

EC309 ECONOMETRIC THEORY

Dr. MYUNG HWAN SEO
LONDON SCHOOL OF ECONOMICS

March 13, 2012

Contents

1	Efficient Estimation	5
1.1	Maximum Likelihood Estimation	6
1.2	Asymptotic Theory for M-Estimation	11
1.3	Generalized Least Squares	25
2	System of Equations	43
2.1	Multivariate regression	45
2.2	SURE	50
2.3	Simultaneous Equation System	53
3	Time Series	64
3.1	Stationary Processes	65
3.2	Process with Deterministic Trend	84
3.3	Vector Autoregressions	88
4	Unit Root & Cointegration	96
4.1	Unit root test	105
4.2	Cointegration	115

5	Asymptotic Tests	124
5.1	Hypothesis testing	124
5.1.1	Trinity of Tests	130
5.1.2	t & Wald Tests	135
5.1.3	GMM Distance statistic	140
5.2	Specification Testing	141
5.2.1	Overidentification Test in GMM	141
5.2.2	Hausman type tests	143
A	Problems	152
A.1	Regression	152
A.2	System of Equations	166
A.3	Time Series	168

Syllabus

- Teacher Responsible: Dr. Myung Hwan Seo (S580), email: m.seo@lse.ac.uk.
- Office hours: Wednesday 1:30 – 2:30 PM;
- Availability and Restrictions: This course is for B.Sc. degrees in Econometrics and Mathematical Economics and is also available to other students as permitted by the regulations. Students should have taken the course Probability, Distribution Theory and Inference (or equivalent) and/or Principles of Econometrics (or equivalent). Good knowledge of linear algebra, calculus and statistical theory (at least the first year college level) is required.
- Description of the Course: This is an advanced course designed to help students develop firm understandings on large sample theory for commonly used parametric estimators and statistics in econometrics. Although econometric tools are best motivated by various empirical applications, econometric theory enables us to understand the strengths and limitations of such tools. Of our emphasis is an in-depth examination of the regression model.
- In the Michaelmas Term, we intend to study: (i) Basics of large sample theory; (ii) Estimation of linear regression models (OLS, GMM, GLS); (iii)

Testing hypotheses and model specifications. In the Lent Term, we shall cover: (i) Estimation of nonlinear models (MLE, Nonlinear least squares); (ii) systems of equations; (iii) time series analysis. The classes are designed to answer your questions and provide solutions to your homework and extra exercises related to lectures.

- Textbooks:
 1. B. Hansen (2008), *Econometrics*, electronic version can be found at <http://www.ssc.wisc.edu/~bhansen>;
 2. R. Davidson and J. MacKinnon (2004), *Econometric Theory and Methods* or earlier version;
 3. Greene (2005) *Econometric Analysis*, for a general reference.
- Exam: One three-hour written exam in the Summer Term.
- Homeworks: The students are expected to present their solutions in the class.

1 Efficient Estimation

1. Finite sample efficiency among unbiased estimators:
 - (a) Gauss-Markov theorem: among linear estimators—linear in Y , in the linear regression framework.
 - i. under homoskedasticity, the OLS estimator is BLUE.
 - ii. w/o homoskedasticity¹, the GLS estimator is BLUE.
 - (b) UMVUE for parametric models (Cramer-Rao lower bound; sufficient and complete statistic): MLE is often UMVUE.
2. Asymptotic efficiency among consistent estimators: GLS is asymptotically efficient for the regression model and MLE is for parametric models under mild regularity conditions.

Remark 1. MLE is not necessarily linear.

2. there is trade-off between efficiency and robustness.
3. The discussion here is to motivate some of the estimators we will analyze and not complete.

¹The regression model is semi-parametric since the distribution of the data is unknown but the structural equation depends only on the finite number of unknowns.

1.1 Maximum Likelihood Estimation

- Let $x = (x_1, \dots, x_n)'$ be an observation from $X = (X_1, \dots, X_n)'$, whose density is assumed to belong to the family

$$\mathcal{P} = \{p(\cdot, \theta) \mid \theta \in \Theta\}$$

If a density is thought of as a function of the unknown parameter θ , i.e., $p(x, \cdot)$ is called the *likelihood function*.

- The *maximum likelihood estimator* (MLE) of θ is defined by

$$\hat{\theta}_{\text{ML}} = \operatorname{argmax}_{\theta \in \Theta} p(x, \theta)$$

1. Computationally, it is often much easier to maximize the *log-likelihood function*

$$\ell(x, \theta) = \log p(x, \theta)$$

which is legitimate since log function is monotonely increasing.

2. Usually, the function $\ell(x, \cdot)$ is differentiable and globally concave for every x . The maximizer can therefore be found simply by solving the first-order condition (FOC)

$$\frac{\partial}{\partial \theta} \ell(x, \theta) = 0$$

for θ in terms of x .

3. The MLE of a function of θ , say $\pi = f(\theta)$, is given by

$$\hat{\pi}_{\text{ML}} = f\left(\hat{\theta}_{\text{ML}}\right)$$

that is, the ML estimation is invariant with respect to reparametrization.

Example 1 Let X_i , $i = 1, \dots, n$, be *i.i.d.* $\mathbf{N}(\mu, \sigma^2)$. Then the log-likelihood function is given by

$$\ell(x, \mu, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

and solving the FOC's yields

$$\begin{aligned}\hat{\mu}_{\text{ML}} &= \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\ \hat{\sigma}_{\text{ML}}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\end{aligned}$$

simultaneously.

As desirable properties for a good estimator, we introduce the concepts of *unbiasedness* and *minimum mean squared error*. We first define unbiasedness. Here and elsewhere, we denote by \mathbf{P}_θ the probability in Ω . Expectation with respect to \mathbf{P}_θ is denoted by \mathbf{E}_θ .

Definition 1 *An estimator $T = \tau(X)$ is called unbiased if*

$$\mathbf{E}_\theta(T) = \theta$$

for all $\theta \in \Theta$.

The *mean squared error* (MSE) of an estimator $T = \tau(X)$ can be decomposed as the sum of the variance and the squared bias, as shown below

$$\mathbf{E}_\theta(T - \theta)^2 = \mathbf{E}_\theta(T - \mathbf{E}_\theta(T))^2 + (\mathbf{E}_\theta(T) - \theta)^2.$$

Remarks

- (a) For an unbiased estimator, mean squared error reduces to variance.
- (b) MSE is, in general, a function of the unknown parameter θ . An estimator that has the smallest MSE for all $\theta \in \Theta$ does not exist. This is because the trivial estimator $\hat{\theta} = \theta_1$ for some fixed value θ_1 has zero MSE at $\theta = \theta_1$. The MSE of any other estimator is strictly positive at $\theta = \theta_1$, since it takes other values with positive probabilities. For example, the estimator $\hat{\theta} = 1$ cannot be beaten in MSE sense at $\theta = 1$ but is a terrible estimator otherwise.

- Some useful definitions and results regarding MLE.

Definition 2 *We define*

(a) (score function) $s(x, \theta) = \frac{\partial}{\partial \theta} \ell(x, \theta)$

(b) ((Fisher) information) $I(\theta) = \mathbf{E}_{\theta} \left(s(X, \theta) s(X, \theta)' \right)$

(c) (Hessian) $H(x, \theta) = \frac{\partial^2}{\partial \theta \partial \theta'} \ell(x, \theta)$

(d) (expected Hessian) $H(\theta) = \mathbf{E}_{\theta} H(X, \theta)$

Proposition 1 *Suppose that Assumption (a) holds. Then*

$$\mathbf{E}_{\theta} s(X, \theta) = 0.$$

Proposition 2 *Suppose that Assumption (b) holds. Then*

$$I(\theta) = -H(\theta)$$

Theorem 3 (Cramer-Rao Bound) Let $T = \tau(X)$ be an unbiased estimator of θ , and suppose that Assumption (c) holds. Then

$$\text{var}_\theta T \geq I(\theta)^{-1}$$

Example 2 Let X_1, \dots, X_n be i.i.d. $\mathbf{N}(\mu, \sigma^2)$. Suppose $\sigma^2 = \sigma_0^2$ is known. Then

$$\begin{aligned}\ell(x_i, \mu, \sigma_0^2) &= -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{(x_i - \mu)^2}{2\sigma_0^2} \\ s(x_i, \mu, \sigma_0^2) &= \sigma_0^{-2} (x_i - \mu) \\ H(x_i, \mu, \sigma_0^2) &= -\sigma_0^{-2}\end{aligned}$$

We may easily deduce that

$$I(\mu, \sigma_0^2) = -H(\mu, \sigma_0^2) = n\sigma_0^{-2}.$$

It can now be easily seen that the estimator \bar{X} for μ achieves the Cramer-Rao bound.

1.2 Asymptotic Theory for M-Estimation

M-estimation stands for MLE like estimation in the sense that the estimator is obtained by maximizing the sample mean of random functions:

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n m(w_i, \theta).$$

Example 3 *Nonlinear regression:*

$$\begin{aligned} y_i &= f(x_i, \beta) + \varepsilon_i \\ \mathbb{E}(\varepsilon_i | x_i) &= 0, \end{aligned}$$

where the functional form of f is known up to unknown parameter β . For instance,

$$f(x_i, \beta) = x_i \theta_1 \left(1 + e^{-(x_i - \theta_2)}\right)^{-1} + x_i \theta_3,$$

is called the smooth transition model. Sometimes, the model is linear in parameters even though f is nonlinear in x_i . For example, the polynomials are written linear in parameters. These are treated as the same way as the linear models.

Some are continuous while others are discontinuous, as in e.g. threshold regression. The nonlinear least squares (NLS) estimator is then obtained by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i, \beta))^2.$$

Example 4 *Limited dependent variable models: For example, we may consider the binary choice model (or random utility model)*

$$\begin{aligned}y_i^* &= x_i' \beta - \varepsilon_i, \\y_i &= 1 \{y_i^* > 0\},\end{aligned}$$

where we only observe $(y_i, x_i)_{i=1}^n$. Thus,

$$y_i = 1 \{\varepsilon_i < x_i' \beta\}.$$

This model is often estimated by MLE. Depending on the distributional assumption imposed on ε_i (it is assumed to be independent of x_i), the model is called probit or logit model. Let ε_i has a distribution Φ . Then, the log-likelihood function is given as

$$\sum_{i=1}^n y_i \ln \Phi(x_i' \beta) + (1 - y_i) \ln (1 - \Phi(x_i' \beta)).$$

Let

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^n m(w_i, \theta) = \frac{1}{n} \sum_{i=1}^n m_i(\theta).$$

The first derivative

$$z_n(\theta) = \frac{1}{n} \sum_{i=1}^n z_i(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} m_i(\theta),$$

is called the score and the second derivative

$$H_n(\theta) = \frac{\partial}{\partial \theta'} z_n(\theta) = \frac{1}{n} \sum_{i=1}^n h_i(\theta),$$

is called the Hessian.

To develop the asymptotic theory for $\hat{\theta}$ in general non-linear models, we need several new ingredients. First of all, the model should be well specified so that the true parameter should be determined uniquely in the limit. This is called (asymptotic) identification condition. Let

$$M(\theta) = p \lim_{n \rightarrow \infty} M_n(\theta).$$

Then, $M(\theta)$ should have a unique maximizer. As we can expect from the LLN,

$$M(\theta) = \text{Em}_i(\theta).$$

Then, when $M(\theta)$ is smooth in θ , we should have from the FOC that

$$\text{E}z_i(\theta_0) = 0.$$

Indeed, in case of MLE, it is well-known that the expected score is zero at θ_0 .

We also need to introduce a new convergence concept. So far, we have concentrated on the convergences of the sequence of random variables or random vectors. We extend them to the sequence of random functions $M_n(\theta)$ indexed by $\theta \in \Theta$. That is, M_n is the mapping

$$M_n : \Omega \rightarrow \mathcal{M} = \{m : \Theta \rightarrow \mathbb{R}\}.$$

Or, it can be viewed as

$$M_n : \Omega \times \Theta \rightarrow \mathbb{R}.$$

Now, we define the uniform convergence in probability.

Definition 3 *We say*

$$M_n(\theta) \xrightarrow{p} M(\theta) \text{ uniformly in } \theta \in \Theta,$$

if

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0.$$

A theorem that yields such a convergence is called a uniform law of large numbers (ULLN). They not only require that the LLN holds for each fixed θ but also that the function $m(x_i, \theta)$ is smooth. I leave more thorough discussion on this to advanced courses like ec484 but I present one simple version given by Newey and McFadden (Handbook of Econometrics IV, Ch. 36).

Lemma 4 *If (i) the data are i.i.d.; (ii) Θ is compact; (iii) $m(z_i, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; and (iv) there is $d(z)$ such that $|m(z, \theta)| \leq d(z)$ for all $\theta \in \Theta$ and $E(d(z)) < \infty$, then $E(m(z, \theta))$ is continuous and*

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n m(z_i, \theta) - E(m(z, \theta)) \right| \xrightarrow{p} 0.$$

The following theorem is out of the scope of this course but is presented for the sake of completeness.

Theorem 5 *Suppose that (i) M_n converges to M uniformly in Θ ; (ii)*

$$M(\theta_0) > \sup_{\theta \notin G} M(\theta)$$

for every open set G that contains θ_0 . Let $M_n(\hat{\theta}) \geq \sup_{\theta} M_n(\theta) + o_p(1)$. Then, $\hat{\theta} \xrightarrow{p} \theta_0$.

SUMMARY OF LAST LECTURE

1. In a likelihood setup, the information matrix provides a lower bound on the variance of any unbiased estimator. And the MLE achieves it asymptotically under mild regularity conditions.
2. The M-estimation means that the estimator is obtained by maximizing an objective function (or criterion function) in the form of

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^n m(w_i, \theta).$$

And the estimator may not have a closed-form expression.

3. In the M-estimation, the true parameter value θ_0 is defined as

$$\theta_0 = \underset{\theta}{\operatorname{argmax}} M(\theta) := \operatorname{E} m(w_i, \theta),$$

and thus

$$\operatorname{E} z(w_i, \theta_0) = \operatorname{E} \left[\frac{\partial}{\partial \theta} m(w_i, \theta_0) \right] = 0.$$

4. ULLN (uniform law of large numbers):

$$\sup_{\theta} \left| \frac{1}{n} \sum_{i=1}^n m(w_i, \theta) - \mathbb{E}m(w_i, \theta) \right| \xrightarrow{P} 0,$$

one of the conditions of the theorem is the continuity of $m(w, \theta)$ at each θ with probability one.

Now, I describe how we get the asymptotic normality of $\hat{\theta}$ with iid observations assuming consistency of $\hat{\theta}$ (Oftentime, the proof of consistency is more demanding in general discussion). We start from the FOC and its expansion using the mean value theorem given by

$$\begin{aligned} 0 &= z_n(\hat{\theta}) \\ &= z_n(\theta_0) + H_n(\tilde{\theta})(\hat{\theta} - \theta_0). \end{aligned}$$

We need that

1. $Ez_i(\theta_0) = 0$ and $Ez_i^2(\theta_0) < \infty$;
2. $H_n(\theta) \xrightarrow{p} H(\theta)$ uniformly and H is nonsingular and continuous at $\theta = \theta_0$. The condition (ii) ensures the convergence of $H_n(\tilde{\theta})$ to $H(\theta_0)$ in probability since

$$|H_n(\theta_n) - H(\theta_0)| \leq \sup_{\theta} |H_n(\theta) - H(\theta)| + |H(\theta_n) - H(\theta_0)|,$$

where the first term goes to zero in probability due to the uniform convergence and so does the second term due to the continuity.

Then,

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) &= H_n(\tilde{\theta})^{-1} \sqrt{n}z_n(\theta_0) \\ &= H(\theta_0)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i(\theta_0) + o_p(1) \\ &\xrightarrow{d} \mathcal{N}\left(0, H^{-1} \mathbf{E}(z_i z_i') H^{-1'}\right).\end{aligned}$$

Remarks

1. If $\hat{\theta}$ were an MLE, the asymptotic variance is the inverse of the information matrix, thus achieving the Cramer-Rao bound.

2. In case of nonlinear least squares, $z_i = 2\varepsilon_i \frac{\partial}{\partial \theta} f(x_i, \theta_0)$ and $H = 2\mathbf{E}\left(\frac{\partial}{\partial \theta} f(x_i, \theta_0) \frac{\partial}{\partial \theta'} f(x_i, \theta_0)\right)$. Thus, it is like a linear regression of y_i on $\frac{\partial}{\partial \theta} f(x_i, \theta)$. If f were linear, then

$$h_i(\theta) = x_i x_i' \text{ and } z_i(\theta) = x_i \varepsilon_i.$$

Thus, the linear regression is embedded in this general discussion.

3. If the error is conditionally homoskedastic, then, $\mathbf{E}(z_i z_i') = \sigma^2 H$, and thus

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2 H^{-1}).$$

4. In case of GMM, the FOC yields that

$$\begin{aligned} 0 &= z_n(\hat{\theta})' W_n \bar{m}_n(\hat{\theta}) \\ &= z_n(\hat{\theta})' W_n \bar{m}_n(\theta_0) + z_n(\hat{\theta})' W_n z_n(\tilde{\theta}) (\hat{\theta} - \theta_0), \end{aligned}$$

for which $E m_i(\theta_0) = 0$.

- **Numerical minimization:** (i) grid search; (ii) gradient methods; (iii) random search such as *simulated annealing*.

1. Gradient method: recall that

$$\begin{aligned} 0 &= z_n(\hat{\theta}) \\ &= z_n(\theta) + H_n(\tilde{\theta})(\hat{\theta} - \theta). \end{aligned}$$

This suggest iteration based on

$$\theta_{i+1} = \theta_i - H_n(\theta_i)^{-1} z_n(\theta_i),$$

with some initial value θ_0 . This is the *Newton-Raphson* method.

– Drawbacks: (i) rank condition for H_n ; (ii) step size

2. *Gauss-Newton* method: in case of MLE, the hessian $H_n(\theta)$ can be replaced by

$$\frac{1}{n} \sum_{i=1}^n z_i(\theta) z_i(\theta)',$$

and thus we do not have to compute the second derivative of M_n .

SUMMARY OF LAST LECTURE

1. The M-estimator, which is defined as

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n m(w_i, \theta),$$

exhibits the asymptotic normality under certain regularity conditions, that is,

$$\sqrt{n} (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1} \Omega H^{-1}),$$

where

$$\Omega = \mathbb{E} \left[\frac{\partial}{\partial \theta} m_i(\theta_0) \frac{\partial}{\partial \theta'} m_i(\theta_0) \right] \text{ and } H = \mathbb{E} \left[\frac{\partial^2}{\partial \theta \partial \theta'} m_i(\theta_0) \right].$$

2. If the estimator is efficient, the sandwich formula for the asymptotic variance becomes simpler.
 - (a) In case of the MLE, it becomes H^{-1} .
 - (b) In case of the (non-) linear regression with conditional homoskedasticity, it becomes $\sigma^2 H^{-1}$.

1.3 Generalized Least Squares

Write the linear regression model

$$Y = X\beta + U,$$

where $E(U|X) = 0$, and let

$$D = E(UU'|X).$$

Then, the (infeasible) GLS estimator is

$$\hat{\beta}_{GLS} = (X'D^{-1}X)^{-1} (X'D^{-1}Y).$$

And its variance is

$$\text{var}(\hat{\beta}_{GLS}|X) = (X'D^{-1}X)^{-1}.$$

- Gauss-Markov Theorem (BLUE): the GLS estimator exhibits the minimum conditional variance among the class of the linear unbiased estimator, that is,

1. $\tilde{\beta} = A(X)Y$

2. $E(\tilde{\beta}|X) = E(A(X)Y) = A(X)X\beta = \beta$, i.e., $A(X)X = I$.

