Lecture 1
Review I

CLM - Assumptions

• Typical Assumptions

(**A1**) DGP: $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ is correctly specified.

(**A2**) $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$

(**A3**) $Var[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\, \mathbf{I}_T$

(**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) $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$ => $E[\mathbf{y}|\mathbf{X}] = f(\mathbf{X}, \theta) + E[\boldsymbol{\varepsilon}|\mathbf{X}] = f(\mathbf{X}, \theta)$

(ii) Using the Law of Iterated Expectations (LIE):

$E[\boldsymbol{\varepsilon}] = E_{\mathbf{X}}[E[\boldsymbol{\varepsilon}|\mathbf{X}]] = E_{\mathbf{X}}[0] = 0$

Least Squares Estimation - Assumptions

• From Assumption (**A3**) we get

$Var[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\mathbf{I}_T$ => $Var[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}_T$

This assumption implies

(i) *homoscedasticity* => $E[\varepsilon_i^2|\mathbf{X}] = \sigma^2$ for all i.

(ii) *no serial/cross correlation* => $E[\varepsilon_i\varepsilon_j|\mathbf{X}] = 0$ for i≠j.

• 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.

Least Squares Estimation – f.o.c.

• Objective function: $S(x_i, \theta) = \Sigma_i\, \varepsilon_i^2$

• We want to minimize w.r.t to θ. The f.o.c. deliver the normal equations:

$-2\, \Sigma_i\, [y_i - f(x_i, \theta_{LS})]\, f'(x_i, \theta_{LS}) = -2\, (\mathbf{y} - \mathbf{X}\mathbf{b})'\, \mathbf{X} = 0$

• Solving for **b** delivers the OLS estimator:

$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\, \mathbf{X}'\mathbf{y}$

Note: (i) $\mathbf{b} = \beta_{OLS}$. (Ordinary LS. *Ordinary*=linear)

(ii) **b** is a (linear) function of the data (y_i ,x_i).

(iii) $\mathbf{X}'(\mathbf{y}-\mathbf{X}\mathbf{b}) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{e} = \mathbf{0} \Rightarrow \mathbf{e} \perp \mathbf{X}$.

OLS Estimation - Properties

Under the typical assumptions, we can establish properties for **b**.

1) $E[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta}$

2) $Var[\mathbf{b}|\mathbf{X}] = E[(\mathbf{b}-\boldsymbol{\beta})\, (\mathbf{b}-\boldsymbol{\beta})'|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\, \mathbf{X}'E[\boldsymbol{\varepsilon}\, \boldsymbol{\varepsilon}'|\mathbf{X}]\, \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$
$= \sigma^2\, (\mathbf{X}'\mathbf{X})^{-1}$

3) **b** is BLUE (or MVLUE) => The Gauss-Markov theorem.

(4) If (**A5**) $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$ => $\mathbf{b}|\mathbf{X} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\, \mathbf{X})^{-1})$

=> $b_k|\mathbf{X} \sim N(\beta_k, \sigma^2(\mathbf{X}'\, \mathbf{X})_{kk}^{-1})$

(the marginals of a multivariate normal are also normal.)

• Estimating σ^2

Under (**A5**), $E[\mathbf{e}'\mathbf{e}|\mathbf{X}] = (T-k)\sigma^2$

The unbiased estimator of σ^2 is $s^2 = \mathbf{e}'\mathbf{e}/(T-k)$.

=> there is a *degrees of freedom* correction.

Goodness of Fit of the Regression

• After estimating the model, we judge the adequacy of the model. There are two ways to do this:
- Visual: plots of fitted values and residuals, histograms of residuals.
- Numerical measures: R^2, adjusted R^2, AIC, BIC, etc.

• Numerical measures. We call them *goodness-of-fit* measures. Most popular: R^2.

$R^2 = SSR/TSS = \mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b}/\mathbf{y}'\mathbf{M}^0\mathbf{y} = 1 - \mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{M}^0\mathbf{y}$

Note: R^2 is bounded by zero and one only if:

(a) There is a constant term in **X** --we need $\mathbf{e}'\, \mathbf{M}^0\mathbf{X}=\mathbf{0}$!

(b) The line is computed by linear least squares.

(c) R^2 never falls when regressors are added to the regression.

Adjusted R-squared

- R^2 is modified with a penalty for number of parameters: Adjusted R^2
  $\bar{R}^2 = 1 - [(T-1)/(T-k)](1 - R^2) = 1 - [(T-1)/(T-k)]\ RSS/TSS$
  $= 1 - [RSS/(T-k)]\ [(T-1)/TSS]$
  => maximizing adjusted R^2 <=> minimizing $[RSS/(T-k)] = s^2$
- *Degrees of freedom* --i.e., $(T-k)$-- adjustment assumes something about "unbiasedness."
- Adjusted-R^2 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*: $[\mathbf{e}'\mathbf{e}/(T-K)] \times (1 + k/T)$
- *Akaike Information Criterion* (AIC)
  AIC = $-2/T(\ln L - k)$ — L: Likelihood
  => if normality AIC = $\ln(\mathbf{e}'\mathbf{e}/T) + (2/T)\,k$ (+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.

<u>Example</u>: 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 **y** as

$$f(y_t) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2\sigma^2}(y_t - x_t'\beta)^2\right]$$

$$L = f(y_1, y_2, \dots, y_T \mid \beta, \sigma^2) = \Pi_{t=1}^{T}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp\left[-\frac{1}{2\sigma^2}(y_t - x_t'\beta)^2\right] = \frac{1}{(2\pi\sigma^2)^{T/2}}\exp\left(-\frac{1}{2\sigma^2}e'e\right)$$

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.:

$$\frac{\partial \ln L}{\partial \beta} = -\frac{1}{2\sigma^2}(-2X'y - 2X'X\beta) = \frac{1}{\sigma^2}(X'y - X'X\beta) = 0$$

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

<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,

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

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

Properties of ML Estimators

(1) *Efficiency*. Under general conditions, we have that $\hat{\theta}_{MLE}$

$$Var(\hat{\theta}_{MLE}) \geq [nI(\theta)]^{-1}$$

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 $L(X_1, X_2, \dots X_n \mid \theta)$. Under general conditions, the MLE of θ is asymptotically distributed as

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

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

$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ -i.e., with two sets of variables.

But, we compute OLS omitting $\mathbf{X}_2$. That is,

$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ <= the "short regression."

