Class Notes for Spring 2021
Advanced Econometrics

AGEC 6213 - ADVANCED ECONOMETRICS

Purpose: Learn to use econometric techniques in applied research. Not just applying techniques to data, etc.

Testing economic theory using real world data.

Economists get their data from the real world rather than a controlled experiment. Since real world data are from a poorly designed experiment, we need econometric techniques. (Hard sciences use designed experiments where the statistics are easy.)

Prerequisite: AGEC 5213 or ECON 6213

Econometrics II in Econ department (ECON 6243) is even better. It at least used to include

Cross-section time series/ Seemingly unrelated regression

Instrumental variables/Simultaneous equations

limited dependent variables

regression discontinuity designs

propensity score matching

ECON 6243 and AGEC 6213 are designed to not overlap

Math Stat is strongly recommended, but not required

How does econometrics compare to other techniques?

Two types of techniques:

1. Normative model: has an explicit objective (preference) function - prescribe
2. Positive model: without an explicit objective (preference) function - describe

Normative

1. Linear programming
2. Quadratic programming MOTAD, TARGET MOTAD
3. Nonlinear programming

primarily static

4. Discrete stochastic programming]
5. Dynamic programming, Optimal control

dynamic

6. Comparative statics

Positive

1. Input-output models
2. Budgeting
3. Computable general equilibrium models

static

4.Simulation	dynamic
5.Econometric models (structural)	either
6.Time series models	dynamic
7.Spectral analysis	dynamic

Two main purposes of econometric modeling

1. Policy analysis → tests of hypotheses and estimates of coefficients→ statistical properties

   We will concentrate here

2. Forecasting→ does it work? → out-of-sample forecasts

Will cover both - *they are related.*

- ✓ A model with poor statistical properties is likely a poor forecaster and vice versa.
- ✓ Choice of method depends both on your objective (optimal) and data availability.
- ✓ In this course, we will learn to select among econometric techniques, but you will find cases where techniques outside of econometrics are appropriate.

I. Review of Econometrics

Based on the General Linear Model

$$Y = X\beta + e$$

where

Y is a T x 1 vector of observations on the dependent variable,

X is a T x k matrix of observations on k exogenous (independent) variables

β is a k x 1 parameter vector, and

e is a T x 1 vector of random disturbances.

Estimate: A number selected according to some choice criteria that is used to represent an unknown parameter value (β).

Problem: Need estimates of the unknowns β and $\sigma^2 = \mathrm{var}(e) \forall i$

Need some criteria to select among possible estimators.

I. a. Desirable Properties

Small Sample Properties

1. Unbiased

   The estimator $\boldsymbol{\hat{\beta}}$ is unbiased if $E\ (\hat{\beta}) = \beta$

2. BLUE (Best Linear Unbiased Estimator)

   (i) *Unbiased*

   (ii) *Linear* $\hat{\beta} = \sum_{i=1}^{T} a_i y_i$

   (Notice that it is the estimator that is linear.)

   (iii) Have *smaller variance* than any other linear unbiased estimator

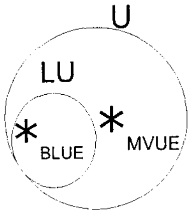
   ($\boldsymbol{COV}(\boldsymbol{\widetilde{\beta}}) - \boldsymbol{COV}(\boldsymbol{\hat{\beta}})$) is p.s.d., where $\boldsymbol{\tilde{\beta}}$ is any other linear unbiased estimator)

3. Efficient (MVUE: Minimum Variance Unbiased Estimator)

   (i) Unbiased

   (ii) Have smaller variance than any other unbiased estimator

Note: Helpful way to remember the meaning of BLUE
U is the set of unbiased estimators
LU is the set of linear unbiased estimators; therefore, LU estimators are a subset of U estimators.
BLUE is the estimator in LU with minimum variance, and MVUE is the minimum variance estimator in the broader set U.
Which one has a 'smaller' variance? Can they be equal?



Large Sample Properties

4. Consistency

An estimator $\hat{\beta}_T$ is consistent if

$$\lim_{T\to\infty} P(|\hat{\beta}_T - \beta| < \delta) = 1 \quad \text{(convergence in probability)}$$

where T is sample size, $\hat{\beta}_T$ is the estimate of β based on a sample of size T, and $\delta > 0$ is some arbitrarily small number.

*****_Consistency_ is _the minimum for a classical econometrician._
****** _Note that MS.E. consistency_ is _a stronger condition and that the above_ is _a probability limit_ (plim($\hat{\beta}_T$)= β).

5. Asymptotic Efficiency

(i) Consistent

(ii) $\hat{\beta}$ has smaller asymptotic variance than any other consistent estimator

Under the central limit theorem, the asymptotic distribution of the asymptotically efficient estimator is

$$\sqrt{T}\left(\hat{\beta}-\beta\right)\xrightarrow{d} N\left(0, \lim_{T\to\infty}\left[\frac{1}{T}I\left(\beta\right)\right]^{-1}\right)$$

All of these properties seem desirable,

When do ordinary least squares (OLS) estimates have these properties?

I. b. Assumptions of OLS

Existence assumptions

(i) y and X are observed

(ii) rank (X) = k => *no perfect multicollinearity*

(iii) k < T => more observations than explanatory variables

If any of these 3 assumptions are violated, OLS estimates either do not exist or are not unique.

Assumptions of classical linear regression model (Kmenta level)

1) The errors are distributed normally. (Used in the proof of MVUE and in hypothesis testing)

2) The expected value of the error term is zero.

3) Homoskedasticity (constant variance of the error term)
$\sigma_i^2 = \sigma_j^2 \forall i, j$
4) No autocorrelation
$\text{cov}(e_t, e_s) = 0 \quad \text{if } t \neq s$

same problem

(missing lagged error term as an explanatory variable)

5) x's are non-stochastic - 3 levels
↗ *no problem*
→ *lagged endogenous*
↘ *simultaneity*

Under these assumptions, OLS estimates have all 5 desirable properties.

- Which assumptions are most important? (Only 2 and 5 are needed to prove consistency and unbiasedness)
- Which are least important?

Econometrics deals with what should be done when these assumptions do not hold What your quiz is about.

I. c. Violation of Assumptions (Handout)

I. d. Asymptotic convergence and some mathematical concepts
Need this to read Econometrica - (high level)

Supremum (sup): Lowest upper bound

Maximum (max): Largest member of a set

example:

A = (0, 1)	max A does not exist	sup A = 1
B = (0, 1]	max B=1	sup B = 1

Similar relationship for minimum and infimum.

Often see sup of likelihood instead of max - assures existence even with strict inequalities.

Measures

✓ Counting measure (μ_c): The number of elements in a set - discrete dist.

✓ Lebesgue measure (μ_L) Length (or area for 2 dimensions) covered by the set. *Neat notation to write mixed (discrete and continuous) distribution functions with integrals.*

Example (continued) $\mu_c(A) = \infty,\ \mu_L(A) = 1,\ \mu_L(B) = 1$

We say that *A* equals *B* everywhere except on a set of Lebesgue measure zero, and write:

A = *B* a.e. $[\mu_L]$ (a.e.=almost everywhere)

$$C = [c: c \in A \cup B, c \notin A \cap B] = \{1\}$$

$$\mu_{L\,(C)} = 0.$$

Probability

A probability is a special kind of measure. Articles often refer to a probability measure.

Sample Space Ω , the set of all possible outcomes.

A commonly used sample space is Ω = R.

Event *A*, subset of Ω for which a probability is defined.

Sigma algebra F, a set of events closed under union, intersection, and complementation.

A probability is a function on events, P: F →[0, 1], such that:

- ✓ P(Ω) = 1
- ✓ P(*A)*≥ 0 for A ∈ F
- ✓ Countably additive $P\left(\bigcup_{i=1}^{\infty} A_1\right) = \sum_{i=1}^{\infty} P\left(A_{i)}\right)$ if $A_1 \cap A_j = \emptyset \; \forall \; i \neq j$

Modes of Convergence

Uniform Convergence

$\hat{\theta}_T(X) \to \theta$ as T→∞, ∀ X

(X, set of random observations)

$P\ [\lim_{T\to\infty}|\ \hat{\theta}_T(X) - \theta\ \ |<\varepsilon]$

$\frac{1}{T}\sum_{i=1}^{T} X_{i\to\mu}$ example

∞

Almost Sure Convergence

$\hat{\theta}_T(X) \to \theta$ a.s. as T→∞

$P\ [\lim_{T\to\infty}|\ \hat{\theta}_T(X) - \theta\ \ |>\varepsilon]$

means the set of outcomes X for which the estimators do not converge has zero probability of occurrence ⟹

- based on SLLN

⟹

Convergence in L_p (If p>1) ⟹

$$\lim_{T\to\infty} E\left|\hat{\theta}T(X) - \ \theta\right|^p = 0$$

- p=1⇒absolute value
- p=2 mean squared error

$$\Longrightarrow \lim_{T\to\infty}\left\{bias^2 + \ var\left(\hat{\theta}_T(X)\right)\right\} = 0$$

Most intuitive definition of consistency

Convergence in Probability ⟹

$$\hat{\theta}_T(X) \xrightarrow{p} \theta$$

$$\lim_{T\to\infty} P\left[\left|\hat{\theta}_T(X) - \theta\right| > \varepsilon\right] = 0$$

means the probability of getting a set of observations X that does not converge to the true parameter goes to zero as the same size increases.

- based on WLLN
- This is our def. of consistency.
- Econometrica uses a.s. Why? Easier to prove and it is a strong result.

Convergence in Distribution

$$\hat{\theta}_T(X) \xrightarrow{d} \theta \sim \mathrm{F}(\theta)$$

The estimator converges in distribution to a random variable with CDF F(θ) if

$\hat{\theta}_T(X) \sim F_T\theta, T = 1, 2$ ...and

$\lim_{T\to\infty} F_T(\theta) = F(\theta), \forall\theta$

- will use for hypothesis testing

II. Estimation Methods

II. a. Minor Methods:

1) Minimize total effort or cost in choosing estimates.

Assume β = b and $\sigma^2 = 0$ **-** sometimes used when pushed for time and used to estimate parameters in LP models, simulation models, many policy analyses.

2) Minimize $\sum_i \left|\hat{e}_j\right|$ - Mean Absolute Deviation (MAD) estimator

Least Absolute Value (LAV) estimator; book calls it (ℓ_1) sometimes called L_1- norm

Robust estimator: based on median, no distributional assumption for error term

Mathematical programming problem - difficult computationally

- Not necessarily a unique solution - serious problem
- Less sensitive to outliers than least squares - Why?
- Hypothesis tests are only valid asymptotically (asymptotic normality of parameter estimates is assumed for hypothesis tests).

Proc Robustreg in SAS has several substitute methods. If you face a problem with severe nonnormality, Proc Robustreg has approaches that you should consider.

M, LTS, S, and MM; Call LAV in PROC IML

3) Methods of Moments

Estimate a moment of the distribution by the moment of the sample. What is a sample moment?

if X_1 -~ uniform (0,*b*), mean = *b*/2, variance = $b^2/12$

$$\hat{b}_1 = 2\bar{X}, \qquad \hat{b}_2 = \left(12 \sum_j (X_i - \bar{X})^2/N\right)^{1/2}$$

- Not unique.
- Consistent - can be used as starting values for nonlinear maximum likelihood

II. b. Least Squares

Minimize $\sum_{t=1}^{T} \hat{e}_t^2$

$$\min S(b) = (Y - Xb)^{/}(Y - Xb) = Y^{/}Y - 2\, b^{'}X^{'}Y + b^{'}X^{'}X\, b$$

$$\frac{\partial\, S}{\partial\, b} = -2X^{'}Y + 2X^{'}Xb = 0$$

$$b = (X'X)^{-1}\, X^{'}Y = \text{argmin } (S)$$

II. c. Maximum Likelihood - important

Assume we know the density function (data generating process)

$$f(y_t|X_t, \beta, \sigma^2) \; t = 1, \dots, T$$

Or joint density function

$$f(y|X, \beta, \sigma^2)$$

What is a density function?

If observations are independent - *(meaning?)* - we have:

$$f\; (y \mid X_t, \beta, \sigma^2) = \prod_{t=1}^{T} f\; (y_t|X_t, \beta, \sigma^2)$$

What if not independent?

Likelihood function: let β, σ^2 be variables and take *y* and *X* as given

$$\ell(\beta, \sigma^2|y, X) = f(y|X, \beta, \sigma^2)$$

Parameters selected are those values of β and σ^2 that maximize the probability of having obtained the sample (*X*, *y*); i.e., maximize

$\ell(\beta, \sigma^2|y, X)$

How do you find the maximum of a function?

