

CLM - Assumptions

Typical Assumptions
(A1) DGP: y = Xβ + ε is correctly specified.
(A2) E[ε|X] = 0
(A3) Var[ε|X] = σ² I_T
(A4) X has full column rank – rank(X)=k-, where T ≥ k.
Assumption (A1) is called *correct specification*. We know how the DGP.
Assumption (A2) is called *regression*. From (A2) we get:
(i) E[ε|X] = 0 => E[y|X] = f(X, θ) + E[ε|X] = f(X, θ)
(ii) Using the Law of Iterated Expectations (LIE): E[ε] = E_X[E[ε|X]] = E_X[0] = 0

Least Squares Estimation - Assumptions

• From Assumption (A3) we get $Var[\boldsymbol{\varrho} | \mathbf{X}] = \sigma^2 I_T \qquad => Var[\boldsymbol{\varrho}] = \sigma^2 I_T$

This assumption implies (i) homoscedasticity

(ii) no serial/cross correlation

• From Assumption (A4) => the *k* independent variables in X are linearly independent. Then, the *kxk* matrix X'X will also have full rank –i.e., rank(X'X) = *k*.

 $= E[\varepsilon_i^2 | \mathbf{X}] = \sigma^2$

 $= \ge E[\mathbf{\epsilon}_i \mathbf{\epsilon}_i | \mathbf{X}] = 0$

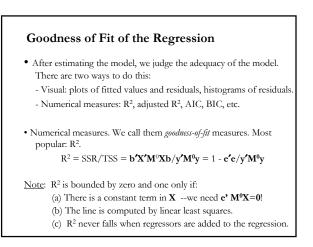
for all i.

for i≠j.

Least Squares Estimation – f.o.c. • Objective function: $S(x_p, \theta) = \sum_i \epsilon_i^2$ • We want to minimize w.r.t to θ . The f.o.c. deliver the normal equations: $-2 \sum_i [y_i - f(x_p, \theta_{LS})] f'(x_p, \theta_{LS}) = -2 (y - Xb)' X = 0$ • Solving for b delivers the OLS estimator: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ <u>Note:</u> (i) $\mathbf{b} = \beta_{OLS}$. (Ordinary LS. *Ordinary*=linear) (ii) \mathbf{b} is a (linear) function of the data (y_i, x_i) . (ii) $\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{e} = \mathbf{0} => \mathbf{e} \perp \mathbf{X}.$

OLS Estimation - Properties Under the typical assumptions, we can establish properties for **b**. 1) $E[\mathbf{b} | \mathbf{X}] = \mathbf{\beta}$ 2) $Var[\mathbf{b} | \mathbf{X}] = E[(\mathbf{b} \cdot \mathbf{\beta}) (\mathbf{b} \cdot \mathbf{\beta})' | \mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'E[\mathbf{\epsilon} \cdot \mathbf{\epsilon}' | \mathbf{X}] \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ $= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ 3) **b** is BLUE (or MVLUE) => The Gauss-Markov theorem. (4) If $(\mathbf{A5} \cdot \mathbf{\epsilon} | \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_T) => \mathbf{b} | \mathbf{X} \sim N(\mathbf{\beta}, \sigma^2 (\mathbf{X}' \mathbf{X})^{-1})$ $=> \mathbf{b}_k | \mathbf{X} \sim N(\mathbf{\beta}_k, \sigma^2 (\mathbf{X}' \mathbf{X})_{kk}^{-1})$ (the marginals of a multivariate normal are also normal.) • Estimating σ^2

Under (A5), $E[\mathbf{e'e} | \mathbf{X}] = (T-k)\sigma^2$ The unbiased estimator of σ^2 is $s^2 = \mathbf{e'e}/(T-k)$. => there is a *degrees of freedom* correction.



Adjusted R-squared

- $\begin{array}{l} R^2 \text{ is modified with a penalty for number of parameters: Adjusted R^2 \\ \overline{\mathsf{R}}^2 = 1 [(T-1)/(T-k)](1-R^2) = 1 [(T-1)/(T-k)] \text{ RSS/TSS} \\ = 1 [\text{RSS}/(T-k)] [(T-1)/\text{TSS}] \\ => \text{maximizing adjusted } R^2 <=> \text{minimizing } [\text{RSS}/(T-k)] = s^2 \end{array}$
- Degrees of freedom --i.e., (T-k)-- adjustment assumes something about "unbiasedness."
- Adjusted-R² includes a penalty for variables that do not add much fit. Can fall when a variable is added to the equation.
- It will rise when a variable, say z, is added to the regression if and only if the t-ratio on z is larger than one in absolute value.

Other Goodness of Fit Measures

- There are other goodness-of-fit measures that also incorporate penalties for number of parameters (degrees of freedom).
- Information Criteria
 Amemiya: [e'e/(T K)] × (1 + k/T)
 Akaike Information Criterion (AIC)
 AIC = -2/T(ln L k)
 L: Likelihood
 => if normality AIC = ln(e'e/T) + (2/T) k (+constants)

 Bayes-Schwarz Information Criterion (BIC)
 - BIC = $-(2/T \ln L [\ln(T)/T] k)$ => if normality AIC = $\ln(\mathbf{e}^{*}\mathbf{e}/T) + [\ln(T)/T] k$ (+constants)

Maximum Likelihood Estimation

• We assume the errors, $\boldsymbol{\varepsilon}$, follow a distribution. Then, we select the parameters of the distribution to maximize the likelihood of the observed sample. Example: The errors, $\boldsymbol{\varepsilon}$, follow the normal distribution: (A5) $\boldsymbol{\varepsilon} | \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_T)$ • Then, we can write the joint pdf of \mathbf{y} as $f(y_t) = (\frac{1}{2\pi\sigma^2})^{1/2} \exp[-\frac{1}{2\sigma^2}(y_t - x_t; \boldsymbol{\beta})^2]$ $L = f(y_1, y_2, ..., y_T | \boldsymbol{\beta}, \sigma^2) = \prod_{t=1}^T (\frac{1}{2\pi\sigma^2})^{1/2} \exp[-\frac{1}{2\sigma^2}(y_t - x_t; \boldsymbol{\beta})^2] = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp(-\frac{1}{2\sigma^2}e'e)$ Taking logs, we have the log likelihood function