Some easily proved results:

(1) $E[\mathbf{b}_1|\mathbf{X}] = E[(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}] = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_1$.

=> Unless $\mathbf{X}_1'\mathbf{X}_2 = 0$, $\mathbf{b}_1$ is *biased*. The bias can be huge.

(2) $Var[\mathbf{b}_1|\mathbf{X}] \leq Var[\mathbf{b}_{1.2}|\mathbf{X}]$ => smaller variance when we omit $\mathbf{X}_2$.

(3) MSE => $\mathbf{b}_1$ may be more "precise."

Specification Errors: Irrelevant Variables

• Irrelevant variables

Suppose the correct model is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$

But, we estimate $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$

Let's compute OLS with $\mathbf{X}_1$, $\mathbf{X}_2$. This is called "long regression."

Some easily proved results:

(1) Since the variables in $\mathbf{X}_2$ are truly irrelevant, then $\boldsymbol{\beta}_2 = \mathbf{0}$,

so $E[\mathbf{b}_{1.2}|\mathbf{X}] = \boldsymbol{\beta}_1$ => No bias

(2) Inefficiency: Bigger variance

Linear Restrictions

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

Model (DGP): $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Theory (information): $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0}$

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

1. Unbiased? YES. $E[\mathbf{b}^*|\mathbf{X}] = \boldsymbol{\beta}$
2. Efficiency? NO. $Var[\mathbf{b}^*|\mathbf{X}] < Var[\mathbf{b}|\mathbf{X}]$
3. $\mathbf{b}^*$ may be more "precise."

   Precision = MSE = variance + squared bias.
4. Recall: $\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \leq \mathbf{e}^{*\prime}\mathbf{e}^* = (\mathbf{y} - \mathbf{Xb}^*)'(\mathbf{y} - \mathbf{Xb}^*)$

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

The General Linear Hypothesis: H_0: $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = 0$

• We have J joint hypotheses. Let $\mathbf{R}$ be a Jxk matrix and $\mathbf{q}$ be a Jx1 vector.

• Two approaches to testing (unifying point: OLS is unbiased):

(1) Is $\mathbf{Rb} - \mathbf{q}$ close to $\mathbf{0}$? Basing the test on the discrepancy vector: $\mathbf{m} = \mathbf{Rb} - \mathbf{q}$. Using the Wald statistic:

$W = \mathbf{m}'(Var[\mathbf{m}|\mathbf{X}])^{-1}\mathbf{m}$ $\quad Var[\mathbf{m}|\mathbf{X}] = \mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'$.

$W = (\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}\}^{-1}(\mathbf{Rb} - \mathbf{q})$

Under the usual assumption and assuming σ^2 is known, $W \sim \chi_J^2$

In general, σ^2 is unknown, we use $s^2 = \mathbf{e}'\mathbf{e}/(T-k)$

$W^* = (\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}\}^{-1}(\mathbf{Rb} - \mathbf{q})$

$= (\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}\}^{-1}(\mathbf{Rb} - \mathbf{q})/(s^2/\sigma^2)$

$F = W/J \,/\, [(T-k)\,(s^2/\sigma^2)/(T-k)] = W^*/J \sim F_{J,T-K}$.

The General Linear Hypothesis: H_0: $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = 0$

(2) We know that imposing the restrictions leads to a loss of fit. R^2 must go down. Does it go down a lot? -i.e., significantly?

Recall (i) $\mathbf{e}^* = \mathbf{y} - \mathbf{Xb}^* = \mathbf{e} - \mathbf{X}(\mathbf{b}^* - \mathbf{b})$

(ii) $\mathbf{b}^* = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$

=> $\mathbf{e}^{*\prime}\mathbf{e}^* - \mathbf{e}'\mathbf{e} = (\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$

Recall

- $W = (\mathbf{Rb} - \mathbf{q})'\{\mathbf{R}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}\}^{-1}(\mathbf{Rb} - \mathbf{q}) \sim \chi_J^2$ (if σ^2 known)
- $\mathbf{e}'\mathbf{e}/\sigma^2 \sim \chi_{T-k}^2$.

Then,

$F = (\mathbf{e}^{*\prime}\mathbf{e}^* - \mathbf{e}'\mathbf{e})/J \,/\, [\mathbf{e}'\mathbf{e}/(T-k)] \sim F_{J,T-K}$.

Or

$F = \{(R^2 - R^{*2})/J\} \,/\, [(1 - R^2)/(T-k)] \sim F_{J,T-K}$.

Example: Testing H_0: $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = 0$

• In the linear model

$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{\beta}_1 + \mathbf{X}_2\,\boldsymbol{\beta}_2 + \mathbf{X}_3\,\boldsymbol{\beta}_3 + \mathbf{X}_4\,\boldsymbol{\beta}_4 + \boldsymbol{\varepsilon}$

• We want to test if the slopes $\mathbf{X}_3$, $\mathbf{X}_4$ are equal to zero. That is,

$H_0: \beta_3 = \beta_4 = 0$

$H_1: \beta_3 \neq 0$ or $\beta_4 \neq 0$ or both β_3 and $\beta_4 \neq 0$

• We can use, $F = (\mathbf{e}^{*\prime}\mathbf{e}^* - \mathbf{e}'\mathbf{e})/J \,/\, [\mathbf{e}'\mathbf{e}/(T-k)] \sim F_{J,T-K}$.

Define $Y = \beta_1 + \beta_2 X_2 + \varepsilon$ $\quad RSS_R$

$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ $\quad RSS_U$

$$F(\text{cost in } df\text{, unconstr } df) = \frac{RSS_R - RSS_U \;/\; k_U - k_R}{RSS_U \;/\; T - k_U}$$

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 => $RSS_U = RSS_1 + RSS_2$.

(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, $\hat{\mathbf{y}} = \mathbf{Xb}$.

• Then, we add $\hat{\mathbf{y}}^2$ to the regression specification. If $\hat{\mathbf{y}}^2$ is added to the regression specification, it should pick up quadratic and interactive nonlinearity:

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \hat{\mathbf{y}}^2\,\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

• We test H_0 (linear functional form): $\gamma=0$

H_1 (non linear functional form): $\gamma\neq0$

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

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

Prediction Intervals

• Prediction: Given $\mathbf{x}^0$ => predict $\mathbf{y}^0$.

