Algebra of Least Squares

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Review (Student Annotation)

- What is the regression model for observed data?

What function does the OLS estimator minimize?
What is the sample mean of the residuals?
What is the sample covariance between the residuals and the covariates?
How long ago were matrix methods used to solve systems of equations?
at least the 2nd century BCE and perhaps the 10th century BCE

Best Linear Prediction

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Review

- How do we approximate $\mathbb{E}(y|x)$?
- $x^{\mathrm{T}}\beta$

How to do you interpret β?

- the linear projection coefficient, which is not generally equal to $\nabla_x \mathbb{E}(y|x)$

What is required for identification of β?

- $\mathbb{E}\left(xx^{\mathrm{T}}\right)$ is invertible

What is the correlation between x and u?

- 0 by construction!

Causal Effects

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Review

- What are individual causal effects?
- $\nabla_x y$

Can individual causal effects be measured?

- No, because they require a counterfactual

What aggregate causal effect do we focus on?

- average causal effects

What is the first requirement for β_1 to equal the average causal effect?

- $\mathbb{E}(y|x_1, x_2) = x_1^{\mathrm{T}}\beta_1 + x_2^{\mathrm{T}}\beta_2$ so that $\beta_1 = \nabla_1 \mathbb{E}(y|x_1, x_2)$

What is the second requirement for β_1 to equal the average causal effect?

- Conditional Independence Assumption

$$f(u|x_1, x_2) = f(u|x_2)$$

Conditional Expectation Functions

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Overview

Reference: B. Hansen Econometrics Chapters 1 and 2.0 - 2.8

- most commonly applied econometrics tool
  - least-squares estimation (regression)
- tool to estimate
  - approximate conditional mean of dependent variable
  - as a function of covariates (regressors)
  - $(y, x_1, \ldots, x_K) := (y, x^{\mathrm{T}})$
- data is observational *not* experimental
  - causality is difficult to infer
- example - wages
  - random variable before measurement
  - observed wages are outcomes of the random variable
  - underpins the application of statistics to economics

Review

- Implication of observational data?
- causality is difficult to infer

Should we model $\mathbb{E}(y|x)$ as linear in x?

- no

What are the key properties of $e = y - \mathbb{E}(y|x)$?

- $\mathbb{E}(e|x) = 0 \qquad$ (by construction)
- uncorrelated with any function of x

Single-Equation GMM: Endogeneity Bias

Lecture for Economics 241B

Douglas G. Steigerwald

UC Santa Barbara

January 2012

Review

Stochastic Model

$$W_i = \beta_0 + \beta_1 S_i + U_i$$

- What two issues lead to correlation between S_i and U_i?
  - 1) endogeneity - latent ability that impacts both S_i and U_i
  - 2) measurement error in S_i
- What is the impact of the correlation on B_{OLS}?
  - biased and inconsistent
- What is needed to construct a consistent estimator?
  - 1) instrument Z_i $\;Cor(Z_i, S_i) \neq 0$ $\;Cor(Z_i, U_i) = 0$
  - 2) assumption about measurement error

Properties of OLS Estimators

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

OLSE General Setting

- general setting, k explanatory variables

Assumption 1 (Linear Regression Model).
The observations (y_i, x_i) come from a random sample and satisfy the linear regression equation

$$\begin{aligned} y_i &= x_i^{\mathrm{T}}\beta + u_i \\ \mathbb{E}\left(u_i | x_i\right) &= 0. \end{aligned}$$

The variables have finite second moments
$\mathbb{E}\left(y_i^2\right) < \infty$
$\mathbb{E}\left\|x_i\right\|^2 < \infty$,
and an invertible design matrix
$\mathbb{E}\left(x_i x_i^{\mathrm{T}}\right) > 0$ *(which means it is positive definite).*

Unbiasedness of the OLS Estimator

- applying the law of iterated expectations

$$\mathbb{E}\left(\widehat{\beta}\right) = \mathbb{E}\left(\mathbb{E}\left(\widehat{\beta}|X\right)\right) = \beta$$

- we have shown the following theorem

Theorem (Mean of OLSE).

In the linear regression model of Assumption 1:

1. $\mathbb{E}\left(\widehat{\beta}\right) = \beta$;

2. $\mathbb{E}\left(\widehat{\beta}|X\right) = \beta$.

1) says the distribution of $\widehat{\beta}$ is centered at β
2) a stonger result, $\widehat{\beta}$ is unbiased for any realization of X

Error Variance Assumptions

- we consider the general case of heteroskedastic regression

$$\mathbb{E}\left(u_i^2|x_i\right) = \sigma^2\left(x_i\right) = \sigma_i^2$$

- we also consider the specialized case of homoskedastic regression

Assumption 2 (Homoskedastic Linear Regression Model).
In addition to Assumption 1,

$$\mathbb{E}\left(u_i^2|x_i\right) = \sigma^2\left(x_i\right) = \sigma^2,$$

is independent of x_i.

Variance of OLSE

Theorem (Variance of the OLSE).

In the linear regression model of Assumption 1:

$$V := Var\left(\widehat{\beta}|X\right) = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}DX \left(X^{\mathrm{T}}X\right)^{-1}$$

with $D = diag(\sigma_1^2, \ldots, \sigma_n^2)$.

In the homoskedastic linear regression model of Assumption 2 $V = \left(X^{\mathrm{T}}X\right)^{-1} \sigma^2$.

Theorem (Gauss-Markov).