Let $\delta = (\beta, \sigma^2)$ and $\tilde{\delta} = \left(\tilde{\beta}, \tilde{\sigma}^2\right)$ be the M.L.E

$$\Rightarrow \tilde{\delta} = \text{argmax}\,[\ell(\beta, \sigma^2 | y, X)]$$

How do we get standard errors?

If the likelihood function is twice differentiable (weaker assumptions are possible), then the information matrix $\ell(\delta)$ is:

$$I(\delta) = -E\left[\frac{\partial^2 \ln \ell\,(\delta | y, x)}{\partial \delta\, \partial \delta'}\right] = -EH(\delta; y, x)$$

Three methods of estimating $I(\delta)$:

(i) $I_1\left(\tilde{\delta}\right) = -H\left(\tilde{\delta}; y, x\right)$ (Hessian)

(ii) $I_2\left(\tilde{\delta}\right) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \ln \ell(\tilde{\delta} | y_t, x_t)}{\partial \delta} \cdot \frac{\partial \ln \ell(\tilde{\delta} | y_t, x_t)'}{\partial \delta}$

BHHH outer (cross) product of the gradients

$(iii) I_3\left(\tilde{\delta}\right) = -EH\left(\tilde{\delta}\right)$, Estimated Information Matrix used with method of scoring.

What about BFGS? DFP?

Under appropriate regularity conditions (it is sufficient that $\ell \in c^2$),

$$\sqrt{T}\left(\tilde{\delta} - \delta\right) \xrightarrow{d} N\left[0, \lim_{T\to\infty}\left(\frac{I(\delta)}{T}\right)^{-1}\right]$$

Why not $\tilde{\delta} \overset{a}{\sim} N\,(\delta, I(\delta)^{-1})$?
Because $\tilde{\delta}$ converges in probability to δ ($\tilde{\delta}$ is consistent) and so the distribution of X will converge to a degenerate distribution at δ, and not to a normal

Estimates of standard errors are the square root of the diagonal of the inverse of the estimated I(δ)

Example

Consider the normal distribution:

$$f(y_t|X_t,\beta,\sigma^2) = (2\Pi\sigma^2)^{-1/2} exp\left[-\frac{(y_t - X_t'\beta)^2}{2\sigma^2}\right]$$

If independent then:

$$f(y|X,\beta,\sigma^2) = \prod_{t=1}^{T}(2\Pi\sigma^2)^{-1/2}\exp\left[-\frac{(y_t - X_t'\beta)^2}{2\sigma^2}\right]$$

$$= (2\Pi\sigma^2)^{-T/2}\ \exp\left[-\frac{(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta)}{2\sigma^2}\right]$$

$$\ell(\beta,\sigma^2|y,X) = (2\Pi\sigma^2)^{-T/2}\ \exp\left[-\frac{(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta)}{2\sigma^2}\right]$$

Take the natural logarithm (Why?)

$$\max_{\beta,\sigma^2}(\beta,\sigma^2|y,X)\ =\ \ln\ell(\beta,\sigma^2|y,X) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\sigma^2 - \frac{(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta)}{2\sigma^2}$$

Same as least squares

$$\frac{\partial L}{\partial\beta} = -\frac{1}{2\sigma^2}(-2X'y + 2X'X\beta) = 0 \Longrightarrow \tilde{\beta} = \left(X^{/}X\right)^{-1}X'y$$

$$\frac{\partial L}{\partial\sigma^2} = -\frac{T}{2\sigma^2} + \frac{(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta)}{2(\sigma^2)^2} = 0 \Longrightarrow \tilde{\sigma}^2 = \frac{(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta)}{T} \qquad \text{(biased)}$$

(s.o.c.)

$$\begin{bmatrix} \dfrac{-X'X}{\sigma^2} & \dfrac{-X'X\beta}{\sigma^4} = -\dfrac{1}{\sigma^4}(X'y - X'X\beta) \\ -\dfrac{1}{\sigma^4}(X'y - X'X\beta) & \dfrac{T}{2\sigma^4} - \dfrac{1}{\sigma^6}(y - X\beta)'(y - X\beta) \end{bmatrix}$$

Off diagonals are zero based on foc.

$$\Rightarrow I(\delta)^{-1} = \begin{bmatrix} \sigma^2 \, (X'X)^{-1} & 0 \\ 0 & \dfrac{2\sigma^4}{T} \end{bmatrix}$$

Often use concentrated log likelihood function.

$$L(\beta|y,X) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln(y - X\beta)'(y - X\beta)/T + \frac{T}{2}\ln T$$

$$-\frac{(y - X\beta)'(y - X\beta)}{2} \cdot \frac{T}{(y - X\beta)'(y - X\beta)}$$

$$= C - \frac{T}{2}\ln(y - X\beta)'(y - X\beta)$$

Minimizing the sum of squares.

Example- Multiplicative Heteroskedasticity

$$y_t = X_t\beta + e_t$$

$$e_t \sim N(0, \sigma_t^2)$$

$$\sigma_t^2 = \exp(\mathrm{Z}_\mathrm{t}'\alpha)$$

Log-likelihood for multivariate normal is

$$L(\beta, \alpha|y, X, Z) == -\frac{T}{2}\ln 2\,\pi - \frac{1}{2}\ln|\Phi| - \frac{1}{2}(y - X\beta)'\Phi^{-1}(y - X\beta)$$

Multiplicative heteroskedasticity is a special case where off-diagonal elements of Φ are zero and the t^{th} diagonal elements is $\sigma_t^2 = \exp(Z_\mathrm{t}'\alpha)$

By substitution

$$L(\beta, \alpha|y, X, Z) = -\frac{T}{2}\ln 2\,\pi - \frac{1}{2}\sum_{t=1}^{T} Z_t'\alpha - \frac{1}{2}\sum_{t=1}^{t} \exp(-Z_\mathrm{t}'\alpha)(y_t - X_t\beta)^2$$

$$I(\beta, \alpha) = \begin{bmatrix} X'\Phi^{-1}X & 0 \\ 0 & \dfrac{1}{2}Z'Z \end{bmatrix}$$

Greene (p. 524) suggests method of scoring to find the MLE. Why?

Note: EGLS provides asymptotically efficient estimates of β but only consistent estimates of α

∴ should use MLE when interested in α

II. d. Generalized method of moments (GMM)

Objective: Know advantages and disadvantages of GMM vs. MLE

- ✓ Increasingly used in finance and economics
- ✓ Get a unique method of moments estimator
- ✓ For notation, follow Greene.

General nonlinear model

$y_i = h\ (x_i, \theta) + \ \varepsilon_i$, where $x_i\ is\ a\ K\ \times 1$ vector

We assume that

$E(\varepsilon\ ) = 0$

$E(\varepsilon\varepsilon') = \Omega$ Ω is unrestricted

$E(x'\varepsilon) = 0$ These are the moment equations

and no distributional assumptions are made. Makes less assumption- Good or Bad?

The GMM estimator for θ *based on the explanatory variables X's* is such that is satisfies

$m(\theta_{GMM}) = \frac{1}{n}\sum_i x_i\ e(x_i, \theta_{GMM}) = 0$ (Sample moments)

or

$m(\theta_{GMM}) = \frac{1}{n}\ X'\ e(X, \theta_{GMM}) = 0$ (i.e., sample moments are zero)

Based on this definition, OLS is a GMM estimator.

Generalized Instrumental Variables

When dealing with a simultaneous equations system, a set of $J \geq K$ instrumental variables is likely to be used. Then, the sample moments are

$$m(\theta) = \frac{1}{n}\sum_i z_i\, e(x_i, \theta) = \frac{1}{n}\, Z' e(X, \theta)$$

where z_i is a $J\times 1$ vector of instrumental variables orthogonal to e.

When exactly identified, $E(e|z) = 0$ the solution is unique.

If there are more equations than parameters to estimate, it will no longer be possible to solve

$m(\theta) = \frac{1}{n}\, Z' e(X, \theta) = 0$ (for θ) as before. But if θ cannot be chosen such that

$m(\theta)$=0, we would like it to be as close to zero as possible. Thus, θ_{GMM} will be such that it

minimizes a 'standardized' quadratic form of m(θ).

The minimum distance estimator is $\theta_{GMM} = \hat{\theta}$ such that

$$\min_{\hat{\theta}} q = m(\hat{\theta})' W^{-1}\, m(\hat{\theta}) = (1/n^2)\, [e(X, \theta)' Z] W^{-1}\, [Z' e(X, \hat{\theta})]$$

An optimal choice of W is

$$\begin{aligned} W_{GMM} &= asy \text{ var } [m(\theta)] \\ &= \left(\frac{1}{n^2}\right) Z' \Omega\, Z \qquad \text{(for our model)} \\ &= \frac{1}{n^2} \textstyle\sum_i \sum_j z_i z_j' \text{cov}\big[y_i - h(X_i, \theta), y_j - h(X_i, \theta)\big] \quad \text{(seems to require estimate of } \theta) \end{aligned}$$

If observations are independent (White's heteroskedasticity consistent covariance matrix, *White 1980)*

$$w_{GMM} = \frac{1}{n}\left[\frac{1}{n}\sum_i z_i z_i' \left(y_i - h(x_i, \hat{\theta})\right)^2\right]$$

White uses

- ✓ OLS estimates of θ (consistent), available in SAS Proc Reg with WHITE option
- ✓ GMM estimate of covariance (consistent even with heteroskedasticity)

Newey-West estimator

- ✓ If you allow autocorrelation – (assume truncated at some lag length)
- ✓ Consistent estimate of covariance

GMM estimators are available in Proc Model of SAS
GMM estimators are weighted minimum distance estimators

The Newey-West kernel is the same as the Bartlett kernel with bandwidth parameter

$l(n) = L + 1$. That is, if the "lag length" for the Newey-West kernel is L then the Newey-West estimator is obtained using KERNEL = (bart, L+1, 0).

The GMM estimator is efficient in the class of instrumental variable estimators defined by the orthogonality conditions

Cluster robust standard errors are commonly used (sandwich estimator)

Kennedy (p. 417) says "Generalized method of moments is unifying view of econometric estimation that in theory is impressive, but in practice has not lived up to its promise."

Monte Carlo studies show that GMM estimators have poor small-sample properties. Newey-West and White's for example, reject too often in small samples (incorrect size).

Review what I want you to know:

- What it is.
- Why you would use it.
- Properties relative to MLE.
- White's estimator.
- Newey-West estimator.

III. Hypothesis Testing

III. a. Basic Concepts

$$Y = \beta + e$$

$$H_0: \beta = 0$$

$$H_A: \beta \neq 0$$

Concerned about:

Type I Error:

Reject H_0, when H_0 is in fact true. Depends on the distribution of $\hat{\beta}$ under H_0

Type II Error:

Fail to reject H_0 when H_0 is in fact false. Depends on the distribution of $\hat{\beta}$ under H_A, which depends on the true but unknown value of β.

Size of a test (α = 0.05):
probability of making a Type I error. Controlled by the researcher.

P[Type I error] = $P\left[\text{rejecting } H_0 \,\middle|\, H_0 \text{true}\right] \leq \alpha$ (p-value on SAS printout)

Power of a test {f (β)}:
1 − probability of making a Type II error

P[Type II error] = P[failing to reject $H_0 \,|\, H_A$true], function of the true value

Uniformly most powerful (UMP) or uniformly most powerful unbiased (UMPU) tests are only available for simple tests like a test of two means. U stands for *uniformly,* which refers to being most powerful *for all* alternative hypotheses.

Accept H_0 vs. Fail to Reject H_0

- ✓ In this class we only "fail to reject" null hypotheses.
- ✓ We do not accept null hypotheses.
- ✓ Research methodology: scientific paradigms are simply a set of <u>unrejected</u> hypotheses.

To test hypotheses, we must know the distribution of the test statistics under H_0 ∴ need to know
6. distributional theorems.

III. b. Review of Distributional Theorems

Needed to develop test of hypotheses.

If $X_{nx1} \sim N\left(\mu_{nx1,} \Sigma_{nxn}\right)$

Then $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2(n)$

Could be used to test $\mathrm{H}_0\colon \mu = \mu_0$ except we don't know Σ.

