ECON6300/7320/8300

Advanced Microeconometrics

Maximum Likelihood: Important Results

Identification and GMM

Christiern Rose

University of Queensland

Lecture 4

Properties of MLE (1)

Lemma (Expectation of the score)

$$E_{f(y|\mathbf{x},\theta_0)}[\mathbf{s}(\theta_0)] = \mathbf{0}$$

Properties of MLE (2)

Lemma (Information matrix equality)

$$
\begin{aligned}
V_{f(y|\mathbf{x},\theta_0)}[s(\theta_0)] &= E_{f(y|\mathbf{x},\theta_0)}\left[\left.\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta}\frac{\partial \mathcal{L}_N(\theta)}{\partial \theta'}\right|_{\theta_0}\right] \\
&= -E_{f(y|\mathbf{x},\theta_0)}\left[\left.\frac{\partial^2 \mathcal{L}_N(\theta)}{\partial \theta \partial \theta'}\right|_{\theta_0}\right] \\
&\equiv \mathcal{I}(\theta_0)
\end{aligned}
$$

The RHS is often called Fisher information measure. It is a measure of the curvature of the likelihood gradient.

Properties of MLE (3)

- The information matrix is the variance of the score, since the score has mean zero. Large values of $\mathcal{I}(\theta)$ mean that small changes in θ lead to large changes in the log-likelihood, which accordingly contains more information about θ.

- The above lemmas hold if the distribution is correctly specified because then the expectation is with respect to $f(y|\mathbf{x}, \theta_0)$.

- Information matrix is related to the variance of $\widehat{\theta}_{MLE}$.

Distribution of MLE

Let $Q_N(\theta) = N^{-1}\mathcal{L}_N(\theta)$. Under regularity conditions $\partial^2 Q_N(\theta)/\partial\theta\partial\theta'\big|_{\theta^+}$ *converges in probability to the finite nonsingular matrix*

$$\mathbf{A}_0 = \text{ plim } \partial^2 Q_N(\theta)/\partial\theta\partial\theta'\Big|_{\theta_0} \tag{1}$$

for any sequence θ^+ *such that* $\theta^+ \xrightarrow{p} \theta_0$;

$\sqrt{N}\,\partial Q_N(\theta)/\partial\theta|_{\theta_0} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{B}_0]$, *where*

$$\mathbf{B}_0 = \text{ plim } \left[N\,\partial Q_N(\theta)/\partial\theta \times \partial Q_N(\theta)/\partial\theta'\Big|_{\theta_0}\right]. \tag{2}$$

Then the limit distribution of $\widehat{\theta}$ *is*

$$\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}] \tag{3}$$

Asymptotic variance derivation

- Vector version of Taylor's theorem plays a key role. Consider an **exact first-order Taylor expansion.** For the differentiable function $f(\cdot)$ there always exists a point x^+ between x and x_0 such that

$$f(x) = f(x_0) + f'(x^+)(x - x_0),$$

where $f'(x) = \partial f(x)/\partial x$ is the derivative of $f(x)$. This result is also known as the **mean value theorem.**

- Vector version is:

$$\mathbf{f}(\widehat{\theta}) = \mathbf{f}(\theta_0) + \left.\frac{\partial \mathbf{f}(\theta)}{\partial \theta'}\right|_{\theta^+} (\widehat{\theta} - \theta_0), \tag{4}$$

where $\partial \mathbf{f}(\theta)/\partial \theta$ is a matrix, for some unknown θ^+ between $\widehat{\theta}$ and θ_0.

Asymptotic variance derivation (2)

- For the MLE estimator the function $f(\theta) = \partial\mathcal{L}_N(\theta)/\partial\theta$.
- Then an exact first-order Taylor series expansion around θ_0 yields

$$\left.\frac{\partial\mathcal{L}_N(\theta)}{\partial\theta}\right|_{\widehat{\theta}} = \left.\frac{\partial\mathcal{L}_N(\theta)}{\partial\theta}\right|_{\theta_0} + \left.\frac{\partial^2\mathcal{L}_N(\theta)}{\partial\theta\partial\theta'}\right|_{\theta^+}(\widehat{\theta} - \theta_0), \tag{5}$$

where $\partial^2\mathcal{L}_N(\theta)/\partial\theta\partial\theta'$ is a $q \times q$ matrix with $(j,k)^{th}$ entry $\partial^2\mathcal{L}_N(\theta)/\partial\theta_j\partial\theta_k$, and θ^+ is a point between $\widehat{\theta}$ and θ_0.

Asymptotic variance derivation (3)

- The first-order conditions set the left-hand side of (5) to zero. Setting the right-hand side to **0** and solving for $(\widehat{\theta} - \theta_0)$ yields

$$\sqrt{N}(\widehat{\theta} - \theta_0) = -\left(\left. \frac{\partial^2 \mathcal{L}_N(\theta)}{\partial\theta\partial\theta'} \right|_{\theta^+} \right)^{-1} \sqrt{N} \left. \frac{\partial \mathcal{L}_N(\theta)}{\partial\theta} \right|_{\theta_0}, \qquad (6)$$

- If $\widehat{\theta}$ is consistent for θ_0 then the unknown θ^+ converges in probability to θ_0, because it lies between $\widehat{\theta}$ and θ_0.
- We conclude by taking the expectation ($=$ **0**) and variance of the RHS evaluated at θ_0
- Our variance estimator is obtained by replacing θ_0 with $\hat{\theta}$