Proof. Then, $\text{var}(\tilde{\beta}|X) = A(X)DA(X)'$. For GLS, $A(X) = A^*(X) = (X'D^{-1}X)^{-1}X'D^{-1}$. Let $C = A(X) - A^*(X)$, then

$$\begin{aligned} A(X)DA(X)' &= (C + A^*(X))D(C + A^*(X))' \\ &= CDC' + A^*(X)DA^*(X)' \\ &\geq A^*(X)DA^*(X)', \end{aligned}$$

which proves the theorem. ■

- The orthogonality for efficient estimator holds, that is,

$$\begin{aligned}
& \text{cov} \left(\tilde{\beta} - \hat{\beta}_{GLS}, \hat{\beta}_{GLS} \right) \\
&= \text{E} \left[\left(A(X) - (X'D^{-1}X)^{-1} X'D^{-1} \right) U U' D^{-1} X (X'D^{-1}X)^{-1} \mid X \right] \\
&= (X'D^{-1}X)^{-1} - (X'D^{-1}X)^{-1} \\
&= 0.
\end{aligned}$$

- Let \hat{D} denote an estimator of D . Typically, \hat{D} is defined as a function of X and the OLS residuals. Then,

$$\hat{\beta}_{FGLS} = \left(X' \hat{D}^{-1} X \right)^{-1} \left(X' \hat{D}^{-1} Y \right).$$

However, the FGLS is not BLUE anymore since \hat{D} introduces different dependences. The efficiency discussion is now in terms of asymptotic efficiency. Also note that the estimation of D is demanding because D is a function.

We consider two special cases:

1. **Heteroskedasticity** where the data is iid and $u_i = \varepsilon_i$ s.t. $E(\varepsilon_i^2 | x_i) = \sigma_i^2$, thus

$$D = \text{diag}(\sigma_1^2, \dots, \sigma_n^2).$$

2. **Autocorrelation** where $\{x_i\}$ and $\{u_i\}$ are independent each other and $\{u_i\}$ is serially correlated. In particular, we consider AR(1) process for u_i such that

$$u_i = \rho u_{i-1} + \varepsilon_i,$$

where $\{\varepsilon_i\}$ is iid.

1. Heteroskedasticity

1. Asymptotic distribution

$$\begin{aligned}\sqrt{n} \left(\hat{\beta}_{GLS} - \beta \right) &= (X' D^{-1} X)^{-1} (X' D^{-1} Y) \\ &= \left(\frac{1}{n} \sum_{i=1}^n \frac{x_i x_i'}{\sigma_i^2} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{x_i \varepsilon_i}{\sigma_i^2} \\ &\xrightarrow{d} \mathcal{N}(0, V_0),\end{aligned}$$

where $V_0 = \left(E \frac{x_i x_i'}{\sigma_i^2} \right)^{-1}$. Recall that the asymptotic variance of OLS estimator is $V = E(x_i x_i')^{-1} E(x_i x_i' \sigma_i^2) E(x_i x_i')^{-1}$. Under homoskedasticity, $V_0 = V$. In general,

$$V_0^{-1} - V^{-1} = E \left(\frac{x_i x_i'}{\sigma_i^2} \right) - E(x_i x_i') E(x_i x_i' \sigma_i^2)^{-1} E(x_i x_i') \geq 0,$$

since this is the sum of the squares error of the regression of $\frac{x_i}{\sigma_i}$ on to $x_i \sigma_i$.² Therefore, GLS is more efficient than OLS.

²It can be viewed as the probability limit of

$$(X' D^{-1} X) - (X' X) (X' D X)^{-1} (X' X) \geq 0.$$

2. As σ_i^2 is unknown in practice, we need to estimate it, which is difficult since

$$\sigma_i^2 = \mathbb{E}(\varepsilon_i^2|x_i) = f(x_i),$$

is a function. We need to parametrize $f(x_i)$ to facilitate our estimation. We may set up a regression model

$$\begin{aligned}\varepsilon_i^2 &= f(x_i, \theta) + \xi_i \\ \mathbb{E}(\xi_i|x_i) &= 0.\end{aligned}$$

This is called *skedastic* regression. For the regression function f , often used are

$$f(x_i, \theta) = x'_{1i}\theta \quad \text{or} \quad \exp(x'_{1i}\theta),$$

where x_{1i} is a subvector of x_i or its transformations.

To implement the feasible GLS,

1. run the OLS and collect the OLS residuals $\hat{\varepsilon}_i$
2. run the skedastic regression replacing ε_i^2 with $\hat{\varepsilon}_i^2$ and obtain $\hat{\theta}$
3. run the GLS replacing σ_i^2 with

$$\hat{\sigma}_i^2 = f(x_i, \hat{\theta}). \quad (1)$$

Then, the FGLS is defined as

$$\begin{aligned} \hat{\beta}_{FGLS} &= (X' \hat{D}^{-1} X)^{-1} X' \hat{D}^{-1} Y \\ &= \beta + \left(\frac{1}{n} \sum_{i=1}^n \frac{x_i x_i'}{\hat{\sigma}_i^2} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \frac{x_i \varepsilon_i}{\hat{\sigma}_i^2} \right). \end{aligned}$$

In practice, $\hat{\sigma}_i$ should not be too close to zero so that trimming is required.

Two issues of concern are

1. the asymptotic distribution of $\hat{\theta}$ in the skedastic regression
2. the asymptotic equivalence of $\hat{\beta}_{GLS}$ and $\hat{\beta}_{FGLS}$, that is,

$$\sqrt{n} \left(\hat{\beta}_{GLS} - \hat{\beta}_{FGLS} \right) = o_p(1).$$

- It is known that the asymptotic distribution of $\hat{\theta}$ is not affected by the use of the first step residuals and that the FGLS estimator achieves the same asymptotic distribution as the infeasible GLS estimator.
- What we need for this result are

$$\frac{1}{n} \sum_{i=1}^n x_i x_i' (\hat{\sigma}_i^{-2} - \sigma_i^{-2}) = o_p(1) \quad \text{and} \quad \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \varepsilon_i (\hat{\sigma}_i^{-2} - \sigma_i^{-2}) = o_p(1),$$

1. For the first, we can apply the ULLN, that is,

$$\frac{1}{n} \sum_{i=1}^n \frac{x_i x_i'}{\sigma_i^2} = \frac{1}{n} \sum_{i=1}^n g(x_i, \theta) \xrightarrow{p} \text{E}g(x_i, \theta) \quad \text{uniformly in } \theta,$$

in conjunction with the consistency of $\hat{\theta}$, to yield

$$\frac{1}{n} \sum_{i=1}^n \frac{x_i x_i'}{\hat{\sigma}_i^2} \xrightarrow{p} \mathbb{E} \left(\frac{x_i x_i'}{\sigma_i^2} \right).$$

2. The second is more involved and can be discussed in a more advanced course.
- Even if f is misspecified, the FGLS still exhibits asymptotic normality.

SUMMARY OF LAST LECTURE

1. In the linear regression model with *heteroskedasticity*, where

$$\begin{aligned}y_i &= x_i' \beta + \varepsilon_i \\ \mathbf{E}(\varepsilon_i | x_i) &= 0 \\ \mathbf{E}(\varepsilon_i^2 | x_i) &= \sigma_i^2,\end{aligned}$$

the most efficient estimator is the (infeasible) GLS estimator $\hat{\beta}_{GLS}$, whose asymptotic distribution is given by

$$\begin{aligned}\sqrt{n}(\hat{\beta}_{GLS} - \beta) &= \left(\frac{1}{n} \sum_{i=1}^n \frac{x_i x_i'}{\sigma_i^2} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{x_i \varepsilon_i}{\sigma_i^2} \\ &\xrightarrow{d} \mathcal{N} \left(0, \left(\mathbf{E} \frac{x_i x_i'}{\sigma_i^2} \right)^{-1} \right).\end{aligned}$$

Compare this to the asymptotic variance of the OLS estimator, which is

$$(\mathbf{E} x_i x_i')^{-1} (\mathbf{E} x_i x_i' \sigma_i^2) (\mathbf{E} x_i x_i')^{-1}.$$

2. In the feasible GLS,

(a) run the OLS of y_i on x_i and obtain the OLS residuals $\hat{\varepsilon}_i = y_i - x_i' \hat{\beta}$,
 $i = 1, \dots, n$

(b) run the *skedastic regression* with the OLS residuals to estimate σ_i ,
that is,

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n (\hat{\varepsilon}_i^2 - f(x_i, \theta))^2$$

(c) let $\hat{\sigma}_i^2 = f(x_i, \hat{\theta})$ and compute the feasible GLS estimator by

$$\hat{\beta}_{FGLS} = \left(\sum_{i=1}^n \frac{x_i x_i'}{\hat{\sigma}_i^2} \right)^{-1} \sum_{i=1}^n \frac{x_i y_i}{\hat{\sigma}_i^2}.$$

This estimator has the same asymptotic distribution as the (infeasible) GLS estimator.

Consistency of Skedastic Regression

$$\begin{aligned}\varepsilon_i^2 &= f(x_i, \theta) + \xi_i \\ \mathbb{E}(\xi_i | x_i) &= 0,\end{aligned}$$

assuming $f(x_i, \theta) = z_i \theta$ where $z_i = x_{1i}^2$. Then, as we do not observe ε_i we replace it with the OLS residual to get

$$\hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^n z_i^2 \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n z_i \hat{\varepsilon}_i^2 \right)$$

and recall that $\hat{\varepsilon}_i = y_i - x_i' \hat{\beta} = \varepsilon_i - x_i' (\hat{\beta} - \beta_0)$. The plug-in yields that

$$\begin{aligned}\hat{\theta} &= \left(\frac{1}{n} \sum_{i=1}^n z_i^2 \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n z_i \varepsilon_i^2 - 2 \frac{1}{n} \sum_{i=1}^n z_i \varepsilon_i x_i' (\hat{\beta} - \beta_0) + \frac{1}{n} \sum_{i=1}^n z_i \left(x_i' (\hat{\beta} - \beta_0) \right)^2 \right) \\ &= (\mathbb{E} z_i^2)^{-1} (\mathbb{E} z_i \varepsilon_i^2 + O_p(1) o_p(1) + O_p(1) o_p(1)) \\ &\rightarrow \theta.\end{aligned}$$

due to the LLN.

2. Autocorrelation ($u_i = \rho u_{i-1} + \varepsilon_i$)

Here, we study

1. structure of D ;
2. implementation of the FGLS by *Cochrane-Orcutt* transformation.

1. By backsubstitution,

$$u_i = \varepsilon_i + \rho\varepsilon_{i-1} + \rho^2\varepsilon_{i-2} + \cdots .$$

Thus, $E(u_i) = 0$, and

$$E(u_i^2) = \frac{\sigma^2}{1 - \rho^2}, \quad E(u_i u_{i-1}) = \frac{\rho}{1 - \rho^2} \sigma^2, \quad E(u_i u_{i-2}) = \frac{\rho^2}{1 - \rho^2} \sigma^2, \dots,$$

where $\sigma^2 = E(\varepsilon_i^2)$. Then,

$$D = \frac{\sigma^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ & 1 & \cdots & \rho^{n-2} \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix} .$$

2. We need to estimate ρ for the FGLS. σ^2 is not relevant here by cancellation. As

$$u_i = y_i - x_i' \beta,$$

we can write by substitution that

$$y_i = \rho y_{i-1} + x_i' \beta - x_{i-1}' \gamma + \varepsilon_i, \quad (2)$$

where $\gamma = \beta\rho$. This yields a consistent estimator of ρ . (This will be proven later)³

³Another approach is to run

the regression of \hat{u}_i on \hat{u}_{i-1} ,

where \hat{u}_i is the OLS residual of y_i on x_i . This is in the same spirit of the skedastic regression.

3. Suppose that the OLSE of the regression (2) yields a consistent estimator of ρ . Then, the OLS of the regression

$$y_i - \hat{\rho}y_{i-1} = (x_i - \hat{\rho}x_{i-1})' \beta + \varepsilon_i,$$

yields an estimator of β , which is equivalent to $\hat{\beta}_{GLS}$ asymptotically,

$$\begin{aligned} & \left(\hat{\beta}_{FGLS} - \beta \right) \\ = & \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' - 2 \frac{1}{n} \sum_{i=1}^n x_i x_{i-1} \hat{\rho} + \frac{1}{n} \sum_{i=1}^n x_{i-1} x_{i-1}' \hat{\rho}^2 \right)^{-1} \\ & \left(\frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i - \hat{\rho} \frac{1}{n} \sum_{i=1}^n x_{i-1} \varepsilon_i - \hat{\rho} \frac{1}{n} \sum_{i=1}^n x_i \varepsilon_{i-1} + \hat{\rho}^2 \frac{1}{n} \sum_{i=1}^n x_{i-1} \varepsilon_{i-1} \right). \end{aligned}$$

We need a CLT for dependent data for general case. But, if $\{x_i\}$ and $\{\varepsilon_i\}$ are iid sequences and mutually independent, and $\hat{\rho} \xrightarrow{p} \rho$,

$$\left(\hat{\beta}_{GLS} - \hat{\beta}_{FGLS} \right) = o_p \left(n^{-1/2} \right).$$

SUMMARY OF LAST LECTURE

1. The algorithm for FGLS in case of *Heteroskedasticity*.
2. Consistency of the OLSE of the skedastic regression using the first step OLS residuals

$$\hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^n z_i^2 \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n z_i \hat{\varepsilon}_i^2 \right)$$

3. In case of *Autocorrelation*, ($u_i = \rho u_{i-1} + \varepsilon_i$) the matrix D is only a function of ρ up to a scale. And ρ can be estimated by the so-called Cochrane-Orcutt transformation.

Remarks:

1. Is the OLS estimator of β in the original regression consistent? As

$$\hat{\beta} - \beta = \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i u_i,$$

we need limit theorems for dependent data. Since

$$u_i = \rho u_{i-1} + \varepsilon_i = \rho^2 u_{i-2} + \rho \varepsilon_{i-1} + \varepsilon_i = \dots,$$

u_i is not independent of u_j for any $i \neq j$ and the LLN for the independent data does not apply here. But $E(x_i u_i) = E(x_i) E(u_i) = 0$ under the maintained assumption that (x_i) and (u_i) are independent each other. And assuming for simplicity that x_i is a scalar such that $E(x_i^2) = \sigma_x^2$, $E\varepsilon_i^2 = 1$ and $j = i + l$,

$$E(x_i x_j u_i u_j) = [E x_i]^2 E u_i u_j = [E x_i]^2 \frac{\rho^l}{1 - \rho^2}$$

and

$$\sum_{l=1}^{\infty} \rho^l < \infty$$

for $|\rho| < 1$. Thus, an application of the Markov inequality yields the convergence in probability.

2. Compare the GLS estimator $\hat{\beta}_{GLS}$ with the OLS estimator of β in the regression

$$y_i = \rho y_{i-1} + x_i' \beta - x_{i-1}' \gamma + \varepsilon_i,$$

which is an inefficient estimator because it does not impose the restriction that $\gamma = \beta\rho$.

2 System of Equations

1. Kronecker product and vec operator

- (a) The Kronecker product of (possibly rectangular) matrices A and B is the $pr \times qs$ matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & a_{22}B & \cdots & a_{2q}B \\ \vdots & \vdots & & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pq}B \end{bmatrix}.$$

- (b) Associative law: $(A \otimes B) \otimes C = A \otimes (B \otimes C) = A \otimes B \otimes C$.
(c) Distributive law: $(A + B) \otimes C = A \otimes C + B \otimes C$.
(d) $(A \otimes B)(C \otimes D) = AC \otimes BD$ if AC and BD are defined.
(e) $(A \otimes B)' = A' \otimes B'$.
(h) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ if $|A| \neq 0$, $|B| \neq 0$.

(j) Denoting by a_j the j -th column of A ,

$$\text{vec}(A) = \begin{bmatrix} a_1 \\ \vdots \\ a_q \end{bmatrix}.$$

(k) $\text{vec}(ABC) = (C' \otimes A)\text{vec}(B)$.

2.1 Multivariate regression

- *Notational convention:*

1. equation $i = 1, \dots, N$
2. observation in each equation (often time) $t = 1, \dots, T$

$$y_t \quad := \quad \begin{pmatrix} \vdots \\ y_{it} \\ \vdots \end{pmatrix}_{i=1, \dots, N} \quad ; \quad \varepsilon_t := \begin{pmatrix} \vdots \\ \varepsilon_{it} \\ \vdots \end{pmatrix}_{i=1, \dots, N}$$
$$y_i \quad := \quad \begin{pmatrix} \vdots \\ y_{it} \\ \vdots \end{pmatrix}_{t=1, \dots, T} \quad ; \quad \varepsilon_i := \begin{pmatrix} \vdots \\ \varepsilon_{it} \\ \vdots \end{pmatrix}_{t=1, \dots, T}$$

x_t and β_i : k -dimensional vectors.

3. The multivariate regression is a collection of N regression models, which have the **same set of regressors**. That is,

$$y_{it} = \beta_i' x_t + \varepsilon_{it}; \quad i = 1, \dots, N, \quad t = 1, \dots, T.$$

4. Matrix B stacks the regression coefficients (β'_i) s.

$$B = \begin{pmatrix} \beta'_1 \\ \vdots \\ \beta'_N \end{pmatrix} : N \times k$$

Similarly, matrices X, Y , and ϵ are constructed by stacking x'_t, y'_t , and ϵ'_t (t^{th} observations), respectively.

5. Then, the model is written

$$y_t = Bx_t + \epsilon_t : (N \times 1).$$

or

$$y_i = X\beta_i + \epsilon_i : (T \times 1).$$

6. Putting whole system in one formula, we may write

$$Y = XB' + \epsilon : (T \times N).$$

Note that each column of these matrices stands for each equation, whereas each row in B stands for each equation. Another way is

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} X & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_N \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix} : (NT \times 1),$$

which is given by stacking up each equations in sequel. (attaching to side versus attaching to bottom). These two representations can be connected by two operators, kronecker product and vec operator, which stacks columns of a matrix to generate a vector. Let

$$y = \text{vec}(Y), \quad \epsilon = \text{vec}(\epsilon), \quad \text{and} \quad \beta = \text{vec}(B'),$$

then

$$y = (I \otimes X) \beta + \epsilon.$$

since $\text{vec}(ABC) = (C' \otimes A) \text{vec}(B)$.

The model can be estimated by OLS, that is,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{t,j} \varepsilon_{jt}(\beta)^2 = \underset{\beta}{\operatorname{argmin}} \varepsilon(\beta)' \varepsilon(\beta).$$

Let $\dot{X} = I \otimes X$, then,

$$\begin{aligned} \operatorname{vec}(\hat{B}') &= \hat{\beta} = (\dot{X}' \dot{X})^{-1} \dot{X}' y \\ &= (I \otimes X' X)^{-1} (I \otimes X) y \\ &= (I \otimes (X' X)^{-1} X) y \\ &= \operatorname{vec}((X' X)^{-1} X Y I). \end{aligned}$$

Thus,

$$\hat{B}' = (X' X)^{-1} X' Y.$$

The asymptotic properties of the OLS \hat{B} ($\hat{\beta}$) follows the same as the standard OLS estimator. Writing

$$\begin{aligned}\hat{B}' &= \left(\frac{1}{T} \sum_{t=1}^T x_t x_t' \right)^{-1} \frac{1}{T} \sum_{t=1}^T x_t y_t' \\ \hat{\beta} &= \beta + \left(\frac{1}{T} \sum_{t=1}^T (I \otimes x_t x_t') \right)^{-1} \frac{1}{T} \sum_{t=1}^T (I \otimes x_t) \varepsilon_t,\end{aligned}$$

we can easily see that the LLN and CLT for iid data apply.⁴

⁴ $(I \otimes x_t) \varepsilon_t = \varepsilon_t \otimes x_t$. That is, $x_t \varepsilon_t' = x_t \mathbf{1} \varepsilon_t' = x_t \varepsilon_t' I_N$.