1. In the linear regression model of Assumption 1:
The best linear unbiased estimator is

$$\widetilde{\beta} = \left(X^{\mathrm{T}} D^{-1} X\right)^{-1} X^{\mathrm{T}} D^{-1} y.$$

2. In the homoskedastic linear regression model of Assumption 2:
The best linear unbiased estimator is the OLSE

$$\widehat{\beta} = \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}} y.$$

Gauss-Markov Theorem Interpretation

1. for linear (heteroskedastic) regression models
   1. BLUE is $\widetilde{\beta}$ - the Generalized LSE (not OLSE)
   2. infeasible as D is unknown
   3. need an estimator of D to have a practical alternative to $\widehat{\beta}$
2. for linear homoskedastic regression models
   1. BLUE is $\widehat{\beta}$
      1. special case of $\widetilde{\beta}$ with $D = I_n\sigma^2$
   2. limited efficiency result
      1. restricted to homoskedastic models and linear unbiased estimators
      2. could be biased or nonlinear estimators with lower MSE

Proof of Gauss-Markov Theorem

Residuals

- residuals

$$\begin{aligned} \widehat{u}_i &= y_i - x_i^{\mathrm{T}}\widehat{\beta} \\ \widehat{u} &= My = Mu \end{aligned}$$

  - $M = I_n - X\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}$ and $MX = \mathbf{0}$

- conditional mean

$$\mathbb{E}\left(\widehat{u}|X\right) = \mathbb{E}\left(Mu|X\right) = M\mathbb{E}\left(u|X\right) = 0$$

- conditional variance

$$Var\left(\widehat{u}|X\right) = M \cdot Var\left(u|X\right) \cdot M = MDM$$

Conditional Variance of Residuals

- under conditional homoskedasticity (Assumption 2)

$$\begin{aligned} \mathbb{E}\left(u_i^2 | x_i\right) &= \sigma^2 \\ Var\left(\widehat{u} | X\right) &= M\sigma^2 \end{aligned}$$

  - this follows from the fact that M is idempotent (and symmetric)
- u is homoskedastic but $\widehat{u}$ is heteroskedastic
  - conditional variance equals M not $I_n \sigma^2$
- i'th diagonal element of $M\sigma^2$ is $Var\left(\widehat{u}_i | X\right)$

$$Var\left(\widehat{u}_i | X\right) = \mathbb{E}\left(\widehat{u}_i^2 | X\right) = \left(1 - h_{ii}\right)\sigma^2$$

  - $h_{ii} = x_i^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)^{-1} x_i$
    - ★ a function of x_i, therefore residuals are heteroskedastic even if errors are homoskedastic

Standardized Residuals

- rescale to get constant conditional variance

$$\overline{u}_i = (1 - h_{ii})^{-1/2}\,\widehat{u}_i$$

$$Var\left(\overline{u}_i | X\right) = \sigma^2$$

- standardized residuals have conditional mean and conditional variance that are identical to the conditional mean and conditional variance of the errors u

Unbiased Error Variance Estimation

- bias takes scale form, therefore rescale to get unbiased estimator

$$s^2 = \frac{n}{n-k}\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\widehat{u}_i^2$$

  - by the earlier calculation $\mathbb{E}\left(s^2|X\right) = \sigma^2$ so $\mathbb{E}\left(s^2\right) = \sigma^2$
    - ★ bias-corrected estimator, widely used

- bias-corrected estimator can also be constructed from standardized residuals

$$\begin{aligned}\overline{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\overline{u}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}\widehat{u}_i^2 \\ \mathbb{E}\left(\overline{\sigma}^2|X\right) &= \sigma^2 \text{ so } \mathbb{E}\left(\overline{\sigma}^2\right) = \sigma^2\end{aligned}$$

- if n is not large relative to k, use a bias-corrected estimator

Heteroskedasticity-Consistent Standard Errors

Hayashi 2.5

Goal: Accurate standard error estimates with conditional heteroskedasticity

$$Var\left(\hat{\beta}\right) = \mathbb{E}\left[\left(\hat{\beta} - \beta\right)^2\right]$$

Scalar model

$$\left(\hat{\beta} - \beta\right)^2 = \frac{\left(\sum x_t u_t\right)^2}{\left(\sum x_t^2\right)^2}$$

General model

$$\left(\hat{\beta} - \beta\right)^2 = \left(\sum x_t x_t'\right)^{-1} \sum u_t^2 x_t x_t' \left(\sum x_t x_t'\right)^{-1}$$

Asymptotic Variance

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \sim\!\longrightarrow N\left(\mathbf{0}, \mathbf{\Sigma}_{xx}^{-1}\, S\, \mathbf{\Sigma}_{xx}^{-1}\right)$$

$\frac{1}{n}\sum x_t x_t' \xrightarrow{P} \mathbf{\Sigma}_{xx} \equiv \mathbb{E}\left[x_t x_t'\right]$

