ECON-GA 2100 Notes on Weak and Many Instruments

Konrad Menzel
km125@nyu.edu

October 16, 2025

1 Overview

Without much loss of generality, let's restrict our attention to the case with one endogenous right-hand side variable $x_i \in \mathbb{R}$ and no exogenous regressors (potentially after having partialed out the included z_i)

$$\begin{aligned} y_i &= \beta x_i + \varepsilon_i \\ x_i &= z_i\pi + \nu_i \end{aligned}$$

where we assume that the errors (ε_i, ν_i) are jointly i.i.d. with $\mathbb{E}[\varepsilon_i|z_i] = \mathbb{E}[\nu_i|z_i] = 0$, but potentially correlated with each other, i.e.

$$\mathbb{E}[\varepsilon\varepsilon'|z] = \sigma_\varepsilon^2\mathbf{I}_n, \quad \mathbb{E}[\varepsilon\nu'|z] = \sigma_{\varepsilon\nu}\mathbf{I}_n, \quad \mathbb{E}[\nu\nu'|z] = \sigma_\nu^2\mathbf{I}_n$$

and the matrix of instruments z has rank K.
We showed in much greater generality that 2SLS is consistent and asymptotically normal, but did not make any statements about the finite-sample properties of the estimator. In fact, it is now known that in fact there exists no unbiased estimator for the linear IV model (Hirano and Porter, 2015). However, we can characterize the "small sample" properties of 2SLS using stochastic expansions that account for higher-order terms in the asymptotic approximation. For simplicity, we'll regard z_i as fixed, so we take expectations to be conditional on the instrumental variables.

A first approximation to the finite-sample bias of 2SLS - which we'll refine in the next section, but which nevertheless gives the full intuition for the mechanics behind the weak-instruments problem - can be calculated from the closed-form expression for the 2SLS coefficient

$$\begin{aligned} \mathbb{E}[\hat\beta_{2SLS}] - \beta &= \mathbb{E}\left[\frac{x'P_z\varepsilon}{x'P_zx}\right] \approx \frac{\mathbb{E}[x'P_z\varepsilon]}{\mathbb{E}[x'P_zx]} \\ &= \frac{\mathbb{E}[\pi'z'P_z\varepsilon] + \mathrm{tr}(\mathbb{E}[\nu'P_z\varepsilon])}{\pi'z'P_zz\pi + 2\mathbb{E}[\pi'z'P_z\nu] + \mathrm{tr}\ (\mathbb{E}[\nu'P_z\nu])} \\ &= \frac{\mathrm{tr}(\mathbb{E}[P_z\varepsilon\nu'])}{\pi'z'z\pi + \mathrm{tr}(\mathbb{E}[P_z\nu\nu'])} = \frac{\mathrm{tr}(P_z\mathbb{E}[\varepsilon\nu'|z])}{\pi'z'z\pi + \mathrm{tr}(P_z\sigma_\nu^2\mathbf{I}_n)} \\ &= \frac{\sigma_{\varepsilon\nu}\mathrm{tr}(P_z)}{\pi'z'z\pi + \sigma_\nu^2\mathrm{tr}(P_z)} = \frac{\sigma_{\varepsilon\nu}K}{n\pi'M_n\pi + \sigma_\nu^2K} \\ &= \frac{\sigma_{\varepsilon\nu}}{\frac{n}{K}R_*^2\mathrm{Var}(x) + \sigma_\nu^2} \end{aligned}$$

where $M_n := \frac{z'z}{n}$, $R_*^2 := \frac{\pi'z'z\pi}{n\mathrm{Var}(x)}$ is the theoretical R-squared of the first stage, and we used that

$$\mathrm{tr}(P_z) = \mathrm{tr}(z(z'z)^{-1}z') = \mathrm{tr}(z'z(z'z)^{-1}) = \mathrm{tr}(\mathbf{I}_K) = K$$

The approximation of the expectation of a ratio by a ratio of expectations seems plausible in large samples (since that step is exact for the plim), but as we'll see right below, a more sophisticated approach gives a slightly (though not qualitatively) different answer.
Now let's compare this to the bias of OLS - by a similar line of reasoning,

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{LS}] - \beta &= \mathbb{E}\left[\frac{x'\varepsilon}{x'x}\right] \approx \frac{\mathbb{E}[(z\pi+\nu)'\varepsilon]}{\pi' z' z\pi + \mathbb{E}[\nu'\nu]} \\ &= \frac{\sigma_{\varepsilon\nu}}{\pi' M\pi + \sigma_\nu^2} = \frac{\sigma_{\varepsilon\nu}}{\mathrm{Var}(x)}\end{aligned}$$

Comparing this to the 2SLS bias, we can see that

- 2SLS is biased *towards* OLS, though for $n\bar{R}^2_{FS}/K > 1$, it will be smaller.
- The bias of 2SLS goes away as the sample size increases, the OLS bias doesn't depend on sample size at all.
- for both OLS and 2SLS, the bias gets worse if the correlation between ε and ν is stronger, i.e. as the endogeneity problem gets worse.
- The bias of 2SLS gets worse as we increase the number of instruments
- 2SLS is more biased if the first stage is weak, i.e. the explained sum of squares $\pi' M\pi$ of the first stage regression is small relative to the residual variance σ_ν^2
- In particular, if there is no first stage, i.e. $\pi = 0$, $\mathrm{Bias(2SLS)} = \frac{\sigma_{\varepsilon\nu}}{\sigma_\nu^2} = \mathrm{Bias(LS)}$

Intuitively, if the projection on the instruments has a large number of degrees of freedom, 2SLS overfits the first stage, and a lot of noise "passes through" the projection, or in other words, averaging conditional on the instruments is not very effective in smoothing out the correlated disturbances. In a sense we can summarize the weak instruments and the many instruments problems together as the instruments "picking up too much noise" relative to the "signal" contained in the instrumental variables.