2.2 SURE

The multivariate regression is easily generalized to seemingly unrelated regression equations (SURE), in which the regressors need not be the same across all the equations:

$$y_i = X_i\beta_i + \varepsilon_i, \quad i = 1, \dots, N,$$

where y_i and ε_i are T -dimensional column vectors, β_i is a $k \times 1$ vector and X_i is a $T \times k$ matrix. Let X be the block-diagonal matrix whose diagonal elements are X_i s. In the traditional treatment of SURE, the followings are assumed:

$$E(\varepsilon_{it}|X) = 0, \quad E(\varepsilon_t\varepsilon_s|X) = \Omega \cdot I\{t = s\},$$

where $I\{\cdot\}$ is the indicator function. This allows for contemporaneous correlations between equations whereas serial correlation is absent. Then, the covariance between two equation errors are

$$E(\varepsilon_i\varepsilon_j'|X) = \omega_{ij}I_T,$$

where ω_{ij} stands for $(i, j)^{th}$ element of Ω .

Defining y, β , and ε as before and letting X be a block diagonal matrix whose

elements are X_1, \dots, X_N , we have

$$\begin{aligned} y &= X\beta + \varepsilon, \\ \mathbf{E}(\varepsilon|X) &= 0 \\ \mathbf{E}(\varepsilon\varepsilon'|X) &= \begin{pmatrix} \cdots & \vdots & \cdots \\ \cdots & \mathbf{E}(\varepsilon_i\varepsilon_j'|X) = \omega_{ij}I_T & \cdots \\ \cdots & \vdots & \cdots \end{pmatrix}_{i,j=1,\dots,N} = \Omega \otimes I_T. \end{aligned}$$

This setup clearly leads us to GLS

$$\hat{\beta}_{GLS} = \left(X' (\Omega \otimes I_T)^{-1} X \right)^{-1} \left(X' (\Omega \otimes I_T)^{-1} y \right).$$

For a feasible GLS, note that the model is indeed homogeneous in each equation and the source of GLS is the contemporaneous correlation across equations which does not vary over t . Therefore, we can still come up with a method of moment estimator for Ω without imposing any functional form, that is,

$$\hat{\Omega} = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_t \hat{\varepsilon}_t'$$

where $\hat{\varepsilon} = y - X\hat{\beta}$ for the OLS $\hat{\beta}$.

Also understand that $\hat{\beta}_{GLS}$ is indeed the OLS estimator $\hat{\beta}$ in case of the multivariate regression and that the asymptotic distribution of $\hat{\beta}_{FGLS}$ is easier to derive than the previous case.

2.3 Simultaneous Equation System

- More interesting model in Economics would be simultaneous equation systems: (Structural Form)

$$\begin{aligned}Ay_t + Bx_t &= u_t, \\ E(u_t|X) &= 0 \text{ and } E(u_t u_t'|X) = \Sigma.\end{aligned}$$

- (Reduce form)

$$\begin{aligned}y_t &= \Pi x_t + v_t \\ E(v_t|X) &= 0.\end{aligned}$$

The reduce form parameter Π is correlation coefficient and thus identified. Consequently, $\Omega = E(v_t v_t')$ is identified. We can easily see the relation

$$\Pi = -A^{-1}B, \text{ that is, } A\Pi + B = 0 \tag{3}$$

$$v_t = A^{-1}u_t, \text{ so that } \Omega = A^{-1}\Sigma A^{-1'}. \tag{4}$$

We say that the structural parameters are *identified* if there is unique (A, B, Σ) which satisfies this set of equations.

- Clearly, it is not the case without further restrictions since there are

$$NK + N(N + 1)/2 \quad \text{restrictions}$$

whereas there are

$$NK + N^2 + N(N + 1)/2 \quad \text{free parameters.}$$

We need at least N^2 additional restrictions to uniquely determine (A, B, Σ) .

- For instance, suppose (A, B) and Σ satisfy (3) and (4). Then, for any p.d. matrix K , (KA, KB) and $K\Sigma K'$ satisfy the restrictions. An example of identified model is the recursive system, for which A is a lower triangular matrix with diagonal elements being 1 and Σ is a diagonal matrix. One may show that the only K that satisfies the restrictions in the recursive system is the identity matrix, thus there being unique solution.
- We do not impose restrictions on the covariance matrix at the moment and focus on the regression coefficients $C = (A, B)$, that is, the equation (3). Rewrite (3) so that

$$0 = \text{vec}(A\Pi + B) = \text{vec}\left(C \begin{pmatrix} \Pi \\ I \end{pmatrix}\right) = ((\Pi', I) \otimes I) \text{vec}(C) = V\psi,$$

where $\psi = \text{vec}(C)$. Introduce additional r (linear) restrictions:

$$W\psi = w : (r \times 1).$$

Together,

$$\begin{pmatrix} V \\ W \end{pmatrix} \psi = R\psi = r = \begin{pmatrix} 0 \\ w \end{pmatrix}. \quad (5)$$

For the identification, clearly we need that

(i) $r \geq N^2$ (order condition) and

(ii) the matrix $\begin{pmatrix} V \\ W \end{pmatrix}$ is of full column rank (rank condition).

(iii) $w \neq 0$.

If $w = 0$, which is called homogeneous restriction, then $\psi = 0$ is one solution. Apart from the uniqueness of the solution, it is not consistent with our model as we need A^{-1} for instance.

As checking the rank of a matrix of dimension at least $N^2 + NK$ is quite difficult, it would be better if we can check the identification equation by equation. It proceeds similarly for the identification of each equation. Let a'_i , b'_i , and c'_i denote i^{th} rows of A , B , and C , respectively. Then,

$$a'_i \Pi + b'_i = c'_i \begin{pmatrix} \Pi \\ I \end{pmatrix} = 0,$$

is given and to identify c_i we need additional $r_i \geq N$ restrictions

$$W_i c_i = w_i.$$

The following is known:

1. The rank condition holds iff the rank of WD' is N^2 , where $D = \begin{matrix} D \\ N^2 \times N(N+K) \end{matrix} = (I_N \otimes A, I_N \otimes B)$.
2. The whole system is identified iff all the equations are identified.
3. The i^{th} equation is identified iff

$$rank(W_i C') = N.$$

4. If W_i consists of one normalization and $(r_i - 1)$ exclusion restrictions only, then the i^{th} equation is identified iff

$$rank(C^*) = N - 1,$$

where C^* is a $(N - 1) \times (r_i - 1)$ matrix which consists of columns of C where the exclusion restrictions are imposed and whose i^{th} elements (the zeros) are dropped.

SUMMARY OF LAST LECTURE

1. We estimate the SURE by the GLS for which the “ D ” matrix is given by $\Omega \otimes I_T$.
2. In simultaneous equations system $Ay_t + Bx_t = u_t$,
 - (a) Identification is first, for which we need to establish uniqueness of the mapping between reduced form parameter (Π, Ω) and the structural form parameter (C, Σ) by imposing at least N^2 further restrictions.
 - (b) Focusing on Π and C , (*i.e.*, $A\Pi + B = 0$) and linear restrictions,

$$\begin{pmatrix} V \\ W \end{pmatrix} \psi = R\psi = r = \begin{pmatrix} 0 \\ w \end{pmatrix},$$

we obtain the rank condition for identification.

- (c) Important results regarding the rank condition are:

- i. The rank condition holds iff the rank of WD' is N^2 , where $\begin{matrix} D \\ N^2 \times N(N+K) \end{matrix} = (I_N \otimes A, I_N \otimes B)$.
- ii. The whole system is identified iff all the equations are identified.

iii. The i^{th} equation is identified iff

$$rank(W_i C') = N.$$

iv. If W_i consists of one normalization and $(r_i - 1)$ exclusion restrictions only, then the i^{th} equation is identified iff

$$rank(C^*) = N - 1,$$

where C^* is a $(N - 1) \times (r_i - 1)$ matrix which consists of columns of C where the exclusion restrictions are imposed and whose i^{th} elements (the zeros) are dropped.

While we do not provide the proof of all these claims, claim 3 can be shown more easily. We decompose the restriction matrix for the equation i :

$$R_i = \begin{pmatrix} \Pi' & I \\ W_{i1} & W_{i2} \end{pmatrix} = \begin{pmatrix} 0 & I \\ \Delta_1 & W_{i2} \end{pmatrix} \begin{pmatrix} \Delta_2 & 0 \\ \Pi' & I \end{pmatrix} = R_{i1}R_{i2}.$$

Then, it should be the case that

$$\begin{aligned} \Delta_1\Delta_2 &= W_{i1} - W_{i2}\Pi', \\ &= W_i \begin{pmatrix} A'A^{-1'} \\ B'A^{-1'} \end{pmatrix} \\ &= W_i \begin{pmatrix} A' \\ B' \end{pmatrix} A^{-1'}. \end{aligned}$$

Thus we conclude that

$$\Delta_1 = W_i \begin{pmatrix} A' \\ B' \end{pmatrix}, \quad \text{and } \Delta_2 = A^{-1'}.$$

Then, R_2 is of full column rank and thus R is of full column rank iff so is $\Delta_1 = W_iC'$.

Estimation of Simultaneous Equations System

1. First, we can obviously estimate Π by OLS and recover our structural parameters C using the identification conditions. That is, for a just-identified model

$$\hat{\psi} = \begin{pmatrix} \hat{V} \\ W \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ w \end{pmatrix},$$

whereas this is not possible when the model is over-identified ($r > N^2$). This estimator is called indirect least squares (ILS) estimator. The consistency of ILS $\hat{\psi}$ follows directly from that of $\hat{\Pi}$ using the continuous mapping theorem. Asymptotic normality follows using the Δ -method from that of $\hat{\Pi}$.

2. In case of overidentified case, multiplying $(\hat{R}'\hat{R})^{-1}\hat{R}'$ both sides of (5) yields an estimator

$$\hat{\psi} = (\hat{R}'\hat{R})^{-1}\hat{R}'r.$$

Clearly this is not the only way to get $\hat{\psi}$ and the issue of optimality here resembles that of the optimal GMM. This $\hat{\psi}$ does not exploit the cross equation correlation structure given by A and Σ . Even though $\hat{\psi}$ is consistent due to the CMT, the asymptotic variance formula is complicated. We turn to a simple procedure.

3. Assume the normalization restriction is imposed that the diagonal elements of A are 1. Then, we may estimate the system equation by equation. The i^{th} equation is given by

$$\dot{y}_{it} = \delta_i' \dot{x}_{it} + u_{it},$$

where \dot{y}_{it} , \dot{x}_{it} and δ_i are linear transformations of $(y_t', x_t')'$ and of c_i , respectively, which are given by W_i and w_i .

- (a) If all the other restrictions are exclusion restrictions, then \dot{x}_{it} and δ_i are subsets of $(y_t', x_t')'$ and c_i , which are obtained by excluding the elements with the restriction.
- (b) As we impose at least N restrictions, the dimension of \dot{x}_{it} is smaller than equal to k . And they are possibly endogenous.
- (c) Since the elements of \dot{x}_{it} are exogenous, they are valid k instruments. Then, we proceed to estimate each equation by the 2SLS, which is the optimal GMM for each equation under homoskedasticity.

4. Let \dot{X}_i be the matrix stacking \dot{x}'_{it} and

$$\ddot{X}_i = X (X'X)^{-1} X' \dot{X}_i.$$

Then, the 2SLS estimator for i^{th} equation is

$$\hat{\delta}_i = \left(\ddot{X}'_i \ddot{X}_i \right)^{-1} \ddot{X}'_i \dot{y}_i.$$

The 2SLS is equivalent to the OLS with fitted regressors \ddot{X}_i . Exploiting the cross equation correlation to achieve more efficiency, we construct SURE with equations $i = 1, \dots, N$

$$\dot{y}_i = \ddot{X}_i \delta_i + v_i,$$

and then do the FGLS as described in the previous section. Such an estimator is called 3SLSE. The asymptotic variance formula is simple as given in SURE.

5. It is known that 3SLS is identical to ILS if the system is just-identified.

3 Time Series

- Let $\{X_t\}$ be a sequence of random variables, where $t \in T$, an index set. The sequence is called a *time series* if the index set T is time.
- Primarily, we refer to a time series a sequence of *dependent*⁵ random variables. Dependency makes difficult statistical inferences based on a time series. Loosely, it implies reduced marginal information from an additional observation, and therefore the laws of large numbers and the central limit theorems may well break down.
- It, however, also has a good aspect. The dependency among observations, especially that of the future on the past, make a meaningful *prediction* feasible. Prediction is indeed one of the main themes of the time series analysis.

⁵What is *independence*?

3.1 Stationary Processes

- Roughly, stationarity implies *invariance under time shift*. In its strongest form, it implies the invariance of the distribution.
- Consider a time series $\{X_t\}$ and denote by $\Pr(\cdot, \dots, \cdot)$ the joint distribution of any finite selection of X_t 's.

- 1. If

$$\Pr(x_{t_1}, \dots, x_{t_n}) = \Pr(x_{t_1+s}, \dots, x_{t_n+s})$$

for any choice of $t_i \in T$ and s for which $t_i + s \in T$, $i = 1, \dots, n$, then we say that $\{X_t\}$ is *strictly* (or strongly) stationary.

2. We may impose a weaker form of invariance, such as the invariance of the first two moments. If a time series $\{X_t\}$ has finite first two moments satisfying

$$E(X_t) = E(X_{t+r}) \text{ and } E(X_t X_s) = E(X_{t+r} X_{s+r})$$

for all $t, s \in T$ and r such that $t + r, s + r \in T$, then it is called *weakly* (or second-order or covariance) stationary.

- It is clear that a strictly stationary time series with finite first two moments is weakly stationary. However, strict stationarity in general does not imply weak stationarity, since a strictly stationary time series may have no first or second moment.
- the *(auto)covariance function* of a weakly stationary time series $\{X_t\}$ is

$$\gamma(k) = \text{cov}(X_t, X_{t-k})$$

which is a function of only k , due to the weak stationarity. It is easy to see that $\gamma(k) = \gamma(-k)$. The *(auto)correlation function* is then given by

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

Clearly, $-1 \leq \rho(k) \leq 1$.

- A stationary time series is *ergodic* if $\gamma(k) \rightarrow 0$ as $k \rightarrow \infty$, loosely speaking.

The ergodicity enables the LLN for time series data. The stationarity and ergodicity are preserved under general transformations. The following two theorems are important in the time series analysis and stated without proof.

Theorem 6 *If $\{X_t\}$ is strictly stationary and ergodic and $Y_t = f(X_t, X_{t-1}, \dots)$ is a random variable for each t , then $\{Y_t\}$ is strictly stationary and ergodic.*

Theorem 7 (Ergodic Theorem) *If $\{X_t\}$ is strictly stationary and ergodic and $E|X_t| < \infty$, then as $n \rightarrow \infty$*

$$\frac{1}{n} \sum_{t=1}^n X_t \xrightarrow{p} E(X_1).$$

Thus, we have the same LLN as in iid case under the stationarity and ergodicity.

- We will often have to deal with a multiple time series $\{X_t\}$, where

$$X_t = (X_{1t}, \dots, X_{mt})'$$

All the concepts introduced above readily extend to this case of vector processes with some obvious modifications.

- In particular, the covariance function of a vector process $\{X_t\}$ whose mean is zero is given by

$$\Gamma(k) = E X_t X_{t-k}'$$

In contrast to the scalar case, it is anti-symmetric, i.e.,

$$\Gamma(-k) = \Gamma(k)'$$

as one may easily see. The correlation function is given by

$$R(k) = D^{-1/2} \Gamma(k) D^{-1/2}$$

where D is a diagonal matrix consisting of the diagonal elements $\gamma_{ii}(0)$'s of $\Gamma(0)$.

- The autocovariance functions can be estimated by the method of moments and their consistency can be shown by the ergodic theorem.

- Common models to study dynamics of time series are AR, MA (moving average), and ARMA.
- An autoregressive model of order k , $\text{AR}(k)$, is given by

$$y_t = \alpha + \rho_1 y_{t-1} + \cdots + \rho_k y_{t-k} + \varepsilon_t, \quad (6)$$

$$\text{E}(\varepsilon_t | \mathcal{F}_{t-1}) = 0, \quad (7)$$

where \mathcal{F}_t is the information set available up to time t .

- Of course, \mathcal{F}_t contains all the past y_t, y_{t-1}, \dots and we write $x_s \in \mathcal{F}_t$ if x_s is known by time t . Thus, the conditional expectation $\text{E}(y_t | \mathcal{F}_{t-1})$ depends only on finite number of past y_t 's in this model. As (6) contains k -th order lagged terms, it is called autoregression of order k and written as $\text{AR}(k)$.

- We introduce the *lag operator* L such that

$$Ly_t = y_{t-1}.$$

Then, $L^2 y_t = Ly_{t-1} = y_{t-2}$ and $L^{-1} y_t = y_{t+1}$. In general, $L^k y_t = y_{t-k}$. Also define a polynomial in the lag operator

$$\rho(L) = 1 - \rho_1 L - \dots - \rho_k L^k.$$

In particular, this is called the *autoregressive polynomial* of y_t as it defines an autoregressive model for y_t , that is, (6) can be written as

$$\rho(L) y_t = \alpha + \varepsilon_t.$$

- The lag polynomial is treated as an ordinary polynomial. As a k -th order polynomial has k roots, $\rho(z)$ has k roots, $\lambda_1, \dots, \lambda_k$ such that

$$\rho(z) = (1 - \lambda_1^{-1} z) \times \dots \times (1 - \lambda_k^{-1} z).$$

Thus, if $|\lambda_i| > 1$ for all i , in other words, if all roots of the polynomial $\rho(L)$ lie outside the unit circle, then it is known from elementary algebra that $\rho(L)$ is invertible and $\rho(L)^{-1}$ is an infinite order polynomial.

- A moving average model of order q , $MA(q)$ is

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\{\varepsilon_t\}$ is an iid sequence

- A process that is represented by a q -th order lag polynomial $\alpha(L)$ as

$$y_t = \alpha(L) \varepsilon_t$$

is called a moving average (MA) process of order q .

- And the autoregressive moving average process is

$$\rho(L) y_t = \alpha(L) \varepsilon_t.$$

- In the meantime, a time series satisfies (7) is called a *martingale difference sequence (MDS)*. A MDS $\{\varepsilon_t\}$ is serially uncorrelated and uncorrelated with any past information, say, y_{t-k} , for any $k > 0$, since

$$\mathbb{E}(y_{t-k}\varepsilon_t) = \mathbb{E}(\mathbb{E}(y_{t-k}\varepsilon_t|\mathcal{F}_{t-1})) = \mathbb{E}(y_{t-k}\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1})) = 0.$$

Furthermore, it yields a CLT

Theorem 8 (MDS CLT) *If $\{\varepsilon_t\}$ is a strictly stationary and ergodic MDS with $\mathbb{E}(\varepsilon_t\varepsilon_t') = \Omega < \infty$, then*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t \xrightarrow{d} \mathcal{N}(0, \Omega).$$

As an AR model is a dynamic model, the stationarity property of $\{y_t\}$ is characterized by the autoregressive polynomial. In particular,

Theorem 9 *An AR(k) process $\{y_t\}$, as defined in (6) and (7), is strictly stationary and ergodic provided*

- (i) that the sequence $\{\varepsilon_t\}$ is strictly stationary and ergodic and*
- (ii) that all the roots of the autoregressive polynomial lie outside the unit circle.*

- We do not formally prove the theorem but make some heuristic argument. As noted, the condition on the roots enables us to obtain MA representation of the series. Then, y_t is represented by a function of $\{\varepsilon_t\}$, which is strictly stationary and ergodic, and thus the stationarity and ergodicity is preserved. For example, since $(1 - \rho L)^{-1} = 1 + \rho L + \rho^2 L^2 + \dots$,

$$y_t = \rho y_{t-1} + \varepsilon_t = \varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \dots$$

Being $|\rho| < 1$ makes the infinite sum summable.