(1) Estimate: $E[\mathbf{y}|\mathbf{X}, \mathbf{x}^0] = \boldsymbol{\beta}'\mathbf{x}^0$;

(2) Prediction: $y^0 = \boldsymbol{\beta}'\mathbf{x}^0 + \varepsilon^0$

• Predictor: $\hat{y}^0 = \mathbf{b}'\mathbf{x}^0$ + estimate of ε^0. (Est. ε^0=0, but with variance)

• Forecast error. We predict y^0 with $\hat{y}^0 = \mathbf{b}'\mathbf{x}^0$.

$\hat{y}^0 - y^0 = \mathbf{b}'\mathbf{x}^0 - \boldsymbol{\beta}'\mathbf{x}^0 - \varepsilon^0 = (\mathbf{b} - \boldsymbol{\beta})'\mathbf{x}^0 - \varepsilon^0$

=> $\text{Var}[(\hat{y}^0-y^0)|\mathbf{x}^0] = E[(\hat{y}^0-y^0)'(\hat{y}^0-y^0)|\mathbf{x}^0] = \mathbf{x}^{0\prime}\text{Var}[(\mathbf{b} - \boldsymbol{\beta})|\mathbf{x}^0]\mathbf{x}^0 + \sigma^2$

• How do we estimate this? Two cases:

(1) If $\mathbf{x}^0$ is a vector of constants => Form C.I. as usual.

(2) If $\mathbf{x}^0$ has to be estimated => Complicated (what is the variance of the product?). Use bootstrapping.

Forecast Variance

• Variance of the forecast error is

$\sigma^2 + \mathbf{x}^{0\prime}\,\text{Var}[\mathbf{b}|\mathbf{x}^0]\mathbf{x}^0 = \sigma^2 + \sigma^2[\mathbf{x}^{0\prime}\,(\mathbf{X'X})^{-1}\mathbf{x}^0]$

If the model contains a constant term, this is

$$\text{Var}[e^0] = \sigma^2\left[1 + \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}(x_j^0 - \bar{x}_j)(x_k^0 - \bar{x}_k)(\mathbf{Z}'\mathbf{M}^0\mathbf{Z})^{jk}\right]$$

(where $\mathbf{Z}$ is $\mathbf{X}$ without $\mathbf{x}_1$=$\mathbf{i}$). 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.

Forecasting performance of a model: Tests and measures of performance

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

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

• Chow prediction test

Evaluation of forecasts

• Summary measures of out-of-sample forecast accuracy

Mean Error = $\frac{1}{m}\sum_{i=T+1}^{T+m}(\hat{y}_i - y_i) = \frac{1}{m}\sum_{i=T+1}^{T+m}e_i$

Mean Absolute Error (MAE) = $\frac{1}{m}\sum_{i=T+1}^{T+m}|\hat{y}_i - y_i| = \frac{1}{m}\sum_{i=T+1}^{T+m}|e_i|$

Mean Squared Error (MSE) = $\frac{1}{m}\sum_{i=T+1}^{T+m}(\hat{y}_i - y_i)^2 = \frac{1}{m}\sum_{i=T+1}^{T+m}e_i^2$

Root Mean Square Error (RMSE)= $\sqrt{\frac{1}{m}\sum_{i=T+1}^{T+m}(\hat{y}_i - y_i)^2} = \sqrt{\frac{1}{m}\sum_{i=T+1}^{T+m}e_i^2}$

Theil's U-stat = $U = \frac{\sqrt{\frac{1}{m}\sum_{i=T+1}^{T+m}e_i^2}}{\sqrt{\frac{1}{T}\sum_{i=1}^{T}y_i^2}}$

CLM: Asymptotics

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

• New assumptions:

(1) $\{x_i, \varepsilon_i\}$ i=1, 2, ..., T is a sequence of independent observations.

- **X** is stochastic, but independent of the process generating $\boldsymbol{\varepsilon}$.
- We require that **X** have finite means and variances. Similar requirement for $\boldsymbol{\varepsilon}$, but we also require E[$\boldsymbol{\varepsilon}$]=**0.**

(2) Well behaved **X**:

plim $(\mathbf{X}'\mathbf{X}/T)$ = **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: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

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

(**A3**) $\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\, \mathbf{I}_T$

(**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**? $\mathbf{b} \xrightarrow{a} N(\boldsymbol{\beta}, (\sigma^2/T)\mathbf{Q}^{-1})$

Q 3: What about the distribution of the tests?

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

=> $W = (\mathbf{z}_T - \boldsymbol{\mu})'\mathbf{S}_T^{-1}(\mathbf{z}_T - \boldsymbol{\mu}) \xrightarrow{d} \chi^2_{\text{rank}(\mathbf{S}_T)}$

=> $F \xrightarrow{d} \chi^2_{\text{rank}(\text{Var}[\mathbf{m}])}$

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, $\boldsymbol{\varepsilon}$.

(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.

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

The Delta Method

• It is used to obtain the asymptotic distribution of a non-linear function of a RV (usually, an estimator).

Tools: (1) A first-order Taylor series expansion
(2) Slutsky's theorem.

• Let x_n be a RV, with plim $x_n=\theta$ and $\text{Var}(x_n)=\sigma^2 < \infty$.
We use the CLT to obtain $n^{1/2}(x_n - \theta)/\sigma \xrightarrow{d} N(0,1)$

• Goal: $g(x_n) \xrightarrow{a}$? ($g(x_n)$ is a continuous differentiable function, independent of n.)

Steps:

(1) Taylor series approximation around θ:

$g(x_n) = g(\theta) + g'(\theta)\,(x_n - \theta)$ + higher order terms

We assume the higher order terms are $o(n)$ --as n grows, they vanish.

The Delta Method

(2) Use Slutsky theorem: plim $g(x_n) = g(\theta)$
plim $g'(x_n) = g'(\theta)$

Then, as n grows, $g(x_n) \approx g(\theta) + g'(\theta)\,(x_n - \theta)$

=> $n^{1/2}([g(x_n) - g(\theta)]) \approx g'(\theta)\,[n^{1/2}(x_n - \theta)]$.

=> $n^{1/2}([g(x_n) - g(\theta)]/\sigma) \approx g'(\theta)\,[n^{1/2}(x_n - \theta)/\sigma]$.

The asymptotic distribution of $(g(x_n) - g(\theta))$ is given by that of $[n^{1/2}(x_n - \theta)/\sigma]$, which is a standard normal. Then,

$n^{1/2}([g(x_n) - g(\theta)]) \xrightarrow{a} N(0, [g'(\theta)]^2\, \sigma^2)$.