1.1 Higher-Order/"Small-Sample" Asymptotics

The derivation above was very heuristic, and in even slightly more complex problems, we'd probably not get very far with this type of reasoning. Therefore, we need some more refined way of thinking about a more effective way of characterizing the small-sample behavior of our estimator.
"Small-Sample Asymptotics" may sound like a contradiction in itself, but you should first remember that the only reason why we *ever* care about the asymptotic properties of estimators and statistics is that exact finite-sample calculations are generally tedious and often require strong assumptions on the distribution of the errors in our model. Therefore, asymptotics serve mainly the purpose of giving an easy-to-work-with, but credible *approximation* to the statistical properties of the estimator of interest in the (typically finite) sample we have to work with.
However, as you already noticed, if we do the wrong asymptotic experiment, some prominent features of our object of interest may actually wash out as we increase the sample size, and weak instruments are probably the prime example of this. For linear IV problems, the most prominent ways of doing asymptotics are the following

- standard large-sample / first-order asymptotics: we hold the number of instruments K and the first stage coefficient Π fixed, and let the sample size n go to infinity

- second-order asymptotics: same as previous, except that in addition to the first-order terms in the approximation, you keep track of terms of an order of up to $\frac{1}{n}$.[1]
- Staiger & Stock (1997) "weak IV" asymptotics: K is held fixed, n goes to infinity, and we let the first stage coefficient go to zero at a specific rate: $\Pi_n = \Pi/\sqrt{n}$. This thought experiment results in a non-normal limiting distribution.
- Bekker (1994) asymptotics: The size of the first stage is held constant, but the number of instruments increases at the same rate as the sample size, i.e. $\lim \frac{K}{n} = \alpha > 0$. This setup yields a normal limiting distribution.
- Edgeworth expansions of the probability distribution of the estimator (see Rothenberg(1983) or Phillips (1983)): this is a more technical literature in which the expansion is done in terms of the cumulants/the characteristic function of the distribution. The approximating distributions that follow from this approach are non-normal.

The derivation of 2SLS bias in the next section is a standard second-order approximation, whereas the other limiting experiments may be mentioned here and there in the following sections. It should be noted that for "moderate" deviations from the standard scenario the approximations obtained from the different asymptotics aren't too different, but they can differ wildly for "extreme" situations. Monte-Carlo evidence seems to suggest that in the latter case, Bekker asymptotics do best. The asymptotics with respect to n and $\frac{\sigma_{\varepsilon\nu}}{\sigma_\nu \sigma_{\varepsilon\nu}}$ turn out to work quite well, whereas the impact of using many instruments K on 2SLS seems to be grossly overstated by the approximations.

1.2 The Concentration Parameter

In terms of the empirical R-squared, we can rewrite the 2SLS bias as

$$\mathbb{E}[\hat{\beta}_{2SLS}] - \beta \approx \frac{\sigma_{\varepsilon\nu}}{\frac{n}{K} R^2_{FS}} = \frac{\sigma_{\nu\varepsilon}/\sigma^2_\nu}{\mu^2}$$

where we defined the *concentration parameter* as

$$\mu^2 := \frac{nR^2}{K(1-R^2)}$$

where R^2 denotes the (sample) R-squared from the first-stage regression. μ^2 essentially captures how much better our instruments do at predicting the endogenous regressor than any arbitrary instruments which are completely unrelated to x.
The concentration parameter is the most important measure of the quality of the instruments, and we'd speak of a weak instruments problem if the concentration parameter takes values $\mu^2 < 30$. Note that we are still looking at the model without additional exogenous regressors. Therefore, the R-squared refers to the first stage *after* partialing out the exogenous variables included in the second stage. Also, the definition of the parameter is in terms of the theoretical R-squared, so that in the context of estimation, we'd have to correct for the degree of over-identification K, which may well make the *estimated* concentration parameter negative.

[1] We say that a term X_n is of order 1 (or stochastically bounded - denoted $X_n = O_p(1)$) if X_n converges to either a non-zero constant or a stable distribution, and we say that $X_n = O_p(n^{-k})$ if $n^k X_n = O_p(1)$

2 Derivations for Higher-Order Bias of 2SLS

Following the approach in Hahn and Hausman (2003), we can derive a higher-order bias approximation for the distribution of the 2SLS coefficient as follows: Denoting $M_n = \frac{1}{n}z'z$, the 2SLS formula implies