 $\ln L = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} e' e$

Maximum Likelihood Estimation

• Let $\theta = (\beta, \sigma)$. Then, we want to $Max_{\theta} \ln L(\theta \mid y, X) = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \sigma^{2} - \frac{1}{2\sigma^{2}} (y - X\beta)'(y - X\beta)$

• Then, the f.o.c.:

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta} &= -\frac{1}{2\sigma^2} (-2X'y - 2X'X\beta) = \frac{1}{\sigma^2} (X'y - X'X\beta) = 0\\ \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} (y - X\beta)'(y - X\beta) = 0 \end{aligned}$$

<u>Note</u>: The f.o.c. deliver the normal equations for β ! The solution to the normal equation, β_{MLE} , is also the LS estimator, **b**. That is,

$$\hat{\beta}_{MLE} = b = (X'X)^{-1}X'y; \qquad \hat{\sigma}_{_{MLE}}^2 = \frac{e'e}{T}$$

• Nice result for b: ML estimators have very good properties!

Properties of ML Estimators

(1) *Efficiency*. Under general conditions, we have that θ_{MLE} $Var(\hat{\theta}_{MLE}) \ge [nI(\theta)]^{-1}$

The right-hand side is the Cramer-Rao lower bound (CR-LB). If an estimator can achieve this bound, ML will produce it.

(2) Consistency.

 $S_n(X; \theta)$ and $(\hat{\theta}_{MLE} - \theta)$ converge together to zero (i.e., expectation).

(3) Theorem: Asymptotic Normality Let the likelihood function be L(X₁,X₂,...X_n | θ). Under general conditions, the MLE of θ is asymptotically distributed as

 $\hat{\theta}_{MLE} \longrightarrow N\left(\theta, [nI(\theta)]^{-1}\right)$

Properties of ML Estimators

- (4) *Sufficiency*. If a single sufficient statistic exists for θ , the MLE of θ must be a function of it. That is, $\hat{\theta}_{MLE}$ depends on the sample observations only through the value of a sufficient statistic.
- (5) *Invariance*. The ML estimate is invariant under functional transformations. That is, if $\hat{\theta}_{MLE}$ is the MLE of θ and if $g(\theta)$ is a function of θ , then $g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$.

Specification Errors: Omitted Variables

• Omitting relevant variables: Suppose the correct model is $\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ -i.e., with two sets of variables. But, we compute OLS omitting \mathbf{X}_2 . That is, $\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$ <= the "short regression."

 $\begin{array}{ll} \text{Some easily proved results:} \\ (1) & \operatorname{E}[\mathbf{b}_1 | \mathbf{X}] = \operatorname{E}\left[(\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \, \mathbf{y} \right] = \boldsymbol{\beta}_1 + (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2 \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_1. \\ & => \operatorname{Unless} \mathbf{X}_1' \mathbf{X}_2 = 0, \, \mathbf{b}_1 \text{ is biased. The bias can be huge.} \\ (2) & \operatorname{Var}[\mathbf{b}_1 | \mathbf{X}] \leq \operatorname{Var}[\mathbf{b}_{1,2} | \mathbf{X}] & => \operatorname{smaller variance when we omit } \mathbf{X}_2. \\ (3) & \operatorname{MSE} & => \mathbf{b}_1 \text{ may be more "precise."} \end{array}$

Specification Errors: Irrelevant Variables

• Irrelevant variables Suppose the correct model is $y = X_1\beta_1 + \epsilon$ But, we estimate $y = X_1\beta_1 + X_2\beta_2 + \epsilon$ Let's compute OLS with X_1, X_2 . This is called "long regression."

Some easily proved results: (1) Since the variables in X_2 are truly irrelevant, then $\beta_2 = 0$, so $E[\mathbf{b}_{1,2}|\mathbf{X}] = \beta_1 => No$ bias (2) Inefficiency: Bigger variance

Linear Restrictions

• Q: How do linear restrictions affect the properties of the least squares estimator?

Model (DGP): $y = X\beta + \epsilon$ Theory (information): $R\beta - q = 0$

 $\label{eq:Restricted LS estimator: b* = b - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(Rb-q)$

1. Unbiased? YES. $E[\mathbf{b}^* | \mathbf{X}] = \mathbf{\beta}$

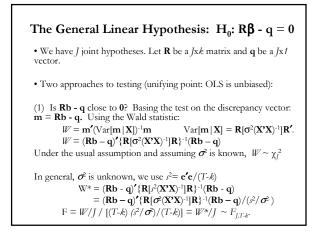
2. Efficiency? NO. $\operatorname{Var}[\mathbf{b}^* | \mathbf{X}] \leq \operatorname{Var}[\mathbf{b} | \mathbf{X}]$

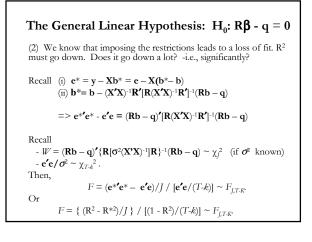
3. **b*** may be more "precise."

Precision = MSE = variance + squared bias.

4. Recall: $\mathbf{e'e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \le \mathbf{e^{*'e^*}} = (\mathbf{y} - \mathbf{Xb^*})'(\mathbf{y} - \mathbf{Xb^*})$