$\frac{1}{n}\sum u_t^2 x_t x_t' \xrightarrow{P} S$

- form of S depends on assumptions
  - martingale difference sequence assumption (no correlation)

    $$S = \mathbb{E}\left[(x_t u_t)^2\right]$$

  - serial correlation yields

    $$S = \sum_{j=0}^{J} \mathbb{E}\left[u_t u_{t-j} x_t x_{t-j}'\right]$$

Scalar to General Model

Scalar

$$Var\hat{\beta} = \frac{1}{\sum x_t^2} \sum x_t^2 u_t^2 \frac{1}{\sum x_t^2}$$

General

$$Var\hat{\beta} = \left(\sum x_t x_t'\right)^{-1} \sum u_t^2 x_t x_t' \left(\sum x_t x_t'\right)^{-1}$$

$$= \frac{1}{n}\left(\frac{1}{n}\sum x_t x_t'\right)^{-1} \frac{1}{n}\sum u_t^2 x_t x_t' \left(\frac{1}{n}\sum x_t x_t'\right)^{-1}$$

$$\equiv \frac{1}{n} S_{xx}^{-1} S \, S_{xx}^{-1}$$

Eicker-White Estimator

Need to estimate

$$S = \frac{1}{n} \sum u_t^2 x_t x_t'$$

Eicker-White estimator

$$\hat{S} = \frac{1}{n} \sum \hat{u}_t^2 x_t x_t'$$

$\hat{u}_t = y_t - x_t' \hat{\beta}$

$\hat{\beta}$ consistent for β (e.g. OLSE)

Heteroskedasticity-consistent standard errors

$$\widehat{se} = \sqrt{\frac{1}{n} S_{xx}^{-1} \hat{S} \, S_{xx}^{-1}}$$

Finite-Sample Accuracy

Test H_0 : $\beta_k = 0$

Statistic

$$\frac{\hat{\beta}_k}{\widehat{se}}$$

Size

$$\Pr\left(\text{reject } H_0 | H_0 \text{ true}\right)$$

in practice: over-rejection problem

nominal size 5%,empirical size $> 5\%$

reason: estimated standard error is too small

$\hat{u}_t^2$ is a downward biased estimator of u_t^2

Modification 1: Degrees-of-Freedom

Replace $\hat{S}$ with

$$\tilde{S} = \frac{1}{n-k} \sum \hat{u}_t^2 x_t x_t'$$

$n - k < n \Rightarrow$ estimated standard errors are larger

$$\widetilde{se} = \sqrt{\frac{1}{n} S_{xx}^{-1} \hat{S}\, S_{xx}^{-1}}$$

Modification 2: Influence

Under homoskedasticity

$$E\hat{u}_t^2 = \sigma^2 (1 - p_t)$$

Therefore, replace $\hat{u}_t^2$ with

$$\frac{\hat{u}_t^2}{1 - p_t}$$

yielding

$$\tilde{S} = \frac{1}{n} \sum \frac{\hat{u}_t^2}{1 - p_t} x_t x_t'$$

Influence Calculation

Observation t influence: p_t

Scalar model

$$p_t = \frac{x_t^2}{\sum x_t^2}$$

General model

$$p_t = x_t' \left(X'X \right)^{-1} x_t$$

Asymptotic Theory

Goal: Establish

$$\hat{S} \xrightarrow{p} S$$

Approach (outline for scalar case)

$$\sum \hat{u}_t^2 x_t^2 - \sum u_t^2 x_t^2 \xrightarrow{p} 0$$

Step 1: Algebra

$$\begin{aligned}
\hat{u}_t^2 &= \left((y_t - \beta x_t) - \left(\hat{\beta} - \beta \right) x_t \right)^2 \\
&= u_t^2 - 2 \left(\hat{\beta} - \beta \right) x_t u_t + \left(\hat{\beta} - \beta \right)^2 x_t^2
\end{aligned}$$

Asymptotic Theory Outline

Step 2: Moment form

$$\frac{1}{n}\sum \hat{u}_t^2 x_t^2 - \frac{1}{n}\sum u_t^2 x_t^2 =$$

$$-2\left(\hat{\beta}-\beta\right)\frac{1}{n}\sum u_t x_t^3 + \left(\hat{\beta}-\beta\right)^2 \frac{1}{n}\sum x_t^4$$

Key issue: convergence of

$$\frac{1}{n}\sum x_t^4 \text{ and } \frac{1}{n}\sum u_t x_t^3$$

Steps for convergence

- Establish moments exist and are finite
- Ergodic stationarity ensures sample moments converge

Asymptotic Theory: Final Step

$$\frac{1}{n}\sum \hat{u}_t^2 x_t^2 - \frac{1}{n}\sum u_t^2 x_t^2 =$$

$$-2\left(\hat{\beta}-\beta\right)\frac{1}{n}\sum u_t x_t^3 + \left(\hat{\beta}-\beta\right)^2 \frac{1}{n}\sum x_t^4$$

Because $\frac{1}{n}\sum x_t^4 \xrightarrow{p} c_1$ and $\frac{1}{n}\sum u_t x_t^3 \xrightarrow{p} c_2$

$$-2\left(\hat{\beta}-\beta\right)\frac{1}{n}\sum u_t x_t^3 + \left(\hat{\beta}-\beta\right)^2 \frac{1}{n}\sum x_t^4 \xrightarrow{p} 0$$

Hence

$$\frac{1}{n}\sum \hat{u}_t^2 x_t^2 - \frac{1}{n}\sum u_t^2 x_t^2 \xrightarrow{p} 0$$
$$\hat{S} - S \xrightarrow{p} 0$$

Economics 241B

Hypothesis Testing: Large Sample Inference

Statistical inference in large-sample theory is based on test statistics whose distributions are known under the truth of the null hypothesis. Derivation of these distributions is easier than in finite-sample theory because we are only concerned with the large-sample approximation to the exact distribution.

In what follows we assume that a consistent estimator of S exists, which we term $\hat{S}$. Recall that $S = E(g_t g_t')$, where $g_t = X_t U_t$.