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{2SLS} - \beta) &= (\frac{1}{n}x'z(z'z)^{-1}z'x)^{-1}\frac{1}{\sqrt{n}}x'z(z'z)^{-1}z'\varepsilon \\
&= \sqrt{n}\frac{(z\pi+\nu)'zM_n^{-1}z'\varepsilon}{(z\pi+\nu)'zM_n^{-1}z'(z\pi+\nu)} \\
&= \frac{\pi'M_nM_n^{-1}\frac{z'\varepsilon}{\sqrt{n}} + \frac{1}{\sqrt{n}}\nu'P_z\varepsilon}{\pi'M_nM_n^{-1}M_n\pi + \frac{2}{\sqrt{n}}\pi'M_nM_n^{-1}\frac{z'\nu}{\sqrt{n}} + \frac{\nu'P_z\nu}{n}} \\
&= \frac{\left(\pi'\frac{z'\varepsilon}{\sqrt{n}}\right) + \frac{1}{\sqrt{n}}\left(\nu'P_z\varepsilon\right)}{(\pi'M_n\pi) + \frac{2}{\sqrt{n}}\left(\pi'\frac{z'\nu}{\sqrt{n}}\right) + \frac{1}{n}\left(\nu'P_z\nu\right)}
\end{aligned}$$

where I put brackets around the terms which are going to be nonzero, but stochastically bounded in the limit by a LLN or a CLT. Now we do the following "trick" (I'm not going to argue why exactly this approximation works - if you're interested in the technical details, you should just look at the paper): since we are interested in what happens for large n, we can do a first-order Taylor expansion of the above expression in $x = \frac{1}{\sqrt{n}}$ around the limit $x_0 := \lim_n \frac{1}{\sqrt{n}} = 0$.

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{2SLS} - \beta) &= \frac{A_1 + xA_2}{B_1 + 2xB_2 + x^2B_3} \\
&\approx \frac{A_1 + x_0A_2}{B_1 + 2x_0B_2 + x_0^2B_3} + \frac{A_2(B_1 + 2x_0B_2 + x_0^2B_3) - (A_1 + x_0A_2)(2B_2 + 2x_0B_3)}{(B_1 + 2x_0B_2 + x_0^2B_3)^2}(x - x_0) \\
&\quad + O((x - x_0)^2) \\
&= \frac{A_1}{B_1} + \frac{A_2B_1 - 2A_1B_2}{B_1^2}x + O(x^2)
\end{aligned}$$

Note that for this type of expansion, I did not account for the terms in n *inside* the brackets. The rationale for this is that for a reasonably large sample, the terms $A_1, A_2, B_1, \ldots$ will converge to either fixed values or random variables with finite variance, so that we can treat the bracketed terms as constants and/or random numbers which aren't affected by a second-order expansion around the limit.
Plugging back in the expressions for the As and Bs, this becomes

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \approx \frac{\pi'\frac{z'\varepsilon}{\sqrt{n}}}{\pi'M_n\pi} + \frac{1}{\sqrt{n}}\left[\frac{\nu'P_z\varepsilon}{\pi'M_n\pi} - \frac{2\pi'\frac{z'\varepsilon}{\sqrt{n}}\pi'\frac{z'\nu}{\sqrt{n}}}{(\pi'M_n\pi)^2}\right]$$

This means that, in addition to the probability limit of that expression (the usual first-order approximations we've been doing all along) we get an adjustment term which should vanish as n increases. Note that this is a non-degenerate random variable, and in order to get the second-order bias, we can just take expectations. By the usual trace argument we can see that

$$\mathbb{E}\left[\frac{\nu'P_z\varepsilon}{n}\right] = \mathbb{E}\left[\mathrm{tr}\left(\frac{\nu'P_z\varepsilon}{n}\right)\right] = \mathrm{tr}\left(\frac{P_z\mathbb{E}[\varepsilon\nu']}{n}\right) = \sigma_{\varepsilon\nu}\mathrm{tr}(z(z'z)^{-1}z) = \sigma_{\varepsilon\nu}\mathrm{tr}(z'z(z'z)^{-1}) = \sigma_{\varepsilon\nu}K$$

Since the expectations of ν and ε conditional on z are zero,

$$\mathbb{E}[\sqrt{n}(\hat{\beta}_{2SLS} - \beta)] \approx \frac{\pi'\mathbb{E}\left[\frac{z'\varepsilon}{\sqrt{n}}\right]}{\pi'M_n\pi} + \frac{1}{\sqrt{n}}\left[\frac{K\sigma_{\nu\varepsilon}}{\pi'M_n\pi} - 2\frac{\pi'Z\mathbb{E}[\varepsilon'\nu|Z]Z'\pi}{(\pi'M_n\pi)^2}\right] = \frac{1}{\sqrt{n}}\frac{(K-2)\sigma_{\nu\varepsilon}}{\pi'M_n\pi}$$

Based on the approximation above, the expectations of the 2SLS estimator for the residual variance, $\hat{\sigma}^2_\varepsilon$, can be derived as follows (note that we now correct for the degrees of freedom because we are interested in the small-sample properties of the estimator):

$$\begin{aligned}
\hat{\sigma}^2_\varepsilon &= \frac{(y - x\hat{\beta}_{2SLS})'(y - x\hat{\beta}_{2SLS})}{n-1} = \frac{(\varepsilon - x(\hat{\beta}_{2SLS} - \beta))'(\varepsilon - x(\hat{\beta}_{2SLS} - \beta))}{n-1} \\
&= \frac{\varepsilon'\varepsilon}{n-1} - 2\frac{\varepsilon' x(\hat{\beta}_{2SLS} - \beta)}{n-1} + \frac{(\hat{\beta}_{2SLS} - \beta)x'x(\hat{\beta}_{2SLS} - \beta)}{n-1} \\
&= \frac{\varepsilon'\varepsilon}{n-1} - 2\frac{(\pi' z'\varepsilon + \nu'\varepsilon)(\hat{\beta}_{2SLS} - \beta)}{n-1} + \frac{(n\pi' M_n \pi + 2\pi' Z'\nu + \nu'\nu)(\hat{\beta}_{2SLS} - \beta)^2}{n-1}
\end{aligned}$$