Can still use as an asymptotic test if know $\hat{\Sigma}$ and $\hat{\Sigma} \overset{\mathrm{p}}{\rightarrow} \Sigma$

Let *z*~N(0,1), and let
$\chi^2(r)$ denote a chi-square random variable with *r* degrees of freedom

z and $\chi^2(r)$ are independent, then

$$T = \frac{z}{\sqrt{\frac{\chi^2(r)}{r}}} \sim t(r)$$

Example:

$\mathrm{H}_0\colon \mu = 0$ $\mathrm{H}_\mathrm{A}\colon \mu \neq 0$

$X \sim N(\mu, \sigma^2)$

$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ where n is sample size $\left(\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\right)$

Standardize

$\frac{(\bar{X}-\mu)}{\sqrt{\sigma^2/n}} \sim N(0,1)$ $\qquad$ $\frac{\hat{\sigma}^2(n-1)}{\sigma^2} \sim \chi^2(n-1)$

Divide

$$\frac{\left[(\bar{X}-\mu)/\sqrt{\sigma^2/n}\right]}{\sqrt{\frac{\hat{\sigma}^2(n-1)/\sigma^2}{n-1}}} = \frac{\bar{X}-\mu}{\sqrt{\hat{\sigma}^2/n}} \sim t(n-1)$$

Let $\chi^2(r_1)$ and $\chi^2(r_2)$ be independent chi-squared random variables with $r_1$and r_2 degrees of freedom.

Then $F = \frac{\chi^2(r_1)/r_1}{\chi^2(r_2)/r_2} \sim F\ (r_1, r_2)$

Aside $\mathrm{t}^2(n) = \mathrm{F}(1, n), t(\infty) = z,\ F(1, \infty) = \chi_1^2, z^2 = \chi_1^2$

- Why can we use a t-test for testing a single restriction, but must use an F-test for multiple restrictions?
- Advantage of a t-test?
- Can do one-tailed tests. Why do we like one-tailed tests?

Distribution of a linear transformation – **important**

If $X \sim N\ (\mu, \Omega)$ and $Y = A\,X + b$

Rank(A) = $n \Longrightarrow n \leq m$ Why needed? So var-cov matrix of Y is not singular

Then $Y \sim N\ (A'\mu + b, A\Omega A'\ )$

Used in stochastic simulation (Monte Carlo studies)

Example: $X \sim N(0, I)\ \ Y = PX + \mu_y$
($X \rightarrow$ vector of standard normal random variables)
$Y \sim N\left(\mu_y, PP'\right)$

This theorem is also used to calculate standard errors of elasticities when the elasticity is a linear transformation of the parameters.

General version of:

$$E[ax + by] = aE(x) + bE(y)$$
$$\mathrm{var}[ax + by] = a^2\mathrm{var}(x) + b^2\mathrm{var}(y) + 2ab\mathrm{cov}(x, y)$$

Nonlinear transformation (delta method)
Used in testing nonlinear hypotheses and obtaining standard errors of nonlinear transformations of parameters.

If $X_{mx1} \sim N(\mu, \Omega)$ $\qquad Y_{nx1} = g(x), n \leq m$

1st order Taylor Series Expansion around μ

$$\begin{aligned} g(x) &\doteq g(\mu) + g'(\mu)(X - \mu) \\ E[g(x)] &\doteq g(\mu) + g'(\mu)\ (E(x) - \mu) = g(\mu) \\ Var[g(x)] &= E\{[g(x) - E(g(x))][g(x) - E(g(x))]^T\} \\ &\doteq E\{[g(\mu) + g'(\mu)(x - \mu) - g(\mu)][\cdot]^T\} \\ &= E\{[g'(\mu)(x - \mu)][g'(\mu)(x - \mu]^T\} \\ &= g'(\mu)\Omega g'(\mu)^T \end{aligned}$$

$\therefore Y \overset{a}{\sim} N(g(\mu), g'(\mu)\Omega g'(\mu)^T)$
When Ω and μ are unknown, substitute their consistent estimates $\widehat{\Omega}$ and $g'(\hat{\mu})$ to get consistent estimates of the variance-covariance matrix.

- ✓ From Slutsky Theorem, if *f* is a continuous function and $\hat{\theta} \xrightarrow{P} \theta$ then $f(\hat{\theta}) \xrightarrow{P} f(\theta)$
- ✓ Important for testing nonlinear hypotheses
- ✓ Since the delta method is sensitive to units of measurement, it is now considered inferior to bootstrap methods, but sometimes it is much easier to calculate.
- ✓ If $g'(\mu)$ is difficult to calculate analytically – can use numerical derivatives (a numerical derivative is calculated as $[g(x+\Delta) - g(x)]/\Delta$ where Δ is a small number).

III. c. <u>General Tests of Restrictions</u>

The Normal GLM

Y= Xβ + *e* Gauss-Markov assumptions plus normality.

Essentially the same as before and may be more familiar
Assume

Rank(X) = k < T
$(e/X) \sim N(0, \sigma^2 I_T) \Rightarrow E(e/X) = 0$ -mean zero,
$E(ee'/X) = \sigma^2 I_T$ -no autocorrelation, no heteroskedasticiy
(conditional on X gives non-stochastic X)

then

(1) $b = (X'X)^{-1}X'y \sim N\left[\beta, \sigma^2 (X'X)^{-1}\right]$
(2) $\frac{(T-k)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{T-k}$
(3) b and $\widehat{\sigma}^2$ are statistically independent

Now assume some prior information

$R_{Jxk}\beta_{kx1} = q_{Jx1}$ R and q are known, R has full row rank, J independent restrictions

Constrained Optimization Problem

$\min_{\beta} \sum_i e_i^2 = (Y - X\beta)'(Y - X\beta)$

s.t. *R*β=*q*

Examples

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \beta_3$

Assume $\beta_1 = 2\beta_3$ or $\beta_1 - 2\beta_3 = 0$

$$\begin{bmatrix} 0 & 1 & 0 & -2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = [0]$$

$\beta_1 = \beta_2 = \beta_3$- two independent restrictions $\beta_1 = \beta_2, \beta_2 = \beta_3$

$$\begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Constrained OLS estimator (b^*)

$b^* = b - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(Rb - q)$

✓ $E(b^*) \neq$ β unless the restriction is true ∴ biased in general

✓ $Var(b^*) \leq \mathrm{Var}(b)$

✓ MVUE (efficient) if restriction is true

✓ Could still be preferred on the basis of mean-squared error

$\tilde{\sigma}_0^2 = \frac{(y - Xb^*)'(y - Xb^*)}{T - k + J}$ (estimates of σ^2can be smaller or greater $\frac{\mathrm{SSE}\uparrow}{\mathrm{df}\uparrow}$

<u>Reasons to Use RLS (or RMLE)</u>

1) Believe restriction is true (from theory)
2) Sacrifice possibility of bias to get a lower variance

Note: Any regression is RLS (example: demand equation without sunspots)
We must derive some restrictions from theory in order to estimate anything.

Test the Truth of Restrictions in 3 Ways

1) Wald – based on unrestricted model
2) Likelihood ratio- based on restricted and unrestricted
3) Lagrange multiplier- based on the restricted model

Wald Test

$\mathrm{H_0}: \mathrm{R}\beta = \mathrm{q} \quad \mathrm{R}$ is $\mathrm{J} \times \mathrm{k}$ q is $\mathrm{J} \times 1$

$\mathrm{H_A}: \mathrm{R}\beta \neq \mathrm{q} \quad \beta$ is $\mathrm{k} \times 1$

Know $\tilde{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$ $\quad \tilde{\beta}$ is the MLE

$R\tilde{\beta} - q$ is a linear transformation

[use theorem if $X \sim N(\mu, \Sigma), Y = \mathrm{A}X + b, Y \sim N(\mathrm{A}\mu + b, \mathrm{A}\Sigma\mathrm{A}')$]

$\therefore (R\tilde{\beta} - q) \sim N\ (R\beta - q, \sigma^2 R (X'X)^{-1} R')$

[use theorem: if $X \sim N(\mu, \Sigma)$ then $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2$]

Under $\mathrm{H_0}: R\beta - \mathrm{q} = 0$

$$\left(R\tilde{\beta} - q\right)' \left[\sigma^2 R \left(X'X\right)^{-1} R'\right]^{-1} \left(R\tilde{\beta} - q\right) \sim \chi^2_J \text{ (Under } \mathrm{H_0}: R\beta - q = 0)$$

**could substitute $\hat{\sigma}^2$ and get an asymptotic test (Slutsky's theorem)

Use another distributional theorem

$\frac{\chi^2(r_1)/r_1}{\chi^2(r_2)/r_2} \sim F(r_1, r_2)$, know $\frac{\hat{\sigma}^2 (T-k)}{\sigma^2} \sim \chi^2_{T-k}$

$$\lambda_w = \frac{\left(R\tilde{\beta} - q\right)' \left[\sigma^2 R \left(X'X\right)^{-1} R'\right]^{-1} \left(R\tilde{\beta} - q\right)/J}{\frac{\hat{\sigma}^2 (T-k)/\sigma^2}{T-k}} \sim F(J, T-k)$$

$\lambda_w = \left(R\tilde{\beta} - q\right)'[\hat{\sigma}^2 R(X'X)^{-1}R']^{-1}\left(R\tilde{\beta} - q\right)/J \sim F(J, T-k)$

Under $\mathrm{H}_0 \colon R\beta = \mathrm{q}$

What if $\mathbf{H_0}$ is not true?

In general, λ_w is distributed by a noncentral F random variable with *J* and *T*-*k* degrees of freedom and noncentrality parameter *a*.

$\lambda_w \sim F(J, T-k, a)$

Where

$$a = \frac{(R\beta - \mathrm{q})'[R(X'X)^{-1}R']^{-1}(R\beta - q)}{2\sigma^2}$$

Which depends on unknowns β and σ^2
The farther *R*β is from *q* the larger the value of a
The larger the value of a, the more likely we are to reject a false null hypothesis
∴ test is more powerful the farther the null hypothesis is from the truth
Nothing new, true for any statistical test

Non linear hypotheses (Wald Test)-*chapter 12 of Judge*

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{e} \qquad \mathrm{E(e)} = 0 \qquad \mathrm{E(ee')} = \sigma^2\Phi(\theta)$$

let $\gamma' = (\beta', \sigma^2, \theta')$ θ is $H \times 1$ γ is $(k + 1 + H) \times 1$

$H_0\colon \mathrm{g}(\gamma) = 0,$ J restrictions

$H_A\colon \mathrm{g}(\gamma) \neq 0$

We know that MLE is distributed as

$\sqrt{T}(\tilde{\gamma} - \gamma) \xrightarrow{d} N\left[0, \lim_{T\to\infty} (I(\gamma)/T)^{-1}\right]$

Apply our formula for a nonlinear transformation (If $Y \sim N(\mu, \Omega)$ and $Z = g(Y)$ then $\sqrt{T}[Z - \mathrm{g}(\mu)] \xrightarrow{d} N[0, T\mathrm{g}'(\mu)\Omega\mathrm{g}'(\mu)^{\mathrm{T}}]$ and we get

$$\sqrt{T}[\mathrm{g}(\tilde{\gamma}) - \mathrm{g}(\mu)] \xrightarrow{d} N\left[0, \lim_{T\to\infty} T\, F' I(\gamma)^{-1} F\right]$$

Where

$$F = \frac{\partial \mathrm{g}(\gamma)'}{\partial \gamma} = \mathrm{g}^{/}(\gamma)^{\mathrm{T}}$$

is the matrix of partial derivatives, F is (*k* +*H* +1) x *J*

Then using the same formula as for the linear case and substituting consistent estimates for the unknown parameters

$$\lambda_w = \mathrm{g}(\tilde{\gamma})'\left[\tilde{F}'\mathrm{I}(\tilde{\gamma})^{-1}\tilde{F}\right]^{-1}\mathrm{g}(\tilde{\gamma}) \xrightarrow{\mathrm{d}} \chi^2(J) \text{ under } \mathrm{H}_0\colon \mathrm{g}(\gamma) = 0$$

Is an asymptotically valid test.

We can also derive an asymptotically valid F-Test:

$$\lambda_{w_f} = \mathrm{g}(\tilde{\gamma})'[\tilde{F}'\mathrm{I}(\tilde{\gamma})^{-1}\tilde{F}]^{-1}\,\mathrm{g}(\tilde{\gamma})/\mathrm{J} \xrightarrow{\mathrm{d}} F(J, T - K), \qquad \lambda_{w_f} = \lambda_w/J$$

So if $I(\gamma)^{-1} = \sigma^2(X'X)^{-1}$ $(\Phi(\theta) = I)$, then

$$\lambda_{w_f} = \mathrm{g}(\tilde{\gamma})'[\tilde{F}'\tilde{\sigma}^2(X'X)^{-1}\tilde{F}]^{-1}\,\mathrm{g}(\tilde{\gamma})/\mathrm{J} \xrightarrow{\mathrm{d}} F(J, T - K),$$

- ✓ A linear restriction is a special case $(\mathrm{g}(\gamma) = R\beta - q)$
- ✓ How do we get standard errors?
- ✓ Wald test is not invariant to nonlinear transformations and therefore it is a poor choice for testing nonlinear hypotheses.