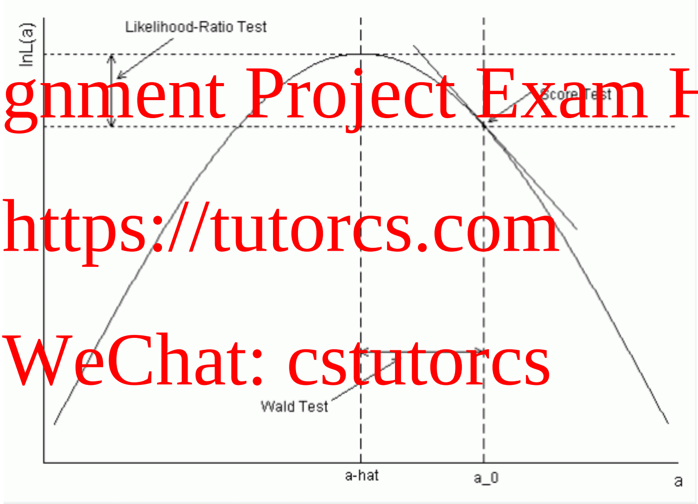
Practical Implications for computing asymptotic variance

- If the IM equality holds, then $\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$, $-\mathbf{A}_0^{-1}$ and $\mathbf{B}_0^{-1}$ are all asymptotically equivalent, as are the corresponding consistent estimates of these quantities.
- The estimate $\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^{-1}$ is called the "robust sandwich estimate", or the Huber estimate, after Huber (1967), or Huber-White estimate who considered the distribution of the MLE under misspecification.
- The sandwich estimate is in theory more robust than $-\widehat{\mathbf{A}}^{-1}$ or $\widehat{\mathbf{B}}^{-1}$, which are default options when the likelihood is assumed to be correctly specified.
- If the IM equality does not hold $\widehat{\theta}_{\text{ML}}$ is inconsistent. Using the robust version does not provide protection against, e.g., misspecified conditional mean

Set-up of Hypothesis Tests

- For hypothesis testing, the key starting point is the asymptotic normality of MLE and knowledge of the asymptotic variance of MLE
- Formulate the null hypothesis generally as a set of constraints on the parameter space Θ so that our hypothesis is $\mathbf{h}(\theta_0) = \mathbf{0}$.
- Tests may be constructed using both restricted and unrestricted models (e.g. LR test), or using unrestricted model only (Wald test), or restricted model only (Score test or Lagrange multiplier (LM) test).
- LM test will not be discussed.

LR, Score and Wald tests



- Define the estimators $\widehat{\theta}_u$: unrestricted MLE, and $\tilde{\theta}_r$: restricted MLE.
- $\widehat{\theta}_u$ maximizes $\ln L(\theta)$; $\tilde{\theta}_r$ maximizes the Lagrangian $\ln L(\theta) - \lambda' \mathbf{h}(\theta)$, where λ is a $h \times 1$ vector of Lagrangian multipliers.
- Consider specific case of exclusion restrictions $\mathbf{h}(\theta) = \theta_2 = \mathbf{0}$, where $\theta = (\theta_1', \theta_2')'$. The $\tilde{\theta}_r = (\tilde{\theta}_{1r}', \mathbf{0}')$ where $\tilde{\theta}_{1r}'$ is obtained simply as the maximum with respect to θ_1 of the restricted likelihood $\ln L(\theta_1, \mathbf{0})$ and $\mathbf{0}$ is a $(q - h) \times 1$ vector of zeroes.

Likelihood Ratio Test

- Want to compare a model before and after imposing restrictions. Are the results significantly different?
- If the null hypothesis is supported by the data then the differences should be small. I.e., if H_0 is true, the unconstrained and constrained maxima of the log-likelihood function should be the same or similar. That is $\ln L(\widehat{\theta}_u) \simeq \ln L(\tilde{\theta}_r)$.
- Test requires obtaining the limit distribution of $[\ln L(\widehat{\theta}_u) - \ln L(\tilde{\theta}_r)]$. It can be shown that $-2\left[\ln L(\tilde{\theta}_r) - \ln L(\widehat{\theta}_u)\right]$ is asymptotically chi-square distributed with degrees of freedom equal to the number of restrictions.

$$\mathrm{LR} = -2\left[\ln L(\tilde{\theta}_r) - \ln L(\widehat{\theta}_u)\right]. \tag{7}$$

Wald Test

- The intuition for the Wald test is that if H_0 is true, the unrestricted MLE $\widehat{\theta}_u$ should satisfy the restrictions of H_0, so $\mathbf{h}(\widehat{\theta}_u)$ should be close to zero.

- To implement the test the asymptotic distribution of $\mathbf{h}(\widehat{\theta}_u)$ is required. The general form of the Wald test is a quadratic form $\mathbf{h}'(\widehat{\theta}_u)[Var\ \mathbf{h}(\widehat{\theta}_u)]^{-1}\mathbf{h}(\widehat{\theta}_u)$, which follows an asymptotic χ^2_h distribution under H_0, where h is the dimension of $\mathbf{h}(\widehat{\theta}_u)$

- The standard t-test or $z-$test (or F-test) can be interpreted as a Wald-test.