Testing Linear Hypotheses

Consider testing a hypothesis regarding the k-th coefficient β_k. Proposition 2.1, which established the asymptotic distribution of the OLS estimator, implies that under $H_0 : \beta_k = \bar{\beta}_k$,

$$\sqrt{n}\left(B_k - \bar{\beta}_k\right) \underset{d}{\rightarrow} N\left(0, Avar\left(B_k\right)\right) \quad \text{and} \quad \widehat{Avar\left(B_k\right)} \underset{p}{\rightarrow} Avar\left(B_k\right).$$

Here B_k is the k-th element of the OLS estimator B and $Avar(B_k)$ is the (k,k) element of the $K \times K$ matrix $Avar(B)$. The key issue here is that we have not assumed conditional homoskedasticity, hence

$$\widehat{Avar\left(B_k\right)} = S_{XX}^{-1}\,\hat{S}\,S_{XX}^{-1},$$

which is the (heteroskedasticity-consistent) robust asymptotic variance. Under the Slutsky result (Lemma 2.4c), the resultant robust t-ratio

$$t_k \equiv \frac{\sqrt{n}\left(B_k - \bar{\beta}_k\right)}{\sqrt{\widehat{Avar\left(B_k\right)}}} = \frac{\left(B_k - \bar{\beta}_k\right)}{SE^*\left(B_k\right)} \underset{d}{\rightarrow} N\left(0,1\right),$$

where the robust standard error is $SE^* = \sqrt{\frac{1}{n}\widehat{Avar\left(B_k\right)}}$. Note this robust t-ratio is distinct from the t-ratio introduced under the finite-sample assumptions in earlier lectures.

To test $H_0 : \beta_k = \bar{\beta}_k$, simply follow these steps:

Step 1: Calculate the robust t-ratio

Step 2: Obtain the critical value from the $N(0,1)$ distribution

Step 3: Reject the null hypothesis if $|t_k|$ exceeds the critical value

There are several differences from the finite-sample test that relies on conditional homoskedasticity.

- The standard error is calculated in a different way, to accommodate conditional heteroskedasticity.
- The normal distribution is used to obtain critical values, rather than the t distribution.
- The actual (or empirical) size of the test is not necessarily equal to the nominal size. The difference between the actual size and the nominal size is the size distortion. Because the asymptotic distribution of the robust t ratio is standard normal, the size distortion shrinks to zero as the sample size goes to infinity.

To summarize these results, together with the behavior of the Wald statistic let us briefly recall the assumptions required for Proposition 2.1:

Assumption 2.1 (linearity):

$$Y_t = X_t'\beta + U_t \qquad (t = 1, \ldots, n,)$$

where X_t is a K-dimensional vector of regressors, β is a K-dimensional vector of coefficients and U_t is the latent error.

Assumption 2.2 (ergodic stationarity): *The $(K+1)$-dimensional vector stochastic process $\{Y_t, X_t\}$ is jointly stationary and ergodic.*

Assumption 2.3 (predetermined regressors): *All regressors are predetermined, in the sense that they are orthogonal to the contemporaneous error: $E(X_{tk}U_t) = 0$ for all t and $k(= 1, 2, \ldots, K)$. This can be written as*

$$E(g_t) = 0 \qquad \text{where } g_t \equiv X_t \cdot U_t.$$

Assumption 2.4 (rank condition): *The $K \times K$ matrix $E(X_tX_t')$ is nonsingular (and hence finite). We denote this matrix by Σ_{XX}.*

Assumption 2.5 (g_t is a martingale difference sequence with finite second moments): *$\{g_t\}$ is a martingale difference sequence (so by definition $E(g_t) = 0$). The $K \times K$ matrix of cross moments, $E(g_tg_t')$, is nonsingular. Let S denote Avar $(\bar{g})$ (the variance of the asymptotic distribution of $\sqrt{n}\bar{g}$, where $\bar{g} = \frac{1}{n}\sum_t g_t$).*

By Assumption 2.2 and the Ergodic Stationary Martingale Differences CLT, $S = E(g_t g_t')$.

Proposition 2.3 (robust t-ratio and Wald statistic): *Given a consistent estimator* $\hat{S}$ *of* S*, if Assumptions 2.1 to 2.5 hold, then*

a) *Under the null hypothesis* $H_0 : \beta_k = \bar{\beta}_k$,

$$t_k \underset{d}{\rightarrow} N(0,1)$$

b) *Under the null hypothesis* $H_0 : R\beta = r$*, where* R *is a* $\#r \times K$ *matrix (where* $\#r$*, the dimension of* r*, is the number of restrictions) of full row rank,*

$$W \equiv n \cdot (Rb - r)' \left\{ R\,\widehat{Avar(B)}\,R' \right\}^{-1} (Rb - r) \underset{d}{\rightarrow} \chi^2(\#r).$$

Proof: We have already established part a. Part b is a straightforward application of Lemma 2.4(d). Write W as

$$W = c_n' Q_n^{-1} c_n \quad \text{where } c_n = \sqrt{n}(Rb - r) \text{ and } Q_n = R\,\widehat{Avar(B)}\,R'.$$

