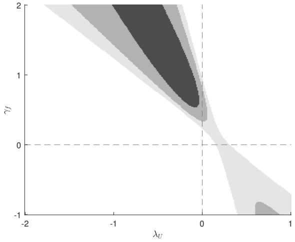
Similar results with the common restriction $\gamma_b + \gamma_f = 1$



(a) AR_a^r with $\chi^2(3)$



(b) AR_a with $\chi^2(3)$



(c) Barnichon and Mesters

- Positive λ_U and negative γ_f are more likely with AR_a^r
- The figure in Barnichon and Mesters (2020) presents much tighter confidence sets because they use $\chi^2(2)$
- Similar results with K_a^r and K_a

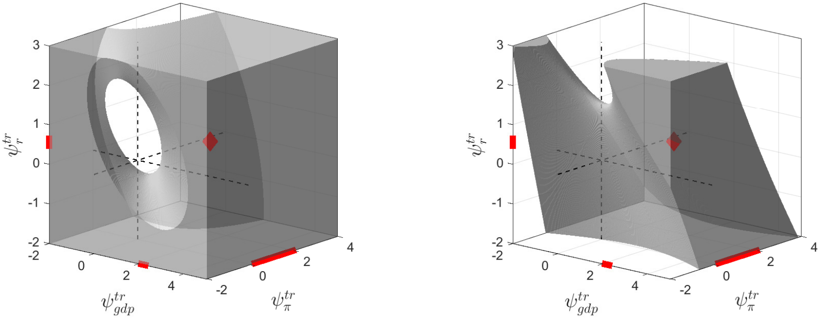
Weak-IV test: The proxies are weak

Tax	Simple		General	
	g	cv	g	cv
TFP	3.2	11.4	0.01	23.3
Oil	-	-	2.5	35.1

Spending	Simple		General	
	g	cv	g	cv
TFP	3.2	11.3	0.01	23.5
Oil	-	-	2.5	35.3

Note: An instrument strength test by Lewis and Mertens (2025), testing the null hypothesis of weak instrument bias less than or equal to 10% of the worst-case benchmark. g is the test statistic, and cv is the critical value at the 5% significance level. Back

General tax rule



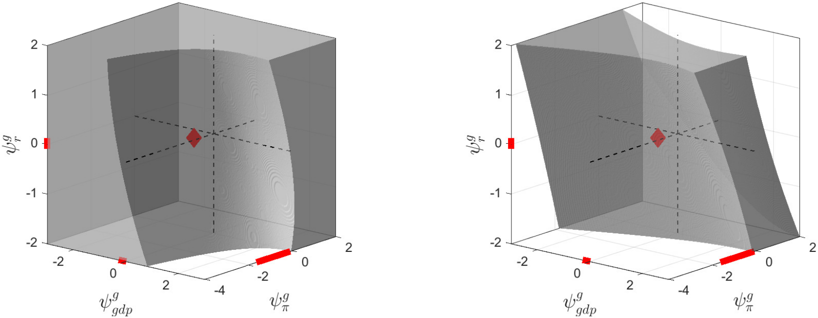
(a) Fernald's TFP shock

(b) Hamilton's oil shock

Figure: 68% confidence sets by AR_a^r and $\chi^2(3)$. The red lines are the 68% credible intervals of CK. A diamond (triangle) indicates that CK's point estimate is (is not) included in the confidence sets.

- Positive ψ_{gdp}^{tr} is not generally rejected, implying a positive multiplier
- CK results are not rejected, but the uncertainty is much larger

General spending rule



(a) Fernald's TFP shock (b) Hamilton's oil shock

Figure: 68% confidence sets. Red lines are the 68% credible intervals of CK. A diamond (triangle) indicates that CK's point estimate is (is not) included in the confidence sets.

- For all cases, ψ_{gdp}^g is more likely to be negative, indicating a positive multiplier
- CK's credible interval is mostly contained in the confidence set