Likelihood Ratio Test

$H_0 \colon R\beta - r$ for convenience, assume normality and $Y = X\beta + e$

$E(ee') = \sigma^2 \Phi(\theta)$

Unconstrained log-likelihood

$$\max L(\beta, \sigma^2, \theta) = \frac{\mathrm{T}}{2}\ln 2\pi - \frac{T}{2}\ln\sigma^2 - \frac{1}{2}\ln|\Phi(\theta)| - \frac{1}{2\sigma^2}(y - X\beta)'\Phi(\theta)^{-1}(\mathrm{y} - \mathrm{X}\beta)$$

denote estimates as $\tilde{\beta}, \tilde{\sigma}^2, \tilde{\theta}$

Constrained log-likelihood

$\max L\,(\beta, \sigma^2, \theta) + \eta'(R\gamma - q)$

Denote solutions as $\tilde{\beta}_0, \tilde{\sigma}_0^2, \tilde{\theta}_0, \eta_0$

Under the sufficient conditions:

- root of likelihood equations are unique
- likelihood function $\in c^2$

we have that $2\left[L\left(\tilde{\beta}, \tilde{\sigma}^2, \tilde{\theta}\right) - L\left(\tilde{\beta}_0, \tilde{\sigma}_0^2, \tilde{\theta}_0\right)\right] \overset{d}{\rightarrow} \chi_J^2$

(Note: this formula is the same when nonlinear or nonnormal)
(Reminder: Maximum likelihood requires the data generating process be known. It does not require normality. Normality is only used here as an example.)

Under normality

$$L\left(\tilde{\beta}, \tilde{\sigma}^2, \tilde{\theta}\right) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\tilde{\sigma}^2 - \frac{1}{2}\ln\left|\Phi(\tilde{\theta})\right| - \frac{T}{2}$$

$$L\left(\tilde{\beta}_0, \tilde{\sigma}_0^2, \tilde{\theta}_0\right) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\tilde{\sigma}_0^2 - \frac{1}{2}\ln\left|\Phi\left(\tilde{\theta}_0\right)\right| - \frac{T}{2}$$

Note that if $\Phi(\tilde{\theta}_0) = I_{\mathrm{T}}, |I_T| = 1, \ln(1) = 0,$ so only need sum of squared errors to calculate log likelihood under i.i.d. normality. The needed information will be on the computer printout.

When restrictions are only on β, model is linear, and residuals are distributed i.i.d. normal:

$\lambda_{LR} = \frac{(SSE_R - SSE_u)/J}{SSE_u/(T-K)} = \frac{SSE_R - SSE_u}{J\tilde{\sigma}^2} \overset{d}{\rightarrow} F(J, T-k)$ is often used and is valid in small samples. SSE_R and SSE_u are the restricted and unrestricted sum of squared errors. This test is often called the sum of squared errors F-test. **Do not attempt to use the sum of squared errors F-test when the assumptions do not hold.**

For linear restrictions with scalar covariance $\lambda_{\mathrm{LR}} = \lambda_W$

Thus the likelihood ratio and Wald tests are the same when the sum-of-squared-errors F-test is valid. That is why some authors call it a Wald test, yet we have called it a likelihood ratio test.

Nonlinear restrictions? use $2\left[L(\tilde{\beta}, \tilde{\sigma}^2, \tilde{\theta}) - L(\tilde{\beta}_0, \tilde{\sigma}_0^2, \tilde{\theta}_0)\right] \overset{d}{\rightarrow} \chi_J^2$

Heteroskedasticity or nonnormality? Use $2\left[L(\tilde{\beta}, \tilde{\sigma}^2, \tilde{\theta}) - L(\tilde{\beta}_0, \tilde{\sigma}_0^2, \tilde{\theta}_0)\right] \overset{d}{\rightarrow} \chi_J^2$

For seemingly unrelated or simultaneous equations some authors use an adjusted likelihood ratio test. Why? The test is only asymptotic, so it can sometimes be made more precise in small samples by making a degrees of freedom adjustment.

Lagrange Multiplier

Based on constrained maximum likelihood (could do constrained LS)

$\varphi(\gamma, \eta) = L(\gamma) + \eta' \mathrm{g}(\gamma)$

where *L* is log-likelihood and η is shadow price

$\gamma = [\beta, \sigma^2, \theta]'$

f.o.c.

$\frac{\partial \varphi}{\partial \gamma} = \frac{\partial L}{\partial \gamma} + \frac{\partial \mathrm{g}'}{\partial \gamma}\eta = 0$

We know

$H_0\colon \mathrm{g}(\gamma) = 0 \Rightarrow H_0\colon \partial L/\partial\gamma = 0$ since $\eta = 0$ (from f.o.c.)

$H_A\colon \mathrm{g}(\gamma) \neq 0 \Rightarrow H_A\colon \partial L/\partial\gamma \neq 0$ since $\eta \neq 0$

Let $S(\gamma) = \partial L/\partial\gamma$ and $F = \partial \mathrm{g}'/\partial\gamma$

$$\lambda_{LM} = \eta_0' \tilde{F}_0' I(\tilde{\gamma}_0)^{-1} \tilde{F}_0 \eta_0 \xrightarrow{d} \chi_J^2$$

In practice, LM tests are most frequently implemented with artificial regressions <u>involving residuals of the restricted model.</u> You will be expected to know the artificial regression form of Lagrange multiplier tests. Why do artificial regressions for the mean equation require including *X* (or a vector of first derivatives in the case of a nonlinear model), but artificial regressions for the variance equation do not?

- ✓ In most cases $\lambda_W \geq \lambda_{LR} \geq \lambda_{LM}$ ∴ Wald test is usually the most powerful
- ✓ But no test is uniformly most powerful (UMP)
- ✓ Use whichever is easiest
- ✓ Likelihood ratio is invariant, ∴ it is preferred for nonlinear hypothesis tests

With PROC REG in SAS

TEST gives Wald=Likelihood Ratio since linear

RESTRICT gives Lagrange Multiplier in SAS-only with J=1

(η is the shadow price in GAMS)

Each restriction reduces the number of independent parameters to estimate ∴ each restriction gives us back one degree of freedom

Five parts of hypothesis testing

1. Null hypothesis
2. Alternative hypotheses
3. Calculated test statistic
4. Critical value of test statistic (p-value is fine)
5. Conclusion in words

Examples – Autocorrelation and Heteroskedasticity

Autocorrelation

$$\boldsymbol{Y}_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{e_t}$$

$$e_t = \rho e_{t-1} + v_t,$$

$$\mathrm{H_0}: \rho = 0, \mathrm{H_A}: \rho \neq 0$$

1. Wald test: Estimate with MLE, use a t-test
2. LR: Use $2\left[L(\tilde{\beta}, \tilde{\sigma}^2, \tilde{\theta}) - L(\tilde{\beta}_0, \tilde{\sigma}_0^2, \tilde{\theta}_0)\right] \overset{d}{\rightarrow} \chi_J^2$ (*J*=1 in this example)
3. LM- can use DW (exact DW is preferred or Godfrey's test:
   $\hat{e}_t = \alpha' X_t + \rho\hat{e}_{t-1} + v_t$
   The LM test with the above artificial regression can use the t-value for ρ, but is more often calculated as $\mathrm{TR}^2 \overset{d}{\rightarrow} \chi_J^2$ (*J*=1 in this example)

*Heteroskedasticity

$Y_t = X_t\beta + e_t$
$h_t^2 = a_0 + a_1 Z_t$
$e_t \sim N(0, h_t^2)$

1. Wald- estimate with MLE $H_0: a_1 = 0$, t-test if a_1 is scalar, F otherwise:

   $$\lambda_w = (R\tilde{\gamma} - 0)'[R\mathrm{I}(\tilde{\gamma})^{-1}R']^{-1}(R\tilde{\gamma} - 0)/J \sim F(J, T-k)$$

2. LR-estimate both, compare likelihood (PROC MIXED)

3. LM-OLS estimate $\hat{e}_t^2 = a_0 + a_1 Z_t$ and test $\mathrm{H_0}: \mathrm{a_1} = 0$ (Breusch-Pagan).

Since we have these tests of restrictions, why don't we use them to help select our model?

III. d. Pretest Estimation

A pretest estimation arises when the investigator, uncertain about the quality of the restrictions, treats *R*β=*r* as a set of hypotheses about β to be tested and then adopts or rejects b^* (restricted OLS estimator) on the basis of a statistical test.

This estimator is $\hat{\upbeta} = \left[\begin{array}{l} b^{*}\text{if } \uplambda < \mathrm{c}_{\upalpha} \\ b \text{ if } \uplambda > \mathrm{c}_{\upalpha} \end{array}\right.$

Where λ is the calculated value of the test statistic (ex. λ_{w}- Wald test) and $\mathrm{c}_{\upalpha}$ is the critical value.

- ✓ Biased in general- non zero probability of accepting a false null hypothesis
- ✓ Unbiased if "true" model is selected
- ✓ Consistent as long as power of the test goes to 1 as sample size approaches infinity
- ✓ No reason to pretest unless you are willing to accept the possibility of bias in order to get a reduced variance.

Need to develop some type of loss function to weight between bias and variance- Can use M.S.E. (mean squared error) norms:

1. Strong (General) M.S.E.

   The estimator $\hat{\upbeta}$ is preferred to b if

   $$E[(b-\upbeta)(b-\upbeta)'] - E\left[(\hat{\upbeta}-\upbeta)(\hat{\upbeta}-\upbeta)'\right] = \Delta$$

   Where Δ is positive semi definite.

   ✘ Problems: Not single valued

   Requires $\hat{\upbeta}$ to be at least as good as b for each element

   Often yields inconclusive results, not of much practical use.

2. Weak M.S.E (Quadratic Loss Criterion)

   (a) Squared error loss

   $\hat{\upbeta}$ is preferred to *b* if

   $$E\left[(b-\upbeta)^{/}(b-\upbeta)\right] - E\left[(\hat{\upbeta}-\upbeta)^{/}(\hat{\upbeta}-\upbeta)\right] > 0$$

   Single valued but

   ✘ Sensitive to units of measurement. Weights largest β's more.

   (b) Squared error loss of prediction

   $\hat{\upbeta}$ is preferred to *b* if

   $$E[(b-\beta')\tilde{X}'\tilde{X}(b-\upbeta)] - E\left[(\hat{\upbeta}-\upbeta)'\tilde{X}'\tilde{X}(\hat{\upbeta}-\upbeta)\right] \geq 0$$

Where $\tilde{X}$ are the observations for which $\min(Y-\tilde{X}\hat{\beta})'(Y-\tilde{X}\hat{\beta})$ is desired

- ✓ It is single valued and not sensitive to units of measurement

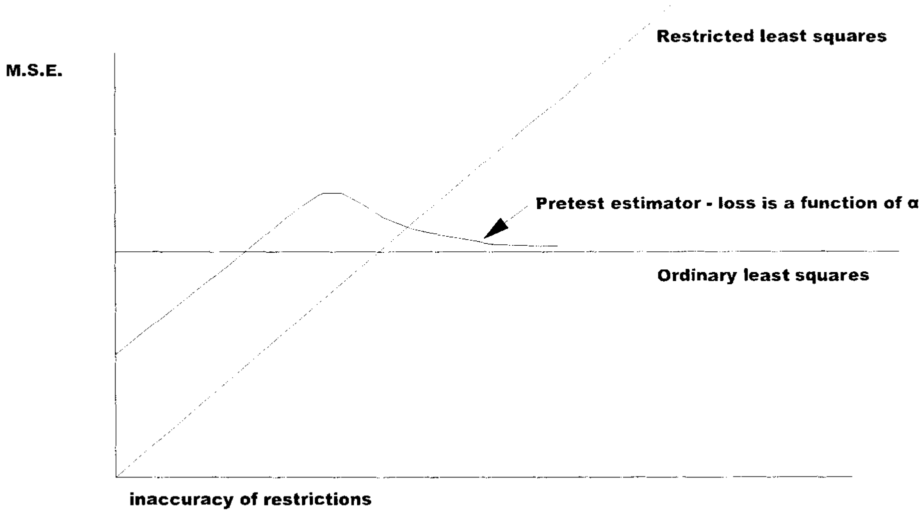
Define a risk function for $\hat{\beta}$

$$\rho\,(\hat{\beta},\beta)=E\left[(\hat{\beta}-\beta)'Q(\hat{\beta}-\beta)\right]$$

- ▸ If Q=I then $\rho(\hat{\beta},\beta)$defines the squared error loss criterion
- ▸ If $\mathbf{Q}=\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}$ then $\rho(\hat{\beta},\beta)$defines the squared error loss of prediction

*Book assumes $X'X=I$ so both are the same.