=> Restrictions cannot increase $R^2 => R^2 \ge R^{2*}$





Example: Testing \mathbf{H}_0 : $\mathbf{R}\mathbf{\beta} - \mathbf{q} = \mathbf{0}$ • In the linear model $\mathbf{y} = \mathbf{X} \mathbf{\beta} + \mathbf{\varepsilon} = \mathbf{\beta}_1 + \mathbf{X}_2 \mathbf{\beta}_2 + \mathbf{X}_3 \mathbf{\beta}_3 + \mathbf{X}_4 \mathbf{\beta}_4 + \mathbf{\varepsilon}$ • We want to test if the slopes \mathbf{X}_3 , \mathbf{X}_4 are equal to zero. That is, $\mathbf{H}_0: \mathbf{\beta}_3 = \mathbf{\beta}_4 = \mathbf{0}$ $\mathbf{H}_1: \mathbf{\beta}_3 \neq \mathbf{0}$ or $\mathbf{\beta}_4 \neq \mathbf{0}$ or both $\mathbf{\beta}_3$ and $\mathbf{\beta}_4 \neq \mathbf{0}$ • We can use, $F = (\mathbf{e}^{**}\mathbf{e}^* - \mathbf{e}^{*}\mathbf{e})/J / [\mathbf{e}^{*}\mathbf{e}/(T \cdot k_2)] \sim F_{J,T \cdot k^*}$ Define $Y = \mathbf{\beta}_1 + \mathbf{\beta}_2 X_2 + \mathbf{\varepsilon}$ RSS_R $Y = \mathbf{\beta}_1 + \mathbf{\beta}_2 X_2 + \mathbf{\beta}_3 X_3 + \mathbf{\beta}_4 X_4 + \mathbf{\varepsilon}$ RSS_U $F(\text{cost in } d_5 \text{ unconstr } d_7) = \frac{RSS_R \cdot RSS_U}{RSS_U} \frac{J}{T \cdot k_U}$

Functional Form: Chow Test

• Assumption (A1) restricts $f(\mathbf{X},\beta)$ to be a linear function: $f(\mathbf{X},\beta) = \mathbf{X} \beta$. But, within the framework of OLS estimation, we can be more flexible: (1) We can impose non-linear functional forms, as long as they are linear in the parameters (*intrinsic linear model*).

(2) We can use qualitative variables (dummies) to create non-linearities (splines, changes in regime, etc.) A Chow test (an F-test) can be used to check for regimes/categories or structural breaks.

(a) Run OLS with no distinction between regimes. Keep RSS_R.

(b) Run two separate OLS, one for each regime (Unrestricted regression). Keep RSS_1 and $RSS_2 => RSS_1 = RSS_1 + RSS_2$.

(3) Run a standard F-test (testing Restricted vs. Unrestricted models):

 $F = \frac{(RSS_R - RSS_U)/(k_U - k_R)}{(RSS_U)/(T - k_U)} = \frac{(RSS_R - [RSS_1 + RSS_2])/k}{(RSS_1 + RSS_2)/(T - 2k)}$

Functional Form: Ramsey's RESET Test

- To test the specification of the functional form, we can use the RESET test. From a regression, we keep the fitted values, $\hat{y} = Xb$.
- Then, we add \hat{y}^2 to the regression specification. If \hat{y}^2 is added to the regression specification, it should pick up quadratic and interactive nonlinearity:

 $y = X \beta + \hat{y}^2 \gamma + \epsilon$

• We test H₀ (linear functional form): γ=0

 H_1 (non linear functional form): $\gamma \neq 0$

=> *t-test* on the OLS estimator of γ .

• If the *t-statistic* for \hat{y}^2 is significant => evidence of nonlinearity.

Prediction Intervals

• Prediction: Given $\mathbf{x}^0 \Rightarrow \text{predict } \mathbf{y}^0$. (1) Estimate: $E[\mathbf{y} | \mathbf{X}, \mathbf{x}^0] = \mathbf{\beta}' \mathbf{x}^0$;

(2) Prediction: y⁰ = β'x⁰ + ε⁰
 Predictor: ŷ⁰ = b'x⁰ + estimate of ε⁰. (Est. ε⁰=0, but with variance)

• Forecast error. We predict
$$y^0$$
 with $\hat{y}^0 = \mathbf{b'x}^0$.
 $\hat{y}^0 - y^0 = \mathbf{b'x}^0 - \mathbf{\beta'x}^0 - \mathbf{\epsilon}^0 = (\mathbf{b} - \mathbf{\beta)'x}^0 - \mathbf{\epsilon}^0$
 $=> \operatorname{Var}[(\hat{y}^0 - y^0) | \mathbf{x}^0] = \operatorname{E}[(\hat{y}^0 - y^0)'(\hat{y}^0 - y^0) | \mathbf{x}^0] = \mathbf{x}^0 \cdot \operatorname{Var}[(\mathbf{b} - \mathbf{\beta}) | \mathbf{x}^0] \mathbf{x}^0 + \sigma^2$

How do we estimate this? Two cases:
(1) If x⁰ is a vector of constants => Form C.I. as usual.
(2) If x⁰ has to be estimated => Complicated (what is the variance of the product?). Use bootstrapping.

Forecast Variance

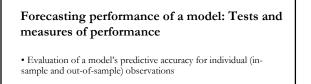
• Variance of the forecast error is
$$\begin{split} & \sigma^2 + \mathbf{x}^0 \cdot \text{Var}[\mathbf{b} \, | \, \mathbf{x}^0 | \mathbf{x}^0 = \sigma^2 + \sigma^2 [\mathbf{x}^0, \, (\mathbf{X}^* \mathbf{X})^{-1} \mathbf{x}^0] \\ & \text{If the model contains a constant term, this is} \end{split}$$

$$\operatorname{Var}[e^{0}] = \sigma^{2} \left[1 + \frac{1}{n} + \sum_{j=1}^{K-1} \sum_{k=1}^{K-1} (x_{j}^{0} - \overline{x}_{j}) (x_{k}^{0} - \overline{x}_{k}) (\mathbf{Z}' \mathbf{M}^{0} \mathbf{Z})^{jk} \right]$$

(where Z is X without $x_i \!=\! \! i).$ In terms squares and cross products of deviations from means.

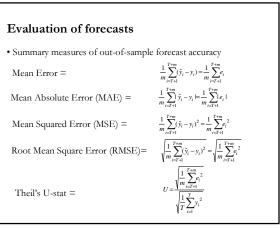
<u>Note</u>: Large σ^2 , small *n*, and large deviations from the means, decrease the precision forecasting error.

 Interpretation: Forecast variance is smallest in the middle of our "experience" and increases as we move outside it.