After some work ("inversion"), we obtain:

$g(x_n) \xrightarrow{a} N(g(\theta), [g'(\theta)]^2\, \sigma^2/n)$.

Delta Method: Example

Let $x_n \xrightarrow{a} N(\theta, \sigma^2/n)$

Q: $g(x_n)=\delta/x_n \xrightarrow{a}$? (δ is a constant)

First, calculate the first two moments of $g(x_n)$:

$g(x_n) = \delta/x_n$ => plim $g(x_n)=(\delta/\theta)$

$g'(x_n) = -(\delta/x_n^2)$ => plim $g'(x_n)=-(\delta/\theta^2)$

Recall delta method formula: $g(x_n) \xrightarrow{a} N(g(\theta), [g'(\theta)]^2\, \sigma^2/n)$.
Then,

$g(x_n) \xrightarrow{a} N(\delta/\theta, (\delta^2/\theta^4)\sigma^2/n)$

The IV Problem

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

• Now, we challenge the assumption that $\{x_i,\varepsilon_i\}$ is a sequence of independent observations.

• Now, we assume plim $(\mathbf{X}'\boldsymbol{\varepsilon}/T) \neq 0$ => This is the IV Problem!

• Q: When might **X** be correlated $\boldsymbol{\varepsilon}$?
- 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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

• Now, we assume plim$(\mathbf{X}'\boldsymbol{\varepsilon}/T) \neq 0.$
plim $(\mathbf{X}'\mathbf{X}/T) = \mathbf{Q}$

• Then, plim $\mathbf{b}$ = plim $\boldsymbol{\beta}$ + plim $(\mathbf{X}'\mathbf{X}/T)^{-1}$ plim $(\mathbf{X}'\boldsymbol{\varepsilon}/T)$
$= \boldsymbol{\beta} + \mathbf{Q}^{-1}$ plim $(\mathbf{X}'\boldsymbol{\varepsilon}/T) \neq \boldsymbol{\beta}$
=> **b** is not a consistent estimator of $\boldsymbol{\beta}$.

• New assumption: we have l instrumental variables, **Z** such that
plim$(\mathbf{Z}'\mathbf{X}/T) \neq \mathbf{0}$ but plim$(\mathbf{Z}'\boldsymbol{\varepsilon}/T) = \mathbf{0}$

Instrumental Variables: Assumptions

• To get a consistent estimator of $\boldsymbol{\beta}$, we also assume:
$\{x_i, z_i, \varepsilon_i\}$ is a sequence of RVs, with:

$E[\mathbf{X}'\mathbf{X}] = \boldsymbol{Q}_{xx}$ (pd and finite) (LLN => plim$(\mathbf{X}'\mathbf{X}/T) = \boldsymbol{Q}_{xx}$)
$E[\mathbf{Z}'\mathbf{Z}] = \boldsymbol{Q}_{zz}$ (finite) (LLN => plim$(\mathbf{Z}'\mathbf{Z}/T) = \boldsymbol{Q}_{zz}$)
$E[\mathbf{Z}'\mathbf{X}] = \boldsymbol{Q}_{zx}$ (pd and finite) (LLN => plim$(\mathbf{Z}'\mathbf{X}/T) = \boldsymbol{Q}_{zx}$)
$E[\mathbf{Z}'\boldsymbol{\varepsilon}] = \mathbf{0}$ (LLN => plim$(\mathbf{Z}'\boldsymbol{\varepsilon}/T) = \mathbf{0}$)

• Following the same idea as in OLS, we get a system of equations:

$$\mathbf{W}'\mathbf{Z}'\mathbf{X}\,\mathbf{b}_{IV} = \mathbf{W}'\mathbf{Z}'\mathbf{y}$$

• 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.

Instrumental Variables: Estimation

• To get the IV estimator, we start from the system of equations:

$$\mathbf{W}'\mathbf{Z}'\mathbf{X}\,\mathbf{b}_{IV} = \mathbf{W}'\mathbf{Z}'\mathbf{y}$$

• **Case 1:** $l = k$ -i.e., number of instruments = number of regressors.
- **Z** has the same dimensions as **X**: Txk => **Z'X** is a kxk matrix
- In this case, **W** is irrelevant, say, **W**=**I**.
- Then,

$$\mathbf{b}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

IV Estimators

• Properties of $\mathbf{b}_{IV}$
(1) Consistent

$$\begin{aligned}\mathbf{b}_{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\\ &= (\mathbf{Z}'\mathbf{X}/T)^{-1}(\mathbf{Z}'\mathbf{X}/T)\,\boldsymbol{\beta} + (\mathbf{Z}'\mathbf{X}/T)^{-1}\mathbf{Z}'\varepsilon/T\\ &= \boldsymbol{\beta} + (\mathbf{Z}'\mathbf{X}/T)^{-1}\mathbf{Z}'\varepsilon/T \xrightarrow{p} \boldsymbol{\beta} \qquad \text{(under assumptions)}\end{aligned}$$

(2) Asymptotic normality

$$\begin{aligned}\sqrt{T}\,(\mathbf{b}_{IV} - \boldsymbol{\beta}) &= \sqrt{T}\,(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\varepsilon\\ &= (\mathbf{Z}'\mathbf{X}/T)^{-1}\sqrt{T}\,(\mathbf{Z}'\varepsilon/T)\end{aligned}$$

Using the Lindberg-Feller CLT $\sqrt{T}\,(\mathbf{Z}'\varepsilon/T) \xrightarrow{d} N(0, \sigma^2\boldsymbol{Q}_{zz})$

Then, $\sqrt{T}\,(\mathbf{b}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2\boldsymbol{Q}_{zx}^{-1}\boldsymbol{Q}_{zz}\boldsymbol{Q}_{xz}^{-1})$

IV Estimators

• Properties of $\hat{\sigma}^2$, under IV estimation:
- We define $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{i=1}^{T} e_{IV}^2 = \frac{1}{T}\sum_{i=1}^{T}(y_i - x'b_{IV})^2$$

where $\mathbf{e}_{IV} = \mathbf{y} - \mathbf{X}\,\mathbf{b}_{IV} = \mathbf{y} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} = [\mathbf{I} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}']\mathbf{y} = \mathbf{M}_{zx}\,\mathbf{y}$
- Then,

$$\begin{aligned}\hat{\sigma}^2 &= \mathbf{e}_{IV}'\mathbf{e}_{IV}/T = \boldsymbol{\varepsilon}'\mathbf{M}_{zx}'\mathbf{M}_{zx}\boldsymbol{\varepsilon}/T\\ &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/T - 2\,\boldsymbol{\varepsilon}'\mathbf{X}\,(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}/T + \boldsymbol{\varepsilon}'\mathbf{Z}\,(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}/T\end{aligned}$$