Plugging in the expansion for $(\hat{\beta}_{2SLS} - \beta)$, and keeping terms of order up to $\frac{1}{n}$,

$$\hat{\sigma}^2_\varepsilon \approx \frac{\varepsilon'\varepsilon}{n-1} - 2\left[\frac{\nu'\varepsilon}{n-1}(\hat{\beta}_{2SLS} - \beta) + \frac{\pi' z'\varepsilon\varepsilon' z\pi}{(n-1)\pi' M_n \pi}\right] + \left[\frac{\nu'\nu}{n-1} + \frac{\pi' M_n \pi}{n-1}\right]\frac{\pi' z'\varepsilon\varepsilon' z\pi}{(\pi' M_n \pi)^2}$$

Taking expectations (and using the bias approximation for the 2SLS coefficient derived above)

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2_\varepsilon] &\approx \frac{n}{n-1}\sigma^2_\varepsilon - 2\left[\frac{\sigma^2_{\nu\varepsilon}(K-2)}{(n-1)\pi' M_n \pi} + \frac{\sigma^2_\varepsilon}{n-1}\right] + \frac{n}{n-1}\left[\frac{\sigma^2_\varepsilon \sigma^2_\nu}{\pi' M_n \pi} + \frac{\sigma^2_\varepsilon \pi' M_n \pi}{n\pi' M_n \pi}\right] \\
&= \frac{n}{n-1}\sigma^2_\varepsilon - \frac{1}{n-1}\left[2\frac{\sigma^2_{\nu\varepsilon}(K-2)}{R^2 \mathrm{Var}(x)} + \sigma^2_\varepsilon - \frac{\sigma^2_\varepsilon \sigma^2_\nu}{R^2 \mathrm{Var}(x)}\right] \\
&= \sigma^2_\varepsilon - \frac{2\varrho^2(K-2)\sigma^2_\varepsilon \sigma^2_\nu}{(n-1)R^2 \mathrm{Var}(x)} + \frac{\sigma^2_\varepsilon \sigma^2_\nu}{(n-1)R^2 \mathrm{Var}(x)} \\
&= \sigma^2_\varepsilon - \frac{2\sigma_{\varepsilon\nu} n}{n-1}\mathrm{Bias}(\hat{\beta}_{2SLS}) + \frac{\sigma^2_\nu}{n-1}\mathrm{asy.Var}(\hat{\beta}_{2SLS})
\end{aligned}$$

where $R^2 := \frac{\pi' z' z\pi}{x'x} = \frac{\frac{1}{n}\pi' z' z\pi}{\mathrm{Var}(x)}$ is the R-squared from the first stage, and $\varrho := \frac{\sigma_{\varepsilon\nu}}{\sigma_\varepsilon \sigma_\nu}$ is the correlation coefficient of ν and ε. The first part of the approximation is the contribution of the finite-sample bias - which is invariably negative for a large degree of overidentification ($K >> 1$) because the 2SLS projection overfits the first stage, and therefore absorbs part of the disturbance ε through its correlation with the first-stage error. This makes the fit in the second stage look better than it actually should be and attributes too much variation in y to x (note that this doesn't have anything to do with the sign of the true coefficient). The second term of this expansion comes from the variance with which the coefficient is estimated - recall that in this setup, the 2SLS variance is $\frac{\sigma^2_\varepsilon}{R^2 \mathrm{Var}(x)}$. Interestingly, the second-order approximation of the variance of the 2SLS estimator coincides with the first-order approximation.

3 Bias-Corrected 2SLS/Nagar

Remember that the 2SLS bias was

$$\mathbb{E}[\hat{\beta}_{2SLS}] - \beta \approx \frac{K\sigma_{\varepsilon\nu}}{x' P_Z x}$$

so in principle, we know everything we need to compute the bias except $\sigma_{\varepsilon\nu}$ and could just subtract everything off our 2SLS estimate. For the true parameter β, it is clearly true that

$$K\sigma_{\varepsilon\nu} = \frac{K}{n-K}\mathbb{E}[x(\mathbf{I}_n - P_Z)\varepsilon] = \frac{K}{n-K}\mathbb{E}[x(\mathbf{I}_n - P_Z)(y - \beta x)]$$

Denoting the vector $\hat{c} := \frac{K}{(n-K)x'P_Zx}(\mathbf{I}_n - P_Z)x$, we can therefore write the bias equation as

$$\mathbb{E}\hat{\beta}_{2SLS} \approx \beta + \hat{c}'(y - \beta x) \Leftrightarrow \beta \approx \frac{\mathbb{E}\hat{\beta}_{2SLS} - \hat{c}'y}{1 - \hat{c}'x}$$