• Evaluation of a model's predictive accuracy for a group of (insample and out-of-sample) observations

· Chow prediction test



CLM: Asymptotics

• To get exact results for OLS, we rely on (A5) $\boldsymbol{\epsilon} | \mathbf{X} \sim iid N(\mathbf{0}, \sigma^2 \mathbf{I}_T)$ But, (A5) in many situations is unrealistic. Then, we study on the behavior of **b** (and the test statistics) when $T \rightarrow \infty$ i.e., *large samples*.

New assumptions:

{x_ρε_j} i=1, 2, ..., T is a sequence of independent observations.
 X is stochastic, but independent of the process generating *e*.
 We require that X have finite means and variances. Similar requirement for *e*, but we also require E[*e*]=0.

(2) Well behaved X:

plim $(\mathbf{X'X}/T) = \mathbf{Q}$ (**Q** a pd matrix of finite elements)

=> (not too much dependence in **X**).

CLM: New Assumptions

```
• Now, we have a new set of assumptions in the CLM:

(A1) DGP: \mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}.

(A2') X stochastic, but \mathbf{E}[\mathbf{X}^* \boldsymbol{\epsilon}] = 0 and \mathbf{E}[\boldsymbol{\epsilon}] = \mathbf{0}.

(A3) \operatorname{Var}[\boldsymbol{\epsilon} | \mathbf{X} ] = \sigma^2 \mathbf{I}_T

(A4') plim (\mathbf{X}^* \mathbf{X} / T) = \mathbf{Q} (p.d. matrix with finite elements, rank= k)

• We want to study the large sample properties of OLS:

Q 1: Is b consistent? s^2? YES & YES

Q 2: What is the distribution of b? b \xrightarrow{a} N(\boldsymbol{\beta}, (\sigma^2 / T) \mathbf{Q}^{-1})

Q 3: What about the distribution of the tests?

= > t_T = [(z_T - \boldsymbol{\mu}) / s_T] \xrightarrow{d} N(0,1)

= > W = (\mathbf{z}_T - \boldsymbol{\mu}) ' \mathbf{S}_T^{-1} (\mathbf{z}_T - \boldsymbol{\mu}) \xrightarrow{d} \chi^2_{\operatorname{rank}(\mathbf{S}T)}

= > T \xrightarrow{d} \chi^2_{\operatorname{rank}(\operatorname{Var}[\mathbf{m}])}
```

Asymptotic Tests: Small sample behavior?

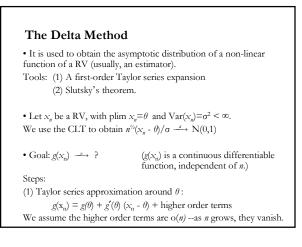
• The p-values from asymptotic tests are approximate for small samples. They may be very bad. Their performance depends on: (1) Sample size, *T*.

- (2) Distribution of the error terms, **ɛ**.
- (3) The number of regressors, k, and their properties
- (4) The relationship between the error terms and the regressors.

• A simulation/bootstrap can help.

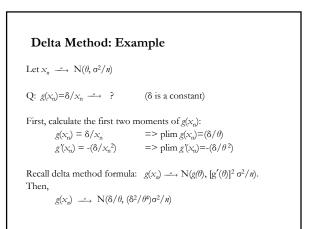
• Bootstrap tests tend to perform better than tests based on approximate asymptotic distributions.

 \bullet The errors committed by both asymptotic and bootstrap tests diminish as T increases.



The Delta Method (2) Use Slutsky theorem: $\underset{\alpha'}{\text{plim } g(\mathbf{x}_n) = g(\theta)}{\underset{\alpha'}{\text{plim } g'(\mathbf{x}_n) = g'(\theta)}}$ Then, as *n* grows, $g(\mathbf{x}_n) \approx g(\theta) + g'(\theta) (\mathbf{x}_n - \theta)$ $=> n^{n'/2} ([g(\mathbf{x}_n) - g(\theta)]) \approx g'(\theta) [n^{n'/2}(\mathbf{x}_n - \theta)].$ $=> n^{n'/2} ([g(\mathbf{x}_n) - g(\theta)]/\sigma) \approx g'(\theta) [n^{n'/2}(\mathbf{x}_n - \theta)/\sigma].$ The asymptotic distribution of $(g(\mathbf{x}_n) - g(\theta))$ is given by that of $[n^{n/2}(\mathbf{x}_n - \theta)/\sigma]$, which is a standard normal. Then, $n^{n'/2} ([g(\mathbf{x}_n) - g(\theta)]) \xrightarrow{a} N(0, [g'(\theta)]^2 \sigma^2).$

After some work ("inversion"), we obtain: $g(x_n) \xrightarrow{a} N(g(\theta), [g'(\theta)]^2 \sigma^2/n).$



The IV Problem

• What makes **b** consistent when **X'** $\boldsymbol{\varepsilon}$ / $T \xrightarrow{p} \boldsymbol{0}$ is that approximating (**X'\boldsymbol{\varepsilon}**/*T*) by **0** is reasonably accurate in large samples.

• Now, we challenge the assumption that $\{x_{\rho}\varepsilon_{i}\}$ is a sequence of independent observations.

- Now, we assume plim $(\mathbf{X}^{*}\mathbf{\epsilon}/T) \neq 0$ => This is the IV Problem!
- Q: When might **X** be correlated **ɛ**?
- Correlated shocks across linked equations
- Simultaneous equations
- Errors in variables
- Model has a lagged dependent variable and a serially correlated error term

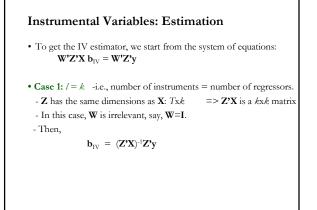
The IV Problem • We start with our linear model $\mathbf{y} = \mathbf{X}\mathbf{\beta} + \mathbf{\epsilon}$ • Now, we assume $\operatorname{plim}(\mathbf{X}^{*}\mathbf{\epsilon}/T) \neq 0$. $\operatorname{plim}(\mathbf{X}^{*}\mathbf{X}/T) = \mathbf{Q}$ • Then, $\operatorname{plim} \mathbf{b} = \operatorname{plim} \mathbf{\beta} + \operatorname{plim}(\mathbf{X}^{*}\mathbf{X}/T)^{-1}\operatorname{plim}(\mathbf{X}^{*}\mathbf{\epsilon}/T)$ $= \mathbf{\beta} + \mathbf{Q}^{-1}\operatorname{plim}(\mathbf{X}^{*}\mathbf{\epsilon}/T) \neq \mathbf{\beta}$ $=> \mathbf{b}$ is not a consistent estimator of $\mathbf{\beta}$. • New assumption: we have / instrumental variables, \mathbf{Z} such that $\operatorname{plim}(\mathbf{Z}^{*}\mathbf{X}/T) \neq \mathbf{0}$ but $\operatorname{plim}(\mathbf{Z}^{*}\mathbf{\epsilon}/T) = \mathbf{0}$