Identification, SEM & GMM

The rest of this lecture:

1. Covers the fundamental issue of *identification*, which precedes estimation and interpretation.
2. Attempts to define causal model and *causal parameter*, which is often the target of empirical work
3. Explains what *structural ("autonomous")* and *reduced form* relationships conventionally mean
4. Reviews the *simultaneous equations model* (SEM)
5. We move on to Instrumental Variable and GMM estimation

Structural Models in the SEM Framework

In the well-established Cowles Commission approach a "**structure**" consists of

1. A set of variables $\mathbf{W}$ ("data") partitioned for convenience as $[\mathbf{Y}_{endog}\ \mathbf{Z}_{exog}]$;
2. A joint probability distribution of $\mathbf{W}$, $f(\mathbf{w}|\theta)$;
3. An a priori ordering of $\mathbf{W}$ according to hypothetical cause and effect relationships and specification of a priori restrictions on the hypothesized model;
4. A parametric, semiparametric or nonparametric specification of functional forms and the restrictions on the parameters of the model.

Identification

- If we could observe the population (instead of a sample), would we know the parameters of our structural model θ_0?
- **Definition:** Two structures are observationally equivalent if $f(\mathbf{w}|\theta_1) = f(\mathbf{w}|\theta_2) \quad \forall \mathbf{w}$ for $\theta_1 \in \Theta, \theta_2 \in \Theta$
- **Definition:** θ_0 is identified if there is no other observationally equivalent parameter in Θ

Identification example

$$y_i = \mathbf{x}_i'\theta_0 + u_i \quad \mathrm{E}[u_i|\mathbf{x}_i] = 0 \tag{8}$$

- $\mathbf{W} = [\mathbf{y}, \mathbf{X}]$. Restrict $f(\mathbf{w}|\theta)$ to satisfy (8). Parameter space $\Theta = \mathbb{R}^K$
- Since $\mathrm{E}[u_i|\mathbf{x}_i] = 0 \Rightarrow \mathrm{E}[\mathbf{x}_i u_i] = \mathbf{0}$, θ_0 is identified if:

$$\mathrm{E}[\mathbf{x}_i(y_i - \mathbf{x}_i'\theta_0)] = \mathrm{E}[\mathbf{x}_i(y_i - \mathbf{x}_i'\theta_1)](= \mathbf{0}) \Rightarrow \theta_0 = \theta_1 \tag{9}$$

- Rearranging

$$\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'](\theta_0 - \theta_1) = \mathbf{0} \tag{10}$$

- So $\theta_0 = \theta_1$ if $\mathrm{E}[\mathbf{x}_i\mathbf{x}_i']$ is invertible (rank K)

$$\mathrm{E}[\mathbf{x}_i\mathbf{x}_i']^{-1}\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'](\theta_0 - \theta_1) = \mathbf{0} \Rightarrow \theta_0 = \theta_1 \tag{11}$$

Exogeneity

- In the decomposition of the joint distribution ("Bayes Theorem")

$$f_J(\mathbf{Y},\mathbf{Z}|\theta) = \; f_C(\mathbf{Y}|\mathbf{Z},\theta_1) \times \; f_M(\mathbf{Z}|\theta_2)$$

θ_2 are uninformative about θ_1.

- Z said to be exogenous because knowledge of θ_2 not essential for statistical inference on θ_1
- θ_2 said to be ancillary in inference on θ_1

- Exogeneity is a strong assumption. *It is a property of random variables relative to parameters of interest*.
- The same variable may be treated as endogenous in one context and exogenous in a different one.

Structural Models (2)

- Modeling objective: Explain the values of observable vector-valued variable $\mathbf{y}$, $\mathbf{y}' = (y_1, \ldots, y_G)$.
- Variables $\mathbf{y}$ are assumed to be interdependent. Interdependence between $\mathbf{z}_i$ is not modelled.

- The i^{th} observation satisfies the set of implicit equations

$$\mathbf{g}\left(\mathbf{y}_i, \mathbf{z}_i, \mathbf{u}_i | \theta\right) = \mathbf{0}, \tag{12}$$

where $\mathbf{g}$ is a known (linear or nonlinear) function.

Reduced Form Models

- Assume a unique solution for $\mathbf{y}_i$ for every $(\mathbf{z}_i, \mathbf{u}_i)$.
- Assume solution additively separable in $(\mathbf{z}_i, \mathbf{u}_i)$.
- Then write an explicit form with $\mathbf{y}$ as function of $(\mathbf{z}, \mathbf{u})$

$$\mathbf{y}_i = \mathbf{f}(\mathbf{z}_i, \mathbf{u}_i | \pi) \tag{13}$$

$\Rightarrow$ the **reduced form** of the structural model, where: π vector of reduced form parameters that are functions of θ.

- Reduced form is obtained by solving the structural model for the endogenous variables $\mathbf{y}_i$, given $(\mathbf{z}_i, \mathbf{u}_i)$.

Linear SEM Example

- G-equation SEM (subject to normalization and exclusion restrictions) is written as

$$\mathbf{y}_i'\mathbf{B} + \mathbf{z}_i'\Gamma = \mathbf{u}_i, \tag{14}$$

where $\mathbf{y}_i, \mathbf{B}, \mathbf{z}_i, \Gamma, \mathbf{u}_i$ have dimensions $(G \times 1)$, $(G \times G)$, $(k \times 1)$, $(k \times G)$, and $(1 \times G)$, respectively. For specified values of $(\mathbf{B}, \Gamma)$ and $(\mathbf{z}_i, \mathbf{u}_i)$ G linear simultaneous equations can in principle be solved for $\mathbf{y}_i$. The $(N \times K)$ matrix $\mathbf{Z}$ is formed by stacking $\mathbf{z}_i'$, $i = 1, \ldots, N$.