=> plim $\hat{\sigma}^2$ = plim$(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/T)$ - 2 plim$[(\boldsymbol{\varepsilon}'\mathbf{X}/T)\,(\mathbf{Z}'\mathbf{X}/T)^{-1}(\mathbf{Z}'\boldsymbol{\varepsilon}/T)]$ +
+ plim$(\boldsymbol{\varepsilon}'\mathbf{Z}\,(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}/T) = \sigma^2$

Est Asy. Var$[\mathbf{b}_{IV}] = E[(\mathbf{Z}'\mathbf{X})^{-1}\,\mathbf{Z}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{Z}\,(\mathbf{Z}'\mathbf{X})^{-1}] = \hat{\sigma}^2(\mathbf{Z}'\mathbf{X})^{-1}\,\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{X})^{-1}$

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 x k system of equations:

$$\mathbf{Z'y} = \mathbf{Z'X}\boldsymbol{\beta} + \mathbf{Z'}\boldsymbol{\varepsilon}$$

Note: We cannot approximate all the $\mathbf{Z'}\boldsymbol{\varepsilon}$ by $\mathbf{0}$ simultenously. There will be at least l-k non-zero residuals. (Similar setup to a regression!)

- From the IV normal equations => $\mathbf{W'Z'X}\ \mathbf{b}_{IV} = \mathbf{W'Z'y}$
- We define a different IV estimator
  - Let $\mathbf{ZW} = \mathbf{Z(Z'Z)}^{-1}\mathbf{Z'X} = \mathbf{P_Z X} = \hat{X}$
  - Then, $\mathbf{X'P_Z X}\ \mathbf{b}_{IV} = \mathbf{X'P_Z y}$

$$b_{IV} = (X'P_Z X)^{-1} X'P_Z y = (X'P_Z P_Z X)^{-1} X'P_Z P_Z y = (\hat{X}'\hat{X})^{-1}\hat{X}'\hat{y}$$

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

• We can easily derive properties for $\mathbf{b}_{IV}$:

$$b_{IV} = (X'P_Z X)^{-1} X'P_Z y = (X'P_Z P_Z X)^{-1} X'P_Z P_Z y$$
$$= (\hat{X}'\hat{X})^{-1}\hat{X}'y = (\hat{X}'\hat{X})^{-1}\hat{X}'\hat{y}$$

(1) $\mathbf{b}_{IV}$ is consistent

(2) $\mathbf{b}_{IV}$ is asymptotically normal.

- This is estimator is also called GIVE (*Generalized IV estimator*)

• Interpretations of $\mathbf{b}_{IV}$

$b_{IV} = b_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$ This is the 2SLS interpretation

$b_{IV} = (\hat{X}'X)^{-1}\hat{X}'y$ This is the usual IV $Z = \hat{X}$

Asymptotic Efficiency

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

(1) OLS is inconsistent.

(2) Mean squared error is uncertain:

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

• IV may be better or worse. Depends on the data: $\mathbf{X}$ and $\boldsymbol{\varepsilon}$.

Problems with 2SLS

- $\mathbf{Z'X}/T$ may not be sufficiently large. The covariance matrix for the IV estimator is Asy. Cov($\mathbf{b}$) = $\sigma^2[(\mathbf{Z'X})(\mathbf{Z'Z})^{-1}(\mathbf{X'Z})]^{-1}$
  – If $\mathbf{Z'X}/T$ goes to 0 (weak instruments), the variance explodes.
- When there are many instruments, $\hat{X}$ is too close to $\mathbf{X}$; 2SLS becomes OLS.

• Popular misconception: "If only one variable in $\mathbf{X}$ is correlated with $\boldsymbol{\varepsilon}$, the other coefficients are consistently estimated." False.
=> The problem is "smeared" over the other coefficients.

- What are the finite sample properties of $\mathbf{b}_{IV}$? We do not have the condition E[$\boldsymbol{\varepsilon}$|$\mathbf{X}$] = 0, we cannot conclude that $\mathbf{b}_{IV}$ is unbiased, or that it has a Var[$\mathbf{b}_{2SLS}$] equal to its asymptotic covariance matrix.
  => In fact, $\mathbf{b}_{2SLS}$ can have very bad small-sample properties.

Endogeneity Test (Hausman)

	Exogenous	Endogenous
OLS	Consistent, Efficient	Inconsistent
2SLS	Consistent, Inefficient	Consistent

• Base a test on $\mathbf{d} = \mathbf{b}_{2SLS} - \mathbf{b}_{OLS}$
- We can use a Wald statistic: $\mathbf{d}'[\text{Var}(\mathbf{d})]^{-1}\mathbf{d}$

Note: Under H_0 (plim $(\mathbf{X'}\boldsymbol{\varepsilon}/T) = 0$) $\mathbf{b}_{OLS} = \mathbf{b}_{2SLS} = \mathbf{b}$
Also, under H_0: Var[$\mathbf{b}_{2SLS}$]= $\mathbf{V}_{2SLS}$ > Var[$\mathbf{b}_{OLS}$]= $\mathbf{V}_{OLS}$
=> Under H_0, one estimator is efficient, the other one is not.

• Q: What to use for Var($\mathbf{d}$)?
- Hausman (1978): $\mathbf{V}$ = Var($\mathbf{d}$) = $\mathbf{V}_{2SLS} - \mathbf{V}_{OLS}$

$$H = (\mathbf{b}_{2SLS} - \mathbf{b}_{OLS})'[\mathbf{V}_{2SLS} - \mathbf{V}_{OLS}]^{-1}(\mathbf{b}_{2SLS} - \mathbf{b}_{OLS}) \xrightarrow{d} \chi^2_{\text{rank}(\mathbf{V})}$$

Endogeneity Test: The Wu Test

• The Hausman test is complicated to calculate

• Simplification: The Wu test.

• Consider a regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, an array of proper instruments $\mathbf{Z}$, and an array of instruments $\mathbf{W}$ that includes $\mathbf{Z}$ plus other variables that may be either clean or contaminated.