Instrumental Variables: Assumptions

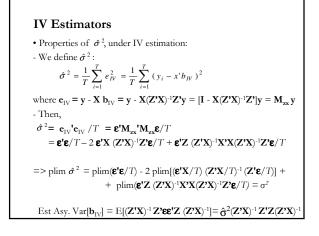
• To get a consistent estimator of $\boldsymbol{\beta}$, we also assume:

 $\begin{aligned} &\{x_{j}, z_{j}, \varepsilon_{j}\} \text{ is a sequence of RVs, with:} \\ & \text{E}[\mathbf{X}^{*}\mathbf{X}] = \boldsymbol{Q}_{xx} \text{ (pd and finite)} & (\text{LLN} => \text{plim}(\mathbf{X}^{*}\mathbf{X}/T) = \boldsymbol{Q}_{xx}) \\ & \text{E}[\mathbf{Z}^{*}\mathbf{Z}] = \boldsymbol{Q}_{zz} \text{ (finite)} & (\text{LLN} => \text{plim}(\mathbf{Z}^{*}\mathbf{Z}/T) = \boldsymbol{Q}_{zz}) \\ & \text{E}[\mathbf{Z}^{*}\mathbf{X}] = \boldsymbol{Q}_{zx} \text{ (pd and finite)} & (\text{LLN} => \text{plim}(\mathbf{Z}^{*}\mathbf{X}/T) = \boldsymbol{Q}_{zx}) \\ & \text{E}[\mathbf{Z}^{*}\mathbf{g}] = \mathbf{0} & (\text{LLN} => \text{plim}(\mathbf{Z}^{*}\mathbf{g}/T) = \mathbf{0}) \end{aligned}$

- Following the same idea as in OLS, we get a system of equations: W'Z'X b_{IV} = W'Z'y
- We have two cases where estimation is possible:
- Case 1: l = k -i.e., number of instruments = number of regressors.
- Case 2: l > k -i.e., number of instruments > number of regressors.



IV Estimators • Properties of \mathbf{b}_{IV} (1) Consistent $\mathbf{b}_{IV} = (\mathbf{Z}^*\mathbf{X})^{-1}\mathbf{Z}^*\mathbf{y} = (\mathbf{Z}^*\mathbf{X})^{-1}\mathbf{Z}^*(\mathbf{X}\mathbf{\beta} + \mathbf{\epsilon})$ $= (\mathbf{Z}^*\mathbf{X}/T)^{-1}(\mathbf{Z}^*\mathbf{X}/T)\mathbf{\beta} + (\mathbf{Z}^*\mathbf{X}/T)^{-1}\mathbf{Z}^*\mathbf{\epsilon}/T$ $= \mathbf{\beta} + (\mathbf{Z}^*\mathbf{X}/T)^{-1}\mathbf{Z}^*\mathbf{\epsilon}/T \xrightarrow{r} \mathbf{\beta}$ (under assumptions) (2) Asymptotic normality $\sqrt{T} (\mathbf{b}_{IV} - \mathbf{\beta}) = \sqrt{T} (\mathbf{Z}^*\mathbf{X})^{-1}\mathbf{Z}^*\mathbf{\epsilon}$ $= (\mathbf{Z}^*\mathbf{X}/T)^{-1}\sqrt{T} (\mathbf{Z}^*\mathbf{\epsilon}/T)$ Using the Lindberg-Feller CLT $\sqrt{T} (\mathbf{Z}^*\mathbf{\epsilon}/T) \xrightarrow{r} \mathbf{N}(0, \sigma^2 \mathbf{Q}_{zz})$ Then, $\sqrt{T} (\mathbf{b}_{IV} - \mathbf{\beta}) \xrightarrow{r} \mathbf{N}(0, \sigma^2 \mathbf{Q}_{zz}^{-1})$



IV Estimators: 2SLS (2-Stage Least Squares)

• Case 2: *l* > *k* −i.e., number of instruments > number of regressors. - This is the usual case. We can throw *l k* instruments, but throwing away information is never optimal.

- The IV normal equations are an $l \ge k$ system of equations:

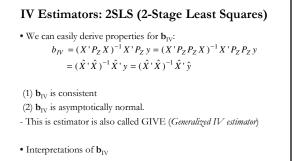
```
Z'y = Z'X\beta + Z'\epsilon
```

<u>Note</u>: We cannot approximate all the $Z^{\prime e}$ by 0 simultenously. There will be at least *l*-k non-zero residuals. (Similar setup to a regression!)

- From the IV normal equations \implies **W'Z'X b**_{IV} = **W'Z'y**

- Let $\mathbf{Z}\mathbf{W} = \mathbf{Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'X} = \mathbf{P}_{\mathbf{Z}}\mathbf{X} = \hat{X}$

- Then, $\mathbf{X'P_Z X} \mathbf{b}_{\text{IV}} = \mathbf{X'P_Z y}$ $b_{IV} = (X'P_Z X)^{-1} X'P_Z y = (X'P_Z P_Z X)^{-1} X'P_Z P_Z y = (\hat{X}'\hat{X})^{-1} \hat{X}'\hat{y}$



$$\begin{split} b_{IV} &= b_{2SLS} = (\hat{X}^{\,\prime} \hat{X})^{-1} \hat{X}^{\,\prime} y \qquad \text{This is the 2SLS interpretation} \\ b_{IV} &= (\hat{X}^{\,\prime} X)^{-1} \hat{X}^{\,\prime} y \qquad \text{This is the usual IV } Z = \hat{X} \end{split}$$

Asymptotic Efficiency

• The variance is larger than that of 0LS. (A large sample type of Gauss-Markov result is at work.)