If J=k i.e. complete restrictions (not a necessary assumption) then loss is



Inaccuracy is a function of *a* for F(J, T-k,*a),* the non-central F.

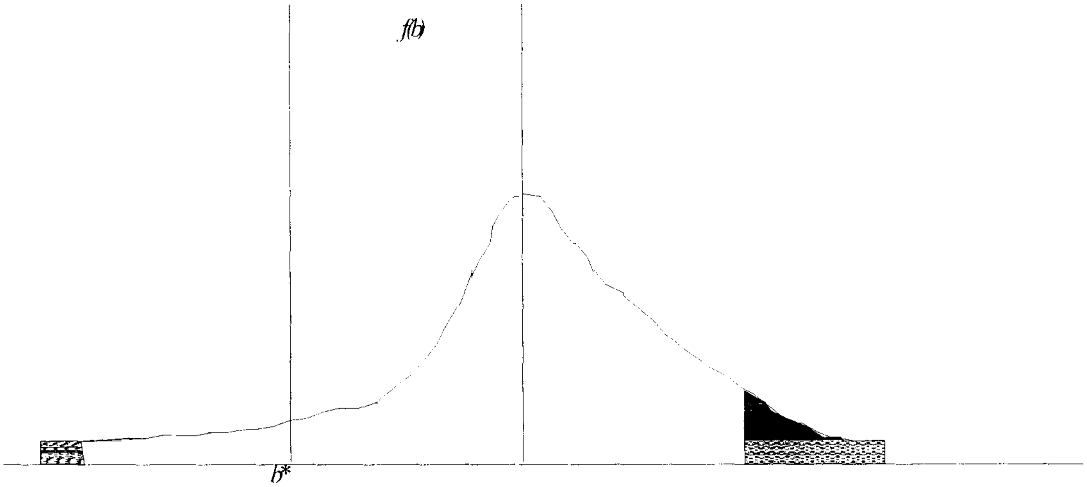
∴ on the basis of M.S.E., pretest estimator can be worse than either OLS or RLS

b is the minimax (among this set of estimators), since it minimizes the maximum risk

How can the pretest estimator be worse than OLS or RLS when it is always equal to one or the other?

- M.S.E. is based on **expectations**∴ mean of repeating the same pretest
- In some cases pretest does a poor job of selecting between OLS and RLS

Example-p.d.f. of b

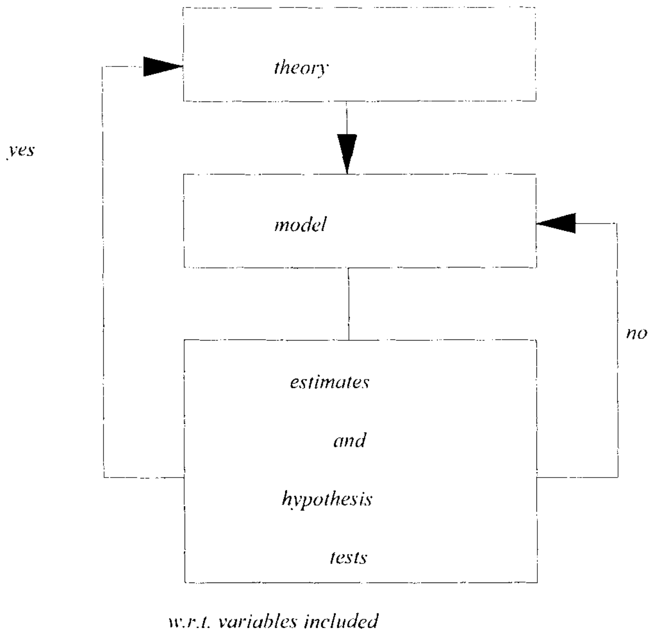


Get all of the OLS extremes and none of the close estimates.

Note: Pretest risk function is not valid for sequential testing as in stepwise regression.
Distribution of pretest estimators unknown.
∴ cannot conduct hypothesis tests
(of course could still use bootstrap or Monte Carlo hypothesis testing)

III. e. Econometric Ethics Methodology (Opinion)

Supply/demand elasticities vary greatly by study.
You will have to develop your own methodology



My Methodology

Classify econometricians in 3 groups

1. Classical- consistency, unbiasedness, purity of hypothesis tests

2. Bayesians- theoretically sound- combines sample information and prior information, diffuse prior-little different, M.S.E.

3. Data snoopers

   Empiricists- numerous specification tests as a substitute for theory

   Hendry- encompassing principle (general-to-specific)

   Spanos- misspecification testing

Ideal approach

Estimate only one model specification- naive

Problems

1. Theory is often not sufficient to fully specify model (ex. number of lags). Alternative theories are available or theory is consistent with several specifications.

2. Results from first model may not be "publishable." First specification may have been stupid.

3. Results are fragile

Solutions

1. (a) Narrow set of models to only those that are consistent with theory.
   (b) Follow conventions
   (c) Estimate alternative specifications and report in footnote or text.
   (d) If not fragile, you're OK. (*Note: even Cox and Snell did this)*

2. (a) Re-examine your theory
   (b) Often thought a sign was wrong until I thought more about it
   Example- short-run supply of cattle is downward sloping
   (c) Maybe data are right and your theory is wrong.
   (d) If necessary, redo theory and specify an alternative model.
   Don't include 2 measures of the same thing.

3. (a) Try to narrow the set of theoretically valid models.
   (b)Report the model that has the strongest theoretical justification.
   Which side would you want to defend in an argument?
   (c) Select a new topic. An ethical researcher never wants to publish misleading results. If you lose integrity you have lost everything
   (d) Try to publish the fragile results.

Reviewing

Insist that the specification be consistent with theory. (Lessens opportunities for pretesting.)

Do not insist on high $R^{2\prime}$s

Do not insist that every variable be significant

Do not force the author to re-specify the model exactly the way you would have. It is not your research.

Do not insist that the conclusions be the same across all possible models.

What I do

Pretest more than I would like

Be truthful-often report results of alternative models in footnotes. (This is becoming more acceptable.)

Sometimes commit sins of omission (not often)

Try to select topics that are publishable regardless of conclusions.
(**Not possible for everybody.)

Use large sample sizes.

III. f. Nonnested hypotheses test

Hypotheses are nonnested if neither model can be written as a special case of the other

Example

$$H_0: y = X\beta_0 + W\beta_1 + \mu_0 \; \mu_0 \sim N(N, \sigma_0^2 I)$$
$$H_1: y = Za_0 + Wa_1 + \mu_1 \; \mu_1 \sim N(N, \sigma_1^2 I)$$

$\mathrm{X\ is\ n \times k_0}$ $\mathrm{y\ is\ n \times 1}$ $\mathrm{Z\ is\ n \times k_1}$ $\mathrm{W\ is\ n \times k_2}$

-----Neither model is a special case of the other----

Example:

- Private vs. public markets
- Measurements of pork carcass quality
- Also shows up in issues of functional form:
  - Beta vs. stable
  - GARCH-t vs. diffusion-jump

Solutions

1. **Orthodox text; Wald test, encompassing test,** nest the two models in a more general model:

$$W: y = X\lambda_0 + Z\lambda_1 + W\lambda_2 + \; e$$

Test

$H_0: \lambda_1 = 0$ and $H_1: \lambda_0 = 0$ independently

		H_0	
		Not Rejected	*Rejected*
	Not rejected	*X and Z contain same information*	*Favors* H_1
H_1	rejected	*Favors* H_0	*X and Z both contain unique info favors the general model*

- Hypotheses can be tested with a Wald F-test
- Possible to reject both $\mathrm{H_0}$ and $\mathrm{H_1}$ or fail to reject both $\mathrm{H_0}$ and $\mathrm{H_1}$ $\therefore$ can be inconclusive
- Always true of non-nested hypotheses
- Orthodox is only test that is valid for small samples

Greene does not consider orthodox test a nonnested test but rather a test of the general model

2. **J-test** Davidson and MacKinnon *Econometrica* 1981 pp. 781-93

   Let $\hat{y}_0$ and $\hat{y}_1$ be predicted values from models estimated under H_0 and H_1

   Estimate

   $J_0: y = Xc_1 + Wc_2 + \hat{y}_1 c_3 + v_0$

   $J_1: y = Zd_1 + Wd_2 + \hat{y}_0 d_3 + v_1$

And use the t-statistic to test

$H_0: c_3 = 0 \qquad H_1: d_3 = 0$ (measuring explanatory power)

Test is only valid asymptotically

Why use *t* instead of *Z*?

Monte Carlo studies- what they tell us

3. **Cox's nonnested test,** based on likelihood ratio but it is no longer χ^2 under $H_0: LR_0 - LR_1$ estimate two additional regressions for each hypothesis (artificial regressions)

   1. Regress the predicted values under H_1 against the exogenous variables in H_0
      $H_2: \hat{y}_1 = Xa_1 + Wa_2 + \mu_2$
   2. Regress residuals from H_2 against the exogenous variables in H_1
      $H_3: \hat{\mu}_2 = Zb_1 + Wb_2 + \mu_3$

   To test H_1:

$$N_0 = \frac{\sqrt{n}}{2}\ln\left[\frac{SSE_0}{SSE_2 + SSE_1}\right] / \frac{[SSE_1 \times SSE_3]^{1/2}}{[SSE_1 + SSE_2]} \xrightarrow{d} N(0,1)$$

Where SSE_j is the sum of squared errors from model $\mathrm{H_j}$; should use a t-test for small samples (Godfrey & Pesaran J. Ecmt. 1983:133-54)

OK for $\begin{matrix} H_2: \hat{y}_0 = Za_1 + Wa_2 + \mu_2 \\ H_3: \hat{\mu}_2 = X_b + Wb_2 + \mu_3 \end{matrix}$

$$N_0 = \frac{\sqrt{T}}{2}\ln\left[\frac{SSE_1}{SSE_2 + SSE_0}\right] + \frac{[SSE_0 x SSE_3]^{1/2}}{[SSE_0 + SSE_2]} \overset{asy}{\sim} Z$$

Asymptotic power of the tests
Pesaran Ecmt Sept. 1982:

$\mathrm{P_{Cox}} = \mathrm{P_J} \geq \mathrm{P_{orthodox}}$ equality only holds if the number of non-overlapping variables is 1.

- ✓ The larger the number of non-overlapping variables, the more powerful would be the two non-nested tests
- ✓ Monte Carlo study suggested size of the Cox test is too large in small samples (20)-OK with 60
- ✓ Godfrey and Pesaran have suggested a modification of Cox's test for small samples
- ✓ Monte Carlo results suggest J-test may have
- ✓ Which to use? Report J and Cox test.
- ✓ J-test is inferior to P-test for nonlinear models (P-test has greater power)

P-test

$$y - \hat{y}_0 = a(\hat{y}_1 - \hat{y}_0) + \hat{F}_i b + \varepsilon_i$$

where $\hat{F}_i$ is a matrix containing the derivatives of $\mathrm{H_0}$ w. r. t β for each observation evaluated at $\hat{\beta}$.

$\mathrm{H_0}: a = 0$
Same as J-test if linear

A Monte Carlo Cox test or some other Monte Carlo method can almost always be used.

With a Bayesian framework, could use marginal likelihoods and Bayes factors.

III. g. <u>Monte Carlo Methods</u>

(deeper than what I really want you to know; avoid being a black box)

Econometricians use them to determine small sample properties when analytical results cannot be obtained. Occasionally used to verify analytical results

Where is Monte Carlo?
Monte Carlo hypothesis testing/bootstrapping
Stochastic simulation

Random numbers (pseudo-random numbers)

A random number is a random variable with a uniform (0,1) density function

$$f(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 \text{ if } 0 \leq x \leq 1 \\ 0 \text{ if } x > 1 \end{cases}$$

Also, since we can use a sequence, we want $\mathrm{cov}(\mathrm{r_t}, \mathrm{r_{t-j}}) = 0$

Slow to repeat

Example: Multiplicative Congruential Generator

$\mathrm{r_{i+1}}$=remainder $\left(\frac{ar_i}{M}\right)$, a and M are given constants

r_0 is called the seed (usually best to pick 5-7 digit odd numbers). Can pick $\mathrm{r_0}$ based on computer's clock by default, but then it is not replicable unless you record starting point.
The best choice of *M* on a binary computer is
$M = 2^{b-1}$

where *b* is the number of digits (bytes) in the computer word. This gives the largest possible integer on the machine (need one byte for the sign). Older computers usually have 32 bytes/word, but many newer ones have 64 bytes/word.