• Wu test for H_0: $\mathbf{X}$ is clean. Setup

(1) Regress $\mathbf{X}$ on $\mathbf{Z}$. Keep fitted values $\hat{X} = \mathbf{Z(Z'Z)}^{-1}\mathbf{Z'X}$

(2) Using $\mathbf{W}$ as instruments, do a 2SLS regression of $\mathbf{y}$ on $\mathbf{X}$, keep RSS_1.

(3) Do a 2SLS regression of $\mathbf{y}$ on $\mathbf{X}$ and a subset of m columns of $\hat{X}$ that are linearly independent of $\mathbf{X}$. Keep RSS_2.

(4) Do an F-test: $F = [(RSS_1 - RSS_2)/m]/[RSS_2/(T-k)]$.

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.

Endogeneity Test: Augmented DWH Test

• Davidson and MacKinnon (1993) suggest an augmented regression test (DWH test), by including the residuals of each endogenous right-hand side variable.

• Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, we suspect **X** is endogenous.

• Steps for augmented regression DWH test:

1. Regress **x** on IV (**Z**) and **U**:

$\mathbf{x} = \mathbf{Z}\,\boldsymbol{\Pi} + \mathbf{U}\,\varphi + \upsilon$ => save residuals $\mathbf{v_x}$

2. Do an augmented regression: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\gamma} + \mathbf{v_x}\,\delta + \boldsymbol{\varepsilon}$

3. Do a *t-test* of δ. If the estimate of δ, say **d,** is significantly different from zero, then OLS is not consistent.

Measurement Error

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

• But, we do not observe or measure correctly **x***. We observe **x, y**:

$\mathbf{x} = \mathbf{x}^* + \mathbf{u}$ $\mathbf{u} \sim iid\ D(\mathbf{0}, \sigma_u^2)$ -no correlation to $\boldsymbol{\varepsilon}$,**v**

$\mathbf{y} = \mathbf{y}^* + \mathbf{v}$ $\mathbf{v} \sim iid\ D(\mathbf{0}, \sigma_v^2)$ -no correlation to $\boldsymbol{\varepsilon}$,**u**

• Let's consider two cases:

CASE 1 - Only **x*** is measured with error (**y**=**y***):

$\mathbf{y} = \beta(\mathbf{x} - \mathbf{u}) + \boldsymbol{\varepsilon} = \beta\mathbf{x} + \boldsymbol{\varepsilon} - \beta\mathbf{u} = \beta\mathbf{x} + \mathbf{w}$

$E[\mathbf{x}'\mathbf{w}] = E[(\mathbf{x}^* + \mathbf{u})'(\boldsymbol{\varepsilon} - \beta\mathbf{u})] = -\beta\sigma_u^2 \neq 0$

=> CLM assumptions violated => OLS inconsistent!

Measurement Error

CASE 2 - Only **y*** is measured with error.

$\mathbf{y^*} = \mathbf{y} - \mathbf{v} = \beta\mathbf{x}^* + \boldsymbol{\varepsilon}$

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

• Q: What happens when **y** is regressed on **x**?

A: Nothing! We have our usual OLS problem since $\boldsymbol{\varepsilon}$ and **v** are independent of each other and $\mathbf{x}^*$. CLM assumptions are not violated!

Finding an Instrument: Not Easy

• The IV problem requires data on variables (**Z**) such that

(1) $\text{Cov}(\mathbf{x},\mathbf{Z}) \neq \mathbf{0}$ -relevance condition

(2) $\text{Cov}(\mathbf{Z},\boldsymbol{\varepsilon}) = \mathbf{0}$ -valid (exogeneity) condition

Then, we do a first-stage regression to obtain fitted values of **X**:

$\mathbf{x} = \mathbf{Z}\boldsymbol{\Pi} + \mathbf{U}\boldsymbol{\delta} + \mathbf{V}$ $-\mathbf{V} \sim N(0, \sigma_V^2\mathbf{I})$

Then, using the fitted values we estimate and do tests on β.

• Finding a **Z** that meets both requirements is not easy.

- The valid condition is not that complicated to meet.
- The relevant condition is more complicated: Finding a **Z** correlated with **X**. But, the explanatory power of **Z** may not be enough to allow inference on β. In this case, we say **Z** is a *weak* instrument.

Weak Instruments: Finance application

• Finance example: The consumption CAPM.

• In both linear and nonlinear versions of the model, IVs are weak, -- see Neeley, Roy, and Whiteman (2001), and Yogo (2004).

• In the linear model in Yogo (2004):

X (endogenous variable): consumption growth

Z (the IVs): twice lagged nominal interest rates, inflation, consumption growth, and log dividend-price ratio.

• But, log consumption is close to a random walk, consumption growth is difficult to predict. This leads to the IVs being weak.

=> Yogo (2004) finds *F-statistics* for H_0: $\boldsymbol{\Pi} = 0$ in the 1st stage regression that lie between 0.17 and 3.53 for different countries.

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 $\mathbf{x}_k$ on **Z**. Staiger and Stock (1997) suggest that $F < 10$ is a sign of problems.
  - Low partial-$R^2_{X,Z}$.
  - Large Var[$\mathbf{b}_{IV}$] 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($\mathbf{Z'}\varepsilon/T$) zero, is close to being violated.
- Detection of instrument exogeneity:
  - Endogenous IV's: Inconsistency of $\mathbf{b}_{IV}$ that makes it no better (and probably worse) than $\mathbf{b}_{OLS}$
  - 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

• An extremum estimator is one obtained as the optimizer of a criterion function, q(**z**,**b**).