- The moments of y_t can be calculated easily.

$$E(y_t) = 0, \quad E(y_t^2) = (1 + \rho^2 + \rho^4 + \dots) E(\varepsilon_t^2),$$

and so on.

- The autocovariance functions are obtained through the relation $\Gamma(k) = \rho\Gamma(k-1)$.
- If one of the roots of the polynomial $\rho(L)$ equals one, the process $\{y_t\}$ is said to “has a unit root.” For an AR(1) process, this means that

$$\begin{aligned}y_t &= y_{t-1} + \varepsilon_t \\ &= y_0 + \sum_{s=1}^t \varepsilon_s.\end{aligned}$$

Thus, an AR process with a unit root is also called an *integrated process*. Or more precisely, it is called an integrated process of order 1 and also called $I(1)$ process compared to $I(0)$ process, which implies a stationary process.

Estimation of AR(k)

- The autoregressive model can be estimated by OLS and it can be shown that the OLS estimator is consistent and asymptotically normal using the ergodic theorem and the MDS CLT provided that $E(\varepsilon_t^4) < \infty$.

- The ARMA model is commonly estimated by MLE assuming the normality of the innovations $\{\varepsilon_t\}$. The asymptotic analysis of the estimator is more demanding. It is plausible to use the sample analogue to estimate the ARMA coefficients.

Lag order selection

- In practice, the lag order k is unknown and needs to be chosen based on the data. This is a model selection issue.
- Most commonly used method is to use the information criteria such as AIC or BIC, that is, choose the model that minimizes the criterion functions, e.g.,

$$AIC(k) = \log \hat{\sigma}^2(k) + 2\frac{k}{n},$$

or

$$BIC(k) = \log \hat{\sigma}^2(k) + \log(n) \frac{k}{n},$$

where $\hat{\sigma}^2(k)$ is the residual variance estimated from an $AR(k)$. The probability that the true k , say, k^* is chosen is

$$\Pr \{AIC(k) - AIC(k^*) > 0\} = \Pr \{n(\log \hat{\sigma}^2(k) - \log \hat{\sigma}^2(k^*)) > -2(k - k^*)\}.$$

If this probability goes to 1, the selection rule is call ‘consistent’. BIC is consistent, while AIC is not.

Testing for remaining serial correlation in the error.

- We may consider a model

$$\begin{aligned}y_t &= \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + u_t \\u_t &= \rho u_{t-1} + \varepsilon_t,\end{aligned}$$

where $\{\varepsilon_t\}$ is a sequence of *iid* random variables. And you want to test for the missing serial correlation in u_t , that is, test the hypothesis

$$H_0 : \rho = 0.$$

This can be tested using *t*-test after the Cochrane-Orcutt transformation.

(Linear Process) A linear process can effectively represent any weakly stationary process owing to the Wold representation theorem.

Theorem 10 (*Wold representation theorem*) Any weakly stationary process $\{y_t\}$ can be decomposed as

$$y_t = x_t + v_t,$$

where v_t is deterministic (predetermined) and x_t is a linear process, that is,

$$x_t = \sum_{i=0}^{\infty} c_i \varepsilon_{t-i},$$

where $\varepsilon_t \sim WN(0, \sigma^2)$.

A process $\{\varepsilon_t\}$ is a *white noise* if $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, and $E\varepsilon_t\varepsilon_s = 0$ for any $t \neq s$.

Some basic results on linear processes

Lemma 11 *If $\sup_t E |\varepsilon_t| < \infty$ and*

$$\sum_{j=-\infty}^{\infty} |c_j| < \infty,$$

then, the series

$$\sum_{j=-n}^n c_j \varepsilon_j$$

converges a.s. If in addition $\sup_t E |\varepsilon_t|^2 < \infty$, the series converges in L^2 to the same limit.

Proof. 1. $\sum_{j=-n}^n |c_j \varepsilon_j|$ always converges (including ∞ as a limit) as it increases monotonically.

2. The monotone convergence theorem and finiteness of $\sup_t \mathbf{E} |\varepsilon_t| < \infty$ imply

$$\begin{aligned} \mathbf{E} \sum_{j=-\infty}^{\infty} |c_j \varepsilon_j| &= \lim_{n \rightarrow \infty} \mathbf{E} \sum_{j=-n}^n |c_j \varepsilon_j| \\ &\leq \sup_t \mathbf{E} |\varepsilon_t| \sum_{j=-\infty}^{\infty} |c_j| \\ &< \infty, \end{aligned}$$

and thus $\sum_{j=-\infty}^{\infty} |c_j \varepsilon_j| < \infty$ a.s.

3. 1 and 2 imply $\sum_{j=-\infty}^{\infty} |c_j \varepsilon_j| - \sum_{j=-n}^n |c_j \varepsilon_j|$ converges to zero a.s. and so does $\left| \sum_{j=-\infty}^{\infty} c_j \varepsilon_j - \sum_{j=-n}^n c_j \varepsilon_j \right|$ by the triangular inequality. That is, $\sum_{j=-n}^n c_j \varepsilon_j$ converges a.s.

The proof of second assertion is omitted. ■

Remark 1 *A linear process converges in L^2 under a weaker condition*

$$\sum_{j=0}^{\infty} |c_j|^2 < \infty.$$

Theorem 12 (CLT) Let $\{X_t\}$ be a linear process, i.e.,

$$X_t = c(L) \varepsilon_t = \sum_{i=0}^{\infty} c_i \varepsilon_{t-i},$$

where $\sum_{i=0}^{\infty} |c_i| < \infty$ and $\varepsilon_t \sim WN(0, \sigma^2)$ such that

$$\sup_t \mathbf{E} |\varepsilon_t|^{2+\delta} < \infty$$

for some $\delta > 0$. Then,

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n X_t \xrightarrow{d} \mathcal{N}(0, \omega^2),$$

where

$$\omega^2 = |c(1)|^2 \sigma^2 = \sigma^2 \left(\sum_{i=0}^{\infty} c_i \right)^2.$$

- The asymptotic variance ω^2 is called a *long-run variance* of X_t . It is also the value of spectral density at frequency zero.

- The LLN for the linear process is a direct corollary of the CLT, that is, under the same condition,

$$\frac{1}{n} \sum_{t=1}^n X_t \xrightarrow{p} 0.$$

- Compared to the MDS CLT, it requires higher moment condition, which is the price to pay for the serial correlation in the linear processes.

3.2 Process with Deterministic Trend

- It is often the case that a time series consists of a stochastic part x_t and a deterministic part s_t :

$$y_t = s_t + x_t.$$

For instance, s_t might be a linear time trend or seasonality dummies and x_t be an AR process, etc. Then, the statistical inference is made after filtering out the deterministic component. Here we consider the case with a linear time trend, which can be readily analysed by what we have covered so far.

- Let

$$\begin{aligned} y_t &= \tilde{\alpha} + \tilde{\mu}t + x_t \\ \rho(L)x_t &= \varepsilon_t, \end{aligned} \tag{8}$$

where $\{\varepsilon_t\}$ is an iid sequence. This can be equivalently written as

$$\rho(L)y_t = \alpha + \mu t + \varepsilon_t, \tag{9}$$

where $\rho(L)(\tilde{\alpha} + \tilde{\mu}t) = \alpha + \mu t$. If the series $\{x_t\}$ is stationary then the series $\{y_t\}$ is called trend stationary.

- The model (9) can be estimated by OLS or y_t can be regressed to $(1, t)$ to collect the residuals, which is called *detrending*, and the residuals are fit to the autoregression. The two procedures are equivalent by Frisch-Waugh-Lovell theorem. It may appear that the presence of the trend complicates the asymptotic analysis. However, to some extent, our previous discussion extends to cover this case.

For simplicity, assume $\rho(L) = 1 - \rho L$ and transform the model to evade the asymptotic collinearity of y_t and t :

$$\begin{aligned} y_t &= \alpha + \mu t + \rho y_{t-1} + \varepsilon_t \\ &= \rho(y_{t-1} - \tilde{\alpha} - \tilde{\mu}(t-1)) + \alpha + \rho\tilde{\alpha} - \rho\tilde{\mu} + (\mu + \rho\tilde{\mu})t + \varepsilon_t \\ &= \rho x_{t-1} + \alpha^* + \mu^* t + \varepsilon_t. \end{aligned}$$

The original parameters (α, μ, ρ) are a linear transformation of (α^*, μ^*, ρ) and thus the asymptotic distribution of the OLS estimator of the original parameters can be obtained from that of the transformed model as if x_t were observed. Then, the OLS estimator is

$$\begin{pmatrix} \hat{\alpha}^* - \alpha^* \\ \hat{\mu}^* - \mu^* \\ \hat{\rho} - \rho \end{pmatrix} = \begin{pmatrix} n & \sum_{t=1}^n t & \sum_{t=1}^n x_{t-1} \\ & \sum_{t=1}^n t^2 & \sum_{t=1}^n t x_{t-1} \\ & & \sum_{t=1}^n x_{t-1}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^n \varepsilon_t \\ \sum_{t=1}^n t \varepsilon_t \\ \sum_{t=1}^n x_{t-1} \varepsilon_t \end{pmatrix}.$$

From elementary algebra, $\sum_{t=1}^n t = \frac{n(n+1)}{2}$ and $\sum_{t=1}^n t^2 = \frac{n(n+1)(2n+1)}{6}$. Thus,

$$\frac{1}{n^2} \sum_{t=1}^n t \rightarrow \frac{1}{2} \quad \text{and} \quad \frac{1}{n^3} \sum_{t=1}^n t^2 \rightarrow \frac{1}{3}.$$

Or, note that

$$\frac{1}{n^3} \sum_{t=1}^n t^2 = \frac{1}{n} \sum_{t=1}^n (t/n)^2 \rightarrow \int_0^1 r^2 dr = 1/3.$$

And, it follows that, for an iid sequence $\varepsilon_t \sim (0, \sigma^2)$,

$$\text{var} \left(\sum_{t=1}^n t \varepsilon_t \right) = \sum_{t=1}^n t^2 \sigma^2 = \frac{n(n+1)(2n+1)}{6} \sigma^2.$$

Then, the Lindeberg-Feller CLT for an independent sequence yields that

$$\frac{1}{n^{3/2}} \sum_{t=1}^n t \varepsilon_t = \frac{1}{\sqrt{n}} \sum_{t=1}^n \frac{t}{n} \varepsilon_t \xrightarrow{d} N \left(0, \frac{\sigma^2}{3} \right).$$

Furthermore, for an iid sequence $\{\xi_t\}$ with nonzero mean, an LLN yields

$$\frac{1}{n^2} \sum_{t=1}^n (t \xi_t - \text{E} t \xi_t) = \frac{1}{n} \sum_{t=1}^n \left(\frac{t}{n} \xi_t - \text{E} \frac{t}{n} \xi_t \right) \xrightarrow{p} 0,$$

that is,

$$\frac{1}{n^2} \sum_{t=1}^n t \xi_t \xrightarrow{p} \frac{\text{E} \xi_t}{2}.$$

3.3 Vector Autoregressions

- We consider an r -dimensional *vector autoregressive (VAR)* process $\{Y_t\}$ generated by

$$Y_t = A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + \varepsilon_t \quad (10)$$

where $\{\varepsilon_t\}$ is MDS. The model is commonly written more compactly as

$$A(L) Y_t = \varepsilon_t$$

where L is the lag operator and

$$A(x) = I - A_1 x - \cdots - A_p x^p$$

Therefore, $A(L)$ becomes a matrix consisting of polynomials in lag operator. The model (10) includes p lags, and for this reason, is called a p -th *order* VAR and denoted by VAR(p).

- When and only when

$$\det A(x) \neq 0 \quad \text{for} \quad |x| \leq 1,$$

which is called invertibility, $\{Y_t\}$, defined by (10), is stationary and has an MA representation

$$Y_t = \Phi(L) \varepsilon_t$$

where

$$\Phi(x) = \sum_{i=0}^{\infty} \Phi_i x^i$$

with Φ_i 's satisfying

$$\sum_{i=1}^{\infty} |\Phi_i| < \infty$$

where $|\cdot|$ denotes the matrix norm given by $|M| = \max |m_{pq}|$ for a matrix $M = (m_{pq})$. The MA coefficients Φ_i 's can be obtained from the power series expansion of $A(x)^{-1}$. This MA representation will be used in the impulse response analysis.

- If $\det A(1) = 0$, then some of the eigenvalues of $A(1)$ are zero, which implies that some elements of Y_t is $I(1)$.

- (SVAR) For r -dimensional time series $\{Y_t\}$, consider the model given by

$$BY_t = B_1Y_{t-1} + \cdots + B_pY_{t-p} + \varepsilon_t \quad (11)$$

with

$$\text{var}(\varepsilon_t) = \Lambda$$

a *diagonal* matrix.

- Compare the model with the usual VAR

$$Y_t = A_1Y_{t-1} + \cdots + A_pY_{t-p} + u_t$$

with $\text{var}(u) = \Omega$. Model (11) is referred to as VAR in structural form (SF) or *structural* VAR (SVAR) and, in contrast, model (10) as VAR in reduced form (RF). The $\{u_t\}$ in RF VAR and the $\{\varepsilon_t\}$ in SF VAR are called, respectively, *reduced form errors* and *structural innovations*, which are related by

$$Bu_t = \varepsilon_t \quad (12)$$

Note the differences between SF and RF VAR's: First, VAR in SF allows for contemporaneous relationships in the components of Y_t , contrary to VAR in RF. Second, Ω for reduced form errors is unrestricted, while Λ for structural innovations is restricted to be diagonal.

- SVAR in (11) can be extended to a more general form

$$CY_t = C_1Y_{t-1} + \cdots + C_pY_{t-p} + D\varepsilon_t \quad (13)$$

where contemporaneous relationships in the components of structural errors ε_t , as well as of Y_t , are permitted. The model may simply be regarded as SVAR (11) with

$$B = D^{-1}C$$

and treated as such in our subsequent exposition.

- Triangular System (Recursive System):
 - B is a lower triangular matrix with unit diagonals.
- As for any p.d. P , $(\tilde{B}, \tilde{B}_1, \tilde{\Lambda}) = (PB, PB_1, P\Lambda P')$ has the same reduced form. Thus, we need to show that if \tilde{B} is lower triangular with unit diagonals and $\tilde{\Lambda}$ is diagonal matrix then $P = I$.
- Since B is lower triangular with unit diagonal and so is PB , P should also be lower triangular with unit diagonal. Similarly, argue that since Λ is diagonal and so is $P\Lambda P'$ and since P is lower triangular with unit diagonal, we must have $P = I$.
- **Estimation** of the recursive system is easy. The Cholesky decomposition of a given Ω produces the corresponding B directly, that is, apply the decomposition to write

$$\Omega = LL'$$

where L is known to be unique. Then we have

$$B = \Lambda^{\frac{1}{2}} L^{-1}.$$

As B has unit diagonals, $\Lambda^{1/2}$ should be the diagonal matrix whose diagonals are the inverses of those of L^{-1} . Thus, let

$$\hat{\Omega} = \frac{1}{n} \hat{u}' \hat{u} = \hat{L} \hat{L}', \quad (14)$$

where \hat{u} is the matrix stacking the OLS residuals \hat{u}_t 's. Then, $\hat{\Lambda}$ is the diagonal matrix of the inverse of the diagonals of \hat{L}^{-1} and

$$\hat{B} = \hat{\Lambda}^{\frac{1}{2}} \hat{L}^{-1}.$$

- Sims (1980)'s Recursive model: the beginning of SVAR modeling

m	money	m	$=$	ε_m
y/p	real GNP	y/p	$=$	$\beta_{21}m + \varepsilon_{y/p}$
u	unemployment	u	$=$	$\beta_{31}m + \beta_{32}y/p + \varepsilon_u$
w	wage level	w	$=$	$\beta_{41}m + \beta_{42}y/p + \beta_{43}u + \varepsilon_w$
p	price level	p	$=$	$\beta_{51}m + \beta_{52}y/p + \beta_{53}u + \beta_{54}w + \varepsilon_p$
pm	import price	pm	$=$	$\beta_{61}m + \beta_{62}y/p + \beta_{63}u + \beta_{64}w + \beta_{65}p + \varepsilon_{pm}$

- **(Impulse Response Analysis and Forecast Error Variance Decomposition)** Write Y_t in the MA representation

$$Y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i} = \sum_{i=0}^{\infty} \Pi_i \varepsilon_{t-i}^*,$$

where $\{\varepsilon_t^*\}$ are normalized $\{\varepsilon_t\}$ to have unit variances, and

$$\Pi_i = \Phi_i B^{-1} \Lambda^{\frac{1}{2}}. \quad (15)$$

Recall that u_t is the error from reduced form and ε_t the one from structural form. Then, the response in period i of the p -th variable to an impulse in the q -th structural innovation is given by ${}_i\pi_{pq}$, where $\Pi_i = ({}_i\pi_{pq})$. Note that the unit shock in a component of $\{\varepsilon_t^*\}$ is identical to one standard deviation shock in the corresponding component of $\{\varepsilon_t\}$. This is impulse response analysis.

4 Unit Root & Cointegration

- We noted in the previous section that the stationarity requires to exclude unit roots in the AR models. However, the presence of such a root is prevalent in the economic data and it distorts the standard inference procedure. Spurious regression is noted by an experiment performed by Granger and Newbold (1974). An interesting real example is given by Hendry who showed that when the money supply in England is regressed to rainfall in Mongolia the t -statistic is highly significant if we based our inference on the standard Normal distribution.
- We look at the tests that examines the presence of unit root in a time series. As the asymptotic behavior of the sample means of a process with a unit root is different from that of a stationary process, we need to introduce a more general convergence concept. And the asymptotic distribution depends on a stochastic process called the Brownian motion (BM).

Definition 4 *The univariate standard Brownian Motion (or Wiener process) W is a stochastic process with continuous sample path satisfying the following three properties:*

(i) $W(0) = 0$ a.s.

(ii) (Independent Increment) For $0 < r_1 < r_2 < \dots < 1$

$$W(r_1), W(r_2) - W(r_1), W(r_3) - W(r_2), \dots$$

are independent.

(iii) (Gaussianity) For $r < s$

$$W(s) - W(r) \sim \mathcal{N}(0, s - r)$$

We have in particular that $W(r) \sim \mathcal{N}(0, r)$. Also, $W(r)$ and $W(s)$ are jointly normal with covariance $r \wedge s$. We may generalize this to any finite selection of $W(r)$'s.

- We can view the stochastic process as a random mapping from the sample space Ω to a function space defined on the interval $[0, 1]$. In case of the Brownian motion, the space consists of continuous functions. As for a time series, it can also be viewed as the mapping from the product space $\Omega \times [0, 1]$ to \mathbb{R} . Thus, for each $r \in [0, 1]$ W is a random variable and for each $\omega \in \Omega$, W is a function, which is called sample path.