OLS is inconsistent.

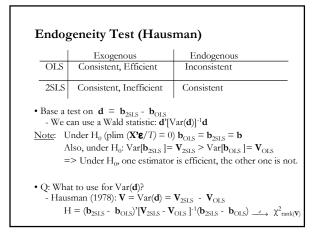
(2) Mean squared error is uncertain:

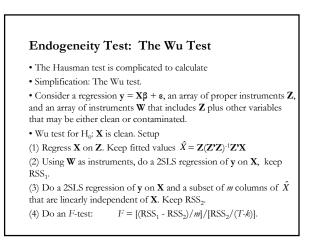
 $MSE[estimator | \beta] = Variance + square of bias.$

 \bullet IV may be better or worse. Depends on the data: X and $\epsilon.$

Problems with 2SLS

- Z'X/T may not be sufficiently large. The covariance matrix for the IV estimator is Asy. Cov(b) = σ²[(Z'X)(Z'Z)⁻¹(X'Z)]⁻¹
 If Z'X/T goes to 0 (weak instruments), the variance explodes.
- When there are many instruments, \hat{X} is too close to **X**; 2SLS becomes OLS.
- <u>Popular misconception</u>: "If only one variable in X is correlated with **ɛ**, the other coefficients are consistently estimated." False.
 => The problem is "smeared" over the other coefficients.
- What are the finite sample properties of b_{IV}? We do not have the condition E[e | X] = 0, we cannot conclude that b_{IV} is unbiased, or that it has a Var[b_{2SLS}] equal to its asymptotic covariance matrix.
 => In fact, b_{2SLS} can have very bad small-sample properties.



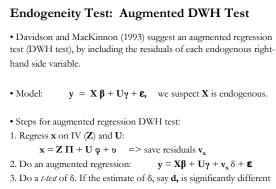


Endogeneity Test: The Wu Test

• Under H_0 : **X** is clean, the *F* statistic has an approximate $F_{m,T-k}$ distribution.

Davidson and MacKinnon (1993, 239) point out that the DWH test really tests whether possible endogeneity of the right-hand-side variables not contained in the instruments makes any difference to the coefficient estimates.

• These types of exogeneity tests are usually known as DWH (Durbin, Wu, Hausman) tests.



from zero, then OLS is not consistent.

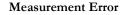
Measurement Error

• DGP: $\mathbf{y}^* = \boldsymbol{\beta} \mathbf{x}^* + \boldsymbol{\varepsilon}$ - $\boldsymbol{\varepsilon} \sim iid D(\mathbf{0}, \sigma_{\varepsilon}^2)$

• But, we do not observe or measure correctly \mathbf{x}^* . We observe \mathbf{x} , \mathbf{y} : $\mathbf{x} = \mathbf{x}^* + \mathbf{u}$ $\mathbf{u} \sim iid \operatorname{D}(\mathbf{0}, \sigma_u^2)$ -no correlation to $\boldsymbol{\varepsilon}_s \mathbf{v}$ $\mathbf{y} = \mathbf{y}^* + \mathbf{v}$ $\mathbf{v} \sim iid \operatorname{D}(\mathbf{0}, \sigma_v^2)$ -no correlation to $\boldsymbol{\varepsilon}_s \mathbf{u}$

• Let's consider two cases:

CASE 1 - Only \mathbf{x}^* is measured with error $(\mathbf{y}=\mathbf{y}^*)$: $\mathbf{y} = \beta(\mathbf{x} \cdot \mathbf{u}) + \mathbf{\varepsilon} = \beta \mathbf{x} + \mathbf{\varepsilon} - \beta \mathbf{u} = \beta \mathbf{x} + \mathbf{w}$ $\mathrm{E}[\mathbf{x}^*\mathbf{w}] = \mathrm{E}[(\mathbf{x}^* + \mathbf{u})^*(\mathbf{\varepsilon} - \beta \mathbf{u})] = -\beta \sigma_u^2 \neq 0$ $=> \mathrm{CLM} \text{ assumptions violated } => \mathrm{OLS} \text{ inconsistent!}$

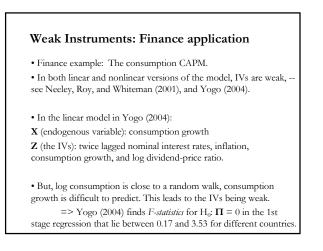


CASE 2 - Only \mathbf{y}^* is measured with error. $\mathbf{y}^* = \mathbf{y} \cdot \mathbf{v} = \beta \mathbf{x}^* + \boldsymbol{\epsilon}$

 $=> \ \mathbf{y} = \beta \mathbf{x}^* + \pmb{\epsilon} + \mathbf{v} = \beta \mathbf{x}^* + (\pmb{\epsilon} + \mathbf{v})$

Q: What happens when y is regressed on x?
 A: Nothing! We have our usual OLS problem since g and v are independent of each other and x^{*}. CLM assumptions are not violated!

inference on β . In this case, we say Z is a *weak* instrument.



Weak Instruments: Summary

- Even if the instrument is "good" –i.e., it meets the relevant condition--, matters can be made far worse with IV as opposed to OLS ("the cure can be worse...").
- Weak correlation between IV and endogenous regressor can pose severe finite-sample bias.
- Even small Cov(Z,e) will cause inconsistency, and this will be exacerbated when Cov(X,Z) is small.
- Large *T* will not help. A&K and Consumption CAPM tests have very large samples!