Hence, even though we couldn't calculate the bias without an unbiased estimator, we can solve the bias approximation for the true coefficient and plug in the 2SLS estimator for its expectation in order to get the bias-corrected 2SLS (B2SLS)/Nagar estimator

$$\begin{aligned}\hat{\beta}_{Nagar} &= \frac{\hat{\beta}_{2SLS} - \hat{c}'y}{1 - \hat{c}'x} = \frac{x'P_Zy - \frac{K}{n-K}x'(I - P_Z)y}{x'P_Zx - \frac{K}{n-K}x'(I - P_Z)x} \\ &= \frac{(n - K + K)x'PZy - Kx'y}{(n - K + K)x'P_Zx - x'x} = \frac{x'P_Z'y - \frac{K}{n}x'y}{x'P_Zx - \frac{K}{n}x'x}\end{aligned}$$

where for the second equality I used that $\hat{\beta}_{2SLS} = \frac{x'P_Zy}{x'P_Zx}$, so that we can plug in for $\hat{c}$ and multiply through by $x'P_Zx$.

4 Limited Information Maximum Likelihood (LIML)

The following is based on the Newey (2004) notes which also derive the higher-order approximation of the distribution of LIML and 2SLS under Bekker (many IV) asymptotics. The LIML estimator is the maximum likelihood estimator for β in the Gaussian homoskedastic IV model

$$\begin{aligned}y_i &= x_i'\beta + \varepsilon_i \\ x_i &= z_i'\pi + \nu_i, \qquad (\varepsilon_i, \nu_i')'|z_i \sim N(0, \Sigma)\end{aligned}$$

For our simplified problem, LIML can be shown to be equivalent to the solution to the Least Variance Ratio problem[2]

$$\hat{\beta}_{LIML} = \arg\min_{\beta} \lambda(\beta) = \arg\min_{\beta} \frac{(y - x\beta)'P_Z(y - x\beta)}{(y - x\beta)'(y - x\beta)}$$

In order to see the relationship with 2SLS, note that

$$\hat{\beta}_{2SLS} = \arg\min_{\beta} \frac{(y - x\beta)'P_Z(y - x\beta)}{n\hat{\sigma}^2_{2SLS}}$$

where $\hat{\sigma}^2_{2SLS} := \frac{1}{n}(y - x\hat{\beta}_{2SLS})'P_Z(y - x\hat{\beta}_{2SLS}))$ is held *fixed* while we minimize over β. Therefore the (crucial) difference is that 2SLS holds the variance term fixed, whereas LIML updates the estimator of σ^2 "continuously" as we look for the optimal β. This matters because, as we'll see in a minute, the simultaneous optimization through the orthogonality condition $P_Z(y - x\beta) = 0$ and the residual variance alters the first-order conditions which pin down the estimator. Since the maximand is a somewhat more involved function of the parameter β, there is no closed-form solution for the LIML estimator, however there is a formulation of the estimator as a pencil/eigenvalue problem which allows to compute the estimator in two steps instead of solving the first-order conditions iteratively.

[2] In fact, one can rewrite the concentrated Limited Information Likelihood in a way such that the only term which depends on β is the log of the variance ratio.

4.1 Higher-Order Bias of LIML

In order to check for finite sample bias, we'll state the first-order conditions for LIML and show that they will in fact be satisfied (up to approximation error) in expectation at the true parameter value. Remember that for 2SLS, the first-order condition can be written as

$$\frac{1}{n}x'P_Z(y - x\hat{\beta}_{2SLS}) = 0$$

In contrast, taking first-order conditions for the LIML estimator with respect to β, we get

$$0 = \frac{x'P_Z(y - x\beta)}{(y - x\beta)'(y - x\beta)} - \frac{(y - x\beta)'P_Z(y - x\beta)}{[(y - x\beta)'(y - x\beta)]^2}x'(y - x\beta)$$

which becomes, after rearranging terms and denoting $\hat{\varepsilon} := y - x\hat{\beta}_{LIML}$

$$\frac{1}{n}x'P_Z(y - x\hat{\beta}_{LIML}) = \frac{\hat{\varepsilon}'P_Z\hat{\varepsilon}}{\hat{\varepsilon}'\hat{\varepsilon}}\frac{x'\hat{\varepsilon}}{n}$$

Note that this differs from 2SLS only by the term on the right-hand side. Now we can approximate the expectations of both sides of the equation. By the same token as in the derivation of 2SLS bias, at the true parameter value β

$$\mathbb{E}\left[\frac{1}{n}x'P_Z(y - x\beta)\right] \approx \frac{K}{n}\sigma_{\nu\varepsilon}$$

Since we can see from its first-order condition that 2SLS sets the left-hand side to zero, it will be biased unless $\sigma_{\nu\varepsilon} = 0$. Similarly, can verify that

$$\mathbb{E}\left[\frac{1}{n}\varepsilon'P_Z\varepsilon\right] = \frac{K}{n}\sigma_\varepsilon^2, \quad \mathbb{E}\left[\frac{1}{n}\varepsilon'\varepsilon\right] = \sigma_\varepsilon^2, \quad \mathbb{E}\left[\frac{1}{n}x'\hat{\varepsilon}\right] = \mathbb{E}\left[\frac{1}{n}\pi'z'\varepsilon + \frac{1}{n}\nu'\varepsilon\right] = \sigma_{\nu\varepsilon}$$

Plugging this back into the first-order condition for LIML, we can see that the expectation of the left-hand side is (approximately) equal to that of the right-hand side so that LIML in fact solves the right first-order condition.