Under H_0, $c_n = R\sqrt{n}(b - \beta)$, so Proposition 2.1 implies

$$c_n \underset{d}{\rightarrow} c \quad \text{where } c \sim N(0, RAvar(B)R').$$

Also by Proposition 2.1

$$Q_n \underset{p}{\rightarrow} Q \quad \text{where } Q \equiv RAvar(B)R'.$$

Because R is full row rank and $Avar(B)$ is positive definite, Q is invertible. Therefore, Lemma 2.4(d) implies

$$W \underset{d}{\rightarrow} c'Q^{-1}c.$$

Because c is normally distributed with dimension $\#r$, and because Q equals the variance of c, $c'Q^{-1}c = \chi^2(\#r)$. *QED*

The statistic W is a Wald statistic because it is constructed from unrestricted estimators (B and $\widehat{Avar(B)}$) that are not constrained by the null hypothesis. To test $H_0 :: R\beta = r$, simply follow these steps:

Step 1: Construct W.

where o_P here is $\left(\bar{Y}_n - \mu\right)' X_n$.

To test nonlinear hypotheses, given the asymptotic distribution of the estimator, we need

Lemma (delta method): *Let $\{B_n\}_{n\geq 1}$ be a sequence of K-dimensional random variables such that*

$$B_n \xrightarrow{P} \beta$$
$$n^{\frac{1}{2}} (B_n - \beta) \xrightarrow{D} N(0, V).$$

Let $g(\cdot) : R^K \to R^Q$ have continuous first derivatives, with $G(\beta)$ denoting the $Q \times K$ matrix of first derivatives evaluated at β:

$$G(\beta) = \frac{\partial g(\beta)}{\partial \beta'}.$$

Then

$$n^{\frac{1}{2}} (g(B_n) - g(\beta)) \xrightarrow{D} N\left(0, G(\beta) V G(\beta)'\right).$$

Proof: A mean-value expansion yields $n^{\frac{1}{2}} g(B_n) = n^{\frac{1}{2}} g(\beta) + n^{\frac{1}{2}} G(B_n^*) (B_n - \beta)$, where B_n^* is the mean value between B_n and β. Because $B_n \xrightarrow{P} \beta$, $B_n^* \xrightarrow{P} \beta$ and, by continuity, $G(B_n^*) \xrightarrow{P} G(\beta)$. Thus

$$n^{\frac{1}{2}} [g(B_n) - g(\beta)] = n^{\frac{1}{2}} G(B_n^*) (B_n - \beta) \xrightarrow{D} G(\beta) N(0, V).$$

Viewing Estimators as Sequences of Random Variables

Let B_n be an estimator of β. If $B_n \xrightarrow{P} \beta$, then B_n is consistent for β.
The asymptotic bias of an estimator is not an agreed upon concept. The most common definition is

$$\lim_{n\to\infty} E(B_n) - \beta,$$

so consistency and asymptotic unbiasedness are distinct concepts. Hayashi uses $p\lim_{n\to\infty} B_n - \beta$, so consistency and asymptotic unbiasedness are identical for him.
If , in addition, $n^{\frac{1}{2}} (B_n - \beta) \xrightarrow{D} N(0, V)$, then B_n is asymptotically Gaussian (and $n^{\frac{1}{2}}$ consistent, hence CAN). The asymptotic variance is V.

OLS Regression

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Covariance Matrix Estimation: Homoskedasticity

- to estimate $V_{\widehat{\beta}} = \left(X^{\mathrm{T}}X\right)^{-1}\sigma^2$

$$\widehat{V}^0_{\widehat{\beta}} = \left(X^{\mathrm{T}}X\right)^{-1} s^2$$

- unbiased

$$\mathbb{E}\left(\widehat{V}^0_{\widehat{\beta}}|X\right) = \left(X^{\mathrm{T}}X\right)^{-1}\mathbb{E}\left(s^2|X\right) = V_{\widehat{\beta}}$$

- substantial bias if the error is heteroskedastic
- suppose $\sigma_i^2 = x_i^2$ and $k = 1$

$$\frac{V_{\widehat{\beta}}}{\mathbb{E}\left(\widehat{V}^0_{\widehat{\beta}}|X\right)} = \frac{\sum_{i=1}^n x_i^4}{\sigma^2\sum_{i=1}^n x_i^2} \approx \frac{\mathbb{E}\left(x_i^4\right)}{\left(\mathbb{E}\left(x_i^2\right)\right)^2} = \kappa$$

- κ is the kurtosis (standardized fourth moment) of x_i
    - if x_i is $\mathcal{N}(0,1)$, $\kappa = 3$
        - true variance is 3 times larger than the expected $\widehat{V}^0_{\widehat{\beta}}$

Covariance Matrix Estimation: Heteroskedasticity

- to estimate $V_{\widehat{\beta}} = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}} D X \left(X^{\mathrm{T}}X\right)^{-1}$
  - $D = diag(\sigma_1^2, \ldots, \sigma_n^2)$
  - $\widehat{D}^{ideal} = diag(u_1^2, \ldots, u_n^2)$
- $\widehat{V}_{\widehat{\beta}}^{ideal} = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}} \widehat{D}^{ideal} X \left(X^{\mathrm{T}}X\right)^{-1}$ is unbiased

$$
\begin{aligned}
\mathbb{E}\left(\widehat{V}_{\widehat{\beta}}^{ideal} | X\right) &= \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}} \mathbb{E}\left(\widehat{D}^{ideal} | X\right) X \left(X^{\mathrm{T}}X\right)^{-1} \\
\mathbb{E}\left(u_i^2 | X\right) &= \sigma_i^2 \quad \Rightarrow \quad \mathbb{E}\left(\widehat{D}^{ideal} | X\right) = D
\end{aligned}
$$