Weak Instruments: Detection and Remedies

- Symptom: The relevance condition, $plim(\mathbf{Z'X}/T)$ not zero, is close to being violated.
- Detection of weak IV:
- Standard F test in the 1st stage regression of x_k on Z. Staiger and Stock (1997) suggest that F < 10 is a sign of problems.
- $\quad {\rm Low} \ {\rm partial-} R^2_{X,Z}.$
- Large $\mathrm{Var}[\boldsymbol{b}_{\mathrm{IV}}]$ as well as potentially severe finite-sample bias.
- Remedy:
 - Not much most of the discussion is about the condition, not what to do about it.
 - Use LIML? Requires a normality assumption. Probably, not too restrictive. (Text, 375-77)

Weak Instruments: Detection and Remedies

- Symptom: The valid condition, plim(Z'e/T) zero, is close to being violated.
- Detection of instrument exogeneity:
 - -~ Endogenous IV's: Inconsistency of $b_{\rm IV}$ that makes it no better (and probably worse) than $b_{\rm OLS}$
 - Durbin-Wu-Hausman test: Endogeneity of the problem regressor(s)
- Remedy:
 - Avoid endogeneous weak instruments. (Also avoid weak IV!)
 - General problem: It is not easy to find good instruments in theory and in practice. Find *natural experiments*.

M-Estimation

 \bullet An extremum estimator is one obtained as the optimizer of a criterion function, $q({\boldsymbol{z}},{\boldsymbol{b}}).$

- Examples:
 - OLS: $\mathbf{b} = \arg \max \left(-\mathbf{e'e}/T\right)$

MLE: $\mathbf{b}_{\text{MLE}} = \arg \max \ln L = \sum_{i=1,...,T} \ln f(\mathbf{y}_i \mathbf{x}_i, \mathbf{b})$ GMM: $\mathbf{b}_{\text{GMM}} = \arg \max - \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{b})$ **W** $\mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{b})$

- There are two classes of extremum estimators:
- M-estimators: The objective function is a sample average or a sum.
 Minimum distance estimators: The objective function is a measure
- of a distance.
- "M" stands for a maximum or minimum estimators --Huber (1967)

M-Estimation

• The objective function is a sample average or a sum. For example, we want to minimize a population (first) moment:

 $\textit{min}_{b} \operatorname{E}[q(\pmb{z},\pmb{\beta})]$

– Using the LLN, we move from the population first moment to the sample average:

 $\sum_{i} q(\mathbf{z}_{i}, \mathbf{b}) / T \xrightarrow{P} E[q(\mathbf{z}, \boldsymbol{\beta})]$

- We want to obtain: $\mathbf{b} = \operatorname{argmin} \sum_{i} q(\mathbf{z}_{i}, \mathbf{b})$ (or divided by T)

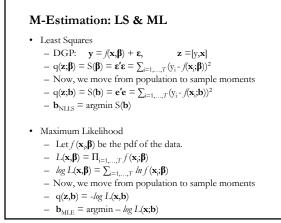
– In general, we solve the f.o.c. (or zero-score condition):

Zero-Score: $\sum_{i} \partial q(\mathbf{z}_{i}, \mathbf{b}) / \partial \mathbf{b}' = \mathbf{0}$

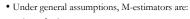
– To check the s.o.c., we define the (pd) Hessian: $\mathbf{H} = \sum_i \partial^2 q(\mathbf{z}_i, \mathbf{b}) / \partial \mathbf{b} \partial \mathbf{b'}$

M-Estimation

- If $\mathbf{s}(\mathbf{z}, \mathbf{b}) = \partial q(\mathbf{z}, \mathbf{b}) / \partial \mathbf{b}'$ exists (almost everywhere), we solve $\sum_{\mathbf{i}} \mathbf{s}(\mathbf{z}_{\mathbf{i}}, \mathbf{b}_{\mathbf{M}}) / T = 0 \qquad (*)$
- If, in addition, $E_X[s(z,b)] = \partial/\partial b' E_X[q(z,b)]$ -i.e., differentiation and integration are exchangeable-, then $E_X[\partial q(z,\beta)/\partial \beta'] = 0.$
- Under these assumptions, the M-estimator is said to be of ψ-type (ψ= s(z,b)=score). Often, b_M is taken to be the solution of (*) without checking whether it is indeed a minimum).
- Otherwise, the M-estimator is of ρ -type. (ρ = q($\mathbf{z}, \boldsymbol{\beta}$)).

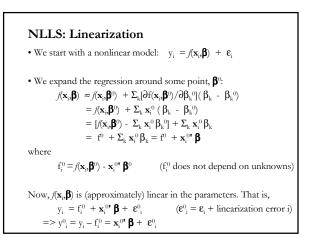


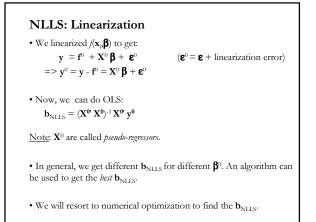
M-Estimators: Properties

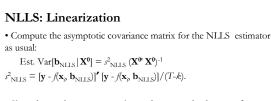


- $-\mathbf{b}_{\mathbf{M}} \xrightarrow{p} \mathbf{b}_{\mathbf{0}}$
- $\mathbf{b}_{\mathbf{M}} \xrightarrow{a} N(\mathbf{b}_{\mathbf{0}}, \operatorname{Var}[\mathbf{b}_{\mathbf{0}}])$
- $\operatorname{Var}[\mathbf{b}_{\mathbf{M}}] = (1/T) \mathbf{H}_{\mathbf{0}}^{-1} \mathbf{V}_{\mathbf{0}} \mathbf{H}_{\mathbf{0}}^{-1}$ - If the model is correctly specified: -**H** = **V**.
- Then, $Var[b] = V_0$
- $-~\mathbf{H}$ and \mathbf{V} are evaluated at $b_0\!\!:$
 - **H** = $\sum_{i} \left[\frac{\partial^2 q(\mathbf{z}_i, \mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}'} \right]$
 - $\mathbf{V} = \sum_i [\partial q(\mathbf{z}_i, \mathbf{b}) / \partial \mathbf{b}] [\partial q(\mathbf{z}_i, \mathbf{b}) / \partial \mathbf{b'}]$

Nonlinear Least Squares: Example Example: Min $_{\beta} S(\beta) = \{\frac{1}{2} \sum_{i} [y_{i} - f(X\beta)]^{2} \}$ • From the f.o.c., we cannot solve for β explicitly. But, using some steps, we can still minimize RSS to obtain estimates of β . • Nonlinear regression algorithm: 1. Start by guessing a plausible values for β , say β^{0} . 2. Calculate RSS for $\beta^{0} => \text{get RSS}(\beta^{0})$ 3. Make small changes to β^{0} , $=> \text{get RSS}(\beta^{1})$ 4. Calculate RSS for $\beta^{1} => \text{get RSS}(\beta^{1})$ 5. If RSS(β^{1}) < RSS(β^{0}) $=> \beta^{1}$ becomes your new starting point. 6. Repeat steps 3-5 until you RSS(β) cannot be lowered. $=> \text{get } \beta^{1}$.