5 Jackknife 2SLS

Jackknife is a resampling technique, closely related to the bootstrap. We can use the Jackknife to estimate, and correct for, the bias the 2SLS estimator without having to derive an analytical expression for that bias. Under regularity conditions, the second-order bias of an estimator $\hat{\beta}$ can be expanded as

$$\hat{\beta} = \beta + \frac{B}{n} + O\left(\frac{1}{n^2}\right)$$

where the (normalized) approximate bias B does not depend on sample size. Therefore, if we re-estimate the coefficient using all n observations except the ith, we'd get

$$\hat{\beta}_{-i} = \beta + \frac{B}{n-1} + O\left(\frac{1}{n^2}\right)$$

Therefore, we have potentially T different (although not independent) estimators with bias $\frac{B}{n-1}$, and one with bias $\frac{B}{n}$. If we take the average $\bar{\beta}_{-1} := \frac{1}{n}\sum_{i=1}^n \hat{\beta}_{-i}$ over the "leave-one-out" estimators, we can form

$$\hat{\beta}_{JK} = n\hat{\beta} - (n-1)\bar{\beta}_{-1} = n\beta + B - (n-1)\beta - B + O\left(\frac{1}{n^2}\right) = \beta + O\left(\frac{1}{n^2}\right)$$

Applying this idea to 2SLS, we get the Jackknife 2SLS described by Hausman and Kuersteiner,

$$\hat{\beta}_{J2SLS} = n\frac{\hat{\pi}'\sum_i z_i y_i}{\hat{\pi}\sum_i z_i x_i} - \frac{n-1}{n}\sum_i \frac{\hat{\pi}_{-i}\sum_{j\neq i} z_j y_j}{\hat{\pi}_{-i}\sum_{j\neq i} z_j x_j}$$

There are computationally simpler Jackknife estimators, most prominently Angrist and Imbens and Krueger (1999)'s JIVE, which basically takes the diagonal out of the projection matrix P_Z in 2SLS, however the finite-sample performance of the different Jackknife variants may vary significantly under weak instruments.

6 The Moment Problem

Earlier in the semester we showed that, subject to validity of the instrumental variables, all IV estimators are consistent and asymptotically normal with an easy-to-derive variance matrix. This suggests that, at least in relative large samples, we should be relatively confident that our estimator has all desirable properties of a normal random variable. However, it turns out that this is not so if instruments are weak, or we are using too many, and in particular LIML - which is median unbiased - and Nagar - which is mean-unbiased - turn out not to have any finite sample moments in the sense that the integral

$$\mathbb{E}[(\hat{\beta})^r] = \int_{\mathbb{R}} b^r F_{\hat{\beta}}(db)$$

is not defined (in most cases not finite) for any value of $r = 1, 2, ...$
The underlying reason for this is that these estimators are ratios of random variables for which, in particular if instruments are very weak, the denominator is very close to zero with high probability, which will move probability mass of the distribution of the estimator to the far tails. To illustrate this, let's first look at the denominator of 2SLS which turns out *not* to have the moment problem (at least up to the degree of over-identification):

$$\mathbb{E}\left[x'P_Z x\right] = \mathbb{E}\left[\pi'(Z'Z)\pi + \mathrm{tr}(\nu' P_Z \nu)\right] = n\pi' M_n \pi + K\sigma_\nu^2$$

so that even if there is no first stage, i.e. $\pi = 0$, this expectation is strictly positive. However looking e.g. at Nagar's estimator, the denominator becomes

$$\begin{aligned}\mathbb{E}[x'P_Z x - \frac{K}{n}x'x] &= \mathbb{E}\left[\frac{n-K}{n}\pi'(Z'Z)\pi + \mathrm{tr}(\nu' P_Z \nu) - \frac{K}{n}\nu' n u\right] \\ &= (n-K)\pi' M_n \pi + (K-K)\sigma_\nu^2 = (n-K)\pi' M_n \pi\end{aligned}$$

Therefore, if the instruments are very weak, i.e. $\pi \approx 0$, the distribution of the denominator of the bias-corrected 2SLS estimator is centered at zero. This motivates why for weak instruments, we would not necessarily expect the no-moments problem to go away even for reasonably large samples. Comparing this to 2SLS, we can in fact see that it is exactly the adjustment we did in order to get rid of the bias which killed the term that gave us something like a first stage even in the absence of (strong) instruments in 2SLS.
An analogous argument for the LIML is a little more involved, but looking at the first-order conditions of LIML derived above, it's easy to see that the underlying reason for the moment problem is basically the same as for B2SLS.
This suggests that you shouldn't use an estimator which has the moment problem (Nagar, LIML) in a weak instruments setting.