- feasible estimators replace u_i^2 with $\widehat{u}_i^2$ (Eicker 1963, White 1980)

Feasible Covariance Matrix Estimators

Heteroskedasticity-Robust Estimators

- no bias correction

$$\widehat{V}_{\widehat{\beta}}^{W} = \left(X^{\mathrm{T}}X\right)^{-1}\left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \widehat{u}_i^2\right)\left(X^{\mathrm{T}}X\right)^{-1}$$

  - yet $\widehat{u}_i^2$ is biased toward zero

- bias-correction (termed Eicker-White)

$$\widehat{V}_{\widehat{\beta}} = \left(\frac{n}{n-k}\right)\left(X^{\mathrm{T}}X\right)^{-1}\left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \widehat{u}_i^2\right)\left(X^{\mathrm{T}}X\right)^{-1}$$

  - correction is ad hoc but preferable to $\widehat{V}_{\widehat{\beta}}^{W}$ (default method in Stata)
  - $X^{\mathrm{T}}DX = \sum_{i=1}^{n} x_i^2 \sigma_i^2$
    - weighted version of $X^{\mathrm{T}}X$

Standard Errors

- $\widehat{V}_{\widehat{\beta}}$ is an estimator of the variance of the distribution of $\widehat{\beta}$
- A standard error $s\left(\widehat{\beta}\right)$ for a real-valued estimator $\widehat{\beta}$ is an estimate of the standard deviation of the distribution of $\widehat{\beta}$
- if β is a vector with estimate $\widehat{\beta}$ and covariance matrix estimate $\widehat{V}_{\widehat{\beta}}$
    - standard error for $\widehat{\beta}_j$ is square-root of diagonal element $[j, j]$

$$s\left(\widehat{\beta}_j\right) = \sqrt{\widehat{V}_{\widehat{\beta}_j}} = \sqrt{\left[\widehat{V}_{\widehat{\beta}}\right]_{jj}}$$

Measures of Fit

- classic

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2} = \frac{\widehat{\sigma}^2}{\widehat{\sigma}_y^2}$$

  - estimates $\rho^2 = Var\left(x_i^{\mathrm{T}}\beta\right) / Var\left(y_i\right) = 1 - \sigma^2 / \sigma_y^2$

- $\widehat{\sigma}^2$ and $\widehat{\sigma}_y^2$ are biased estimators, Theil (1961) used unbiased estimators s^2 and $\widetilde{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$

$$\overline{R}^2 = 1 - \frac{s^2}{\widetilde{\sigma}_y^2} = 1 - \frac{(n-1)\sum_{i=1}^{n} \widehat{u}_i^2}{(n-k)\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}$$

Multicollinearity

- *strict* multicollinearity: $X^{\mathrm{T}}X$ is singular
  - columns of X are linearly dependent
    - there exists some $\alpha \neq 0$ such that $Xa = 0$
  - $\left(X^{\mathrm{T}}X\right)^{-1}$ and $\widehat{\beta}$ are not defined
  - arises only through mistakes, include hourly and weekly wages, everyone works 40 hours each week
- more relevant, *near* multicollinearity
  - columns of X are nearly linearly dependent
  - not clear what it means to be near
- affects precision of estimation
- if $\frac{1}{n}X^{\mathrm{T}}X = \frac{1}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

$$Var\left(\widehat{\beta}|X\right) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n\left(1-\rho^2\right)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

  - as $\rho \longrightarrow 1$ variance grows

Test Statistic

if standard errors are calculated using homoskedastic formula

$$
\begin{aligned}
\frac{\widehat{\beta}_j - \beta}{s\left(\widehat{\beta}_j\right)} &= \frac{\widehat{\beta}_j - \beta}{s\sqrt{\left[\left(X^{\mathrm{T}}X\right)^{-1}\right]_{jj}}} \sim \frac{\mathcal{N}\left(0, \sigma^2\left[\left(X^{\mathrm{T}}X\right)^{-1}\right]_{jj}\right)}{\sqrt{\frac{\sigma^2}{(n-k)}\chi^2_{(n-k)}}\sqrt{\left[\left(X^{\mathrm{T}}X\right)^{-1}\right]_{jj}}} \\
&= \frac{\mathcal{N}(0,1)}{\sqrt{\frac{1}{(n-k)}\chi^2_{(n-k)}}} \sim t_{n-k}
\end{aligned}
$$

Finite Sample Distribution

Theorem (Finite Sample Distribution).
In the linear regression model of Assumption 1, if u_i is independent of x_i and distributed $\mathcal{N}\left(0, \sigma^2\right)$ then

- $\widehat{\beta} - \beta \sim \mathcal{N}\left(0, \sigma^2 \left(X^{\mathrm{T}} X\right)^{-1}\right)$
- $\frac{n\widehat{\sigma}^2}{\sigma^2} = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k)}$
- $\frac{\widehat{\beta}_j - \beta}{s\left(\widehat{\beta}_j\right)} \sim t_{n-k}$

Projections and Influence

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Projection Matrix

- projection (hat) matrix

$$\underset{n\times n}{P} = X\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}$$

- why projection?
  - $PX = X\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}X = X$
    - holds for any matrix in the range space of X
- why hat?
  - $Py = X\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y = X\widehat{\beta} := \widehat{y}$
    - creates fitted values
    - $X = \mathbf{1}$ (n vector of ones) $P = \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$
    - $Py = \mathbf{1}\overline{y}$ (fitted value is the sample mean)