- We generalize the sample mean to the *partial sum process* on $r \in [0, 1]$, which is defined as

$$W_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nr \rfloor} \varepsilon_t,$$

for a sequence of random variables $\{\varepsilon_t\}$. Note that $W_n(1)$ is the sample mean of $\{\varepsilon_t\}$. The weak convergence of the partial sum process is signified by “ \Rightarrow ”. Conceptually, weak convergence is the same as convergence in distribution but defined in more general spaces.

Theorem 13 (Functional CLT; Invariance Principle; Donsker’s Theorem)

Let $W_n(r)$ be a partial sum process with $\{\varepsilon_t\}$, which is an iid sequence with mean zero and variance σ^2 . Then

$$W_n(r) \Rightarrow \sigma W(r).$$

- This is a generalization of the standard CLT. For each fixed r , the CLT yields that

$$\frac{1}{\sqrt{[nr]}} \sum_{t=1}^{[nr]} \varepsilon_t \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Compare this with

$$W_n(r) = \frac{\sqrt{[nr]}}{\sqrt{n}} \left(\frac{1}{\sqrt{[nr]}} \sum_{t=1}^{[nr]} \varepsilon_t \right) \xrightarrow{d} \sqrt{r} \mathcal{N}(0, \sigma^2) = \mathcal{N}(0, r\sigma^2),$$

It is difficult to fully appreciate the meaning of the *functional CLT (FCLT)* at this stage. We focus on how this theorem can be used to obtain the representation of the asymptotic distribution of sample means of AR processes with a unit root. The following theorem sheds some light on it.

Hereafter, $\{\varepsilon_t\}$ is an iid sequence with mean zero and variance σ^2 .

Proposition 14 *Let $y_t = y_{t-1} + \varepsilon_t$ and $y_0 = 0$. Then,*

$$\begin{aligned} \frac{1}{n^{1+k/2}} \sum_{t=1}^n y_{t-1}^k &\Rightarrow \sigma^k \int_0^1 W^k(r) dr \\ \frac{1}{n} \sum_{t=1}^n y_{t-1} \varepsilon_t &\Rightarrow (1/2) \sigma^2 (W(1)^2 - 1). \end{aligned}$$

Proof. Sketch) Note that

$$\frac{1}{n^{1+k/2}} \sum_{t=1}^n y_{t-1}^k = \int_0^1 \left(\frac{y_{[nr]}}{\sqrt{n}} \right)^k dr \Rightarrow \sigma^k \int_0^1 W^k(r) dr,$$

by the CMT. Furthermore, since

$$y_t^2 = y_{t-1}^2 + \varepsilon_t^2 + 2y_{t-1}\varepsilon_t,$$

we have

$$\begin{aligned}2 \frac{1}{n} \sum_{t=1}^n y_{t-1} \varepsilon_t &= \frac{1}{n} \sum_{t=1}^n (y_t^2 - y_{t-1}^2) - \frac{1}{n} \sum_{t=1}^n \varepsilon_t^2 \\ &= \frac{1}{n} y_n^2 - \frac{1}{n} y_0^2 - \frac{1}{n} \sum_{t=1}^n \varepsilon_t^2 \\ &\Rightarrow \sigma^2 (W(1)^2 - 1).\end{aligned}$$

■

- Consider an AR(1) process

$$y_t = \alpha y_{t-1} + \varepsilon_t,$$

and the OLS estimator

$$\hat{\alpha} = \alpha + \left(\sum_{t=2}^n y_{t-1}^2 \right)^{-1} \sum_{t=2}^n y_{t-1} \varepsilon_t.$$

We saw in the previous section that $\hat{\alpha}$ is consistent and asymptotically normal when $|\alpha| < 1$. If $\alpha = 1$, then $\hat{\alpha}$ is still consistent but not asymptotically normal. From the proposition and the CMT,

$$n(\hat{\alpha} - \alpha) \Rightarrow \left(\int_0^1 W^2(r) dr \right)^{-1} \left(W(1)^2 - 1 \right) \frac{1}{2}.$$

- Due to the so-called Ito's lemma, we can write

$$\left(W(1)^2 - 1 \right) \frac{1}{2} = \int_0^1 W dW.$$

- In practice, we always include the constant 1 in the regression. Then, by the FWL theorem,

$$n(\hat{\alpha} - 1) = \left(\frac{1}{n^2} \left(\sum_{t=2}^n y_{t-1}^2 - n\bar{y}^2 \right) \right) \left(\frac{1}{n} \left(\sum_{t=2}^n y_{t-1}\varepsilon_t - n\bar{y}\bar{\varepsilon} \right) \right).$$

And from the FCLT and CLT

$$\begin{aligned} \frac{\bar{y}}{\sqrt{n}} &= \frac{1}{n^{3/2}} \sum_{t=1}^n y_t \implies \int_0^1 W, \\ \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_t &\xrightarrow{d} W(1). \end{aligned}$$

Thus,

$$n(\hat{\alpha} - 1) \implies \left(\int_0^1 \bar{W}^2 \right)^{-1} \int_0^1 \bar{W} d\bar{W},$$

where $\bar{W} = W - \int_0^1 W$. This result can be rewritten as

$$\hat{\alpha} = 1 + O_p\left(\frac{1}{n}\right).$$

Thus, the estimation error diminishes faster than the standard \sqrt{n} rate, which is called super-consistency of $\hat{\alpha}$. The inclusion of the constant term in the regression changes the asymptotic distribution of $\hat{\alpha}$ in a significant way. Furthermore, the asymptotic distribution of $\hat{\alpha}$ is not normal if $\alpha = 1$. Therefore, determining if the process contains a unit root is important. We do not consider the case where $\alpha > 1$.

4.1 Unit root test

Testing for the presence of a unit root for economic time series has become a routine procedure. However, the testing is rather complex due to the factors we explain below and requires careful treatment. First, the unit root test is one-sided test as the unit root hypothesis lies at the boundary of the stationarity hypothesis. Second, the asymptotic distribution of the t -test is not normal. Third, the asymptotic distribution of the statistic changes depending on the presence of a constant and/or a linear trend in the estimation. We study two most common cases here.

- Two unit root tests in AR(1) model under iid error u_t :
We distinguish the cases by the presence of the linear time trend. The models are written in differenced forms by subtracting y_{t-1} from both sides, that is,

$$y_t = \alpha y_{t-1} + u_t \rightarrow \Delta y_t = (\alpha - 1) y_{t-1} + u_t,$$

where $\Delta = 1 - L$ is the difference operator.

1. Test 1 (with no deterministic time trend):

The null and alternative hypothesis are formulated as follows.

$$\begin{aligned} \mathcal{H}_0 & : \Delta y_t = u_t \\ \text{vs. } \mathcal{H}_1 & : \Delta y_t = \mu + \rho y_{t-1} + u_t, \text{ and } \rho < 0. \end{aligned} \quad (16)$$

Then, one estimate the alternative model (16), typically, by OLS. Let the OLS estimator $(\hat{\mu}, \hat{\rho})$. Then,

$$\begin{aligned} n\hat{\rho} & \Rightarrow \left(\int \bar{W}^2 \right)^{-1} \int \bar{W} d\bar{W} \\ t_{\rho} & = \frac{\hat{\rho}}{\sqrt{\hat{\sigma}^2 (\sum_{t=1}^n \bar{y}_{t-1}^2)^{-1}}} \Rightarrow \left(\int \bar{W}^2 \right)^{-1/2} \int \bar{W} d\bar{W}, \end{aligned}$$

where $\bar{W} = W - \int_0^1 W$, $\bar{y}_t = y_t - \frac{1}{n} \sum_{t=1}^n y_t$ and $\hat{\sigma}^2 = n^{-1} \sum_{t=1}^n \hat{u}_t^2$. The consistency of $\hat{\sigma}^2$ can be derived in a straightforward way. As this test is one-sided, you reject the null if the sample statistic is smaller than, say, 5 percentile of the limit distribution, which is often called the Dickey-Fuller distribution. This is skewed to the left and has a negative mean. The estimator $\hat{\sigma}^2$ is the standard homoskedastic

variance estimator. It can be shown that even when the error is heteroskedastic the limit distribution is valid. Thus, the unit root test is robust to the heteroskedasticity.

Strictly speaking we need to test the joint hypothesis that $\mu = 0$ and $\rho = 0$ but typically we do as above treating $\mu = 0$ as an auxiliary assumption under the null.

2. Test 2 (with a deterministic time trend):

$$\begin{aligned} \mathcal{H}_0 & : \Delta y_t = \delta + u_t \\ \text{vs. } \mathcal{H}_1 & : \Delta y_t = \mu_0 + \mu_1 t + \rho y_{t-1} + u_t \text{ and } \rho < 0. \end{aligned}$$

In this case, y_t has an deterministic time trend under both hypotheses. Then, we estimate the alternative model and construct the standard t -statistic for testing the null of $\rho = 0$. It can be shown that

$$t_\rho \Rightarrow \left(\int \tilde{W}^2 \right)^{-1/2} \int \tilde{W} d\tilde{W},$$

where $\tilde{W} = W(r) - a_0 - b_0 r$ and $(a_0, b_0) = \arg \min_{(a,b)} \int (W(r) - a - br)^2 dr$. NB. y_t is collinear to t asymptotically as $y_t = \delta t + \xi_t$ under the \mathcal{H}_0 , where $\xi_t = y_0 + \sum_{s=1}^t u_s$. But, the asymptotic distribution can still be derived for ρ after some transformation and new parametrization. See Hamilton (1994, p. 498).

- AR(p) model

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t,$$

is transformed to

$$\Delta y_t = \rho_0 y_{t-1} + \rho_1 \Delta y_{t-1} + \cdots + \rho_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

where $\alpha_1 = 1 + \rho_0 + \rho_1, \alpha_2 = -\rho_1 + \rho_2, \dots$ By solving it for ρ s,

$$\rho_0 = -1 + \alpha_1 + \cdots + \alpha_p.$$

In terms of the lag polynomial $\alpha(z) = 1 - \alpha_1 z - \alpha_2 z^2 - \cdots$,

$$\rho_0 = -\alpha(1).$$

Thus, we can test for the presence of a unit root using ρ_0 , i.e.,

$$\mathcal{H}_0 : \rho_0 = 0 \quad vs. \quad \mathcal{H}_1 : \rho_0 < 0.$$

Implicit here is an assumption that the lag polynomial $\alpha(z)$ contains at most one unit root and the others lie outside unit circle. The FCLT can be generalized to cover this case. In particular, see the subsequent discussion.

- Linear Processes:

Consider a linear process $\{u_t\}$ such that

$$u_t = \phi(L) \varepsilon_t = \sum_{s=0}^{\infty} \phi_s \varepsilon_{t-s},$$

where $\sum_{s=0}^{\infty} s |\phi_s| < \infty$ and $\{\varepsilon_t\}$ is an iid sequence with mean zero and variance σ^2 . Then, we can rewrite u_t as

$$\begin{aligned} u_t &= \left(\sum_{s=0}^{\infty} \phi_s - \sum_{s=1}^{\infty} \phi_s \right) \varepsilon_t + \left(\sum_{s=1}^{\infty} \phi_s - \sum_{s=2}^{\infty} \phi_s \right) \varepsilon_{t-1} + \cdots \\ &= \sum_{s=0}^{\infty} \phi_s \varepsilon_t - \sum_{s=1}^{\infty} \phi_s (\varepsilon_t - \varepsilon_{t-1}) - \sum_{s=2}^{\infty} \phi_s (\varepsilon_{t-1} - \varepsilon_{t-2}) - \cdots \\ &= \phi(1) \varepsilon_t + \tilde{\phi}(L) (\varepsilon_t - \varepsilon_{t-1}), \end{aligned}$$

where $\tilde{\phi}_i = \sum_{s=1+i}^{\infty} \phi_s$ and $\tilde{\phi}(L) = \sum_{j=0}^{\infty} \tilde{\phi}_j L^j$. This is called Beverage-

Nelson decomposition. Then, the sum of u_t s becomes

$$\begin{aligned}\sum_{t=1}^l u_t &= \sum_{t=1}^l \left(\phi(1) \varepsilon_t + \tilde{\phi}(L) (\varepsilon_t - \varepsilon_{t-1}) \right) \\ &= \phi(1) \sum_{t=1}^l \varepsilon_t + \tilde{\phi}(L) (\varepsilon_l - \varepsilon_0).\end{aligned}$$

Therefore, thus the partial sum process of u_t

$$B_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t \Rightarrow B(r),$$

where $B(r) = \omega W(r)$ and $\omega^2 = \phi(1)^2 \sigma^2$, which is called the long-run variance. Note that $\omega^2 \neq \mathbf{E}u_t^2 = \sigma^2 \sum_{j=0}^{\infty} \phi_j^2$. Let $\gamma(j) = \mathbf{E}u_t u_{t+j}$ and note that

$$\omega^2 = \lim_{n \rightarrow \infty} \text{var} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n u_t \right) = \sum_{j=-\infty}^{\infty} \gamma(j).$$

Furthermore, letting $y_t = y_{t-1} + u_t$, we obtain

$$\begin{aligned}
 2\frac{1}{n}\sum_{t=1}^n y_{t-1}u_t &= \frac{1}{n}\sum_{t=1}^n (y_t^2 - y_{t-1}^2) - \frac{1}{n}\sum_{t=1}^n u_t^2 \\
 &= \frac{1}{n}y_n^2 - \frac{1}{n}y_0^2 - \frac{1}{n}\sum_{t=1}^n u_t^2 \\
 &\Rightarrow \left(\omega^2 W(1)^2 - \sigma_u^2\right). \tag{17}
 \end{aligned}$$

- Unit Root Tests with serially correlated errors:

$$y_t = \alpha y_{t-1} + u_t,$$

where $\{u_t\}$ is now a linear process. The major difficulty it causes in inference is that it introduces a nuisance parameter that need to be estimated. As seen in (17), the long-run variance cannot be cancelled out by normalization in the standard way. There are two commonly used methods.

1. First is to estimate it directly (Newey and West 1987) using nonpara-

metric method, e.g.,

$$\hat{\omega}^2 = \sum_{j=1}^n k\left(\frac{j}{m}\right) \hat{\gamma}(j),$$

where $k(\cdot)$ is a weight function, such as $k\left(\frac{j}{m}\right) = \left(1 - \frac{j}{m}\right)$, and $\hat{\gamma}(j) = \frac{1}{n} \sum \hat{u}_t \hat{u}_{t-j}$, where \hat{u} is the OLS residual. This sort of methods in general are called Heteroskedasticity-Autocorrelation Consistent (HAC) estimation.

2. Another approach is based on an AR approximation of the linear process (or an ARMA). That is, fit the regression

$$\Delta y_t = \mu + \rho_0 y_{t-1} + \rho_1 \Delta y_{t-1} + \cdots + \rho_p \Delta y_{t-p} + \varepsilon_t,$$

or the regression with the linear time trend depending on case, but here p also increases to infinity as $n \rightarrow \infty$. In this case, the standard t -statistic to test the hypothesis $\rho_0 = 0$ converges to the Dickey-Fuller distribution. The lag order p is often chose by AIC or BIC. This is known as the Augmented Dickey-Fuller test (ADF test).

If a regression contains a/some nonstationary variable/s, the standard inference is not valid. We may use it in a first differenced form. If we consider a VAR

model of nonstationary variables, then we may construct VAR model with first difference form. Or, cointegration should be considered.

4.2 Cointegration

The idea of cointegration was proposed by Granger (1981), and further developed by Engle and Granger (1987).

Definition 5 *The $r \times 1$ series Y_t is cointegrated if Y_t is $I(1)$ yet there exists $r \times h$ matrix β whose rank is h and $z_t = Y_t' \beta$ is $I(0)$. The h vectors in β are called the cointegrating vectors.*

If the series Y_t is not cointegrated, then $h = 0$. If $h = r$, then $Y_t = I(0)$. For $0 \leq h < r$, Y_t is $I(1)$ and cointegrated, and shares $r - h$ common stochastic trends.

In some cases, it may be believed that β is known a priori. Often, $\beta = (1, -1)$. For example, if $Y = (\log(\text{Consumption}), \log(\text{Income}))$, then $\beta = (1, -1)$ specifies that $\log(\text{Consumption}/\text{Income})$ is stationary. Thus, while the consumption and income behave randomly and are not predictable, the relation between the two is stable over time. In other words, they share a common stochastic trend, which can be eliminated by differencing the two series. The mean of the log difference is understood as *long-run equilibrium* and the deviation from the mean as the deviation from the equilibrium, or *equilibrium error*. It is also called the *error-correction* term. Another example is the term structure of interest rates. Here we plot the federal fund rate and 10 year interest rates.

In other cases, β may not be known and need to be estimated. A typical estimation method is based on the following representation of the cointegrating system.

Granger Representation Let $\{Y_t\}$ be an r -dimensional VAR(p) process. More explicitly, we write

$$\Phi(L)Y_t = \epsilon_t,$$

where $\Phi(L)$ is a matrix of lag polynomials given by

$$\Phi(z) = I - \Phi_1 z - \dots - \Phi_p z^p$$

Lemma 15 *Let*

$$\det \Phi(z) = 0$$

have m -roots at $z = 1$, where $m \leq r$, and all the other roots be outside the unit circle. Moreover, in the representation

$$\Phi(z) = -z\Phi(1) + (1-z)\Gamma(z),$$

we assume that $\text{rank } \Phi(1) = h$, where $h = r - m$, and $\Gamma(z)$ is nonsingular for all $|z| \leq 1$. Then we may represent the long run impact matrices $\Phi(1)$ as

$$\Phi(1) = \alpha\beta'$$

where β and α are $h \times r$ matrices of full column ranks.

If the rank of $\Phi(1)$ is r , then $\det(\Phi(1)) \neq 0$ which may imply that Y_t is $I(0)$. On the other hand, the rank is 0, then there is no cointegration.

Johansen's Cointegration Test Its objective is testing general cointegrating relations using LR test. That is, testing $H_0 : h = h_0$ against $H_1 : h = h_0 + h_1$ where h stands for the number of the cointegrating vector.