Examples:

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

MLE: $\mathbf{b}_{MLE}$ = arg max $ln\, L = \sum_{i=1,\ldots,T} ln\, f\,(y_i,\mathbf{x}_i,\mathbf{b})$

GMM: $\mathbf{b}_{GMM}$ = arg max $- \mathbf{g}(y_i,\mathbf{x}_i,\mathbf{b})'\, \mathbf{W}\, \mathbf{g}(y_i,\mathbf{x}_i,\mathbf{b})$

• 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:

$$\min_{\mathbf{b}} E[q(\mathbf{z},\boldsymbol{\beta})]$$

– 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: $\mathbf{b}$ = 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

$$\sum_i \mathbf{s}(\mathbf{z}_i,\mathbf{b}_M)/T = 0 \qquad (*)$$

• If, in addition, $E_X[\mathbf{s}(\mathbf{z},\mathbf{b})] = \partial/\partial\mathbf{b}'\, E_X[q(\mathbf{z},\mathbf{b})]$ -i.e., differentiation and integration are exchangeable-, then

$$E_X[\partial q(\mathbf{z},\boldsymbol{\beta})/\partial\boldsymbol{\beta}'] = \mathbf{0}.$$

• Under these assumptions, the M-estimator is said to be of *ψ-type* (ψ= $\mathbf{s}(\mathbf{z},\mathbf{b})$=score). Often, $\mathbf{b}_M$ 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-Estimation: LS & ML

- Least Squares
  - DGP: $\mathbf{y} = f(\mathbf{x},\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$, $\qquad \mathbf{z} = [\mathrm{y},\mathbf{x}]$
  - $\mathrm{q}(\mathbf{z};\boldsymbol{\beta}) = \mathrm{S}(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \sum_{i=1,\ldots,T}(y_i - f(\mathbf{x}_i;\boldsymbol{\beta}))^2$
  - Now, we move from population to sample moments
  - $\mathrm{q}(\mathbf{z};\mathbf{b}) = \mathrm{S}(\mathbf{b}) = \mathbf{e}'\mathbf{e} = \sum_{i=1,\ldots,T}(y_i - f(\mathbf{x}_i;\mathbf{b}))^2$
  - $\mathbf{b}_{\mathrm{NLLS}} = \mathrm{argmin}\ \mathrm{S}(\mathbf{b})$
- Maximum Likelihood
  - Let $f(\mathbf{x}_i,\boldsymbol{\beta})$ be the pdf of the data.
  - $L(\mathbf{x},\boldsymbol{\beta}) = \Pi_{i=1,\ldots,T}\, f(\mathbf{x}_i;\boldsymbol{\beta})$
  - $\log L(\mathbf{x},\boldsymbol{\beta}) = \sum_{i=1,\ldots,T} \ln f(\mathbf{x}_i;\boldsymbol{\beta})$
  - Now, we move from population to sample moments
  - $\mathrm{q}(\mathbf{z},\mathbf{b}) = -\log L(\mathbf{x},\mathbf{b})$
  - $\mathbf{b}_{\mathrm{MLE}} = \mathrm{argmin} - \log L(\mathbf{x};\mathbf{b})$

M-Estimators: Properties

• Under general assumptions, M-estimators are:

- $\mathbf{b_M} \xrightarrow{p} \mathbf{b_0}$
- $\mathbf{b_M} \xrightarrow{a} N(\mathbf{b_0},\mathrm{Var}[\mathbf{b_0}])$
- $\mathrm{Var}[\mathbf{b_M}] = (1/T)\ \boldsymbol{H_0}^{-1}\boldsymbol{V_0}\,\boldsymbol{H_0}^{-1}$
- If the model is correctly specified: $-\mathbf{H} = \mathbf{V}$.
  Then, $\mathrm{Var}[\mathbf{b}] = \boldsymbol{V_0}$

– $\mathbf{H}$ and $\mathbf{V}$ are evaluated at $\mathbf{b_0}$:

- $\mathbf{H} = \sum_i [\partial^2 \mathrm{q}(\mathbf{z}_i,\mathbf{b})/\partial\mathbf{b}\partial\mathbf{b}']$
- $\mathbf{V} = \sum_i [\partial \mathrm{q}(\mathbf{z}_i,\mathbf{b})/\partial\mathbf{b}][\partial \mathrm{q}(\mathbf{z}_i,\mathbf{b})/\partial\mathbf{b}']$

Nonlinear Least Squares: Example

Example: $\mathrm{Min}_{\boldsymbol{\beta}}\ \mathrm{S}(\boldsymbol{\beta}) = \{\tfrac{1}{2}\, \Sigma_i\, [y_i - f(\mathbf{X}\beta)]^2\}$

• From the f.o.c., we cannot solve for $\boldsymbol{\beta}$ explicitly. But, using some steps, we can still minimize RSS to obtain estimates of $\boldsymbol{\beta}$.

• Nonlinear regression algorithm:

1. Start by guessing a plausible values for $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^0$.
2. Calculate RSS for $\boldsymbol{\beta}^0$ => get RSS($\boldsymbol{\beta}^0$)
3. Make small changes to $\boldsymbol{\beta}^0$, => get $\boldsymbol{\beta}^1$.
4. Calculate RSS for $\boldsymbol{\beta}^1$ => get RSS($\boldsymbol{\beta}^1$)
5. If RSS($\boldsymbol{\beta}^1$) < RSS($\boldsymbol{\beta}^0$) => $\boldsymbol{\beta}^1$ becomes your new starting point.
6. Repeat steps 3-5 until you RSS($\boldsymbol{\beta}^j$) cannot be lowered. => get $\boldsymbol{\beta}^j$.

=> $\boldsymbol{\beta}^j$ is the (nonlinear) least squares estimates.

NLLS: Linearization

• We start with a nonlinear model: $\quad y_i = f(\mathbf{x}_i,\boldsymbol{\beta}) + \varepsilon_i$

• We expand the regression around some point, $\boldsymbol{\beta}^0$:

$$\begin{aligned} f(\mathbf{x}_i,\boldsymbol{\beta}) &\approx f(\mathbf{x}_i,\boldsymbol{\beta}^0) + \Sigma_k[\partial f(\mathbf{x}_i,\boldsymbol{\beta}^0)/\partial\beta_k^0](\beta_k - \beta_k^0) \\ &= f(\mathbf{x}_i,\boldsymbol{\beta}^0) + \Sigma_k\, \mathbf{x}_i^0\,(\beta_k - \beta_k^0) \\ &= [f(\mathbf{x}_i,\boldsymbol{\beta}^0) - \Sigma_k\, \mathbf{x}_i^0\,\beta_k^0] + \Sigma_k\, \mathbf{x}_i^0\,\beta_k \\ &= f^0 + \Sigma_k\, \mathbf{x}_i^0\,\beta_k = f^0 + \mathbf{x}_i^{0\prime}\,\boldsymbol{\beta} \end{aligned}$$

where

$$f_i^0 = f(\mathbf{x}_i,\boldsymbol{\beta}^0) - \mathbf{x}_i^{0\prime}\,\boldsymbol{\beta}^0 \qquad (f_i^0 \text{ does not depend on unknowns})$$