- Other notation:

$$\begin{array}{ccccc}
\mathbf{B}\mathbf{y}_i & + & \Gamma\mathbf{z}_i & = & \mathbf{u}_i, \\
(G \times G)(G \times 1) & & (G \times K)(K \times 1) & = & (G \times 1) \\
\mathbf{y}_i'\beta_g & + & \mathbf{z}_i'\gamma_g & = & u_{ig} \\
(1 \times G)(G \times 1) & & (1 \times K)(K \times 1) & = & (1 \times 1)
\end{array}$$

Linear SEM Assumptions

1. $\mathbf{B}$ is nonsingular and has rank G.
2. $\text{rank}[\mathbf{Z}] = K$ and
3. $\text{E}[\mathbf{u}_i] = 0$ and $\text{E}[\mathbf{u}_i \mathbf{u}_i'] = \Sigma = [\sigma_{ij}]$, where Σ is a symmetric positive definite matrix (stronger version: $\mathbf{u}_i \sim \mathcal{N}[\mathbf{0}, \Sigma]$)
4. The errors in each equation are serially independent.

- In this model the structure consists of $(\mathbf{B}, \Gamma, \Sigma)$.
- Structural model has primacy over reduced form for several reasons:

1. Equations themselves have interpretations as economic relationships such as demand or supply relations, production functions, and so forth, and

2. $(\mathbf{B}, \Gamma, \Sigma)$ are subject to **restrictions** of economic theory.
3. $\mathbf{B}$ embodies "causal" or **direct** connections between endogenous variables and are often the key target of identification and estimation, e.g. demand and supply elasticities

Linear SEM: Reduced form

Solve for all the endogenous variables in terms of all the exogenous variables,

$$
\begin{aligned}
\mathbf{y}_i' + \mathbf{z}_i'\Gamma\mathbf{B}^{-1} &= \mathbf{u}_i'\mathbf{B}^{-1} \\
\mathbf{y}_i' &= \mathbf{z}_i'\Pi + \mathbf{v}_i', \qquad (15)
\end{aligned}
$$

where $\Pi = -\Gamma\mathbf{B}^{-1}$ and $\mathbf{v}_i' = \mathbf{u}_i'\mathbf{B}^{-1}$. In the SEM framework the reduced form is also of interest because

1. It captures the **direct and indirect** effects
2. It is always identified given sufficient sample variation (with infinite data we can recover Π)
3. It permits conditional prediction of endogenous outcomes
4. It also embodies the restrictions on the structural specification
5. It offers the potential of identifying structural parameters

Identification

- The structural parameters are **point identified** if there is a unique ($\mathbf{B}$,Γ,Σ) which satisfy

$$\Pi = -\Gamma\mathbf{B}^{-1} \tag{16}$$

$$\Omega_v \equiv E(\mathbf{v}_i\mathbf{v}_i') = \mathbf{B}^{-1'}\Sigma\mathbf{B}^{-1} \tag{17}$$

- Structural models typically try to impose enough restrictions on ($\mathbf{B}$,Γ,Σ) so that they are point identified

Causal Interpretation in SEM

- A very simple example of a linear structural model:

$$
\begin{aligned}
y_1 &= \gamma_1 + \beta_1 y_2 + u_1, \; 0 < \beta_1 < 1 \\
y_2 &= y_1 + z_1.
\end{aligned}
$$

- z_1 is exogenous and therefore its variation is induced by external sources that we may regard as interventions. whose impact is measured by reduced form equations

$$
\begin{aligned}
y_1 &= \frac{\gamma_1}{1-\beta_1} + \frac{\beta_1}{1-\beta_1} z_1 + \frac{1}{1-\beta_1} u_1 \\
y_2 &= \frac{\gamma_1}{1-\beta_1} + \frac{1}{1-\beta_1} z_1 + \frac{1}{1-\beta_1} u_1
\end{aligned}
$$

Causal Interpretation in SEM

$$
\begin{aligned}
\frac{\partial y_1}{\partial y_2} &= \beta_1 = \frac{\beta_1}{1-\beta_1} \div \frac{1}{1-\beta_1} \\
&= \frac{\partial y_1}{\partial z_1} \div \frac{\partial y_2}{\partial z_1}.
\end{aligned}
$$

- In what sense does β_1 measure the causal effect of y_2 on y_1?
- IV estimator of β_1 based on $cov(y_1, z_1)/cov(y_2, z_1)$
- Presence of z_1 (or **exclusion restriction**) essential for identification of β_1

Nonlinear structural model example

$$
\begin{aligned}
\ln y_1 &= \gamma_1 + \beta_1 y_2 + u_1, \quad 0 < \beta_1 < 1 \\
y_1 &= \exp(\gamma_1 + \beta_1 y_2 + u_1) \\
&= \exp(\gamma_1 + \beta_1 y_2) \exp(u_1) \\
&= \exp(\gamma_1 + \beta_1 y_2) v_1 \text{ where } v_1 = \exp(u_1) \\
y_2 &= y_1 + z_1.
\end{aligned}
$$

- No exact closed form reduced form available

Limitations of SEM

- Strong restrictions (Manski's "Law of Decerasing Credibility") required for point identification.
- Possibility that relationships are specified arbitrarily and lack a convincing theoretical foundation
- Simultaneous estimation of the full model may be difficult, so all information might not be employed.
- Researcher may be interested only in a narrow set of questions (e.g. related to demand side) for which it may not be necessary to specify a full-fledged simultaneous model (with a supply equation and an inventory equation).