6.1 k-Class Estimators, Fuller

k-class estimators nest 2SLS, LIML, and other linear estimators for the IV model. In order to motivate the k-class framework, let's first look at 2SLS: Denoting the first-stage residuals with $\hat{V} := (I - P_Z)Y_1$, we can rewrite 2SLS as follows:

$$\begin{bmatrix} \hat{\beta}_{2SLS} \\ \hat{\gamma}_{2SLS} \end{bmatrix} = \begin{bmatrix} \hat{Y}_1'\hat{Y}_1 & \hat{Y}_1'Z_1 \\ Z_1'\hat{Y}_1 & Z_1'Z_1 \end{bmatrix} \begin{bmatrix} \hat{Y}_1'\hat{y}_1 \\ Z_1'\hat{y}_1 \end{bmatrix} = \begin{bmatrix} Y_1'Y_1 - \hat{V}'\hat{V} & Y'Z_1 \\ Z_1'Y & Z_1'Z_1 \end{bmatrix} \begin{bmatrix} (Y_1 - \hat{V})'y \\ Z_1'y \end{bmatrix}$$

k-class estimators are a generalization of 2SLS in that they vary the "amount of error" from the first stage which is subtracted off, and they are defined as

$$\begin{bmatrix} \hat{\beta}_k \\ \hat{\gamma}_k \end{bmatrix} = \begin{bmatrix} Y_1'Y_1 - k\hat{V}'\hat{V} & Y'Z_1 \\ Z_1'Y & Z_1'Z_1 \end{bmatrix} \begin{bmatrix} (Y_1 - k\hat{V})'y \\ Z_1'y \end{bmatrix}$$

Without further covariates, this reduces to

$$\begin{aligned} \hat{\beta}_k &= (x'x - kx'(I - P_Z)x)^{-1} (x'y - kx'(I - P_Z)y) \\ &= \left(x'P_Z x - \frac{k-1}{k}x'x\right)^{-1} \left(x'P_Z y - \frac{k-1}{k}x'y\right) \end{aligned}$$

where the last formulation is only valid for $k \neq 0$.
This nests a couple of estimators we've already seen:

- OLS corresponds to $k = 0$ (just check the formula)
- 2SLS sets $k = 1$, as you can see above
- LIML sets $k = \hat{\lambda}$, where $\hat{\lambda}$ is the variance ratio minimized by the estimator which we defined above
- B2SLS/Nagar's estimator picks $k = \frac{n}{n-K} = 1 + \frac{n}{n-K}$ and is therefore a modification of 2SLS
- Fuller's estimator modifies LIML by replacing $\hat{\lambda}$ with $\tilde{\lambda} = \hat{\lambda} - \frac{\kappa}{n-K}$ for some $\kappa > 0$, where Fuller with $\kappa = 1$ is second-order unbiased, and $\kappa = 4$ minimizes the MSE.

You can see that consistency requires that $k_n \overset{n\to\infty}{\longrightarrow} 1$, and this can be shown to be true for LIML, while it obviously holds for 2SLS and Nagar.

7 Mean Square Error of the Estimators

The following is just a short summary of the Mean-squared error of the estimators we looked at, just copied from the presentation slides without further derivations. You should remember that the mean-squared error of an estimator equals its variance plus its squared bias.

- 2SLS has MSE
$$MSE = \frac{1 - R^2}{nR^2} + K^2 \frac{\varrho^2}{n^2} \left(\frac{1 - R^2}{R^2}\right)^2 + O\left(\frac{1}{n^2}\right)$$
- JN2SLS, and JIVE have MSE
$$MSE = \frac{1 - R^2}{nR^2} + K \frac{1 + \varrho^2}{n^2} \left(\frac{1 - R^2}{R^2}\right)^2 + O\left(\frac{1}{n^2}\right)$$
- LIML and Fuller have MSE
$$MSE = \frac{1 - R^2}{nR^2} + K \frac{1 - \varrho^2}{n^2} \left(\frac{1 - R^2}{R^2}\right)^2 + O\left(\frac{1}{n^2}\right)$$

8 Robust Inference with Potentially Weak Instruments

Usually, we put confidence intervals around β by inverting a t-test, i.e. a 95% confidence interval can be defined as all parameter values $\tilde{\beta}$ such that a t-test at significance level 5% doesn't reject the null hypothesis $H_0(\tilde{\beta}) : \beta_0 = \tilde{\beta}$. For weak instruments or many instruments, inference based on the t-test is generally not valid. Therefore, we should look out for test statistics which are valid under both regular and weak-instruments asymptotics. As a cautionary note, Dufour (1997) showed that there exists no confidence interval for β that is both bounded and achieves correct coverage in finite samples.

8.1 The Anderson-Rubin (AR) Statistic

The Anderson-Rubin (1949) statistic is given by

$$AR(\beta) = \frac{n-K}{K} \frac{(\mathbf{y} - \mathbf{X}\beta)' \mathbf{P}_{\mathbf{Z}} (\mathbf{y} - \mathbf{X}\beta)}{(\mathbf{y} - \mathbf{X}\beta)' (\mathbf{I} - \mathbf{P}_{\mathbf{Z}}) (\mathbf{y} - \mathbf{X}\beta)} \xrightarrow{d} \chi^2_{(K)}$$

which is in fact very similar to the LIML objective function. Hypothesis tests on β based on this statistic are uniformly valid so that confidence intervals based on this statistic are going to have correct size (i.e. it covers the true parameter with the pre-specified coverage probability).