Write the model as

$$\Delta Y_t = \Gamma_0 + \Phi(1) Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \cdots + \Gamma_{p-1} \Delta Y_{t-p+1} + \varepsilon_t$$

where ε_t is iid $N_r(0_r, \Sigma)$. Here $\Phi(1)$ is the one in Beverage Nelson decomposition so that it includes all the informations regarding the stationarity of the series. Under the null hypothesis, we can write using the Granger Representation Theorem

$$\Phi(1) = \alpha\beta'$$

where α and β are $r \times h_0$ matrices. Note that this representation is not unique so that we need a kind of normalization which will be mentioned below. Since ε_t

follows iid Normal, the log-likelihood is

$$\mathcal{L}(\Sigma, \Gamma) = -\frac{T}{2} \log 2\pi - \frac{T}{2} \log |\Sigma| - \frac{1}{2} \sum_{t=1}^T \varepsilon_t' \Sigma^{-1} \varepsilon_t$$

First note that for a fixed β MLE is OLS under Normality and $\hat{\Sigma} = T^{-1} \sum_{t=1}^T \hat{\varepsilon}_t \hat{\varepsilon}_t'$ where $\hat{\varepsilon}_t$ is the OLS residual. Since

$$\sum_{t=1}^T \hat{\varepsilon}_t \hat{\Sigma}^{-1} \hat{\varepsilon}_t = \text{tr} \left(\sum_{t=1}^T \hat{\varepsilon}_t \left(T^{-1} \sum_{t=1}^T \hat{\varepsilon}_t \hat{\varepsilon}_t' \right)^{-1} \hat{\varepsilon}_t' \right) = rT,$$

we write the concentrated likelihood as $\mathcal{L}(\beta)$,

$$\arg \max_{\beta} \mathcal{L}(\beta) = \arg \min_{\beta} \left| \sum_{t=1}^T \hat{\varepsilon}_t(\beta) \hat{\varepsilon}_t(\beta)' \right|$$

where $\hat{\varepsilon}_t(\beta)$ is the OLS residual of ΔY_t on $(1, \Delta Y_{t-1}, \dots, \Delta Y_{t-p+1}, \beta' Y_{t-1})$. Let $\tilde{\Delta Y}_t$ and \tilde{Y}_{t-1} be the regression residuals of

$$\Delta Y_t \text{ and } Y_{t-1} \text{ on } 1, \Delta Y_{t-1}, \dots, \Delta Y_{t-p+1}, \quad (18)$$

respectively. Let

$$S_{00} = \sum_t^n \Delta \tilde{Y}_t \Delta \tilde{Y}'_t, S_{01} = \sum_t^n \Delta \tilde{Y}_t \tilde{Y}'_{t-1}, S_{11} = \sum_t^n \tilde{Y}_{t-1} \tilde{Y}'_{t-1}.$$

By FWL Theorem,

$$\begin{aligned} & \min_{\beta} \left| \sum_{t=1}^T \hat{\varepsilon}_t(\beta) \hat{\varepsilon}_t(\beta)' \right| \\ &= \min_{\beta} \left| S_{00} - S_{01} \beta (\beta' S_{11} \beta)^{-1} \beta' S_{10} \right| \\ &= |S_{00}| \min_{\beta} \left| I - S_{00}^{-1} S_{01} \beta (\beta' S_{11} \beta)^{-1} \beta' S_{10} \right|, \end{aligned} \tag{19}$$

where the last equality follows from the fact that $|AB| = |A| |B|$.

Theorem 16

$$\begin{aligned} & \min_{\beta} \left| I - (S_{00})^{-1} (S_{01} \beta) (\beta' S_{11} \beta)^{-1} (\beta' S_{10}) \right| \\ &= \prod_{i=1}^{h_0} (1 - \lambda_i) \end{aligned}$$

where $\lambda_1, \dots, \lambda_{h_0}$ are the largest h_0 eigenvalues of the sample canonical correlations between ΔY_t and Y_{t-1} after controlling the constant and $(\Delta Y_{t-1}, \dots, \Delta Y_{t-p+1})$, that is, those of $(S_{00})^{-1} (S_{01}) (S_{11})^{-1} (S_{10})$. And the MLE of β , which is normalized by $\hat{\beta}' S_{11} \hat{\beta} = I$, consists of the corresponding eigenvectors.

NB As $\sum_{t=1}^T \hat{\varepsilon}_t(\beta) \hat{\varepsilon}_t(\beta)'$ is p.s.d, $(1 - \lambda_i) \geq 0$ for all i .

Proof. If we impose the normalization assumption of β , then, since $|I_k - X'X| = |I_n - XX'|$ and the determinant of a matrix equals to the product of its eigenvalues,

$$\begin{aligned}
& \left| I - (S_{00})^{-1} (S_{01}\beta) (\beta' S_{11}\beta)^{-1} (\beta' S_{10}) \right| \\
&= \left| I_r - (S_{00})^{-1/2} (S_{01}\beta) (\beta' S_{10}) (S_{00})^{-1/2} \right| \\
&= \left| I_{h_0} - \beta' S_{10} (S_{00})^{-1} S_{01}\beta \right| \\
&= \prod_{i=1}^{h_0} (I - \gamma_i), \tag{20}
\end{aligned}$$

where γ_i 's are eigenvalues of $\beta' S_{10} (S_{00})^{-1} S_{01}\beta$. That is, γ_i s solve the following eigenvalue problem

$$\left| \beta' S_{10} S_{00}^{-1} S_{01}\beta - \gamma_i I \right| = 0. \tag{21}$$

But, using the normalization assumption that $I = \beta' S_{11} \beta$,

$$\beta' (S_{10} S_{00}^{-1} S_{01} - \gamma_i S_{11}) \beta = \beta' S_{11} [(S_{11}^{-1} S_{10} S_{00}^{-1} S_{01} - \gamma_i I_r) \beta],$$

whose rank is smaller than h_0 by (21). As β and S_{11} are full column rank, $(S_{11}^{-1} S_{10} S_{00}^{-1} S_{01} - \gamma_i I_r)$ should have a deficient rank, i.e., $|S_{11}^{-1} S_{10} S_{00}^{-1} S_{01} - \gamma_i I_r| = 0$, which in turn implies that γ_i is one of eigenvalues of the canonical correlation. Then, γ_i should be one of the h_0 largest eigenvalues of the canonical correlation to minimize (20), which can be obtained if β are the corresponding eigenvectors. If one of the columns of β is the corresponding eigenvector of γ_i , then a column of the matrix in the braces $[\]$ is zero, i.e., we can say its determinant is zero. ■

Therefore the achieved maximum likelihood is

$$\mathcal{L}(\hat{\Sigma}, \hat{\Gamma}) = -\frac{T}{2} \log 2\pi |S_{00}| - \frac{1}{2} r T - \frac{T}{2} \log \left| \prod_{i=1}^{h_0} (1 - \lambda_i) \right|$$

and the MLE of β is the matrix of the corresponding eigenvectors.

Remark 2 *If we use only h_0 informations of the sample canonical correlations, it is natural to pick out its largest eigen values to minimize the determinant in (19). Or to minimize the prediction error we use the linear combinations which show the highest correlations to $\Delta \bar{Y}$.*

Remark 3 *If the null hypothesis is correct, the rank of the sample canonical correlations should be equal to h_0 in large enough sample. Hence*

$$\mathcal{H}_0 : \lambda_{h_0+1} = \cdots = \lambda_{h_0+h_1} = 0.$$

We can directly test this finding using LR test statistic.

From the above we can write

$$\begin{aligned} LR &= 2(\mathcal{L}_1^* - \mathcal{L}_0^*) & (22) \\ &= -T \sum_{j=h_0+1}^{h_0+h_1} \log(1 - \lambda_j) \\ &= T \sum_{j=h_0+1}^{h_0+h_1} \lambda_j + o_p(1) \end{aligned}$$

where \mathcal{L}_1^* and \mathcal{L}_0^* are the attained maximum likelihood under H_1 and H_0 respectively. The last equality holds because the λ_j 's, $j = h_0 + 1, \dots, h_0 + h_1$ should be close to 0 under H_0 . Equation(22) shows why the LR statistic is sometimes called the "Maximal eigenvalue statistic" when $h_0 = 0$ and $h_1 = 1$ and the "Trace statistic" when $h_0 = 0$ and $h_1 = r$.

The critical values differ depending on whether we include a constant in the regression (18) as in the unit root testing. If no constant is included, the critical value can be obtained from the distribution of the corresponding trace or eigenvalues of the form:

$$\left(\int_0^1 W dW' \right)' \left(\int_0^1 W W' \right)^{-1} \int_0^1 W dW',$$

where W is the r -dimensional standard vector BM. If a constant is included, then we need to replace the BM with the demeaned BM. The tables can be found in *e.g.* Hamilton (1994).

5 Asymptotic Tests

We consider two types of testing: hypothesis testing and specification testing.

5.1 Hypothesis testing

We first discuss the classical hypothesis testing theory. Let a partition of the parameter space Θ be made and given by

$$\Theta = \Theta_0 \cup \Theta_1$$

The *null hypothesis* is given by

$$\mathcal{H}_0 : \theta \in \Theta_0$$

and is tested against the *alternative hypothesis*

$$\mathcal{H}_1 : \theta \in \Theta_1$$

The null hypothesis H_0 is maintained unless it is rejected in favor of the alternative hypothesis H_1 . When Θ_0 and Θ_1 are singleton sets, we say that the hypothesis is *simple*. Otherwise, they are *composite*.

The statistical hypothesis testing is usually based on a *test statistic* τ . According to the value of τ , the state space \mathcal{X} is partitioned as the disjoint union of the *critical region* C and *acceptance region* A , i.e.,

$$\mathcal{X} = C \cup A$$

If $x \in C$, then H_0 is rejected in favor of H_1 . If, on the other hand, $x \in A$, then H_0 is continued to be maintained. A ‘test’ is thus completely synonymous to a ‘critical region’. We will therefore refer to a test with its critical region.

Let $X = (X_1, \dots, X_n)'$ be a random sample, and suppose the distribution of X is given by a parametric family $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$. The *power function* $\pi(\theta)$ of the test C is defined by

$$\pi(\theta) = P_\theta(C)$$

Moreover,

$$\max_{\theta \in \Theta_0} \pi(\theta)$$

is called the *size* of the test, while the values of π at $\theta \in \Theta_1$ are called the *power* of the test.

We define

Definition 6 *The test C^* is Uniformly Most Powerful (UMP) if $\forall \theta \in \Theta_1$*

$$P_{\theta}(C^*) \geq P_{\theta}(C)$$

for any test represented by C of the same size.

When both the null and alternative hypotheses are simple, we may write $H_0 = \theta_0$ and $H_1 = \theta_1$. Since $\Theta = \{\theta_0, \theta_1\}$ in this case, \mathcal{P} consists of two distributions, which we write as

$$P_{\theta_0} = P_0 \quad \text{and} \quad P_{\theta_1} = P_1$$

for the null and alternative distributions, respectively. Clearly, $P_0(C)$ and $P_1(C)$ are the size and power of the test. Notice that $P_0(C)$ is the probability of rejecting H_0 when it is true. On the other hand, $P_1(A)$ is the probability of accepting H_0 when H_0 is false (and H_1 is true). Both of $P_0(C)$ and $P_1(A)$ are the probabilities of making errors, which we refer to as the *type I* and *type II* errors, respectively.

Assume that both the null and alternative hypotheses are simple, and the distributions P_0 and P_1 are given by the likelihood functions $p(x, \theta_0)$ and $p(x, \theta_1)$.

Lemma 17 (Neyman-Pearson) *The test which rejects \mathcal{H}_0 when*

$$\lambda(x) = \frac{p(x, \theta_1)}{p(x, \theta_0)} \geq c \quad \text{or} \quad \lambda(x) = \frac{p(x, \theta_0)}{p(x, \theta_1)} \leq c$$

for a constant c is most powerful. In words, for any size α , the likelihood ratio critical region is the best critical region.

Proof Let $C^* = \{x \mid \lambda(x) \geq c\}$, and suppose C is any test other than C^* with the same size, i.e. $P_0(C) = P_0(C^*)$. We need to show $P_1(C) \leq P_1(C^*)$. Assume without loss of generality that C and C^* are disjoint. Then

$$p(x, \theta_1) \geq c p(x, \theta_0) \quad \text{over } C^*$$

and

$$p(x, \theta_1) < c p(x, \theta_0) \quad \text{over } C$$

We thus have

$$P_1(C^*) = \int_{C^*} p(x, \theta_1) dx \geq c \int_{C^*} p(x, \theta_0) dx = c P_0(C^*)$$

and

$$P_1(C) = \int_C p(x, \theta_1) dx < c \int_C p(x, \theta_0) dx = c P_0(C).$$

The stated result then follows immediately from the fact that $c P_0(C^*) = c P_0(C)$.

■

Remark The above lemma guarantees the existence of a best test under simple null and alternative hypotheses. The criterion used in the construction of the likelihood ratio (LR) test $\lambda(x) \geq c$ indeed tells us about the form of the partition of \mathcal{X} . We can view $\lambda(x)$ as a ratio of marginal power to marginal size. Then the LR test includes only those points in \mathcal{X} that have significant enough power increase per unit of size increase.

The LR test is generalized as

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_1} p(x, \theta)}{\sup_{\theta \in \Theta_0} p(x, \theta)}$$

for composite hypotheses. The generalized LR test rejects \mathcal{H}_0 when $\tau(x) \geq c$, where $\tau(x)$ is any monotone increasing function of $\lambda(x)$ and c is given for a prescribed size. Note that the Neyman-Pearson lemma does not apply to the generalized LR test. Optimality properties of the generalized LR tests are much harder to show.

The regression model is a semi-parametric model, in which the parameter of interest β is parametric while the distribution family is nonparametric in the sense that it depends on infinite-dimensional unknowns. In this setting, the size cannot be calculated exactly but is calculated asymptotically. Here comes the “asymptotic test.”

The rejection/acceptance dichotomy is associated with the Neyman-Pearson approach to hypothesis testing. An alternative approach, associated with Fisher, is to report an asymptotic p-value. The asymptotic p-value for a statistic is constructed as follows. Let a statistic t_n converge in distribution to Z . Define the tail probability, or asymptotic p-value function

$$p(t) = \Pr\{|Z| \geq |t|\}$$

Then the asymptotic p-value of the statistic t_n is

$$p_n = p(t_n).$$

Sometimes the asymptotic p-value function is defined as

$$p(t) = \Pr\{Z \geq t\}.$$

Another helpful observation is that the p-value function has simply made a unit-free transformation of the test statistic. That is, under the null, $p_n \rightarrow^d U[0, 1]$, so the “unusualness” of the test statistic can be compared to the easy-to-understand uniform distribution, regardless of the complication of the distribution of the original test statistic.

In applications, our model is often given in a form other than likelihood function, as in the linear regression. Even in the MLE framework, other test statistics than LR statistic are employed in practice from various reasons. The t/Wald test and score (LM) test are most common and constitute the trinity of tests including the LR test.

5.1.1 Trinity of Tests

We now introduce the likelihood ratio (LR), Lagrange multiplier (LM), and Wald (W) tests, which are based on MLE. It will be shown that their limiting distributions are all chi-square, and that they are asymptotically equivalent. The LM test is also called the score test (Rao 1948) and ideally suited for specification testing. Let the MLE of θ be denoted as $\hat{\theta}$ and the score function as $s(X_i, \theta)$. Let

$$I(\theta) = E [s(X_i, \theta) s(X_i, \theta)'],$$

which is the information matrix.

For simplicity, we consider the hypothesis

$$H_0 : \theta = \theta_0$$

which is to be tested against $\theta \neq \theta_0$. Assume θ is m -dimensional.

Definition 7 *Define*

$$\begin{aligned} \text{LR} &= 2 \left(\sum_{i=1}^n \ell(X_i, \hat{\theta}) - \sum_{i=1}^n \ell(X_i, \theta_0) \right) \\ \text{W} &= \sqrt{n}(\hat{\theta} - \theta_0)' I(\hat{\theta}) \sqrt{n}(\hat{\theta} - \theta_0) \\ \text{LM} &= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n s(X_i, \theta_0) \right)' I(\theta_0)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n s(X_i, \theta_0) \right) \end{aligned}$$

The LR, W and LM statistics are based, respectively, on the ratio of restricted and unrestricted maximum likelihoods, the difference between the estimated and hypothesized values of the parameter, and the first derivative of the likelihood function at the hypothesized value of the parameter. If the hypothesized value is equal to the true value, all three must be small. We have

Theorem 18 *Under suitable regularity conditions, LR, W and LM are asymptotically equivalent, and*

$$\text{LR, W, LM} \rightarrow^d \chi_m^2$$

Proof To get the result for the Wald statistic, assume that I is continuous at

θ_0 so that $I(\hat{\theta}_n) = I(\theta_0) + o_p(1)$, and notice that

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d N(0, I(\theta_0)^{-1}).$$

Next, we have

$$\ell(x, \theta_0) = \ell(x, \hat{\theta}) + s(x, \hat{\theta})'(\theta_0 - \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})' H(x, \tilde{\theta})(\theta_0 - \hat{\theta}),$$

for $\tilde{\theta}$ between θ_0 and $\hat{\theta}$. Note that $\sum_{i=1}^n s(x, \hat{\theta}) = 0$. It follows that

$$\begin{aligned} & \sum_{i=1}^n \ell(X_i, \hat{\theta}_n) - \sum_{i=1}^n \ell(X_i, \theta_0) \\ &= -\frac{1}{2} \sqrt{n}(\hat{\theta}_n - \theta_0)' \left(\frac{1}{n} \sum_{i=1}^n H(X_i, \tilde{\theta}) \right) \sqrt{n}(\hat{\theta}_n - \theta_0), \end{aligned}$$

where $-\frac{1}{n} \sum_{i=1}^n H(X_i, \tilde{\theta}) \rightarrow_p I(\theta_0)$ by the uniform law of large numbers. Also

$$s(X_i, \theta_0) = s(x, \hat{\theta}) + H(X_i, \theta')(\theta_0 - \hat{\theta})$$

for θ' between θ_0 and $\hat{\theta}$. Then

$$\begin{aligned} \text{LM} &= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n s(x_i, \theta_0) \right)' I(\theta_0)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n s(x_i, \theta_0) \right) \\ &= \sqrt{n}(\hat{\theta}_n - \theta_0)' \left(\frac{1}{n} \sum_{i=1}^n H(X_i, \theta') \right)' I(\theta_0)^{-1} \left(\frac{1}{n} \sum_{i=1}^n H(X_i, \theta') \right) \sqrt{n}(\hat{\theta}_n - \theta_0). \end{aligned}$$

■

Remark 4 Different versions of W are possible with the replacement of $I(\hat{\theta}_n) = E_{\theta=\hat{\theta}_n} s(X_i, \theta)s(X_i, \theta)'$ by any one of the following:

- (a) $\frac{1}{n} \sum_{i=1}^n s(X_i, \hat{\theta}_n)s(X_i, \hat{\theta}_n)'$
- (b) $-\frac{1}{n} \sum_{i=1}^n H(X_i, \hat{\theta}_n)$
- (c) $-E_{\theta=\hat{\theta}_n} H(X_i, \theta)$

And, similar argument is possible for LM with $\hat{\theta}_n$ replaced with θ_0 .

Example For the linear regression model with independent normal errors, the log-likelihood is

$$\ell(Y, X; \beta, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{\sum_{i=1}^n (y_i - x_i' \beta)^2}{2\sigma^2}.$$

Then, the score for β is

$$s(y_i, x_i; \beta, \sigma^2) = \frac{x_i(y_i - x_i'\beta)}{\sigma^2} = \frac{x_i(e_i - x_i'(\beta - \beta_0))}{\sigma^2}$$

while the information matrix for β is

$$I(\beta; \sigma^2) = E\left(x_i x_i' (e_i - x_i'(\beta - \beta_0))^2\right) / \sigma^4 = E(x_i x_i') / \sigma^2 \text{ if } \beta = \beta_0.$$

It also follows from the FOC that $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - x_i'\hat{\beta})^2$ and $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - x_i'\beta_0)^2$.

We can easily see that

$$\begin{aligned} LR &= n (\log \tilde{\sigma}^2 - \log \hat{\sigma}^2) = n \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \\ LM &= \sum_{i=1}^n e_i x_i' \left(\sum_{i=1}^n x_i x_i' \right)^{-1} \tilde{\sigma}^{-2} \sum_{i=1}^n x_i e_i, \end{aligned}$$

where $\hat{\sigma}^2$ lies between $\tilde{\sigma}^2$ and $\hat{\sigma}^2$. It follows from this that $LM \leq LR \leq W$.