Useful references

1. J. Angrist and J-S Pischke, *Mostly harmless econometrics* (PUP, 2009)
2. J. Angrist and J-S Pischke, *Mastering metrics: The path from cause to effect* (PUP, 2014)

Introduction to Instrumental Variables

$$y = \beta x + \varepsilon. \tag{18}$$

- The fundamental assumption for consistency of least squares estimators: $\mathrm{E}[\varepsilon|x] = 0$.
- OLS is inconsistent if this assumption fails.

- Then the OLS estimator can no longer be given a causal interpretation because β is then not identified. β_j is not the marginal effect of an exogenous change in x_j.

- Example: y :(log) earnings; x : years of schooling; β : rate of return to schooling
- IV/GMM is a family of estimation methods that avoid the inconsistency property of OLS

Inconsitency of OLS in SEM

- A fundamental problem as such marginal effects are a key input to economic policy.
- IV/GMM estimator provides a consistent estimator under the very strong assumption that valid instruments exist, i.e. instruments z are correlated with the regressors x and satisfy $\mathrm{E}[\varepsilon|z] = 0$.
- What would be a valid IV for schooling? What determines whether an instrument is valid?
- IV approach is the original and leading approach for estimating models with endogenous regressors.

Assumptions

- To identify the parameter(s) of interest, need to have valid instruments
- Two conditions must be satisfied for a valid IV z - noncorrelation with ε and high correlation with x
- Practically, it can be very difficult to obtain valid instruments. Also instruments may be weakly correlated with endogenous regressors.
- Suppose y measures earnings and x measures years of schooling and ε is the error term.
- The OLS exogeneity assumption implies we have the following path diagram.

$$\begin{tikzcd} x \arrow[r] & y \\ \varepsilon \arrow[ur] & \end{tikzcd}$$

Endogeneity

- Error ε embodies all factors other than schooling that determine earnings, e.g. ability.
- However, high ability will induce correlation between x and ε as high (low) ability will on average be associated with high (low) years of schooling. So a more appropriate schematic diagram is the following:

$$\begin{tikzcd} x \arrow[r] & y \\ \varepsilon \arrow[u] \arrow[ur] & \end{tikzcd}$$

  where now there is an association between x and ε
- OLS estimator $\widehat{\beta}$ is then inconsistent for β, because $\widehat{\beta}$ combines the desired direct effect of schooling on earnings (β) with the indirect effect that people with high schooling are likely to have high ability, high ε, and hence high y

IV approach

- Obvious solution to the endogeneity problem: include as regressors controls for ability (control function approach).
  - But such regressors may not be available.
- IV provides an alternative solution. Introduce a (new) instrumental variable z that has the property that changes in z are associated with changes in x but do not lead to change in y (except indirectly via x). This leads to the following path diagram

$$\begin{array}{ccccc} z & \longrightarrow & x & \longrightarrow & y \\ & & \uparrow & \nearrow & \\ & & \varepsilon & & \end{array}$$

  For example, proximity to college (z) may determine college attendance (x) but not directly determine earnings (y).

Model set-up

- Consider the more general regression model with scalar dependent variable y_1 that depends on m endogenous regressors, denoted $\mathbf{y}_2$, and K_1 exogenous regressors (including an intercept), denoted $\mathbf{x}_1$: "a structural equation"

$$y_{1i} = \mathbf{y}_{2i}'\beta_1 + \mathbf{x}_{1i}'\beta_2 + \varepsilon_i, \quad i = 1, \ldots, N,$$

  The regression errors ε_i are assumed to be uncorrelated with $\mathbf{x}_{1i}$, but are correlated with $\mathbf{y}_{2i}$. This correlation leads to the OLS estimator being inconsistent for β.

- To obtain a consistent estimator we assume the existence of at least m instrumental variables $\mathbf{x}_2$ for $\mathbf{y}_2$ that satisfy the assumption that $\mathrm{E}[\varepsilon_i|\mathbf{x}_{2i}] = 0$. The instruments $\mathbf{x}_2$ need to be correlated with $\mathbf{y}_2$. Motivate this is using the first-stage equation (also called a "reduced form" model)

$$y_{2ji} = \mathbf{x}_{1i}'\pi_{1j} + \mathbf{x}_{2i}'\pi_{2j} + v_{ji}, \quad j = 1, \ldots, m.$$

Model set-up (2)

- The first-stage equations have only exogenous variables on the right-hand side.
- The model can be more simply written as $y_i = \mathbf{x}_i'\beta + \varepsilon_i$, where the regressor vector $\mathbf{x}_i' = (\mathbf{y}_{2i}' \; \mathbf{x}_{1i}')$ combines endogenous and exogenous variables, and the dependent variable is denoted y rather than y_1.
- Then the vector of instrumental variables (or, more simply, instruments) is $\mathbf{z}_i' = (\mathbf{x}_{1i}' \; \mathbf{x}_{2i}')$, where $\mathbf{x}_1$ serves as the (ideal) instrument for itself and $\mathbf{x}_2$ is the instrument for $\mathbf{y}_2$, and the instruments $\mathbf{z}$ satisfy the conditional moment restriction $\mathrm{E}[\varepsilon_i|\mathbf{z}_i] = 0$.

