1 Andrews, Stock, and Sun (2019)

This article focuses on outlining the weak instruments problem and discussing the most common strategies for ensuring inference is robust to weak instruments. They consider the linear IV model:

$$Y_i = X_i'\beta + \varepsilon_i \tag{1}$$

$$X_i' = Z_i'\pi + \upsilon_i' \tag{2}$$

where Y_i is our scalar outcome, X_i is a p-vector, and Z_i is a q-vector. We refer to (1) as the *structural form* and (2) as the *first stage*. To get the *reduced form* we plug (2) into (1) and have:

$$Y_i = Z_i'\delta + \nu_i \tag{3}$$

The basic approach for estimating β is to observe that $\delta = \pi\beta$. Thus, assuming that π is full-rank, we have that $\beta = (\pi'\pi)^{-1}\pi'\delta$. In practice, if π is "close" to not being full rank, then the behavior of estimates of β using the inverse $\pi'\pi$ will be erratic. Notably, we expect confidence intervals and hypothesis tests to have generally have incorrect coverage and size.

The classic approach to this problem in the context of testing $H_0 : \beta = \beta_0$, is to use the OLS estimators of π and δ, and base tests on how far $\hat{\delta} - \hat{\pi}\beta_0$ is from 0. This is the idea of the Anderson-Rubin test (Anderson and Rubin, 1949).

The Anderson-Rubin test has excellent theoretical optimality properties in the just-identified case, but is under-powered in the case that $q > p$. There are a couple of proposed solutions to this. The first is the conditional likelihood ratio test (CLR) of Moreira (2003). This test is valid under homoskedasticity. Another procedure based on score tests is due to Kleibergen (2005). This test will have low power relative to conditional tests. Under heteroskedastic and autocorrelated errors, there is still no consensus procedure for weak instruments in over-identified settings.

2 Lee, McCrary, Moreira, and Porter (2022)

Given the prevalence of weak instruments in empirical work, another strategy is to "meet practitioners where they are:" develop valid inference procedures based on the two stage least squares estimator, and corresponding t-ratio statistic. Stock and Yogo (2005) encouraged testing the first-stage, and if the first-stage F (or heteroskedastic-robust version) is sufficiently large ($>$ 10 being the cutoff they propose) proceed with standard inference procedures. Lee et al. (2022) find that the necessary critical value is over 100 for this kind of two-step testing procedure to sufficiently control for type-I error for tests on the structural parameter β. Focusing on the just-identified case, they propose using a tF procedure, where the critical values for the second stage test using the t-ratio are adjusted using the first-stage F-statistic.

3 Key Ideas

- Why weak instruments are problematic with 2SLS: delta-method approximation of the distribution of $\hat{\beta}$ will be poor when $\hat{\pi}$ is close to singular.
- Simple case: $p = q = 1$. Then $\hat{\beta} = \hat{\delta}/\hat{\pi}$. The delta method:

$$\hat{\beta} - \beta_0 = g(\hat{\delta}, \hat{\pi}) - g(\delta, \pi) \tag{4}$$

$$= \begin{pmatrix} g_\delta(\tilde{\delta}, \tilde{\pi}) \\ g_\pi(\tilde{\delta}, \tilde{\pi}) \end{pmatrix}' \begin{pmatrix} \hat{\delta} - \delta \\ \hat{\pi} - \pi \end{pmatrix} \tag{5}$$

$$= \begin{pmatrix} \frac{1}{\tilde{\pi}} \\ -\frac{\tilde{\delta}}{\tilde{\pi}^2} \end{pmatrix}' \begin{pmatrix} \hat{\delta} - \delta \\ \hat{\pi} - \pi \end{pmatrix} \tag{6}$$

- When π is close to 0, in large samples $\tilde{\pi}$ (which is between $\hat{\pi}$ and π) will be close to 0, the approximation error from using $\hat{\pi}$ instead of $\tilde{\pi}$ will be quite large.
- Obviously when $\pi = 0$ the mean-value theorem does not apply.
- Confidence intervals and hypothesis tests based on the normality of $\hat{\beta} - \beta_0$ will perform poorly.
- Stock and Yogo (2005) promote testing if $\hat{\pi} \neq 0$, rejecting if the F-stat from the first-stage is larger than 10.
- Why 10? Original: Staiger and Stock (1997), approximate 5% test that bias is less than 10%.
- Stock and Yogo (2005): worst-case size distortion: 5% tests can be as bad as 15% tests.
- Lee et al. (2022): for first-stage F screening to ensure correct size: $F > 104.7$.
- Idea of Lee et al. (2022): use first-stage F to smoothly adjust the critical values for tests on β based on identification strength.
- Leads to confidence intervals strictly longer than standard intervals based on the normal model.
- Intervals will tend to be shorter (in terms of expected length) than Anderson-Rubin intervals.

References

ANDERSON, T. W. AND H. RUBIN (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of mathematical statistics*, 20, 46–63.

ANDREWS, I., J. H. STOCK, AND L. SUN (2019): "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 11, 727–753.

KLEIBERGEN, F. (2005): "Testing parameters in GMM without assuming that they are identified," *Econometrica*, 73, 1103–1123.

LEE, D. S., J. MCCRARY, M. J. MOREIRA, AND J. PORTER (2022): "Valid t-ratio Inference for IV," *American Economic Review*, 112, 3260–90.

MOREIRA, M. J. (2003): "A conditional likelihood ratio test for structural models," *Econometrica*, 71, 1027–1048.

STAIGER, D. AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica: Journal of the Econometric Society*, 557–586.

STOCK, J. AND M. YOGO (2005): "Asymptotic distributions of instrumental variables statistics with many instruments," *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 6, 109–120.