Random Number Generator Function in SAS

RANUNI (seed)
Where seed is an integer $< 2^{31} - 1$. If the seed $\leq$0, the time of day is used to initialize the seed.

The RANUNI function returns a number generated from the uniform distribution on the interval (0,1) using a prime modulus multiplicative generator with modulus 2^{31}-1 and multiplier 397204094 (see Fishman and Moore 1982). The CALL RANUNI routine, an alternative to the RANUNI function, gives greater control of the seed and random number streams.

According to our notation, $M = 2^{31} - 1, a = 397204094, \text{and } seed = r_0$

Random Deviate Generation
How do we get random deviates for the desired distribution from the random number?

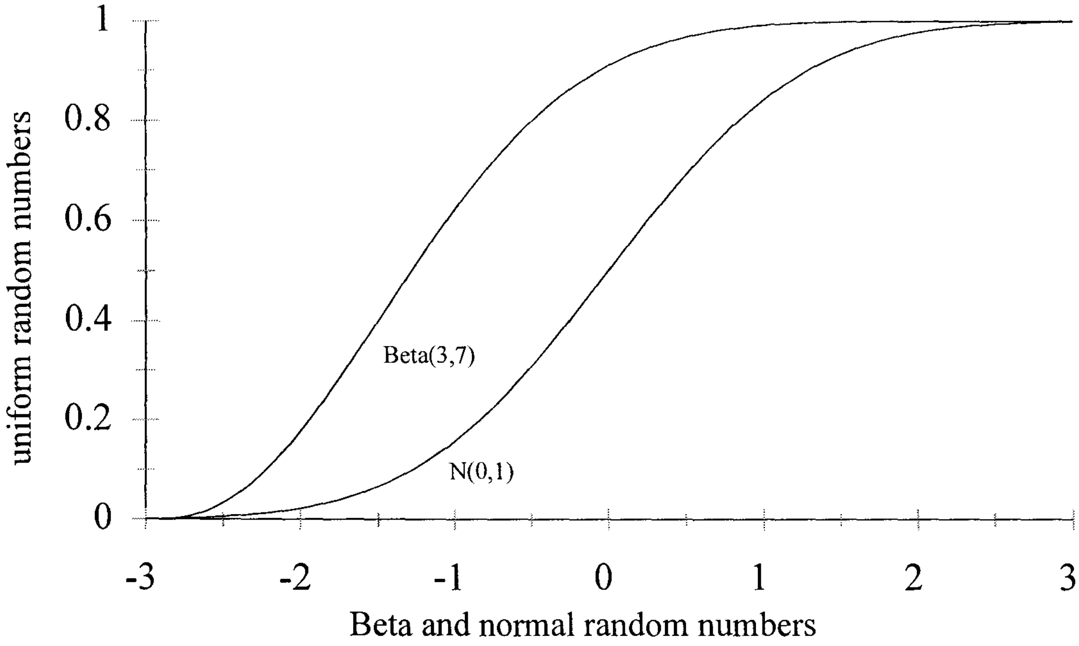
Inverse Transformation Method

Can be used whenever f(x) can be integrated or F(x) determined by some other means.

- We know $0 \leq F(x) \leq 1 \qquad 0 \leq RN \leq 1$
- Seek inverse transformation function such as $\mathrm{X} = \Phi[\mathrm{F(x)}]$
- Replace F(x) by RN: X=$\Phi(\mathrm{RN})$

Inverse Transformation Method

For Beta and Normal random numbers



Uniform distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{if x} < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & \text{if x} > b \end{cases}$$

Let r = F(x) then solve for x

$r = \frac{x-a}{b-a}$ $x = r(b-a) + a$

$U(a, b) = r(b-a) + a$

Uniform (10, 20)

$r_1 = .6678$ 16.678

$$r_2 = .4932 \;\; 14.932$$

Normal (0, 1) (not an I.T.M)

$X_1 = (-2 \ln r_1)^{1/2} \sin(2\pi r_2)$

In SAS use RANNOR (seed) to get N(0,1) *(What is a seed?)*

RANNOR(0) will use computer's clock to pick a seed, but then your work is not replicable.

Normal N($\boldsymbol{\mu}$, V)

DEV= $\mu + \sqrt{V} * X_1 \quad X_1 \sim N(0,1)$

Based on formula for linear transformation

What is multivariate normal?

$e_{Gx1} \sim N(\mu_{Gx1}, \Omega_{GxG})$

Ω is a positive definite symmetric matrix

∴ we can use the Cholesky decomposition (kind of a √ of a matrix

Ω=$P'P$ in SAS PROC IML: P=ROOT (Ω);

where *P* is a non singular matrix

P is triangular $\begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot \\ 0 & 0 & 0 & \cdot \end{bmatrix}$

To generate deviates: *e* =$P'Z + \mu$

Z~N(0,I) ∴ Z is a vector of N(0,1) random deviates

To generate any normal random vectors

1. Need random number generator for N(0,1)
2. Formula $e^* = P'Z + \mu$

Example:

Let $\mu = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$, $\Omega = \begin{bmatrix} 9 & 0 \\ 0 & 16 \end{bmatrix}$, and $\mathrm{Z} = \begin{bmatrix} 0.6 \\ -1 \end{bmatrix}$. So $P = \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}$, and

$e = P'Z + \mu = \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 0.6 \\ -1 \end{bmatrix} + \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 3.8 \\ -1 \end{bmatrix}.$

<u>Monte Carlo Study</u>

Know whether assumptions are true or not
Simulation when really true
Inductive logic

1. Select the model-Note: the model is known exactly
   $\mathrm{Y = X\beta + e \qquad e{\sim}f(\theta)}$
   $\mathrm{X, \beta, and\ f(\theta) are\ all\ known}$

2. Create sets of data
   Say 100 samples of size N (1,000 or 10,000 are more common now)
   100 replications-repeated samples
   N observations

   $Y_i = X\beta + e_i^*$ Where $e^*$are generated from a random deviate generator for f(θ)

3. Calculate estimates

   $\hat{\beta}_i \quad i = 1, \dots, 100$ 100 estimates of β all from the same model

4. Estimate Sampling Distribution Properties

   (a) Estimated expected value = $\bar{\beta} = \sum_{i=1}^{100} \hat{\beta}_i /100$

   (b) Estimated bias = $\bar{\beta} - \beta$

   (c) Estimated variance-covariance $\sum_{i=1}^{100} \left(\hat{\beta}_i - E(\hat{\beta}_i)\right)\left(\hat{\beta}_i - E(\hat{\beta}_i)\right)^{/} /99$

(d) Estimated $\sum_{i=1}^{100}(\hat{\beta}_i - \beta)\,(\hat{\beta}_i - \beta)^{/}/100$

-diagonal will give M.S.E. for each element of $\hat{\beta}_i$

MSE = $\text{bias}^2 + \text{ variance}$

Can also test hypotheses, calculate % rejected and compare to size or power
-size (H_0: true) calculate % reject
-power (H_0: false) calculate % reject

Usually test under different assumptions about β, *X*, and *N* (sample size)

Note: If doing Monte Carlo integration (e.g. calculating an expected value) there are many variance reduction techniques such as antithetic variates (use DEV and –DEV) and Halton draws.

III. h. Bootstrapping Techniques

Parametric Bootstrapping for Hypotheses Testing

- Called parametric because sampling is done from a parametric distribution
- Can be used when other techniques fail
- Often has superior properties- faster rate of convergence than conventional tests.

Sample data

$\tilde{X}_1, \dots, \tilde{X}_n$
$\mathrm{H}_0: X_1, \dots, X_n \sim \mathrm{F}(\theta)$
$\mathrm{H}_\mathrm{A}: X_1, \dots, X_n \text{ not} \sim \mathrm{F}(\theta)$

- *for example:* GARCH
- if θ must be estimated, then you only have large sample properties.

Steps

1. Select test statistic
   t = g(X)
   $\hat{t} = g(\tilde{X})$

   For example*:* Kurtosis, skewness, technical trading returns, likelihood ratio statistic with nonnested models. Best if *t* is an asymptotically pivotal quantity.
   ⇒ asymptotic distribution does not depend on any unknown parameters. (e.g. $\hat{t} \xrightarrow{d} \chi_J^2$)

2. Estimate $\hat{\theta}$ if necessary

3. Generate NS random samples of size *n* from distribution *F*($\hat{\theta}$)

4. Calculate $\hat{t}^i = g\left(X^i\right)$

   Significance level= $\left(\#\left[\hat{t}^i \geq \hat{t}\right] + 1\right)/(NS + 1)$

   (one-tailed test) (multiply by ½ for two-tailed test)

Example:

H_0: normal distribution
Test statistic relative kurtosis-asymptotically pivotal

$\hat{\theta} = (\bar{x}, \hat{\sigma}^2)$
$\hat{t} = \hat{\mu}_4(\tilde{x})$ Fourth moment

Generate NS random samples of size n from N($\bar{x}, \hat{\sigma}^2$)

Calculate $\hat{t}^i = \hat{\mu}_4\left(X^i\right)$

Say $\hat{t} = 4$
NS=999
$\#\left[\hat{t}^i > 4\right] = 2$

Obtaining distribution of $\hat{t}$ under H_0: by Monte Carlo methods

Asymptotically pivotal quantities-relative kurtosis

$2 + \frac{1}{1000} = .003$

valid in small samples- with large values of NS when θ *is given*

Non-Parametric Bootstrap-Resampling Procedure

- Sampling done from the empirical distribution function
- Draws repeated random samples <u>with replacement</u> from the data (e.g. Urn problem)

- Applied in situations where distributions of test statistics cannot be derived analytically.

Sampling with replacement- **example**

A={6, 10, 5, 7, 23, 2}

We want 2 samples **with replacement** of size 8 from A.
Note that when sampling with replacement, any size is possible.

Roll dice (random generator) and follow this rule:

If dice	☞	Choose
1		6
2		10
3		5
4		7
5		23
6		2

{1, 1, 5, 6, 3, 5, 2, 2} ☞ {6, 6, 23, 2, 5, 23, 10, 10}

{2, 1, 6, 3, 6, 1, 4, 3} ☞ {10, 6, 2, 5, 2, 6, 7, 5}

Data-n observations

$x_{1,} \dots, x_n \sim \mathrm{F}$

Where <u>F is an unknown probability distribution function</u>

Interested in some estimator $\hat{t} = g(x_{1,} \dots, x_n)$

$H_0: \hat{t} = \theta_0$ Want to test, for example that skewness is zero.

Steps

1. Construct $\hat{f}(X) = \begin{cases} 1/n & \text{if } X = X_i \\ 0 & \text{otherwise} \end{cases}$ $\quad$ empirical density function $\quad \}$

   Where X_i are the observed sample values

2. Draw a sample **of size *n*** from distribution $\hat{f}$ (1 bootstrap sample). Note that a bootstrap sample will always have the same size as the original sample (*n*)

   $x_i^*, x_2^*, \dots, x_n^* \overset{iid}{\sim} \hat{f}$

   Independently repeat step 2 a large number (B) of times.

3. Calculate
   $\hat{t}_b = g(x_{1b}^*, x_{2b}^*, \dots, x_{nb}^*)$
   $b = 1, \dots, B$

   B estimates of $\hat{t}$

   Then, the significance level is $(\#\{\hat{t}_b \geq \theta_0\} + 1)/(B+1)$ (**percentile method)**

   $S\widehat{D} = \left\{\frac{1}{B-1}\sum_{b=1}^{B}\left[\hat{t}_b - \bar{\hat{t}}\right]^2\right\}^{1/2}$ usual formula for standard deviation

   Where $\bar{\hat{t}} = \sum_{b=1}^{B} \hat{t}_b / B$ usual formula for mean

   - Can use $\widehat{SD}$ to compute confidence intervals **(standard deviation method)**
   - Under normality of $\hat{t}$ can compute t-test.

Example

Given the following set of observations
(*n*=3)

OBS	X	Y
1	0.6	1.2
2	0.9	2.0
3	1.4	3.3

(i) Define an appropriate random deviate generator
(ii) Select two bootstrap samples

Random Numbers	x	y	
0.5441	0.9	2.0	sample 1(*n*=3)
0.08573	0.6	1.2	
0.81067	1.4	3.3	
0.39737	0.9	2.0	sample 2(*n*=3)
0.34958	0.9	2.0	
0.61417	0.9	2.0	

Double bootstrapping

Consists of sampling in two steps.

