Heteroscedasticity-Based Identification (Lewbel, 2012 and Extensions)

June 12, 2025

0 Preliminaries and Notation

Throughout the paper $Z = g(X)$ denotes a *centered* transformation of the exogenous variables, i.e. $\mathbb{E}[Z] = 0$ by construction.[1] All instruments in what follows are therefore understood to be mean-zero.

Shorthands. We pre-define the operators $\mathbb{E}$, Cov, Corr, and Var and the probability limit plim to streamline formulas.

1 Structural Forms and Reduced-Form Residuals

Let the endogenous vector be $(Y_1, Y_2)'$, X a vector of exogenous variables (including a constant), and $Z = g(X)$ as above.

$$
\begin{aligned}
\textbf{Triangular:} \quad & Y_1 = X'\beta_1 + \gamma_1 Y_2 + \varepsilon_1, & Y_2 = X'\beta_2 + \varepsilon_2, \\
\textbf{Simultaneous:} \quad & Y_1 = X'\beta_1 + \gamma_1 Y_2 + \varepsilon_1, & Y_2 = X'\beta_2 + \gamma_2 Y_1 + \varepsilon_2.
\end{aligned}
$$

Projecting each Y_j on X yields *reduced-form residuals* $W_j \triangleq Y_j - X'(\mathbb{E}[XX'])^{-1}\mathbb{E}[XY_j]$. A short calculation gives

$$
W_1 = \frac{\varepsilon_1 + \gamma_1 \varepsilon_2}{1 - \gamma_1 \gamma_2}, \qquad W_2 = \frac{\varepsilon_2 + \gamma_2 \varepsilon_1}{1 - \gamma_1 \gamma_2},
$$

with the triangular case obtained by setting $\gamma_2 = 0$.

Remark 1. *The algebra follows directly from solving the two-equation system for the reduced form and collecting the error terms; see Lewbel (2012, App. B) for details.*

2 Core Assumptions for Point Identification

(A1) **Strict exogeneity.** $\mathbb{E}[\varepsilon_j \mid X] = 0$ for $j = 1, 2$. *Time-series note:* in Section **??** we weaken this to a martingale-difference assumption and add HAC inference.

(A2) **Covariance restriction.** $\mathrm{Cov}(Z, \varepsilon_1\varepsilon_2) = 0$. This holds automatically if the errors have a common factor $u \perp X$; then $\varepsilon_j = a_j u + \eta_j$ with u independent of Z implies the restriction.

(A3) **Instrument relevance via heteroscedasticity.**

- *Triangular:* $\mathrm{Cov}(Z, \varepsilon_2^2) \neq 0$.

[1] In practice we demean Z in sample; writing Z keeps formulas uncluttered. The sample mean $\bar{Z}$ appears only where an explicit finite-sample expression is needed (e.g. generated instruments).

- *Simultaneous:* the $r \times 2$ matrix $\Phi_W = [\,\mathrm{Cov}(Z, W_1^2) \quad \mathrm{Cov}(Z, W_2^2)\,]$ has rank 2; equivalently each column is linearly independent in Z. See Lewbel (2012, Prop. 4).

(A4) **Normalization (simultaneous case).** As in Lewbel (2012), the parameter space precludes the observationally equivalent pair $(1/\gamma_2, 1/\gamma_1)$ unless $\gamma_1\gamma_2 = 1$.