Leverage

- i^{th} diagonal element of P

$$h_{ii} = x_i^{\mathrm{T}} \left(X^{\mathrm{T}} X \right)^{-1} x_i$$

  - *leverage* of observation i
  - property 1: $0 \leq h_{ii} \leq 1$
  - property 2: $\sum_{i=1}^{n} h_{ii} = k$

Proof of Property 2

Orthogonal Projection

- orthogonal projection matrix (annihilator matrix)

$$M = I_n - P$$

- why orthogonal projection?
  - $MX = 0$ therefore M and X are orthogonal
- why annihilator matrix?
  - for any matrix Z in the range space of X
    - ★ $MZ = Z - PZ = 0$
  - examples
    - ★ $MX_1 = 0$
    - ★ $MP = 0$
- M creates least squares residuals

$$My = y - Py = y - \widehat{y} = \widehat{u}$$

An Interesting Fact Regarding the Variance Estimator

consider

$$\begin{aligned} \tilde{\sigma}^2 - \hat{\sigma}^2 &= n^{-1} u^{\mathrm{T}} u - n^{-1} \hat{u}^{\mathrm{T}} \hat{u} \\ &= n^{-1} u^{\mathrm{T}} \left(I_n - M \right) u \\ &= n^{-1} u^{\mathrm{T}} P u \\ &\geq 0 \end{aligned}$$

- the last inequality holds because
  - P is positive semidefinite
  - $u^{\mathrm{T}} P u$ is a quadratic form
- feasible estimator is numerically smaller than ideal estimator

Analysis of Variance Formula

- subtracting $\overline{y}$ from both sides of the decomposition

$$y - 1\overline{y} = (\widehat{y} - 1\overline{y}) + \widehat{u}$$

- orthogonal decomposition when X contains a constant: $1^{\mathrm{T}}\widehat{u} = 0$
- $(y - 1\overline{y})^{\mathrm{T}}(y - 1\overline{y}) = (\widehat{y} - 1\overline{y})^{\mathrm{T}}(\widehat{y} - 1\overline{y}) + \widehat{u}^{\mathrm{T}}\widehat{u}$
- analysis of variance formula for LS regression

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i-1}^{n}\widehat{u}_i^2$$

- coefficient of determination (algebraic measure of fit, we have better measures that require statistical derivation)

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i-1}^{n}\widehat{u}_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Regression Components

- partition $\quad X = [X_1 \quad X_2]$
- OLS regression of y on X yields
  - $y = X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + \widehat{u}$
- algebraic expressions for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ identical to algebra for population coefficients

$$
\begin{aligned}
\widehat{\beta}_1 &= \left(X_1^{\mathrm{T}} M_2 X_1\right)^{-1} \left(X_1^{\mathrm{T}} M_2 y\right) \\
\widehat{\beta}_2 &= \left(X_2^{\mathrm{T}} M_1 X_2\right)^{-1} \left(X_2^{\mathrm{T}} M_1 y\right)
\end{aligned}
$$

  - $M_1 = I_n - X_1 \left(X_1^{\mathrm{T}} X_1\right)^{-1} X_1^{\mathrm{T}}$
  - $M_2 = I_n - X_2 \left(X_2^{\mathrm{T}} X_2\right)^{-1} X_2^{\mathrm{T}}$
  - $\widehat{\beta}_1$ - projection onto M_2 removes component correlated with X_2
    - in essence, "holding X_2 constant"

Matrix Algebra Derivation

Residual Regression

First recognized by Frisch and Waugh (1933)

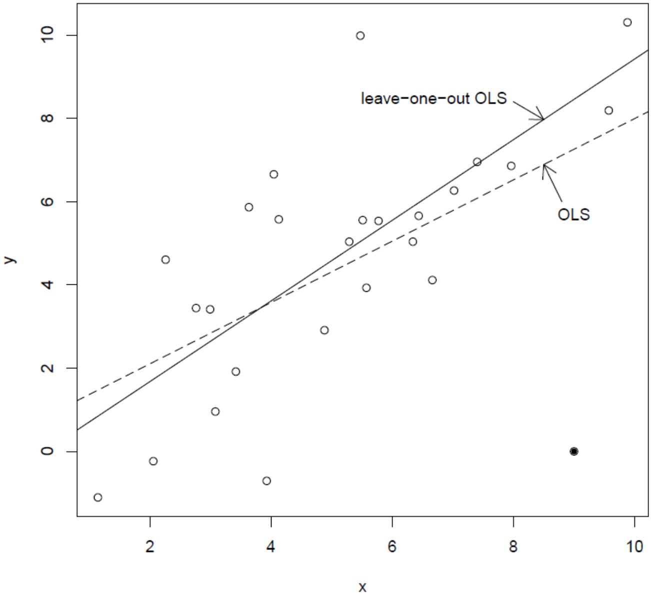
- because $M_1 = M_1 M_1$

$$\widehat{\beta}_2 = \left(X_2^{\mathrm{T}} M_1 M_1 X_2\right)^{-1} \left(X_2^{\mathrm{T}} M_1 M_1 y\right)$$
$$= \left(\widetilde{X}_2^{\mathrm{T}} \widetilde{X}_2\right)^{-1} \left(\widetilde{X}_2^{\mathrm{T}} \overline{u}_1\right)$$

