Examiner Designs and First-Stage $F$ Statistics: A Caution

High-dimensional instrumental variable (IV) regressions can be cumbersome to implement. To ease their computational burden, applied researchers often reduce the dimensionality of a many-IV first stage in a manual first step. For example, in the quasi-experimental “examiner” design, a researcher observes draws of an outcome $Y$, an endogenous variable $X$, and a vector of $K$ mutually-exclusive and exhaustive binary variables, $Z$, which indicate as-good-as-random assignment to examiner groups.\(^1\) Rather than directly estimating a two-stage least squares (2SLS) regression of $Y$ on $X$ with the $K$ instruments, a researcher may compute the equivalent IV coefficient by first constructing the examiner-level average of the endogenous variable, $\bar{X}$, and then instrumenting $X$ by $\bar{X}$. Often researchers use leave-one-out averages to form $\bar{X}$, in which case the two-step constructed IV coefficient matches that of the Angrist, Imbens, and Krueger (1999) jackknife IV estimator.\(^2\)

Computing group-level averages is typically much simpler than inverting a high-dimensional instrument design matrix in 2SLS. Since the two approaches produce numerically identical coefficients, it seems natural to prefer the use of “constructed instruments” $\bar{X}$ in these cases. Nevertheless, a researcher should never forget in doing so that the dimensionality of her true identifying variation is $K$, not one. If she does forget, she may, for example, fall into the common trap of using the $F$ statistic from a regression of $X$ on $\bar{X}$ to gauge the first-stage strength of her identification.

$$
\hat{F}_1 = \frac{(N - 2) \hat{R}^2}{1 - \hat{R}^2},
$$

where $\hat{R}^2$ denotes the sample $R$-squared from this regression and $N$ is the sample size. The “true” first-stage $F$ statistic from the regression of $X$ on $Z$ is, by contrast,

$$
\hat{F}_K = \frac{(N - K - 1) \hat{R}^2}{K(1 - \hat{R}^2)} = \frac{N - K - 1}{K(N - 2)} \hat{F}_1,
$$

which is approximately $K$ times smaller than $\hat{F}_1$. In practice, therefore, researchers run the risk of greatly overstating their first-stage $F$-statistics when using constructed instruments in examiner designs – estimators suffering from severe many-weak IV bias may go undetected. Researchers should always be sure to apply the above degree-of-freedom correction when using $\hat{F}_1$ to diagnose the strength of examiner instruments. Alternatively, they may prefer more formal dimension-reduction techniques for IV, such as the LASSO approach in Belloni et al. (2012), which may prove useful in many-weak designs.\(^3\)

\(^1\) See Kling (2006), Maestas, Mullen, and Strand (2013), and Doyle et al. (2015) for three recent examples of this approach.

\(^2\) The result demonstrated here is easily extended to cases with multiple endogenous regressors and auxiliary controls.

\(^3\) The Angrist, Imbens, and Krueger (1999) estimator also has favorable properties in many-weak environments, but also many drawbacks; see, e.g., Davidson and MacKinnon (2006).
References