Recap of Linear GMM with Instruments

- Consider the linear regression model

$$y_i = \mathbf{x}_i'\beta + u_i, \tag{19}$$

where each component of $\mathbf{x}$ is viewed as being an **exogenous regressor** if it is uncorrelated with the error in model (19) or an **endogenous regressor** if it is correlated. If any components of $\mathbf{x}$ are endogenous then LS estimators are inconsistent for β.

- Consistent estimates can be obtained by IV estimation. The key assumption is the existence of a vector of **instruments** $\mathbf{z}$ that satisfies

$$\mathrm{E}\left[u_i | \mathbf{z}_i\right] = \mathbf{0}\ . \tag{20}$$

Exogenous regressors are valid instruments so may form a subset of $\mathbf{z}$.

- Minimally need dim $(\mathbf{z}) = \dim(\mathbf{x})$.
- dim $(\mathbf{z}) > \dim(\mathbf{x}) \Rightarrow$ "overidentified" model.

A Simple Example ("Just-identified" case)

- One endogenous variable; one instrumental variable; one moment restriction

$$y = x\beta + \varepsilon. \tag{21}$$

- Moment condition

$$E(\varepsilon|z) = 0 \rightarrow E[y - x\beta|z] = 0 \rightarrow E[z(y - x\beta)] = 0. \tag{22}$$

- MM/IV estimator

$$\begin{aligned} \widehat{\beta}_{IV} &= \frac{\sum z_i y_i}{\sum z_i x_i} \\ &= \beta + \frac{\sum z_i \varepsilon_i / N}{\sum z_i x_i / N} \xrightarrow{p} \beta \end{aligned}$$

Variance of IV Estimator under Homoskedasticity

$$
\begin{aligned}
\widehat{\beta}_{IV} - \beta &= \frac{\sum z_i \varepsilon_i / N}{\sum z_i x_i / N} \\
\mathrm{V}\,[\widehat{\beta}_{IV} | \mathbf{X}, \mathbf{Z}] &= \mathrm{V}[\sum w_i \varepsilon_i] \text{ where } w_i = \frac{z_i}{\sum z_i x_i} \\
&= \sigma^2 \frac{\sum z_i^2}{(\sum z_i x_i)^2} \\
&= \sigma^2 [\mathbf{X}'\mathbf{Z}]^{-1} [\mathbf{Z}'\mathbf{Z}] [\mathbf{Z}'\mathbf{X}]^{-1}
\end{aligned}
$$

- Interpretation in terms of weak instruments
- Impact of weak IV on variance

Large sample properties of IV

- In applied work the just identified case is very common
- IV estimator is consistent (asymptotically unbiased)
- In finite sample the IV would in general be biased (as would be the OLS)
- The IV/GMM estimator is asymptotically normal distributed. That means that we can do inference using large sample theory,
- That is we can apply the usual z-test, t-test, etc; construct confidence interval in the usual way as long as we use the appropriate estimates of the asymptotic variance.
- IV/GMM are semi-parametric methods. They do not require as detailed a model specification as MLE.

Overidentified Case (1)

- "Structural equation"

$$y_i = x_i\beta + \varepsilon_i, \quad i = 1, \ldots, N,$$

The regression errors ε_i are assumed to be correlated with x_i. This correlation leads to the OLS estimator being inconsistent for β.

- Given multiple instruments, we want to combine (use) them optimally.
- Optimality property refers to the variance of the estimator.

- To combine two instruments (z_1, z_2), use the first-stage regression equation (also called a "reduced form" model)

$$x_i = z_{1i}\pi_1 + z_{2i}\pi_2 + v_i.$$

- The reduced form predicted value is the "optimally weighted" linear combination of the available instruments.

Overidentified Case (2)

- Generate a predicted or fitted value of x_i, denoted $\widehat{x}_i$, and use it as an instrument at the second stage. The moment condition is:

$$E(\varepsilon|\widehat{x}) = 0 \;\rightarrow\; E[y - x\beta|\widehat{x}] = 0 \rightarrow\; E[\widehat{x}(y - x\beta)] = 0. \qquad (23)$$

- MM/IV efficient estimator

$$\begin{aligned}\widehat{\beta}_{2SLS} &= \frac{\sum \widehat{x}_i y_i}{\sum \widehat{x}_i x_i} \\ &= \beta + \frac{\sum \widehat{x}_i \varepsilon_i / N}{\sum \widehat{x}_i x_i / N} \xrightarrow{p} \beta\end{aligned}$$

General Overidentified Case

- Above approach generalizes to the case of two or more endogenous regressors.

$$y_{1i} = \mathbf{y}_{2i}'\beta_1 + \mathbf{x}_{1i}'\beta_2 + \varepsilon_i, \quad i = 1, \ldots, N,$$

The regression errors ε_i are assumed to be uncorrelated with $\mathbf{x}_{1i}$, but are correlated with $\mathbf{y}_{2i}$. This correlation leads to the OLS estimator being inconsistent for β.

$$\begin{aligned} \mathbf{X} &= [\mathbf{Y}_2 \mid \mathbf{X}_1] \\ \widehat{\mathbf{X}} &= \mathbf{Z}\widehat{\Pi} = \mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{P}_Z\mathbf{X} \end{aligned}$$

- Using $\widehat{\mathbf{X}}$ as instruments generates the 2SLS estimator for the overidentified case.