1. First obtain N samples from the data.
2. Then, from each of the N samples, sample N more times

In this way, one obtains a total of N^2 samples. Can offer slight improvement in some cases. Often required to get asymptotically valid tests when the test statistic is not asymptotically pivotal.

III. i. <u>Misspecification Testing</u> (in Linear Regression Models)

Objective: Test the assumptions of the classical linear regression model.

Methodolgy

Theory ===> Preliminary Model ===> Misspecification tests===> Model

↑____|

Assumptions of the Classical Linear Regression Model

Consider the linear regression model (henceforth, LRM) $\mathrm{y_t} = \boldsymbol{\beta}'\mathbf{x_t} + \varepsilon_t\ \mathrm{t} = 1, \dots, \mathrm{T}$. Where $\mathrm{y_t}$ and ε_t are 1 x 1 and **β** and $\mathbf{x_t}$ are K x 1. The assumptions underlying the LRM are

1) **Normality**: $\mathrm{D}(\mathrm{y_t}|\mathrm{X_t}; \theta)$ is normal: The distribution of $\mathrm{y_t}$ conditional on $\mathrm{X_t}$ is normal, and $\theta = (\beta, \sigma^2)$;

2) **Functional form**: $\mathrm{E}(\mathrm{y_t}|\mathrm{X_t} = \mathrm{x_t}) = \beta'\mathrm{x}_t$: the functional form of the conditional mean is known and linear in the parameters. ===> E($\varepsilon_t|\mathrm{X_t} = \mathrm{x_t}$)=0

3) Homoskedasticity

   Static:$\mathrm{Var}(\mathrm{y_t}|\mathrm{X_t} = \mathrm{x_t}) = \sigma^2$: conditional variance does not depend on $\mathrm{x_t}$ ===> $\mathrm{Var}(\varepsilon_t) = \sigma^2$
   Dynamic: The conditional variance does not depend on the past history of $\varepsilon_t, y_t, \mathrm{or}\ \mathrm{x_t}$. ===> no autoregressive conditional heteroskedasticity (ARCH) effects.

4) **Parameter Stability**: $\theta = (\beta, \sigma^2)$: The parameters of the conditional mean and conditional variance do not vary with time (t).

5) **Independence:** $\mathrm{Y} = \mathrm{Y_1}, \dots, \mathrm{Y_T}$, is an independent sample drawn sequentially from $\mathrm{D}(\mathrm{y_t}|\mathrm{X_t}; \theta), \mathrm{t} = 1, \dots, \mathrm{T}$.

6) **Weak exogeneity:** $\mathrm{X_t}$ is weakly exogenous with respect to θ, t=1, …, t. The marginal distribution of $\mathrm{X_t}$ does not contain relevant information for estimation of θ. Thus, it can be ignored. While tests of exogeneity have been suggested such as the Hausman test, the tests generally require additional assumptions and this makes the test difficult to apply. We therefore follow McGuirk et al. (AJAE 1993) and test it indirectly by testing 1-3.

7) **No perfect collinearity**: Rank (X)=k with T>k, failure of this assumption indicates that the sample information in X is inadequate for the estimation of the statistical parameters β and σ^2.

> ===> When assumptions 1-7 hold, OLS estimators of β and σ^2 are BLU; $\hat{\beta}$ is normally distributed, and $(T-K)\hat{\sigma}^2/\sigma^2$ is distributed χ^2_{T-K}.

Definition and causes of misspecification

Misspecification arises from the violation of the assumptions underlying the linear regression model. Hence, causes of misspecification are:

Departure from normality,
Departure from linearity,
Departure from homoskedasticity,
Departure from parameter stability,
Departure from independence

Consequences of misspecification

Biased estimators,
Inconsistent estimators.

===> Inappropriate inferences and policy recommendations.

Testing for misspecification

There are two types of misspecification tests:

Individual tests: They are tests of a single model assumption. These tests should not be used in isolation to identify sources of misspecification because:

1) An accurate test is ensured only when the model contains no other misspecification.

2) The alternative hypotheses are too broad ===> an appropriate modification is usually not evident if the null is rejected.

Joint tests: They are a comprehensive set of individual tests. They can check simultaneously for parameter stability, appropriateness of functional form and independence. However, the validity of these tests requires that the assumptions of normality and stable variance be met.

Individual tests

Normality

Tests of normality are statistical inference procedures designed to test that the underlying distribution of a random variable is normally distributed. There are several (a plethora) of them. Among those tests, the chi-squared test and Kolmogorov test have poor power properties and should not be used when testing for normality. The Shapiro-Wilk W test is good for testing when the sample size T<50. The third sample moment $(\sqrt{b_1})$ and the fourth sample moment (b_2) tests and the D'Agostino-Pearson K^2 are excellent tests. D'Agostino-Pearson K^2 test is an omnibus test in that it can detect deviations from normality due to either skewness or kurtosis. Let μ be a random variable $\mu = (\mu_1, \dots, \mu_\mathrm{T})'$ and

$$\bar{\mu} = \frac{1}{T}\sum_{t=1}^{T} \mu_t$$

Then the *k*th moment m_k around $\bar{\mu}$ is given by

$$m_k = \frac{1}{T}\sum_{t=1}^{T} (\mu_t - \bar{\mu}) = E(\mu_t - \bar{\mu})^k$$

It follows that the third and fourth moments are: $m_3 = E(\mu_t - \bar{\mu})^3$

and $m_4 = E(\mu_t - \bar{\mu})^4$. The standardized third and fourth moments are known as skewness and kurtosis (respectively) are given by:

$$\sqrt{\beta_1} = \frac{E(\mu_t - \bar{\mu})^3}{\sigma^3} = \frac{m_3}{(\sigma^2)^{3/2}} = \frac{m_3}{(m^2)^{3/2}}$$

Since $m_2 = E(\mu_t - \bar{\mu})^2 = \sigma^2$

and

$$\beta_2 = \frac{E(\mu_t - \bar{\mu})^4}{\sigma^4} = (m_4/m_2^2)$$

Sample estimates of $\sqrt{\beta_1}$ and β_2 could be used to describe nonnormal distributions and used as the bases for tests of normality. Let $\sqrt{b_1}$ and b_2 be the estimates of $\sqrt{\beta_1}$ and β_2. Values of $\sqrt{b_1}$ and b_2 close to 0 and 3 respectively, indicate normality. More precisely, under normality $\mathrm{E}[\sqrt{b_1}] = 0$ and $\mathrm{E}[\mathrm{b}_2] = 3(\mathrm{T}-1)/(\mathrm{T}+1)$.

Tests of skewness ($\sqrt{b_1}$)

H_0: μ is normal ===> $\sqrt{b_1} = 0$

H_a: μ is not normal due to skewness ===> $\sqrt{b_1} \neq 0$ for two sided test ($\sqrt{b_1} > 0$ or $\sqrt{b_1}$ <0 for one sided tests).

Test statistic

1) Compute $\sqrt{b_1}$ from the sample data

2) Compute

$$Y = \sqrt{b}_1 \left[\frac{(T+1)(T+3)}{6(T-2)}\right]^{\frac{1}{2}}$$

$$\beta_2(\sqrt{b}_1) = \frac{3(T^2+27T-70)(T+1)(T+3)}{(T-2)(T+5)(T+7)(T+9)}$$

$$W^2 = -1 + \left[2\left(\beta_2\left(\sqrt{b}_1\right)-1\right)\right]^{\frac{1}{2}}$$

$$\delta = \frac{1}{\sqrt{\ln{(W)}}}$$

$$a = \left[\frac{2}{W^2-1}\right]^{\frac{1}{2}}$$

3) Compute

$$Z\left(\sqrt{b}_1\right) = \delta \ln\left(Y/a + [(Y/a)^2+1]^{\frac{1}{2}}\right)$$

$Z\left(\sqrt{b}_1\right)$ is approximately standard normal with mean zero and variance one. The above is just a transformation of the distribution of $\left(\sqrt{b}_1\right)$ to that of the standard normal. For a two-sided test with a test size of 0.05, reject if

$$\left|Z\left(\sqrt{b}_1\right)\right| > 1.96$$

Test of kurtosis (b_2)

H_0: μ is normal ===>β_2**=**3

H_a: μ is not normal due to kurtosis ===>$\beta_2 \neq 3$ for two-sided test ($\beta_2 > 3$ or < 3 for one-sided test).

For T>=20 the following approximation is valid

1) Compute b_2 from the sample data

2) Compute the mean and variance of b_2 under H_0:normality.

$$E(b_2) = \frac{3(T-1)}{(T+1)}$$

$$\mathrm{var}(b_2) = \frac{24T(T-2)(T-3)}{(T+1)^2(T+3)(T+5)}$$

3) Compute the standardized version of (b_2)

$$x = \frac{\left(b_2 - E(b_2)\right)}{\sqrt{\mathrm{var}(b_2)}}$$

4) Compute the third standardized moment of b_2

$$\sqrt{\beta_1(b_2)} = \frac{6(T^2-5T+2)}{(T+7)(T+9)}\sqrt{\frac{6(T+3)(T+5)}{T(T-2)(T-3)}}$$

5) Compute

$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}}\left[\frac{2}{\sqrt{\beta_1(b_2)}} + \sqrt{1 + \frac{4}{\left(\beta_1(b_2)\right)}}\right]$$

6) Compute

$$Z(b_2) = \left(\left(1 - 2/(9A)\right) - \left[\frac{(1-2/A)}{1+x\sqrt{2/(A-4}}\right]^{\frac{1}{3}}\right)/\sqrt{2(9A)}$$

$Z(b_2)$ is asymptotically standard normal under H_0

Omnibus test K^2

It combines both the skewness and kurtosis tests.

$\mathrm{H}_0 : \mu$ is normal ===> $\sqrt{\beta_1} = 0$ and $\beta_2 = 3$

$\mathrm{H}_\mathrm{a} : \mu$ is not normal ===> $\sqrt{\beta_1} \neq 0$ and/or $\beta_2 \neq 3$

The test statistic is

$$K^2 = Z^2\left(\sqrt{b_1}\right) + Z^2(b_2) \overset{d}{\rightarrow} \chi^2_{(2)}$$

Under H_0.

Note: Skewness and kurtosis statistics given in SAS (PROC UNIVARIATE) and SHAZAM (OLS…/GF) cannot be used because they are not $\sqrt{b_1}$ and b_2. What SAS and SHAZAM give is the Fisher g statistic defined as:

Skewness: $g_1 = \frac{T\sum_{t=1}^{T}(\mu_t - \bar{\mu})^3}{(T-1)(T-2)S^3}$

Kurtosis: $g_2 = \frac{T(T+1)T\sum_{t=1}^{T}(\mu_t - \bar{\mu})^4}{(T-1)(T-2)(T-3)S^4} - \frac{3(T-1)^2}{(T-2)(T-3)}$

Where

$$S^2 = \frac{1}{(T-1)}\sum_{t=1}^{T}(\mu_t - \bar{\mu})^2$$

However, g_1 and g_2 are related to $\sqrt{b_1}$ and b_2. via the following

$$\sqrt{b_1} = \frac{(T-2)}{\sqrt{T(T-1)}}\mathrm{g}_1$$

and

$$b_2 = \frac{(T-2)(T-3)}{(T+1)(T-1)}\mathrm{g}_2 + \frac{3(T-1)}{(T+2)}$$

The Bera-Jarque normality test

Bera and Jarque derived a skewness-kurtosis test as a Langrange multiplier test. Their test statistic under the null hypothesis of normality is:

$$J_N(\mathrm{\mu}) = \frac{T}{6}\hat{a}_3^2 + \frac{T}{24}(\hat{a}_4 - 3)^2 \overset{d}{\rightarrow} \chi^2_{(2)}$$

where

$$\hat{a}_3 = \left[\frac{\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathrm{\mu}}_{\mathrm{t}}^3\right)}{\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathrm{\mu}}_{\mathrm{t}}^2\right)^{\frac{3}{2}}}\right]$$

$$\hat{a}_4 = \left[\frac{\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathrm{\mu}}_{\mathrm{t}}^4\right)}{\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathrm{\mu}}_{\mathrm{t}}^2\right)^{2}}\right]$$

Note about the Bera-Jarque test:

i) It is an asymptotic test

ii) It is sensitive to outliers (if $\mathrm{H}_0$is rejected look for outliers)

iii) Rejection of H_0 gives no information as to the nature of the departure from normality

What to do when normality assumption is invalid?

a) Postulate a new model (Go back to theory).

b) Use a normalizing transformation (eg. logarithm and square root).

===> apply a transformation to $\mathrm{y_t}$ and/or $\mathrm{X_t}$ to induce normality. A common transformation is the Box-Cox transformation.