Now, $f(\mathbf{x}_i,\boldsymbol{\beta})$ is (approximately) linear in the parameters. That is,

$$y_i = f_i^0 + \mathbf{x}_i^{0\prime}\,\boldsymbol{\beta} + \varepsilon_i^0 \qquad (\varepsilon_i^0 = \varepsilon_i + \text{linearization error i})$$

$$=> y_i^0 = y_i - f_i^0 = \mathbf{x}_i^{0\prime}\,\boldsymbol{\beta} + \varepsilon_i^0$$

NLLS: Linearization

• We linearized $f(\mathbf{x}_i,\boldsymbol{\beta})$ to get:

$$\mathbf{y} = \mathbf{f}^0 + \mathbf{X}^0\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}^0 \qquad (\boldsymbol{\varepsilon}^0 = \boldsymbol{\varepsilon} + \text{linearization error})$$

$$=> \mathbf{y}^0 = \mathbf{y} - \mathbf{f}^0 = \mathbf{X}^0\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}^0$$

• Now, we can do OLS:

$$\mathbf{b}_{\mathrm{NLLS}} = (\mathbf{X}^{0\prime}\,\mathbf{X}^0)^{-1}\,\mathbf{X}^{0\prime}\,\mathbf{y}^0$$

Note: $\mathbf{X}^0$ are called *pseudo-regressors*.

• In general, we get different $\mathbf{b}_{\mathrm{NLLS}}$ for different $\boldsymbol{\beta}^0$. An algorithm can be used to get the *best* $\mathbf{b}_{\mathrm{NLLS}}$.

• We will resort to numerical optimization to find the $\mathbf{b}_{\mathrm{NLLS}}$.

NLLS: Linearization

• Compute the asymptotic covariance matrix for the NLLS estimator as usual:

$$\text{Est. Var}[\mathbf{b}_{\mathrm{NLLS}}\,|\,\mathbf{X}^0] = s^2_{\mathrm{NLLS}}\,(\mathbf{X}^{0\prime}\,\mathbf{X}^0)^{-1}$$

$$s^2_{\mathrm{NLLS}} = [\mathbf{y} - f(\mathbf{x}_i,\mathbf{b}_{\mathrm{NLLS}})]'\,[\mathbf{y} - f(\mathbf{x}_i,\mathbf{b}_{\mathrm{NLLS}})]/(T-k).$$

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

Gauss-Newton Algorithm

• $\mathbf{b}_{NLLS}$ depends on $\boldsymbol{\beta}^0$. That is,

$\mathbf{b}_{NLLS}\ (\boldsymbol{\beta}^0) = (\mathbf{X}^{0\prime}\,\mathbf{X}^0)^{-1}\,\mathbf{X}^{0\prime}\,\mathbf{y}^0$

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

$\beta_{k+1} = \beta_k + (\mathbf{J}^T\,\mathbf{J})^{-1}\,\mathbf{J}^T\,\boldsymbol{\varepsilon}$ -- $\mathbf{J}$: Jacobian = $\delta f(xi;\boldsymbol{\beta})/\delta\boldsymbol{\beta}$.

• Given a $\mathbf{b}_{NLLS}$ at step m, $\mathbf{b}(j)$, we find the $\mathbf{b}_{NLLS}$ for step j+1 by:

$\mathbf{b}(j+1) = \mathbf{b}(j) + [\mathbf{X}^0(j)'\mathbf{X}^0(j)]^{-1}\mathbf{X}^0(j)'\boldsymbol{e}^0(j)$

Columns of $\mathbf{X}^0(j)$ are the derivatives: $\partial f(\mathbf{x}_i,\mathbf{b}(j))/\partial\mathbf{b}(j)'$

$\mathbf{e}^0(j) = \mathbf{y} - f[\mathbf{x},\mathbf{b}(j)]$

• 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.

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

$\mathbf{x}_k^{(\lambda)} = (\mathbf{x}_k^{\lambda} - 1)/\lambda$ $\quad\lim_{\lambda\to 0}\ (\mathbf{x}_k^{\lambda} - 1)/\lambda = \ln\,\mathbf{x}_k$

• 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:

$\mathbf{y}^{(\lambda 1)} = \Sigma_k\,\mathbf{x}_k^{(\lambda 2)}\,\beta_k + \boldsymbol{\varepsilon}$

Testing non-linear restrictions

• Testing linear restrictions as before.

• Non-linear restrictions change he usual tests. We want to test:

H_0: $R(\boldsymbol{\beta}) = 0$

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

• Let $\mathbf{m} = R(\mathbf{b}_{NLLS}) - \mathbf{0}$.

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

But, we do not know the distribution of $R(\mathbf{b}_{NLLS})$. We know the distribution of $\mathbf{b}_{NLLS}$. Then, we linearize $R(\mathbf{b}_{NLLS})$ around $\boldsymbol{\beta}$:

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

Testing non-linear restrictions

• Linearize $R(\mathbf{b}_{NLLS})$ around $\boldsymbol{\beta}$ $(=\mathbf{b}_0)$

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

• Recall $\sqrt{T}\,(\mathbf{b}_M - \mathbf{b}_0) \xrightarrow{d} N(\mathbf{0}, \text{Var}[\mathbf{b}_0])$

where $\text{Var}[\mathbf{b}_0] = \boldsymbol{H}(\boldsymbol{\beta})^{-1}\,\boldsymbol{V}(\boldsymbol{\beta})\,\boldsymbol{H}(\boldsymbol{\beta})^{-1}$

$\Rightarrow \sqrt{T}\,[R(\mathbf{b}_{NLLS}) - R(\boldsymbol{\beta})] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{G}(\boldsymbol{\beta})\,\text{Var}[\mathbf{b}_0]\,\boldsymbol{G}(\boldsymbol{\beta})')$

$\Rightarrow \text{Var}[R(\mathbf{b}_{NLLS})] = (1/T)\ \boldsymbol{G}(\boldsymbol{\beta})\,\text{Var}[\mathbf{b}_0]\,\boldsymbol{G}(\boldsymbol{\beta})'$

• Then,

$W = T\,R(\mathbf{b}_{NLLS})'\{\boldsymbol{G}(\mathbf{b}_{NLLS})\,\text{Var}[\mathbf{b}_{NLLS}]\,\boldsymbol{G}(\mathbf{b}_{NLLS})'\}^{-1}\,R(\mathbf{b}_{NLLS})$

$\Rightarrow W \xrightarrow{d} \chi_J^2$