3 Triangular System: Closed-Form Identification and 2SLS

3.1 Closed-form

$$\gamma_1 = \frac{\mathrm{Cov}(Z, W_1 W_2)}{\mathrm{Cov}(Z, W_2^2)}. \tag{1}$$

Remark 2 (Why (**??**) identifies γ_1)**.** *Under (A2) the numerator simplifies to* $\mathrm{Cov}(Z, \varepsilon_1\varepsilon_2) + \gamma_1 \mathrm{Cov}(Z, \varepsilon_2^2) = \gamma_1 \mathrm{Cov}(Z, \varepsilon_2^2)$*. Because the denominator is non-zero by (A3),* γ_1 *cancels on the right-hand side, yielding point identification.*

3.2 Feasible two-step 2SLS

1. **Generate residuals.** Regress Y_2 on X via OLS and store $\hat{\varepsilon}_2 = Y_2 - X'\hat{\beta}_2^{\mathrm{OLS}}$.
2. **Construct the heteroscedasticity-based instrument.** $IV = (Z - \bar{Z})\hat{\varepsilon}_2$.
3. **First stage.** Regress Y_2 on $[X, IV]$, obtain fitted values $\hat{Y}_2$.
4. **Second stage.** Regress Y_1 on $[X, \hat{Y}_2]$ to estimate (β_1, γ_1).

Practical guidance. Weak-instrument concerns apply because IV is generated: report the first-stage F-statistic on IV and, where necessary, use weak-IV-robust inference (Anderson–Rubin or Kleibergen–Paap r_k tests). Standard errors should be obtained from a two-step GMM covariance matrix or via bootstrap to account for first-stage estimation.

4 GMM Moment Conditions

Let θ collect structural parameters and let $\mu = \mathbb{E}[Z] = 0$ by centering. The identifying moments are

(i) Triangular system.

$$Q_{\mathrm{TRI}}(\theta) = \begin{pmatrix} X(Y_1 - X'\beta_1 - Y_2\gamma_1) \\ X(Y_2 - X'\beta_2) \\ (Z)(Y_1 - X'\beta_1 - Y_2\gamma_1)(Y_2 - X'\beta_2) \end{pmatrix}. \tag{2}$$

(ii) Simultaneous system.

$$Q_{\mathrm{SIM}}(\theta) = \begin{pmatrix} X(Y_1 - X'\beta_1 - Y_2\gamma_1) \\ X(Y_2 - X'\beta_2 - Y_1\gamma_2) \\ (Z)(Y_1 - X'\beta_1 - Y_2\gamma_1)(Y_2 - X'\beta_2 - Y_1\gamma_2) \end{pmatrix}. \tag{3}$$

Equations (**??**)–(**??**) satisfy $\mathbb{E}[Q(\theta)] = 0$ under (A1)–(A3).

5 Set Identification Under a Relaxed Covariance Restriction

When assumption (A2) is weakened to $|\operatorname{Corr}(Z, \varepsilon_1\varepsilon_2)| \leq \tau|\operatorname{Corr}(Z, \varepsilon_2^2)|$, $\tau \in [0,1)$, γ_1 in the triangular model is set-identified.

Theorem 1 (Bounds for γ_1 with $\tau > 0$)**.** *γ_1 is contained in the closed interval whose endpoints are the (real) roots of*

$$\frac{\operatorname{Cov}(Z, W_1W_2)^2}{\operatorname{Cov}(Z, W_2^2)^2} - \frac{\operatorname{Var}(W_1W_2)}{\operatorname{Var}(W_2^2)}\tau^2 + 2\left(\frac{\operatorname{Cov}(W_1W_2, W_2^2)}{\operatorname{Var}(W_2^2)}\tau^2 - \frac{\operatorname{Cov}(Z, W_1W_2)}{\operatorname{Cov}(Z, W_2^2)}\right)\gamma_1 + (1-\tau^2)\gamma_1^2 = 0.$$

The interval collapses to a point when $\tau = 0$ and widens monotonically as $\tau \to 1$.

6 Two Illustrative Extensions

6.1 Conditional heteroscedasticity (Prono, 2013)

Let $\operatorname{Var}(\varepsilon_{2t} \mid \mathcal{F}_{t-1}) = \sigma_{2t}^2$ follow a GARCH model; set $Z_t = \sigma_{2t}^2$. Replacing population variances by the fitted $\hat{\sigma}_{2t}^2$ yields the generated instrument $(\hat{Z}_t - \bar{Z})\hat{\varepsilon}_{2t}$, and estimation proceeds exactly as in Section3, with HAC-robust standard errors.