• Since the results are asymptotic, we do not need a degrees of freedom correction. However, a *df* correction is usually included.

Gauss-Newton Algorithm

•
$$\mathbf{b}_{\text{NLLS}}$$
 depends on $\mathbf{\beta}^0$. That is,
 $\mathbf{b}_{\text{NLLS}} (\mathbf{\beta}^0) = (\mathbf{X}^0 \mathbf{X}^0)^{-1} \mathbf{X}^0 \mathbf{y}^0$

 \bullet We use a Gauss-Newton algorithm to find the $\mathbf{b}_{\text{NLLS}}.$ Recall GN:

 $\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \boldsymbol{\epsilon} \qquad - -\mathbf{J}: \text{Jacobian} = \delta f(xi; \boldsymbol{\beta}) / \delta \boldsymbol{\beta}.$

• Given a \mathbf{b}_{NLLS} at step \mathbf{m} , $\mathbf{b}_{(j)}$, we find the \mathbf{b}_{NLLS} for step j+1 by: $\mathbf{b}_{(j+1)} = \mathbf{b}_{(j)} + [\mathbf{X}^0(j)]^* \mathbf{X}^0(j)]^{-1} \mathbf{X}^0(j)' \mathbf{c}^0(j)$

Columns of $\mathbf{X}^{0}(j)$ are the derivatives: $\frac{\partial f(\mathbf{x}_{ij}, \mathbf{b}(j))}{\partial \mathbf{b}(j)'} = \mathbf{y} - f[\mathbf{x}, \mathbf{b}(j)]$

• The *update* vector is the slopes in the regression of the residuals on $\mathbf{X}^0.$ The update is zero when they are orthogonal. (Just like OLS)

Box-Cox Transformation

```
• A simple transformation that allows non-linearities in the CLM.
```

$$\mathbf{y} = f(\mathbf{x}_i, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} = \sum_k \mathbf{x}_k^{(\lambda)} \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}$$

- For a given λ , OLS can be used. An iterative process can be used to estimate λ . OLS s.e. have to be corrected. Not a very efficient method.
- NLLS or MLE will work fine.
- We can have a more general Box-Cox transformation model: $\mathbf{y}^{(0,1)} = \boldsymbol{\Sigma}_k \; \mathbf{x}_k^{(0,2)} \; \boldsymbol{\beta}_k + \; \boldsymbol{\epsilon}$

Testing non-linear restrictions

- Testing linear restrictions as before.
- Non-linear restrictions change he usual tests. We want to test: $H_0; R(\pmb{\beta}) = 0$

where $R(\boldsymbol{\beta})$ is a non-linear function, with rank $[\partial R(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = G(\boldsymbol{\beta})] = J$.

• Let $\mathbf{m} = R(\mathbf{b}_{\text{NLLS}}) - \mathbf{0}$. Then, $W=\mathbf{m}'(\text{Var}[\mathbf{m} | \mathbf{X}])^{-1}\mathbf{m} = R(\mathbf{b}_{\text{NLLS}})'(\text{Var}[R(\mathbf{b}_{\text{NLLS}}) | \mathbf{X}])^{-1}R(\mathbf{b}_{\text{NLLS}})$

But, we do not know the distribution of $R(b_{\text{NLLS}})$. We know the distribution of b_{NLLS} . Then, we linearize $R(b_{\text{NLLS}})$ around $\pmb{\beta}$:

 $R(\mathbf{b}_{\text{NLLS}}) \approx R(\boldsymbol{\beta}) + \mathbf{G}(\mathbf{b}_{\text{NLLS}}) (\mathbf{b}_{\text{NLLS}} - \boldsymbol{\beta})$

Testing non-linear restrictions

- Linearize $R(\mathbf{b}_{\text{NLLS}})$ around $\boldsymbol{\beta} (=\mathbf{b}_0)$ $R(\mathbf{b}_{\text{NLLS}}) \approx R(\boldsymbol{\beta}) + G(\mathbf{b}_{\text{NLLS}}) (\mathbf{b}_{\text{NLLS}} - \boldsymbol{\beta})$
- Recall $\sqrt{T(\mathbf{b}_{\mathbf{M}} \mathbf{b}_{0})} \xrightarrow{d} N(\mathbf{0}, \operatorname{Var}[\mathbf{b}_{0}])$ where $\operatorname{Var}[\mathbf{b}_{0}] = H(\mathbf{\beta})^{-1} V(\mathbf{\beta}) H(\mathbf{\beta})^{-1}$
 - $=> \sqrt{T} \left[\mathbf{R}(\mathbf{b}_{\mathrm{NLLS}}) \mathbf{R}(\mathbf{\beta}) \right] \xrightarrow{d} N(\mathbf{0}, \mathbf{G}(\mathbf{\beta}) \operatorname{Var}[\mathbf{b}_{\mathbf{0}}] \mathbf{G}(\mathbf{\beta})')$
 - $= \operatorname{Var}[\mathbf{R}(\mathbf{b}_{\mathrm{NLLS}})] = (1/T) \ \boldsymbol{G}(\boldsymbol{\beta}) \operatorname{Var}[\mathbf{b}_{\mathbf{0}}] \ \boldsymbol{G}(\boldsymbol{\beta})'$

• Then,

 $W = T \operatorname{R}(\mathbf{b}_{\operatorname{NLLS}})' \{ G(\mathbf{b}_{\operatorname{NLLS}}) \operatorname{Var}[\mathbf{b}_{\operatorname{NLLS}}] G(\mathbf{b}_{\operatorname{NLLS}})' \}^{-1} \operatorname{R}(\mathbf{b}_{\operatorname{NLLS}}) \\ => W \stackrel{d}{\longrightarrow} \chi_{j}^{2}$