Since $\hat{\theta}_n$ is chosen such that k linear combinations of the $r \times 1$ vector $g_n(\hat{\theta}_n; Y^n)$ are set to zero (see FOC for $\hat{\theta}_n$ above), we get

$$T_J = n g_n(\hat{\theta}_n; Y^n) \hat{S}^{-1} g_n(\hat{\theta}_n; Y^n) \xrightarrow{d} X^2_{r-k} \, .$$

If $r = k$, there is nothing to test, but by construction $\hat{\theta}_n$ satisfies all $r = k$ moment conditions. The J-specification test is usually run as a justification to use GMM on a particular set of moment conditions derived from some theoretical model.

6.4 Instrumental Variable Estimation

Section 3.4 introduced the endogeneity problem in the context of the linear regression model. It occurs if we want to estimate β in $y_i = x_i'\beta + u_i$, but $\mathbb{E}[x_i u_i] \neq 0$ is suspected. As discussed in Section 3.4, this leads to an inconsistent OLS estimator. Instrumental variable (IV) estimation is a potential remedy.

Suppose we have a variable $z_i \in \mathbb{R}^r$. The r variables in z_i are valid IVs if z_i satisfies two conditions:

1. Relevance: $\mathbb{E}[z_i x_i'] \neq 0$.
2. Exogeneity: $\mathbb{E}[z_i u_i] = 0$, i.e. z_i is uncorrelated with the error term u_i.

The first condition is often stated as requireing that z_i is correlated with the regressor x_i.

The idea is to use z_i to extract the part of the information in x_i that is uncorrelated with u_i. This is best illustrated in the two-stage least squares (2SLS) estimation procedure. For illustration purposes, suppose x_i is a scalar and derive $\hat{\beta}_{2SLS}$ in two steps:

1. Estimate $x_i = z_i'\gamma + e_i$ to get $\hat{\gamma} = (Z'Z)^{-1}Z'X$ and $\hat{X} = Z\hat{\gamma} = P_Z X$.
2. Estimate $y_i = \hat{x}_i'\beta + u_i^*$ to get $\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'P_Z X)^{-1}X'P_Z Y$.

Provided that our model is correctly specified and the above two assumptions are satisfied, $\hat{\beta}_{2SLS}$ is consistent: by WLLN,

$$\begin{aligned}
\hat{\beta}_{2SLS} &= \left[X'P_Z X\right]^{-1} X'P_Z Y \\
&= \beta + \left[X'P_Z X\right]^{-1} X'P_Z U \\
&= \beta + \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'U \\
&\xrightarrow{p} \beta + Q_{xz}^{-1}\mathbb{E}[x_i z_i']\mathbb{E}[z_i z_i']\mathbb{E}[z_i u_i] = \beta \, ,^{9}
\end{aligned}$$

where $Q_{xz} = \mathbb{E}[x_i z_i'] \mathbb{E}[z_i z_i'] \mathbb{E}[x_i z_i']'$. To compute $\hat{\beta}_{2SLS}$, we need $Z'Z$ to be of full rank. Based on the analogous discussion of $X'X$ in Section 3.1, $rank(Z'Z) = r$ requires us to have $n > r$, i.e. more observations than IVs, and it prevents perfect multicollinearity of IVs.[10]

Ideally, z_i should be as highly correlated as possible with x_i in order to preserve as much variation of x_i in $\hat{x}_i$ as possible. To see this, note that the conditional variance of $\hat{\beta}_{2SLS}$ is

$$\mathbb{V}[\hat{\beta}_{2SLS}|X, Z] = (X'P_Z X)^{-1} X'P_Z \mathbb{E}[UU'|X, Z] P_Z X (X'P_Z X)^{-1} .$$

Under homoskedasticity, $\mathbb{E}[UU'|X, Z] = \sigma^2 I$ and we obtain $\mathbb{V}[\hat{\beta}_{2SLS}|X, Z] = \sigma^2 (X'P_Z X)^{-1}$. This variance is larger than that of $\hat{\beta}_{OLS}$, $\mathbb{V}[\hat{\beta}_{OLS}|X] = \sigma^2 (X'X)^{-1}$. To see this, note that

$$\Delta = X'X - X'P_Z X = X'M_Z X = (M_Z X)'(M_Z X)$$

is p.d. It is the sum of squared residuals from the first-stage regression of X on Z. Therefore, the more variation in X is explained by Z, i.e. the higher the R^2 in the first-stage regression, the smaller is the efficiency loss of 2SLS compared to OLS.