5.1.2 t & Wald Tests

Suppose

$$\sqrt{n} \left(\hat{\theta} - \theta_0 \right) \rightarrow^d N(0, V),$$

and we want to test the hypothesis

$$h(\theta_0) = 0,$$

where $h : R^k \rightarrow R^m$, $m < k$. Then, it follows from the delta method that

$$\sqrt{n} \left(h \left(\hat{\theta} \right) - h \left(\theta_0 \right) \right) \rightarrow^d N(0, V_h),$$

where $V_h = h_{\theta_0} V h'_{\theta_0}$ and $h_{\theta} = \partial h(\theta) / \partial \theta$. In other words,

$$\hat{V}_h^{-1/2} \sqrt{n} \left(h \left(\hat{\theta} \right) - h \left(\theta_0 \right) \right) \rightarrow^d N(0, I_m),$$

where $\hat{V}_h \rightarrow_p V_h > 0$.

If $m = 1$, we reject the null hypothesis for a large deviation from zero of the t -statistic

$$t_n = \hat{V}_h^{-1/2} \sqrt{n} \left(h \left(\hat{\theta} \right) - h \left(\theta_0 \right) \right),$$

where the critical value or the asymptotic p-value is given by the standard Normal distribution. Otherwise, we employ the Wald statistic

$$W_n = nh \left(\hat{\theta} \right)' \hat{V}_h^{-1} h \left(\hat{\theta} \right) \rightarrow^d \chi_m^2.$$

Remark 5 *In practice, there can be more than one possible choice of \hat{V}_h which is consistent.*

Remark 6 *The Wald statistic is not invariant to how to impose a nonlinear restriction due to the adopted delta method. For example, consider the hypothesis $\theta = 1$ and $\phi = \log(\theta) = 0$. The standard deviation of $\hat{\phi}$, $sd(\hat{\phi}) = \hat{\theta}^{-1} sd(\hat{\theta})$. Then*

$$t_\phi = \hat{\phi}/sd(\hat{\phi}) = \log(\hat{\theta}) \hat{\theta}/sd(\hat{\theta}) \neq (\hat{\theta} - 1)/sd(\hat{\theta}) = t_\theta.$$

Example Consider two independent samples $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$, which are generated independently and identically from distributions x and y , respectively. Let μ_x and μ_y denote the mean of x and y , respectively. The hypothesis of interest is

$$H_0 : |\mu_x| = |\mu_y|.$$

The μ_x and μ_y can be estimated by the sample means $\hat{\mu}_x = \frac{1}{n} \sum_{i=1}^n x_i$ and $\hat{\mu}_y = \frac{1}{n} \sum_{i=1}^n y_i$. Let

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad z_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}, \quad \text{and} \quad \hat{\theta} = \begin{pmatrix} \hat{\mu}_x \\ \hat{\mu}_y \end{pmatrix} = \frac{1}{n} \sum_{i=1}^n z_i.$$

Then, it follows from the CLT for iid data that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (z_i - \theta) \xrightarrow{d} \mathcal{N}(0, V),$$

where $V = E(z_i - \theta)(z_i - \theta)'$. Let $h(\theta) = |\theta_1| - |\theta_2|$ then

$$H_0 : h(\theta) = 0.$$

However, $h(\cdot)$ is not differentiable and thus the delta method is not applicable. The null hypothesis can equivalently stated as

$$H'_0 : \mu_x^2 = \mu_y^2.$$

and thus let

$$h(\theta) = \theta_1^2 - \theta_2^2,$$

which is now differentiable.

Example (F test) Take the linear regression model

$$Y = X_1\beta_1 + X_2\beta_2 + e$$

where β_1 and β_2 are k_1 and k_2 dimensional vector with $k = k_1 + k_2$. The null hypothesis is

$$H_0 : \beta_2 = 0.$$

The Wald statistic takes the form

$$W = \hat{\beta}_2' (X_2' M_1 X_2) \hat{\beta}_2 / \hat{\sigma}^2,$$

under the homoskedasticity assumption. Recall from the inversion formula of the partitioned matrix that

$$(X'X)_{2,2}^{-1} = (X_2' M_1 X_2)^{-1}.$$

It is useful to observe that

$$W = n \left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \right), \quad (23)$$

where

$$\tilde{\sigma}^2 = \tilde{e}'\tilde{e}/n, \quad \tilde{e} = Y - X_1\tilde{\beta}_1, \quad \text{and} \quad \tilde{\beta}_1 = (X_1'X_1)^{-1} X_1'Y,$$

and

$$\hat{\sigma}^2 = \hat{e}'\hat{e}/n, \quad \hat{e} = Y - X\hat{\beta}, \quad \text{and} \quad \hat{\beta} = (X'X)^{-1} X'Y.$$

This statistic (23) is a constant multiple of the typically reported as an “F-statistic” which is defined as

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2) / k_2}{\hat{\sigma}^2 / n - k} = \frac{n - k}{k_2} \frac{W}{n}.$$

While it should be emphasized that equality (23) only holds for the specific type of variance estimator that is designed under homoskedasticity, still this formula often finds good use in reading applied papers.

We now derive expression (23). Since $M_1M = M$, we have $\tilde{e} = M_1Y = M_1X_2\hat{\beta}_2 + \hat{e}$ ⁶ and taking squares both sides yields

$$\hat{\beta}_2'(X_2'M_1X_2)\hat{\beta}_2 = \tilde{e}'\tilde{e} - \hat{e}'\hat{e}.$$

⁶This result is known as FWL (Frisch-Waugh-Lovell) theorem.

5.1.3 GMM Distance statistic

If the hypothesis is non-linear, a better approach than the Wald is to directly use the GMM criterion function. This is sometimes called the GMM Distance statistic, and sometimes called a LR-like statistic (the LR is for likelihood-ratio). The idea was first put forward by Newey and West (1987).

Suppose the null hypothesis of interest is

$$H_0 : h(\theta) = 0,$$

where $h : R^k \rightarrow R^m$. Then define the estimates under the null

$$\tilde{\theta} = \arg \min_{h(\theta)=0} J(\theta)$$

compared to $\hat{\theta} = \arg \min_{\theta} J(\theta)$. The distance statistic is then defined as

$$D = J(\tilde{\theta}) - J(\hat{\theta}).$$

Theorem 19 *Under some regularity conditions,*

$$D \rightarrow^d \chi_m^2.$$

If h is non-linear, the Wald statistic can work quite poorly. In contrast, current evidence suggests that the D statistic appears to have quite good sampling properties, and is the preferred test statistic.

Newey and West (1987) suggested to use the same weight matrix A_n for both null and alternative, as this ensures that $D \geq 0$. This reasoning is not compelling, however, and some current research suggests that this restriction is not necessary for good performance of the test.

This test shares the useful feature of LR tests in that it is a natural by-product of the computation of alternative models.

5.2 Specification Testing

5.2.1 Overidentification Test in GMM

Overidentified models ($l > k$) are special in the sense that there may not exist a parameter value θ such that the moment condition

$$Eg(z_t; \theta) = 0$$

holds. Thus the model—the overidentifying restriction—is testable. Since $\frac{1}{n} \sum_t g(z_t; \theta) \rightarrow_p Eg(z_t; \theta)$, it can be used to assess whether or not the hypothesis $Eg(z_t; \theta) = 0$

is true. Define

$$J(\theta) = \left(\frac{1}{\sqrt{n}} \sum_t g(z_t; \theta) \right)' A_n \left(\frac{1}{\sqrt{n}} \sum_t g(z_t; \theta) \right),$$

where A_n converges in probability to $(Eg(z_t; \theta)g(z_t; \theta)')^{-1}$. Then,

$$J = \min_{\theta} J(\theta) = J(\hat{\theta}_{GMM}) \rightarrow^d \chi_{l-k}^2.$$

Proof. (sketch) Apply the Taylor series expansion to $\frac{1}{\sqrt{n}} \sum_t g(z_t; \hat{\theta})$ at θ , and plug in the expression for $(\hat{\theta} - \theta)$ from the first order condition. Factor out $A_n^{1/2} \frac{1}{\sqrt{n}} \sum_t g(z_t; \theta)$, the term premultiplied to which can be written in the form of projection matrix of rank $(l - k)$. Check with Q6 in problem set 1. ■

The degrees of freedom of the asymptotic distribution are the number of overidentifying restrictions. If the statistic J exceeds the chi-square critical value, we can reject the model. Based on this information alone, it is unclear what is wrong, but it is typically cause for concern. The GMM overidentification test is a very useful by-product of the GMM methodology, and it is advisable to report the statistic J whenever GMM is the estimation method. When over-identified

models are estimated by GMM, it is customary to report the J statistic as a general test of model adequacy.

5.2.2 Hausman type tests

- orthogonality of efficient estimators
- regression based test
- generated regressor issue
- e.g. for endogeneity or heteroskedasticity

An interesting observation made in Hausman (1978) is an orthogonality result of efficient estimators. This can also be noted proving the Gauss-Markov theorem, that is, for any unbiased linear estimator $\tilde{\beta} = A(X)Y$, it was shown that

$$\text{cov}(\hat{\beta}, \tilde{\beta} - \hat{\beta}) = 0,$$

or equivalently,

$$\text{cov}(\hat{\beta}, \tilde{\beta}) = \text{var}(\hat{\beta}), \quad \text{or} \quad \text{var}(\hat{\beta} - \tilde{\beta}) = \text{var}(\tilde{\beta}) - \text{var}(\hat{\beta}).$$

This result can be formally generalized in the following theorem.

Theorem 20 (Orthogonality of efficient estimators) *Let $\hat{\theta}$ and $\tilde{\theta}$ be unbiased estimators of real parameter vector θ_0 with finite variances. Then, $\hat{\theta}$ is efficient relative to the set of unbiased estimators $\hat{\theta} + A(\tilde{\theta} - \hat{\theta})$ for any real A iff*

$$\text{cov}(\tilde{\theta} - \hat{\theta}, \hat{\theta}) = 0.$$

Note that the set of estimators is the collection of all the linear combinations of $\hat{\theta}$ and $\tilde{\theta}$.

Proof. Let $\dot{\theta}$ denote such an estimator, i.e., $\dot{\theta} = \hat{\theta} + A(\tilde{\theta} - \hat{\theta})$ for some A . First, assume that $\text{cov}(\tilde{\theta} - \hat{\theta}, \hat{\theta}) = 0$. Then, it follows that

$$\text{cov}(\dot{\theta} - \hat{\theta}, \hat{\theta}) = \text{cov}(A(\tilde{\theta} - \hat{\theta}), \hat{\theta}) = 0.$$

Thus,

$$\text{var}(\dot{\theta}) = \text{var}(\dot{\theta} - \hat{\theta} + \hat{\theta}) = \text{var}(\dot{\theta} - \hat{\theta}) + \text{var}(\hat{\theta}) \geq \text{var}(\hat{\theta}).$$

Now assume that $\hat{\theta}$ is efficient among $\dot{\theta}$. Then,

$$\begin{aligned} 0 &= \underset{A}{\operatorname{argmin}} \operatorname{var} \left(\hat{\theta} + A \left(\tilde{\theta} - \hat{\theta} \right) \right) \\ &= \underset{A}{\operatorname{argmin}} \left[\operatorname{var} \left(\hat{\theta} \right) + A \operatorname{var} \left(\tilde{\theta} - \hat{\theta} \right) A' + 2A \operatorname{cov} \left(\hat{\theta}, \tilde{\theta} - \hat{\theta} \right) \right]. \end{aligned}$$

Furthermore, it should follow from the FOC that

$$A = - \operatorname{var} \left(\tilde{\theta} - \hat{\theta} \right)^{-1} \operatorname{cov} \left(\hat{\theta}, \tilde{\theta} - \hat{\theta} \right).$$

Thus, combining the two yields that

$$- \operatorname{var} \left(\tilde{\theta} - \hat{\theta} \right)^{-1} \operatorname{cov} \left(\hat{\theta}, \tilde{\theta} - \hat{\theta} \right) = 0,$$

which is equivalent to $\operatorname{cov} \left(\hat{\theta}, \tilde{\theta} - \hat{\theta} \right) = 0$. ■

Example 1. To test a specification of the model, we may compare the efficient estimator with a consistent and robust estimator following Hausman (1978), which compared the 2SLS estimator with the OLS estimator, which is often called the *Hausman test* for endogeneity.

To be concrete, let $\hat{\beta}$ and $\tilde{\beta}$ denote the OLS and 2SLS estimators, respectively, and V and V_2 their asymptotic variances. Assume homoskedasticity so that the 2SLS estimator is the optimal GMM estimator and $V = \sigma^2 \mathbf{E}(x_i x_i')^{-1}$, and $V_2 = \sigma^2 \left(\mathbf{E}(x_i z_i') \mathbf{E}(z_i z_i')^{-1} \mathbf{E}(z_i x_i') \right)^{-1}$. Then, we can show that under the null of no endogeneity

$$\sqrt{n} \left(\hat{\beta} - \tilde{\beta} \right) \xrightarrow{d} \mathcal{N}(0, V_2 - V),$$

from the orthogonality of efficient estimators. Thus, the Wald test converges to the chi-square asymptotic distribution with consistent estimators

$$\hat{V}_2 = \hat{\sigma}^2 \left(X'Z (Z'Z)^{-1} Z'X \right)^{-1} \quad \text{and} \quad \hat{V} = \hat{\sigma}^2 (X'X)^{-1}.$$

However, if $z_i = (x'_{1i}, z'_{2i})'$ where $x_i = (x'_{1i}, x'_{2i})'$, then

$$\left(X'Z (Z'Z)^{-1} Z'X \right) = \begin{pmatrix} X'_1 X_1 & X'_1 X_2 \\ X'_2 X_1 & X'_2 P_Z X_2 \end{pmatrix}.$$

Therefore, $-\hat{V}_2^{-1} + \hat{V}^{-1}$ is not of full column rank, nor is $\hat{V}_2 - \hat{V}$. In other words, the asymptotic distribution of $\hat{\beta} - \tilde{\beta}$ degenerates and a meaningful test cannot be constructed.

The regression-based test is found useful. Write

$$y_t = x'_{1t}\beta_1 + x_{2t}\beta_2 + u_t, \quad (24)$$

and want to test whether x_{2t} is endogeneous or not. Let z_{2t} be a valid instrument for x_{2t} . Let $z_t = (x'_{1t}, z_{2t})'$ and

$$v_t = x_{2t} - z'_{2t}\pi,$$

where $\pi = (Ez_t z'_t)^{-1} E z_t x_{2t}$. Then, v_t is uncorrelated to u_t if and only if x_{2t} is. Therefore, considering the regression

$$\begin{aligned} u_t &= v_t \rho + e_t, \\ E(v_t e_t) &= 0, \end{aligned}$$

we note that $\rho = 0$ iff x_{2t} is exogeneous. Replacing u_t in (24) by the above relation yields

$$y_t = x'_{1t}\beta_1 + x_{2t}\beta_2 + v_t \rho + e_t.$$

Then, the endogeneity can be tested by testing the hypothesis $\rho = 0$ from the above regression. In practice, however, v_t is unobservable and should be replaced by $\hat{v}_t = x_{2t} - z'_{2t}\hat{\pi}$. It can be shown that the conventional t-statistic for $\rho = 0$ in the regression

$$y_t = x'_{1t}\beta_1 + x_{2t}\beta_2 + \hat{v}_t \rho + e_t, \quad (25)$$

converges in distribution to the standard normal distribution.

Remark

1. \hat{v}_t in the regression (25) is called a “*generated regressor*”, as it is a fitted value.

2. the OLS estimator for $\beta = (\beta'_1, \beta'_2)'$ in (25) is identical to the 2SLS estimator for β in (24) with the instrument z_t . (X_1 is orthogonal to \hat{v} ($= M_Z X_2$); and the projection of X_2 on \hat{v} is \hat{v} as $X_2 = P_Z X_2 + M_Z X_2$).

Example 2. Another example of Hausman type test is given by White (1980), who proposed a test for heteroskedasticity based on the observation that the null hypothesis of homoskedasticity leads to

$$E x_t x_t' e_t^2 = \sigma^2 E x_t x_t',$$

where $\sigma^2 = E(e_t^2 | x_t)$. Let $x_t = (x_{1t}, \dots, x_{kt})'$ and ψ_t be a $k(k+1)/2$ dimensional vector whose elements are $x_{it}x_{jt}, i = 1, \dots, k, j = 1, \dots, i$. Also let \hat{e}_t be the OLS residual and $\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^n \hat{e}_t^2$. Then, the test can be based on

$$\Psi_n = \frac{1}{n} \sum_{t=1}^n \psi_t (\hat{e}_t^2 - \hat{\sigma}^2),$$

which converges to zero under the null. In particular, the Wald statistic can be constructed as

$$W_n = n\Psi_n \left(\frac{1}{n} \sum_{t=1}^n \psi_t \psi_t' (\hat{e}_t^2 - \hat{\sigma}^2)^2 - \Psi_n \Psi_n' \right)^{-1} \Psi_n,$$

which can be shown to converge to $\mathcal{X}_{k(k+1)/2}^2$.

- Asymptotics with generated regressors
Let $\{(y_i, x_i, z_i')\}_{i=1}^n$ be a random sample and generated from the following model

$$\begin{aligned} y_i &= x_i \beta + \xi_i \rho + \varepsilon_i \\ z_{1i} &= z_{2i}' \gamma + \xi_i, \end{aligned}$$

where $E(\varepsilon_i | x_i, z_i) = 0$ and $E(\xi_i | z_{2i}) = 0$. Then, estimation of $\theta = (\beta, \rho)'$ is done in two steps, for which

1. estimate γ by OLS and collect residuals $\hat{\xi} = M_2 Z_1 = M_2 \xi$, where M_2 is the projection matrix onto Z_2 .

2. estimate θ by OLS of Y onto $\hat{W} = (X, \hat{\xi})$, that is,

$$\begin{aligned}\hat{\theta} &= (\hat{W}'\hat{W})^{-1} \hat{W}'Y \\ &= \theta + (\hat{W}'\hat{W})^{-1} \hat{W}' [(W - \hat{W})\theta + \varepsilon].\end{aligned}$$

Note that

$$\hat{W}'(W - \hat{W})\theta = \begin{pmatrix} X' \\ \hat{\xi}' \end{pmatrix} (0, \xi - \hat{\xi}) \begin{pmatrix} \beta \\ \rho \end{pmatrix} = \begin{pmatrix} X'(\xi - \hat{\xi})\rho \\ 0 \end{pmatrix},$$

since $\hat{\xi}'(\xi - \hat{\xi}) = \xi' M_2 (-P_2) \xi = 0$. Furthermore, we can show that

$$\begin{aligned}\frac{1}{n} \hat{W}'\hat{W} &= \frac{1}{n} W'W + o_p(1) \\ \frac{1}{\sqrt{n}} \hat{W}'\varepsilon &= \frac{1}{\sqrt{n}} W'\varepsilon + o_p(1).\end{aligned}$$

Therefore,

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta) &= \left(\frac{W'W}{n} \right)^{-1} \\ &\times \begin{pmatrix} \frac{X'\varepsilon}{\sqrt{n}} - \rho(X'Z_2)(Z_2'Z_2)^{-1} \left(\frac{Z_2'\xi}{\sqrt{n}} \right) \\ \frac{1}{\sqrt{n}}\xi'\varepsilon \end{pmatrix} + o_p(1). \end{aligned}$$

A Problems

A.1 Regression

1. Consider the classical linear regression model

$$y_t = \beta_0' x_t + u_t.$$

Show that the BLUE of $w'\beta_0$ is $w'\hat{\beta}$ for any vector w .