A downside of this statistic is that the corresponding test in fact tests the joint hypothesis that β_0 is equal to a given value *and* that the model is correctly specified. Therefore, if we have mis-specification, this test would reject the true parameter for large samples. This means in particular that the confidence intervals based on the AR-test *shrink* if some of the instruments are in fact not valid. In many cases, the resulting confidence interval will in fact be empty with a relatively high probability for exactly this reason. In sum, if we have doubts about the validity of some instruments, confidence intervals based on the AR-statistic may have particularly bad coverage properties.

Since the test has an asymptotic chi-squared distribution with K degrees of freedom, so that if K is large, the test is going to have very little power. Also, typically the resulting confidence intervals need not be connected.

8.2 Kleibergen's K-Statistic

Kleibergen (2002)'s K-statistic fixes these problems. It is defined as

$$K(\beta) = (n-K) \frac{(\mathbf{y} - \mathbf{X}\beta)' \mathbf{P}_{\mathbf{Z}\hat{\pi}(\beta)} (\mathbf{y} - \mathbf{X}\beta)}{(\mathbf{y} - \mathbf{X}\beta)' (\mathbf{I} - \mathbf{P}_{\mathbf{Z}}) (\mathbf{y} - \mathbf{X}\beta)} \xrightarrow{d} \chi^2_{(m)}$$

where $\hat{\pi}(\beta)$ is the LIML-estimate for the coefficients in a linear first stage under the assumption that the actual structural parameter equals β, and $m = \dim(\beta)$.

Using a projection onto the LIML-fit of the first stage, $Z\hat{\pi})\beta)$ instead of a projection on the instruments may drastically reduce the degrees of freedom of the test, and therefore make the corresponding confidence intervals less conservative (because the underlying test is now more powerful).

One problem with confidence intervals based on Kleibergen's K is that $K(\beta)$ has a second minimum under weak identification which may be far out in the tails, so that the resulting confidence interval consists of several non-connected components which may lie very far apart.

8.3 Moreira's Conditional Likelihood Ratio Statistic

The conditional likelihood ratio statistic is given by

$$LR(\beta) = 2 \left[\max_{\tilde{\beta}, \tilde{\pi}} L(\tilde{\beta}, \tilde{\pi}) - \max_{\tilde{\pi}} L(\beta, \tilde{\pi}) \right]$$

where $L(\beta, \pi)$ is the (concentrated) Gaussian log-likelihood for the parameter (β, π) as in the discussion of LIML.

A crucial point is that, like Kleibergen's K-statistic, the conditional LR statistic is based on the maximum-likelihood estimate of π, which is higher-order efficient and independent of $\hat{\beta}_{LIML}$, so there is no interaction at that point. Moreira (2003) shows that a test based on this statistic comes close to attaining the power envelope (this is analogous to estimators attaining the efficiency bound), and therefore has favorable statistical properties beyond giving valid confidence intervals.

References

[1] Anderson, T., and H. Rubin **(1949):** Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations, *The Annals of Mathematical Statistics* **21** 570-82

[2] Angrist, J., G. Imbens, and A. Krueger **(1999):** Jackknife Instrumental Variables Estimation, *Journal of Applied Econometrics* **14** 57-67

[3] Bekker, P. **(1994):** Alternative Approximations to the Distributions of Instrumental Variable Estimators, *Econometrica* **62(3)** 657-81

[4] Donald, S., and W. Newey **(2001):** Choosing the Number of Instruments, *Econometrica* **69(5)** 1161-91

[5] Dufour, J. **(1997):** Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models, *Econometrica*, **65(6)**, 1365–1387

[6] Fuller, W. **(1977):** Some Properties of a Modification of the Limited Information Estimator, *Econometrica* **45(4)** 939-53

[7] Goldberger, A. **(1964):** Econometric Theory, John Wiley & Sons

[8] Hahn, J., and J. Hausman **(2002a):** A new Specification Test for the Validity of Instrumental Variables, *Econometrica* **70(1)** 163-89

[9] Hahn, J., and J. Hausman **(2002b):** Notes on Bias in Estimators for Simultaneous Equation Models, *Economics Letters* **75** 237-41

[10] Hahn, J., and J. Hausman **(2003):** IV Estimation with Valid and Invalid Instruments, manuscript MIT and UCLA Economics

[11] Hirano, K., and J. Porter **(2015):** Location Properties of Point Estimators in Linear Instrumental Variables and Related Models, *Econometric Reviews* **34(6–10)** 720–733

[12] Kleibergen, F. **(2002):** Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression, *Econometrica* **70(5)** 1781-1803

[13] Maddala, G. **(1977):** Econometrics, McGraw-Hill

[14] Moreira, M. **(2003):** A Conditional Likelihood Ratio Test for Structural Models, *Econometrica* **71(4)** 1027-48

[15] Nagar, A. **(1959):** The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations, *Econometrica* **27(4)** 575-95

[16] Newey, W. **(2004):** Many Instrument Asymptotics, Notes, MIT Economics

[17] Phillips, P. **(1983):** Exact Small Sample Theory in the Simultaneous Equations Model, *Handbook of Econometrics*, vol.I ch.8, 449-516

[18] Rothenberg, T. **(1983):** Asymptotic Properties of Some Estimators in Structural Models, *Studies in Econometrics, Time Series, and Multivariate Statistics*, Academic Press

[19] Staiger, D., and J. Stock **(1997):** Instrumental Variables Regression with Weak Instruments, *Econometrica* **65(3)** 557-86