By the usual arguments, the asymptotic analysis reveals that $\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, V_{2SLS})$ with the huge but simple-to-derive expression for V_{2SLS}:

$$V_{2SLS} = Q_{xz}^{-1} \mathbb{E}[x_i z_i'] \mathbb{E}[z_i z_i'] \mathbb{E}[z_i z_i' u_i^2] \mathbb{E}[z_i z_i']' \mathbb{E}[x_i z_i']' Q_{xz}^{-1} .$$

As usual, we can estimate it by replacing u_i with $\hat{u}_i$ and expectation operators with population means. Thereby, it is important to note that $u_i \neq u_i^*$, i.e. o obtain $\hat{u}_i$, we use regressors x_i and not $\hat{x}_i$: $\hat{u}_i = y_i - x_i' \hat{\beta}_{2SLS}$. Under homoskedasticity, V_{2SLS} simplifies to $V_{2SLS} = \sigma^2 Q_{xz}^{-1}$, which we estimate using $\hat{\sigma}^2 = \frac{1}{n} \sum_i u_i^2$.

Note that the 2SLS estimator does not actually have to be carried out in two steps. It simply involves running a regression of Y on $P_Z X$ instead of X. Relatedly, 2SLS can trivially be applied for more than one regressor in x_i. The only thing that changes is that X is not necessarily an $n \times 1$ vector but more generally an $n \times k$ matrix, where $k \geq 1$. Typically, not all variables in x_i are suspected to be endogenous. While one could in principle replace

[9] To see this, note that we can write out

$$[X'Z(Z'Z)^{-1}Z'X]^{-1} X'Z(Z'Z)^{-1}Z'U = \left[\left(\sum_{i=1}^n x_i z_i' \right) \left(\sum_{i=1}^n z_i z_i' \right) \left(\sum_{i=1}^n x_i z_i' \right)' \right]^{-1} \left(\sum_{i=1}^n x_i z_i' \right) \left(\sum_{i=1}^n z_i z_i' \right) \left(\sum_{i=1}^n z_i u_i \right) ,$$

and we can divide each of the sums by n.

[10] We also need and $X'P_Z X = (P_Z X)'(P_Z X)$ to be of full rank. This condition, $rank(X'P_Z X) = k$, is somewhat more subtle. It is related to the relevance condition above and, intuitively, requires Z to explain some variation in X. With $k = 1$ regressor in x_i, it is equivalent to saying that $\hat{X} = P_Z X$ is not zero.

only the endogenous variables $x_{i,m}$ with their predicted values $\hat{x}_{i,m}$ obtained in regressions on IVs, with possibly different IVs for different variables, a more straightforward approach is to let z_i include not only the actual IVs (all of them), but also all exogenous variables from x_i, and construct the above 2SLS estimator.[11] This guarantees that $r \geq k$, i.e. we have at least as many IVs as endogenous variables.

The use of IVs has been popularized by causal inference methods in the context of the estimation of so-called local average treatment effects (LATEs). It is discussed in Section 12.3.1, which contains examples of IV-analyses.

The rest of this section discusses the challenges posed by weak IVs and GMM-based IV estimation. Likelihood-based IV estimation is touched upon in the Appendix.

Weak Identification in IV Models If the the correlation between x_i and z_i is rather low, we speak of weak IVs. Under weak IVs, the finite sample distribution of $\hat{\beta}$ may not resemble the asymptotic one at all. This needs to be taken into account when conducting inference, i.e. testing hypotheses and constructing confidence sets.

The asymptotic analysis can be adjusted to take into account the weak correlation of x_i and z_i. However, this is more interesting from a methodological than an empirical point of view. The approach is sketched in the Appendix.

In absence of an asymptotic distribution that is useful for approximating the finite sample distribution, we can conduct inference using its numerical approximation via bootstrapping (Section 6.1). Alternatively, we can construct a confidence set for β under weak IVs using the inference procedure of Anderson and Rubin (1949). It is based on the insight that, for $\beta = \beta_0$, the auxiliary regression $y_i - x_i'\beta = \delta z_i + v_i$ should yield $\delta = 0$, because $y_i - x_i'\beta_0 = u_i$ and u_i and z_i are uncorrelated. Suppose for simplicity that z_i is a scalar. For a given β_0, we get

$$\sqrt{n}\hat{\delta}(\beta_0) = \sqrt{n}(Z'Z)^{-1}Z'(Y - X\beta_0) = (Z'Z)^{-1}\sqrt{n}Z'U \xrightarrow{d} N\left(0, \frac{\sigma_u^2}{\mathbb{E}(z_i^2)}\right),$$

which allows us to test $\mathscr{H}_0 : \delta = 0$. As δ is a scalar, we can use the t-test $t_\delta(\beta_0) = \hat{\delta}(\beta_0)/\sqrt{\hat{\sigma}_v^2/Z'Z} \xrightarrow{d} N(0,1)$. A confidence set for β is obtained by taking all β_0 for which $\mathscr{H}_0 : \delta = 0$ cannot be rejected. Note that this can give very large, unbounded and even disconnected or empty confidence sets.

[11] It is straightforward to see that for the exogenous variables in x_i, predicted values will equal the actual values, as regressing a variable on itself (and other covariates) generates $R^2 = 1$. Regarding changing IVs for different endogenous variables: if indeed a set of IVs is irrelevant for one of the endgenous variables, their coefficient will be close to zero, not impacting the predicted value much.