Interpretation of 2SLS

- 2SLS approach involves finding at least as many IVs as the number of right-hand side endogenous variables $\mathbf{y}_2$, and estimating fitted values from reduced form OLS regression.
- These fitted values replace the original endogenous variables. We then apply OLS to this new regression.

$$\begin{aligned} y_{1i} &= \mathbf{y}'_{2i}\beta_1 + \mathbf{x}'_{1i}\beta_2 + \varepsilon_i, \\ &= (\widehat{\mathbf{y}}'_{2i} + \widehat{\mathbf{v}}'_i)\beta_1 + \mathbf{x}'_{1i}\beta_2 + \varepsilon_i \\ &= \widehat{\mathbf{y}}'_{2i}\beta_1 + \mathbf{x}'_{1i}\beta_2 + (\varepsilon_i + \widehat{\mathbf{v}}'_i\beta_1) \end{aligned}$$

- This last equation satisfies the requirements for OLS to produce a consistent estimator of $(\beta'_1 \beta'_2)$.
- However, there is a complication when it comes to estimating the variance of the estimates.

Linear GMM Interpretation (1)

- The IV conditional moment restriction and model imply the unconditional population moment restriction

$$E\left[\mathbf{z}_i(y_i - \mathbf{x}_i'\beta)\right] = \mathbf{0}, \tag{24}$$

- The GMM estimator of β is defined by the solution of the sample analog of the above population moment equation, i.e.

$$\frac{1}{N}\sum_i [\mathbf{z}_i(y_i - \mathbf{x}_i'\beta)] = \mathbf{0}.$$

- Define $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, and let $\mathbf{Z}$ denote the $N \times r$ matrix of instruments with i^{th} row $\mathbf{z}_i'$. Then $\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\beta) = \mathbf{Z}'\mathbf{u}$ and we maximize the quadratic objective function

$$Q_N(\beta) = -\left[\frac{1}{N}(\mathbf{y} - \mathbf{X}\beta)'\mathbf{Z}\right] \mathbf{W}_N \left[\frac{1}{N}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)\right], \tag{25}$$

where $\mathbf{W}_N$ is an $r \times r$ full rank symmetric weighting matrix. Hence GMM is interpreted as an M-estimator.

Linear GMM (2)

- In this special case of GMM first-order conditions

$$\frac{\partial Q_N(\beta)}{\partial \beta} = -2\left[\frac{1}{N}\mathbf{X}'\mathbf{Z}\right]\mathbf{W}_N\left[\frac{1}{N}\mathbf{Z}'(\mathbf{y}-\mathbf{X}\beta)\right] = \mathbf{0}$$

  have solution for β : **GMM estimator in the linear IV model**

$$\widehat{\beta}_{\text{GMM}} = \left[\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{y}, \qquad (26)$$

  where the divisions by N have cancelled out.

Different Linear GMM Estimators

- Implementation requires specification of the weighting matrix $\mathbf{W}_N$.
- For just-identified models all choices of $\mathbf{W}_N$ lead to the same estimator.
- Try $\mathbf{W}_N = [V[\mathbf{Z}'\mathbf{u}]|\mathbf{Z}]^{-1} = \mathrm{E}[\mathbf{Z}'\mathbf{u}\mathbf{u}'\mathbf{Z}|\mathbf{Z}]^{-1} = [\sigma^2\mathbf{Z}'\mathbf{Z}]^{-1}$ which assumes that the equation errors are homoskedastic and serially independent.

$$
\begin{aligned}
\widehat{\beta}_{\text{GMM}} &= \left[\mathbf{X}'\mathbf{Z}[\sigma^2\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Z}[\sigma^2\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{y}, \\
&= \left[\mathbf{X}'\mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{y}, \\
&= \left[\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y}, \text{ where } \mathbf{P}_\mathbf{Z} = \mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}' \\
&= \left[\widehat{\mathbf{X}}'\widehat{\mathbf{X}}\right]^{-1}\widehat{\mathbf{X}}'\mathbf{y} \text{ where } \widehat{\mathbf{X}} = \mathbf{P}_\mathbf{Z}\mathbf{X}, \text{ and } \mathbf{P}_\mathbf{Z}\mathbf{P}_\mathbf{Z} = \mathbf{P}_\mathbf{Z}
\end{aligned}
$$

Extending 2SLS to GMM

- 2SLS assumes that equation errors are **homoskedastic and serially independent**.
- Heteroskedasticity is quite common with cross section data and serial correlation with time series data
- Can we extend 2SLS to make it more robust by accommodating these other realities?
- The main idea for doing this comes from the Huber-White method of handling heteroskedasticity.
- H-W proposed a way of handling heteroskedasticity of any arbitrary form without having to know what its functional form is.
- This method will not affect the parameter estimates, but it will affect the variances of the parameter estimates

Extending 2SLS to GMM (2)