6.2 Regime heteroscedasticity (Rigobon, 2003)

If exogenous regimes $s \in \{1, \dots, S\}$ shift $\operatorname{Var}(\varepsilon_j \mid s)$ but leave $\operatorname{Cov}(\varepsilon_1, \varepsilon_2 \mid s)$ constant, the centered regime dummies serve as Z and satisfy (A2)–(A3), delivering identification by the same 2SLS recipe.

7 Diagnostic Tests and Practical Checks

- **Instrument relevance.** Report the first-stage F-statistic on generated instruments; if $F < 10$ adopt weak-IV-robust tests.
- **Instrument validity.** Sargan/Hansen C-tests can be run by dropping the generated instrument and testing over-identifying restrictions.
- **Endogeneity of Y_2.** Difference-in-Sargan (C-statistic) compares the full and restricted instrument sets.
- **Heteroscedasticity of ε_2.** A Breusch–Pagan test of $\hat{\varepsilon}_2^2$ on Z provides evidence for assumption (A3).

8 Time-Series Variant with Log-Linear Conditional Variances

Let (Y_{1t}, Y_{2t}, X_t) be strictly stationary and α-mixing with absolutely summable mixing coefficients. Maintain the structural forms of Section1 and set $Z_t = X_t - \mu = Z_t$. Assume

$$\varepsilon_{jt} \mid \mathcal{F}_{t-1} \sim (0, \sigma_{jt}^2), \quad \log \sigma_{jt}^2 = X_t'\delta_j, \qquad j = 1, 2. \tag{4}$$

Because $(\delta_1, \delta_2) \neq 0$, $\operatorname{Cov}(Z_t, \varepsilon_{jt}^2) \neq 0$, preserving relevance.

8.1 Moment vector

$$Q_t(\theta) = \begin{pmatrix} X_t\varepsilon_{1t} \\ X_t\varepsilon_{2t} \\ Z_t\varepsilon_{1t}\varepsilon_{2t} \\ Z_t\left(\varepsilon_{1t}^2 - e^{X_t'\delta_1}\right) \\ Z_t\left(\varepsilon_{2t}^2 - e^{X_t'\delta_2}\right) \end{pmatrix}, \qquad \mathbb{E}[Q_t(\theta)] = 0,\ \theta = (\beta_1, \beta_2, \gamma_1, \gamma_2, \delta_1, \delta_2)'.$$

8.2 Estimation

1. OLS/IV of the mean equations, obtain residuals $\hat{\varepsilon}_{jt}$.
2. Estimate (**??**) by regressing $\log \hat{\varepsilon}_{jt}^2$ on X_t; set $\hat{\sigma}_{jt}^2 = e^{X_t' \hat{\delta}_j}$.
3. Instruments: $Z_t,\ Z_t \hat{\varepsilon}_{jt},\ Z_t(\hat{\varepsilon}_{jt}^2 - \hat{\sigma}_{jt}^2)$ (6 per t).
4. One-step 2SLS or (if over-identified) GMM with Newey–West weight matrix; a bandwidth such as $\ell_T = \lfloor 4(T/100)^{2/9} \rfloor$ is conventional.

Mixing ensures a serial-correlation-robust LLN/CLT, so $\hat{\theta}$ is consistent and asymptotically normal.

References

[1] Lewbel, A. (2012). Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models. *Journal of Business & Economic Statistics*, 30(1), 67–80.

[2] Prono, T. (2013). The Role of Conditional Heteroscedasticity in Identifying and Estimating Linear Simultaneous Equation Models. *Journal of Applied Econometrics*, 28(2), 338–356.

[3] Rigobon, R. (2003). Identification Through Heteroskedasticity. *Review of Economics and Statistics*, 85(4), 777–792.