$\mathrm{Z}^* = \frac{Z^{\delta}-1}{\delta} \qquad 0 \leq \delta \leq 1$

If $\delta = -1 ===> \mathrm{Z}^* = \mathrm{Z}^{-1}$ (reciprocal)

If $\delta = 0.5 ===> \mathrm{Z}^* = (\mathrm{Z})^{1/2}$ (square root)

If $\delta = 0 ===> \mathrm{Z}^* = \log_e \mathrm{Z}$ (lograrithmic)

The first two cases are not commonly used in econometrics modeling because of the difficulties involved in interpreting Z^* in an empirical econometric model. The square root transformation is variance stabilizing. The logarithmic transformation is also variance stabilizing. If the distribution of $\mathrm{Z_t}$ is close to the log-normal, gamma, or chi-square, the distribution of $\log_e \mathrm{Z_t}$ is approximately normal.

Functional form

If assumption 2 is invalid, ===> $\mathrm{E}(\mathrm{y_t}|\mathrm{X_t} = \mathrm{x_t}) = \mathrm{h}(\mathrm{X_t}) \neq \beta' \mathrm{X_t}$ and $\mathrm{cov}(\mathrm{X_t}, \varepsilon_t) \neq 0$

===>All OLS properties are lost

Testing for non-linearity

$\mathrm{H_0}: \mathrm{E}(\mathrm{y_t}|\mathrm{X_t} = \mathrm{x_t}) = \beta' \mathrm{X_t}$

$\mathrm{H_a}: \mathrm{E}(\mathrm{y_t}|\mathrm{X_t} = \mathrm{x_t}) = \mathrm{h}(\mathrm{X_t})$

1) *Kolmogorov-Gabor (KG) test*

This test is based on the KG polynomial in the x's. Under this approach, this alternative functional from is

$Y_t = \beta_0^{/} X_t + \gamma_2^{/} \psi_{2t} + \gamma_3^{/} \psi_{3t} + \varepsilon_t \equiv> KG(3)$

Where,

ψ_{2t} includes the second order terms $\mathrm{X_{it}X_{jt}}, \mathrm{i} >= j\ i, j = 2, \dots, K$.

If K=2 we will have:

$\psi_{2t} = \mathrm{X}_{2t}^2 + \mathrm{X}_{3t}\mathrm{X}_{2t} + \mathrm{X}_{3t}^2$

ψ_{3t} includes the third order terms $\mathrm{X_{it}X_{jt}, X_{lt}}\ \mathrm{i} >= j\ >= 1\ \mathrm{i, j, l} = 2, \dots, \mathrm{K}.$

If K=3 we will have:

$\psi_{3t} = \mathrm{X}_{2t}^3 + \mathrm{X}_{3t}\mathrm{X}_{2t}^2 + \mathrm{X}_{3t}^2\mathrm{X}_{2t} + \mathrm{X}_{3t}^3$

Assuming T is large enough we can test:

$\mathrm{H}_0: \gamma_2' = 0 \text{ and } \gamma_3' = 0$

$\mathrm{H}_a: \gamma_2' \neq 0 \text{ and } \gamma_3' \neq 0$

Using an F-type test.

2) Lagrange multiplier test

It is an asymptotic test and is based on the R^2 of the artificial regression

$$\hat{\varepsilon}_t = (\beta_0 - \hat{\beta})^{/} X_t + \gamma_2{}^{/}\Psi_{2t} + \Upsilon_3{}^{/}\Psi_{3t} + v_t$$

The test statistic is

LM=$\mathrm{TR}^2 \overset{d}{\rightarrow} X_{(J)}^2$ under the null hypotheses, *J*=number of restrictions. For small samples, the F-test is preferable because of degrees of freedom adjustment.

3) *Regression specification error test (Reset)*

1) Estimate original model (H_0) and get the predicted values of the dependent variable. For RESET(2), run the following regression (H_1)

$$y_t = \beta^{/} X_t + \gamma^{/} \hat{y}_t^2 + v_t$$

Test statistic

$$\frac{(R_1^2 - R_0^2)}{(1 - R_1^2)/(T - K - 1)} \sim F(1, T - K - 1)$$

Under H_0 (the t-statistic can also be used.)

For RESET (3), run the following regression (H_1):

$$y_t = \beta^{/} X_t + \gamma_1^{/} \hat{y}^2{}_t + \gamma_2^{/} \hat{y}^3{}_t + v_t$$

Test statistic:

$$\frac{(R_1^2 - R_0^2)/2}{(1-R_1^2)/(T-K-2)} \sim F(2, T-K-2) \qquad \text{Under } \mathrm{H}_0.$$

Is this test statistic the same as the sum of squared error F-test?

<u>What to do when functional form assumption is invalid?</u>

1) All results of misspecification tests should be considered simultaneously because the assumptions are closely interrelated (ex. Non-linearity can lead to non-normality).

2) Postulate a more general model (Go back to theory). (plots of data may be helpful)

3) Use some normalizing transformation on the original variables $\mathrm{y_t}$ and $\mathrm{X_t}$.

Homoskedasticity

When this assumption is invalid, parameter estimates lose their BLUE, efficiency and asymptotic efficiency properties.

Static Homoskedasticity

Can use White and RESET-type tests. First, run original model and get the residuals (ε_t). Run the following artificial regression:

$$\hat{\varepsilon}_t^2 = a + \Delta^{/} \Psi_t + v_t$$

For the RESET-type test, $\Psi_t = \hat{y}^2{}_t$.

For the White test, Ψ_t includes KG2 polynomial in x’s (and linear terms).

Use the F-test to assess the significance of Δ’.

a) ***Dynamic Homoskedasticity***

Run the same artificial regression as in the static case with $\Psi_t = \hat{\varepsilon}^2_{t-1}$. Use the F-test to assess the significance of Δ'.

===> White's test for heteroskedasticity does not postulate a particular form of heteroskedasticity. It is important to use this test in conjunction with other tests based on particular forms of heteroskedasticity.

<u>What to do when the assumption of homoskedasticity is invalid?</u>

1) Diagnose the source giving rise to it (go back to theory) and respecify the statistical model to include a variance equation.

2) If heteroskedasticity is accompanied by non-normality and/or non-linearity, the obvious way to proceed is to seek an appropriate normalizing, variance-stabilizing transformation.

3) An alternative to 2) is to postulate a non-normal distribution and proceed to derive the conditional mean and conditional variance.

4) Use a robust estimator or use GMM (last choices).

Parameter stability (structural change)

We would like to test whether parameter estimates differ between the first and second half of the sample. Use a Chow test for stable β. Because the Chow test assumes equal variance, we need to conduct an F-test of variance equality, and then conduct another F-test for stable parameter β.

1) F-test for equal variance

$$\left(\frac{RSS_2}{RSS_1}\right)\left(\frac{(T_1-K_1)}{(T_2-K_2)}\right) \sim F(T_2-K_2, T_1-K_1) \text{ under } \mathrm{H}_0.$$

If you fail to reject H_0 then do:

2) F-test for equal β

$$\left(\frac{RSS_T-RSS_1-RSS_2}{RSS_1+RSS_2}\right)\left(\frac{(T-2K)}{K}\right) \sim F(K, T-2K) \text{ Under } \mathrm{H}_0$$

In practice, we usually use dummy variables and LM tests rather than Chow tests.

What to do when parameter stability assumption is invalid?

===> Go back to theory. What caused the structural change? Is there nonlinearity?

Independence

Often (most of the time) autocorrelation is an indication that the independence assumption is invalid. It can be tested using the following artificial regression.

$$\hat{\varepsilon}_t = \beta_0' X_t + \Lambda' \hat{\varepsilon}_{t-1} + v_t$$

Use a t-type test to assess the significance of Λ'. This test can be used to test for autoregressive or moving average errors of any order, whether or not the regressors include lagged dependent variables.

What to do when assumption of independence is invalid?

===> go back to theory (may need a dynamic model, functional form may be incorrect)

===> MLE with autocorrelation and Newey-West GMM are second choices

Joint tests

Conditional mean test $\mathrm{E}(\mathrm{y_t}|\mathrm{X_t} = \mathrm{x_t}) = \beta'\mathrm{X_t}$

It simultaneously tests parameter stability, functional form, and independence. It is based on the following artificial regression:

$$\hat{\varepsilon}_t = \beta_0' X_t + \Gamma_P^{/}\Psi_t^P + \Gamma_F^{/}\Psi_t^F + \Gamma_I^{/}\Psi_t^I + v_t$$

Where,

Ψ_t^P models structural change (binary variable or time trend)

Ψ_t^F models non-linearity (RESET2 or KG2)

Ψ_t^I allows for temporal (as well as spatial) dependence $(\hat{\varepsilon}_{t-1})$

$\mathrm{H_0}: \Gamma_\mathrm{P} = \Gamma_\mathrm{F} = \Gamma_\mathrm{I} = 0$

$\mathrm{H_a}: \Gamma_\mathrm{P} \neq 0 \text{ or } \Gamma_\mathrm{F} \neq \text{ or } \Gamma_\mathrm{I} \neq 0$

Conduct an F-type test under the null. Cause of rejection can be investigated by assessing separately the significance of $\Gamma_\mathrm{P}, \Gamma_\mathrm{F}, \text{and } \Gamma_\mathrm{I}$.

===>The overall test and individual component tests will be accurate if all other assumptions are valid (normality, homoskedasticity, and σ^2 stable)

Conditional variance $\mathrm{Var}(\mathrm{y_t}|\mathrm{X_t} = \mathrm{x_t}) = \sigma^2$

The test checks for dynamic and static heteroskedasticity as well as for stability of σ^2. It is based on the following artificial regression:

$$\hat{\varepsilon}_t^2 = \Gamma_P^{/}\Psi_t^P + \Gamma_S^{/}\Psi_t^S + \Gamma_D^{/}\Psi_t^D + v_t$$

Where

Ψ_t^P models structural change (binary variable or time trend)

Ψ_t^S allows for static heteroskedasticity (RESET2 or KG2 or KG1)

Ψ_t^D allows for dynamic heteroskedasticity (corr$(\hat{\varepsilon}_t^2, \hat{\varepsilon}_{t-1}^2) \neq 0)$

$\mathrm{H_0}: \Gamma_\mathrm{P} = \Gamma_\mathrm{S} = \Gamma_\mathrm{D} = 0$

$\mathrm{H_a}: \Gamma_\mathrm{P} \neq 0 - \text{or}\, \Gamma_\mathrm{S} \neq 0 \text{ or } \Gamma_\mathrm{D} \neq 0$

Conduct an F-type test under the null. Cause of rejection can be investigated by assessing separately the significance of $\Gamma_\mathrm{P}, \Gamma_\mathrm{S}, \text{and } \Gamma_\mathrm{D}$

===>The overall test and individual component tests will be accurate if the normality assumption is valid and the conditional mean is properly specified.

===> For the overall test, McGuirk et al. argue that the nominal test size should be closer to 0.1 than 0.05 since the costs of misspecification can be high. (Less true if sample size is large.)

===> They suggest adjusting individual tests for multiple comparisons. The adjustment following Sidak is $1 - (1-\alpha)^\mathrm{m}$ where m is the total number of tests conducted. Another test size adjustment by Bonferroni is αm which is less accurate than Sidak's for larger actual test sizes. (Sometimes called the multiple comparisons problem) (Note: this is a conservative approach that most users ignore, but probably should not)

Selecting a Useful Testing Strategy

How many and what tests should be conducted?

At least the validity of all testable assumptions should be examined. An incomplete set of tests can be misleading.

* A test regime is a comprehensive set of misspecification tests.

What constitutes significant evidence that a model is misspecified?

1) The size of the p-value that should be taken as significant evidence that a model is misspecified in any particular test depends on two factors:

i) The acceptable overall size of the regime, and

ii) The number of tests in the regime

Should a significance level less than .05 be used if the sample size is large?

References

D'Agostino, R.B, Albert B., and Ralph B. JR. "A Suggestion for Using Powerful and Informative Tests of Normality." *The American Statistician.* 44(Nov. 1990):316-321.

D'Agostino, R.B.. "Transformation to Normality of the Null Distribution of g_1" *Biometrika* 57(1970): 679-681.

Johnson, N.L. "Systems of Frequency Curves Generated by Methods of Translation" *Biometrika* 36(1949):149-176

McGuirk, A.M., Paul D., and Jeffrey A. misspecification Testing: A Comprehensive Approach." *American Journal of Agricultural Economics.* 75(Nov. 1993):1044-1055.

Spanos, A. *Statistical Foundation of Econometric Modelling.* New York: Cambridge University Press, 1986.