- Consider the **heteroskedastic model**

$$y_i = \beta x_i + u_i \tag{27}$$
$$= \beta x_i + \sigma_i \varepsilon_i. \tag{28}$$

where σ_i is an individual specific constant

- $V[u_i] = V[\sigma_i \varepsilon_i] = \sigma_i^2 V[\varepsilon_i] \Rightarrow$ heteroskedasticity assuming that $V[\varepsilon_i] = \sigma_\varepsilon^2$
- Rewrite the regression as a weighted regression as follows:

$$\frac{y_i}{\sigma_i} = \beta \frac{x_i}{\sigma_i} + \varepsilon_i.$$

Heteroskedasticity is not a problem for the weighted regression. We could apply 2SLS

Extending 2SLS to GMM (3)

- The weighted regression is not possible if we do not know the weights which depend on the exact form of heteroskedasticity.
- Under homoskedasticity $V[\mathbf{Z}'\mathbf{u}|\mathbf{Z}] = \sigma^2\mathbf{Z}'\mathbf{Z}$
- Under heteroskedasticity $V[\mathbf{Z}'\mathbf{u}|\mathbf{Z}] = \mathbf{Z}'V[\mathbf{u}|\mathbf{Z}]\mathbf{Z} = \mathbf{Z}'\Omega\mathbf{Z}$; where $V[\mathbf{u}|\mathbf{Z}] = \Omega = \mathbf{diag}[\sigma_i^2]$ and $\widehat{\mathrm{V}}[\widehat{\beta}_{\mathrm{GMM}}] = [\mathbf{X}'\mathbf{Z}\,\widehat{\mathbf{Z}'\Omega\mathbf{Z}}^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$
- H-W method can be applied to obtain an estimate of $\mathbf{Z}'\Omega\mathbf{Z}$ which then yields a robust GMM estimate of $\widehat{\beta}_{\mathrm{GMM}}$
- The method was extended by Newey and White to allow for serial dependence of residuals.

Three leading estimators

References: CT-Stata, Chapters 6.3. and 6.4.

- Table gives the appropriate specialization of the estimated variance matrix formula.
- When errors are allowed to be heteroskedastic and/or serially dependent the term $[\sigma^2 \mathbf{Z}'\mathbf{Z}]$ is replaced by an unknown matrix $\mathbf{S} = \mathbf{Z}'\Omega\mathbf{Z}$.
- Given an assumption of the form of Ω, we can replace $\mathbf{S}$ with a consistent estimator, just like for GLS.

Estimator	Definition and Estimate of the VCE
IV (just-identified)	$\widehat{\beta}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ $\widehat{\text{V}}[\widehat{\beta}_{IV}] = \widehat{\sigma}^2(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}$
2SLS	$\widehat{\beta}_{\text{2SLS}} = \left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Z}[\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{y}$ $\widehat{\text{V}}[\widehat{\beta}_{2SLS}] = \widehat{\sigma}^2\left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}$
Optimal GMM	$\widehat{\beta}_{\text{OGMM}} = \left[\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{y}$ $\widehat{\text{V}}[\widehat{\beta}_{\text{OGMM}}] = \left[\mathbf{X}'\mathbf{Z}\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}$

Testing for endogeneity

- Variable addition test based on:

$$y_{1i} = \beta_1 y_{2i} + \mathbf{x}'_{1i}\beta_2 + \rho v_{1i} + u_i$$

  where v_{1i} are reduced form residuals from the first stage regression.

- Test $H_0 : \rho = 0$
- If H_0 is rejected then evidence favors endogeneity of the regressor.

- After IV regression in STATA, type `estat endogenous`

Testing for IV validity in an overidentified model

- First, we cannot test the validity of the IV if the model is just-identified.
- In an over-identified model, we can test whether additional moment restrictions are valid.
- The key assumption behind IV estimation is $\mathrm{E}[\mathbf{Z}'\mathbf{u}] = \mathbf{0}$.
- A natural test statistic would test whether the sample covariance between the instruments and estimated GMM residuals $\widehat{\mathbf{u}}$ is close to zero.
- Test statistic: $OIR = \widehat{\mathbf{u}}'\mathbf{Z}'\widehat{\mathbf{S}}^{-1}\mathbf{Z}'\widehat{\mathbf{u}}$ has $\chi^2(p)$ distribution where p is the number of overidentifying restrictions.
- In Stata the test is applied by running the command `estat overid` immediately after estimating the overidentified model.

How to test for weak IV?

- Suppose the chosen instrument is valid, so IV estimator is consistent.
- What if the instrument is weak? Then asymptotic theory can provide a poor guide to actual finite-sample distributions.
- How to test for weak instruments after applying `ivregress`?
- Simple method: examine the pairwise correlations between any endogenous regressor and instruments.

How to test for weak IV when more than 1 IV?

- If more than one instrument, consider the joint correlation of the endogenous regressor with the several instruments.
- Possible measures:
  - R^2 from reduced for regression of y_2 on the several instruments x_2,
  - $F-$statistic for test of overall fit in this regression.
- Low values indicative of weak instruments.
- Because the reduced form also has exogenous regressors x_1 in the first-stage regression should test for *additional explanatory power* of the instruments.
- $F-$ statistic for joint significance of the instruments x_2 in first-stage regression
- Second diagnostic: partial R^2 between y_2 and x_2 after controlling for x_1