2. Suppose that x_t has a constant term. Show that the coefficient of multiple correlation can be written as follows:

$$R^2 = \frac{\sum_t (\hat{y}_t - \bar{y})^2}{\sum_t (y_t - \bar{y})^2} = 1 - \frac{\sum_t \hat{u}_t^2}{\sum_t (y_t - \bar{y})^2}$$

where $\hat{y}_t = \hat{\beta}' x_t$ and $\bar{y} = \frac{1}{T} \sum_t y_t$. R^2 lies in the range $0 \leq R^2 \leq 1$.

3. The dependent variable y is regressed on a constant and the $k + 1$ independent variables x_1, x_2, \dots, x_{k+1} using the observations $y_t, x_{1t}, \dots, x_{k+1t}$ ($t = 1, \dots, T$). The coefficient of multiple correlation R_1^2 is calculated from

this regression. Another regression is run, this time using only the first k independent variables x_1, x_2, \dots, x_k . The coefficient of determination R_2^2 . Show that $R_1^2 \geq R_2^2$.

4. Consider the following linear regression model

$$y_t = x_t' \beta + u_t.$$

Suppose that the parameters β satisfy the following linear restriction

$$R\beta = r.$$

- (a) Obtain the asymptotic distribution function of the restricted least squares estimator.
- (b) Show that the asymptotic covariance matrix of the OLS estimator of β exceeds that of the restricted least squares estimator by a p.s.d. matrix.
- (c) Describe how you would proceed with a nonlinear restriction

$$h(\beta) = 0.$$

5. When the regressor matrix has k columns, Theil's adjusted R^2 is

$$\bar{R}^2 = 1 - \frac{\hat{e}'\hat{e}/(n-k)}{\hat{\sigma}_y^2},$$

where $\hat{\sigma}_y^2$ is the sample variance of y . Prove that if an additional regressor x_{k+1} is added to X , then \bar{R}^2 increases if and only if $|t_{k+1}| > 1$, where $t_{k+1} = \hat{\beta}_{k+1}/s(\hat{\beta}_{k+1})$ is the t-statistic for β_{k+1} and

$$s(\hat{\beta}_{k+1}) = \sqrt{s^2 [(X'X)^{-1}]_{k+1,k+1}}.$$

6. Take the true model $Y = X_1\beta_1 + X_2\beta_2 + e$. Suppose that β_1 is estimated by regressing Y on X_1 only. Find the probability limit of this estimator. In general, is it consistent for β_1 ? If not, are the specific conditions under which this estimator will be consistent for β_1 ?
7. In the last question, assume $\beta_2 = 0$. And derive the asymptotic distribution of $\hat{\beta}_1$ from the regression of Y on X_1 only. And compare the variance to that from the regression of Y and $X = (X_1, X_2)$.
8. Prove that R^2 is the square of the sample correlation between Y and \hat{Y} .

9. Let $Y = X\beta + e$ with $E(x_i e_i) = 0$. Define the ridge regression estimator

$$\hat{\beta} = (X'X + \lambda I_k)^{-1} X'Y$$

where $\lambda > 0$ is a fixed constant. Find the probability limit of $\hat{\beta}$ as $n \rightarrow \infty$. Is $\hat{\beta}$ consistent for β ?

10. Suppose

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma),$$

where $\beta = (\beta_1, \beta_2)'$. Obtain the limit distribution of $\hat{\theta} = \hat{\beta}_1 / \hat{\beta}_2$. Suppose $n = 100$, $\Sigma = I$ and $\beta = (1, 2)'$. Obtain a confidence interval for θ .

11. Assume that f has a continuous derivative at 0. If $X_n Y_n$ converges in distribution to Y and $p \lim Y_n = 0$, then $X_n (f(Y_n) - f(0))$ converges in distribution to $f'(0)Y$.
12. Consider the following regression model:

$$y_t = \alpha + \beta' x_t + \varepsilon_t,$$

where x_t is scalar. Suppose, however, that y_t and x_t are not observed, being the observed variables

$$\begin{aligned}y_t^* &= y_t + v_t \\x_t^* &= x_t + u_t,\end{aligned}$$

where ε_t , v_t , u_t and x_t are normally distributed independent of each other, and we consider the regression model

$$y_t^* = \alpha + \beta' x_t^* + \varepsilon_t^*.$$

Then, show that the OLS estimator of α and β are inconsistent. Under what conditions will they be consistent?

13. A dummy variable takes on only the values 0 and 1. It is used for categorical data, such as an individual's gender. Let $D1$ and $D2$ be vectors of 1's and 0's, with the i th element of $D1$ equaling 1 and that of $D2$ equaling 0 if the person is a man, and the reverse if the person is a woman. Suppose that there are n_1 men and n_2 women in the sample. Consider the three

regressions

$$Y = \mu_1 + D_1\alpha_1 + D_2\alpha_2 + e \quad (26)$$

$$Y = D_1\alpha_1 + D_2\alpha_2 + e \quad (27)$$

$$Y = \mu_1 + D_1\phi + e \quad (28)$$

- (a) Can all three regressions (1), (2), and (3) be estimated by OLS? Explain if not.
- (b) Compare regressions (2) and (3). Is one more general than the other? Explain the relationship between the parameters in (2) and (3).
- (c) Compute $\iota'D_1$ and $\iota'D_2$, where ι is an $n \times 1$ vector of ones.
- (d) write equation (2) as $Y = X\alpha + e$. Consider the assumption $E(e | X) = 0$. Is there any content of to this assumption in this setting?
- (e) What is the form of $E(ee' | X)$ in this setting?
- (f) Compute $Var(\hat{\alpha}|X)$

14. Consider the regression

$$Y = X\alpha + Z\beta + u.$$

Suppose we are given an estimator $\tilde{\beta}$ that is independent of X and u , and $\sqrt{n}(\tilde{\beta} - \beta) \rightarrow^d N(0, \sigma^2)$. Define the estimator $\tilde{\alpha}$ by

$$\tilde{\alpha} = (X'X)^{-1} X' (Y - Z\tilde{\beta}).$$

Obtain the limit distribution of $\tilde{\alpha}$.

15. Suppose the model $y_t = \alpha z_t + u_t$ ($t = 1, 2, \dots, N$), where the residuals u_t are $NID(0, \sigma^2)$. Let w_t be a valid Instrumental Variable with the ordinary least squares and IV estimators of α being defined by

$$\hat{\alpha} = \frac{\sum_{t=1}^T z_t y_t}{\sum_{t=1}^T z_t^2} \quad \text{and} \quad \tilde{\alpha} = \frac{\sum_{t=1}^T w_t y_t}{\sum_{t=1}^T z_t w_t}$$

respectively. Find the sampling or asymptotic variances of $\hat{\alpha}$ and $\tilde{\alpha}$ and the covariance or asymptotic covariance between them. Finally, show that the variance of $(\tilde{\alpha} - \hat{\alpha})$ is the variance of $\tilde{\alpha}$ minus the variance of $\hat{\alpha}$.

16. Show that the asymptotic variance of the GMM estimator is minimized with the optimal weighting matrix.

17. Consider two scalar random variables y_i and x_i that are related by the simple linear model with no intercept,

$$y_i = \beta_0 x_i + u_i,$$

which is assumed to satisfy the two moment conditions

$$0 = E[u_i] = E[u_i x_i].$$

For convenience, the data are assumed to be i.i.d. draws; the first two moments of x_i are $E[x_i]=\mu$ and $E[x_i^2]=\tau^2$, the (unconditional) variance of u_i is σ^2 , and $E[u_i^2 x_i] = \delta, E[u_i^2 x_i^2] = \gamma$. (All of these parameters are, in principle, unknown.)

- (a) Find the asymptotic distributions of the method-of-moments estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ which minimize the GMM criterion $\bar{m}(\beta)' A_j \bar{m}(\beta)$ for $j = 1, 2$, where $\bar{m}(\beta)$ is the vector of sample analogues to the moment conditions and the weighting matrices for each estimator are defined as

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

(that is, the estimators which use each moment condition separately). Be sure to cite any needed identification conditions explicitly (but assume all other needed regularity conditions hold implicitly).

- (b) Derive the asymptotic variance of the optimal GMM estimator based on both of the moment conditions.
- (c) Suppose u_i and x_i just happen to have $\delta = \sigma^2\mu$ and $\gamma = \sigma^2\tau^2$ (which would follow if u_i and x_i were in fact independently distributed), but that this fact is unknown (so these extra moment conditions are not imposed in the estimation of β_0). Show that, for this special case, the optimal GMM estimator has the same asymptotic distribution as one of the estimators in part a. above.

18. Show that the White estimator mentioned in class is consistent.

19. Suppose that

$$y_t = \beta x_t + u_t$$

where u_t are iid $(0, \sigma^2)$ and where x_t and u_t are independent. Assume that

$$p \lim T^{-1} \sum x_t^2 = \sigma_x^2.$$

Consider the estimator of β given by $\tilde{\beta} = \frac{1}{d}$ where d is the coefficient from a least squares regression of X on Y . Obtain the probability limit of $\tilde{\beta}$. Show that $\tilde{\beta}$ is inconsistent and discuss how the asymptotic bias depends on σ_u^2/σ_x^2 .

20. Consider the regression, for τ between zero and one,

$$\begin{aligned} y_t &= x_t' \beta_1 + u_t, & \text{if } 1 \leq t \leq \tau n \\ y_t &= x_t' \beta_2 + u_t, & \text{if } \tau n < t \leq n \end{aligned}$$

and you want to test for the presence of structural break, that is, $\beta_1 = \beta_2$. Propose a test statistic and derive its limit distribution on which your inference will be based.

21. Show that the p-value converges in distribution to uniform $[0, 1]$. (Hint: start from $\Pr\{1 - p_n \leq u\}$ for a given $u \in R$, and show that it converges to the CDF of uniform distribution)
22. Let the random variables X_1, \dots, X_n and Y_1, \dots, Y_n be collected from two independent populations. Suppose we are interested in testing whether the means of two populations are same, say $\mu_X = \mu_Y$, where μ_X and μ_Y are means of X and Y respectively. First, propose consistent estimators for μ_X

and μ_Y and show they are consistent. Then, construct a test statistic, derive a limit distribution for it, and describe an appropriate way of inference. (It will typically involve estimation of the asymptotic variance of the statistic)

23. Consider the following regression

$$y_t = x_t' \beta + u_t.$$

Suppose we suspect that x_t may be correlated to u_t . Let z_t be a variable that is not correlated to u_t but to x_t .

- (a) Construct the IV estimator for β , denoted as $\tilde{\beta}$, and derive its limit distribution
- (b) Now we want to test whether x_t is truly endogenous or not. So, you assume that it is not endogenous. Under this assumption, the OLS estimator $\hat{\beta}$ is also consistent. In other words, $\hat{\beta} - \tilde{\beta}$ should converge to zero. Then, what is the limit distribution of $\sqrt{n}(\hat{\beta} - \tilde{\beta})$?
- (c) What is the asymptotic variance? Propose its consistent estimator under homoscedasticity. Using that, construct a test statistic for endogeneity. What is a critical value for 5% nominal size?

24. Now consider the following regression

$$y_t = x'_{1t}\beta_1 + x_{2t}\beta_2 + u_t, \quad (29)$$

and want to test whether x_{2t} is endogenous or not. Let z_{2t} be a valid instrument for x_{2t} . Let $z_t = (x'_{1t}, z_{2t})'$. Let $v_t = x_{2t} - z'_t\pi$, where $\pi = (Ez_t z'_t)^{-1} Ez_t x_{2t}$. Then, v_t is uncorrelated to u_t if and only if x_{2t} is. Therefore, in the regression

$$u_t = v_t\rho + e_t,$$

$\rho = 0$ iff x_{2t} is exogeneous. Replacing u_t by the above relation yields

$$y_t = x'_{1t}\beta_1 + x_{2t}\beta_2 + v_t\rho + e_t.$$

And the endogeneity test is testing the hypothesis $\rho = 0$ from the above regression. In practice, however, v_t is unobservable and should be replaced by $\hat{v}_t = x_{2t} - z'_t\hat{\pi}$.

(a) Show that the conventional t-statistic for $\rho = 0$ in the regression

$$y_t = x'_{1t}\beta_1 + x_{2t}\beta_2 + \hat{v}_t\rho + e_t, \quad (30)$$

converges in distribution to the standard normal distribution.

- (b) Show that the OLS estimator for $\beta = (\beta_1', \beta_2')'$ in (30) is identical to the IV estimator for β in (29) with the instrument z_t .

25. Mike proposed a GLS estimator

$$\tilde{\beta}^f = \left(X' \hat{\Sigma}^{-1} X \right)^{-1} X' \hat{\Sigma}^{-1} Y,$$

where $\hat{\Sigma} = \text{diag}(\hat{u}_1^2, \dots, \hat{u}_n^2)$. Discuss why or why not we should use this estimator.

26. Given that $E(u_i|x_i) = 0$, show that the OLS estimator is less efficient than the infeasible GLS estimator. In other words, show that

$$(X' \Sigma^{-1} X)^{-1} \leq (X' X)^{-1} X' \Sigma X (X' X)^{-1}.$$

But it is also true that the OLS estimator is more robust than the GLS estimator. That is, the infeasible GLS estimator is not consistent if the assumption that $E(u_i|x_i) = 0$ is violated but only $E x_i u_i = 0$ is satisfied, while the OLS estimator is consistent either case. Prove this claim.

27. Propose a test for heteroskedasticity using the skedastic regression. Specify your hypothesis and derive limit distribution of the test statistic which will base your inference.

28. Consider the regression model for iid sample $\{y_i, x_i\}_{i=1}^n$

$$\begin{aligned}y_i &= x_i' \beta + u_i \\ E(u_i | x_i) &= 0.\end{aligned}$$

Suppose u_i is iid mixed Normal, *i.e.*, $u_i = pe_{1i} + (1-p)e_{2i}$, where e_{1i} and e_{2i} are independent each other, independent of x_i and distributed as standard Normal. Derive the MLE of β and its asymptotic distribution. Observe that you do not need Normality to derive the limit distribution.

29. Let $X_i, i = 1, \dots, n$ be iid Bernoulli(p). Obtain the MLE of p . Is it consistent?
30. Let $X_i, i = 1, \dots, n$ be iid uniform $[\theta_1, \theta_2]$. Obtain the MLE of θ_1 and θ_2 . Is it consistent?
31. Show that

$$\theta_0 = \arg \max_{\theta} E \ell(\theta; x_i),$$

where ℓ is the log-likelihood function and the expectation is taken at the true value θ_0 .

32. Consider the following nonlinear regression model with *iid* data

$$\begin{aligned}y_i &= x_i \theta_3 \left(1 + e^{-\theta_1(x_i - \theta_2)}\right)^{-1} + \varepsilon_i \\E(\varepsilon_i | x_i) &= 0.\end{aligned}$$

Assume that there is a unique $\theta_0 = (\theta_{10}, \theta_{20}, \theta_{30})'$ that satisfies the moment condition. Describe the non-linear least squares estimator (NLLS), say, $\hat{\theta}$. Assume that $\hat{\theta}$ is consistent and obtain the asymptotic distribution of $\hat{\theta}$.

A.2 System of Equations

1. Show the followings:

$$\begin{aligned}(A \otimes B)(C \otimes D) &= AC \otimes BD \\(A \otimes B)' &= A' \otimes B' \\(A \otimes B)^{-1} &= A^{-1} \otimes B^{-1}.\end{aligned}$$

2. Consider the linear regression

$$y_i = x_i' \beta + \varepsilon_i,$$

and

$$y_i = z_i' b + \varepsilon_i,$$

where $z_i = Q' x_i$ for some p.d. matrix Q . What is the relation between the OLS estimators of β and b ? What about the GLS estimators where $E(\varepsilon\varepsilon') = D$.

3. Assess the identification of the equations of the following models:

(a)

$$c_t = \alpha_1 y_t + \alpha_2 r_{t-1} + u_{t1}$$

$$\dot{i}_t = \beta_1 y_t + \beta_2 r_{t-1} + u_{t2}$$

$$y_t = c_t + \dot{i}_t.$$

Endogenous variables y , c and i : exogenous variables r .

(b)

$$p_t + \beta_{12} w_t + \gamma_{11} Q_t + \gamma_{13} p_{t-1} = u_{t1}$$

$$\beta_{12} p_t + w_t + \beta_{23} N_t + \gamma_{22} S_t + \gamma_{24} w_{t-1} = u_{t2}$$

$$\beta_{32} w_t + N_t + \gamma_{32} S_t + \gamma_{33} p_{t-1} + \gamma_{34} w_{t-1} = u_{t3}$$

where p_t , w_t and N_t are indices for prices, money wages and trade union membership (endogenous) and Q_t and S_t are indices for productivity and strikes (exogenous).

How would the rank and order conditions be affected if it were known *a priori* that

- (1) $\gamma_{11} = 0$
 - (2) $\beta_{21} = \gamma_{22} = 0$
 - (3) $\gamma_{33} = 0$.
- (c) Describe the ILS, 2SLS, and 3SLS estimators, where appropriate, as explicit as possible.

A.3 Time Series

1. Suppose

$$\begin{aligned}\Delta y_t &= u_t, \\ u_t &= \rho u_{t-1} + \varepsilon_t,\end{aligned}$$

where $|\rho| < 1$ and $\{\varepsilon_t\}$ is independent and identically distributed with mean zero and variance 1. Now you consider regression of Δy_t on y_{t-1} and Δy_{t-1}

and want to test whether the coefficient of y_{t-1} is zero or less than zero. Propose a test statistic and derive its limit distribution.

2. Consider the regression, for τ between zero and one,

$$\begin{aligned}y_t &= x_t' \beta_1 + u_t, & \text{if } 1 \leq t \leq \tau n \\y_t &= x_t' \beta_2 + u_t, & \text{if } \tau n < t \leq n\end{aligned}$$

and you want to test for the presence of structural break, that is, $\beta_1 = \beta_2$. Propose a test statistic and derive its limit distribution on which your inference will be based.

3. Let y_t be a strictly stationary and ergodic time series s.t.

$$\begin{aligned}y_t &= \rho y_{t-2}^2 + e_t \\E(e_t | I_{t-1}) &= 0.\end{aligned}$$

Consider the OLS estimator for ρ . Is it consistent for ρ ? Find its asymptotic distribution.

4. Suppose $\{e_t\}$ is iid sequence, which is independent of another *iid* sequence $\{\varepsilon_t\}$. Let

$$y_t = y_{t-1} + e_t \text{ and } x_t = x_{t-1} + \varepsilon_t$$

and $y_0 = x_0 = 0$. Thus, y_t is independent of x_t . Now suppose you regress y_t onto x_t and denote the OLS estimate as b . Derive its limit distribution. Is it zero?

5. Let

$$\Delta y_t = \rho \Delta y_{t-1} + \varepsilon_t$$

for *iid* sequence $\{\varepsilon_t\}$. What is the limit distribution of $\frac{1}{n\sqrt{n}} \sum_t y_t$ and $\frac{1}{n} \sum_t y_{t-1} \Delta y_t$? Then, in the regression of y_t onto y_{t-1} , what is the limit distribution of the standard t -statistic?

6. Consider the following bivariate simultaneous equations system

$$Ay_t + Bx_t = u_t,$$

where x_t is a 2-dimensional vector of exogenous variables and the diagonal elements of A are 1s.

- (a) Demonstrate that the system is not identified and provide a set of exclusion restrictions on B such that the system is just-identified. Justify your answer. Also provide a set of exclusion restrictions that satisfies the order condition of the identification but fails to identify the system.

- (b) Consider the identification restriction that consists of the normalization restrictions and the set of exclusion restrictions in (a) that you provided for just-identification. Describe the indirect least squares (ILS) and 3 stage least squares (3SLS) estimation procedures under the identifying restriction. Compare the two estimators and show that they are consistent.