  - $\widetilde{X}_2 = M_1 X_2 \qquad \overline{u}_1 = M_1 y$
  - proves the following theorem

Theorem (Frisch-Waugh-Lovell). In the linear model $y = X_1 \beta_1 + X_2 \beta_2 + u$ the OLS estimator of β_2 and the OLS residuals $\widehat{u}$ may be equivalently computed by either the OLS regression or via the following algorithm:

1. *Regress y on X_1, obtain residuals $\overline{u}_1$;*
2. *Regress X_2 on X_1, obtain residuals $\widetilde{X}_2$;*
3. *Regress $\overline{u}_1$ on $\widetilde{X}_2$, obtain OLSE $\widehat{\beta}_2$ and residuals $\widehat{u}$.*


leave−one−out OLS
OLS
y
x
0
2
4
6
8
10
2
4
6
8
10

Calculation of Influence

- for coefficients of interest, calculate for each i
  - $\widehat{\beta} - \widehat{\beta}_{(-i)} = \left(X^{\mathrm{T}}X\right)^{-1} x_i \widetilde{u}_i \qquad \widetilde{u}_i = (1 - h_{ii})\,\widehat{u}_i$
    - ★ DFBETA - post estimation diagnostic in STATA
    - ★ Is there a meaningful change? (no magic threshold)
    - ★ hard to recommend other proposed diagnostics (DFITS, Cook's Distance, Welsch Distance) - not based on statistical theory
- for general assessment, study predicted value
  - $\mathit{Influence} = \max_{1 \le i \le n} \left|\widehat{y}_i - \widetilde{y}_i\right|$
  - $\widehat{y}_i - \widetilde{y}_i = x_i^{\mathrm{T}}\widehat{\beta} - x_i^{\mathrm{T}}\widehat{\beta}_{(-i)} = h_{ii}\widetilde{u}_i$
  - observation i is influential for the predicted value if h_{ii} and $\left|\widetilde{u}_i\right|$ are large
    - ★ h_{ii} large - x_i is far from its sample mean, leverage point
    - ★ leverage points are not necessarily influential

Normal Regression Model

- linear regression model with u_i independent of x_i with a normal distribution
  - $u_i|x_i \sim \mathcal{N}\left(0,\sigma^2\right)$ which implies $y_i|x_i \sim \mathcal{N}\left(x_i^{\mathrm{T}}\beta,\sigma^2\right)$
- log-likelihood function

$$\begin{aligned}\log L\left(\beta,\sigma^2\right) &= \sum_{i=1}^{n}\log\frac{1}{\left(2\pi\sigma^2\right)^{\frac{1}{2}}}\exp\left(\frac{-1}{2\sigma^2}\left(y_i - x_i^{\mathrm{T}}\beta\right)\right)^2 \\ &= -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) + \frac{1}{2\sigma^2}SSE_n\left(\beta\right)\end{aligned}$$

  - β enters only through $SSE_n(\beta)$ thus $\widehat{\beta}_{mle} = \widehat{\beta}_{ols}$
  - $\widehat{\sigma}^2_{mle}$ - maximize $\log L\left(\widehat{\beta}_{mle},\sigma^2\right)$
    - ★ FOC

$$\frac{\partial}{\partial\sigma^2}\log L\left(\widehat{\beta}_{mle},\widehat{\sigma}^2\right) = -\frac{n}{2}\frac{1}{\widehat{\sigma}^2} + \frac{SSE_n\left(\widehat{\beta}_{mle}\right)}{2\left(\widehat{\sigma}^2\right)^2} = 0$$

    - ★ $\widehat{\sigma}^2_{mle} = \frac{1}{n}\sum\widehat{u}_i^2$

Regression

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Coefficient Decomposition

- $x^{\mathrm{T}} = \left(x_1, x_2^{\mathrm{T}}\right)$ with $\dim\left(x_1\right) = 1$

$$y = x_1 \beta_1 + x_2^{\mathrm{T}} \beta_2 + u \quad \textit{Matrix Expression}$$

- $\beta_1 = \left(\mathbb{E}\left(\widetilde{x}\widetilde{x}^{\mathrm{T}}\right)\right)^{-1} \mathbb{E}\left(\widetilde{x}\widetilde{y}\right)$
  - the correlation between x_1 and y after correlation with x_2 is removed
- "Two-Step" Procedure
  - regression 1: x_1 on x_2
    - ★ $x_1 = x_2^{\mathrm{T}} \gamma_2 + u_1$
  - regression 2: y on u_1
    - ★ $\beta_1 = \mathbb{E}\left(u_1 y\right) / \mathbb{E} u_1^2$

Normal Regression

Assume (y, x) are jointly normally distributed, which implies (u, x) are jointly normally distributed

- Best linear projection

$$y = x^{\mathrm{T}}\beta + u \quad \text{where } \beta = \left(\mathbb{E}\left(xx^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(xy\right)$$

- by construction $\mathbb{E}\left(xu\right) = 0$
    - together with jointly normal distribution, implies x and u are independent
        - $\mathbb{E}\left(u|x\right) = \mathbb{E}\left(u\right) = 0$
        - $\mathbb{E}\left(u^2|x\right) = \mathbb{E}\left(u^2\right) = \sigma^2$
- jointly normal distribution yields "classic regression model"

Approximation Error: Omitted Covariates

- long regression (Goldberger coined these terms)

$$y = x_1^{\mathrm{T}}\beta_1 + x_2^{\mathrm{T}}\beta_2 + u \qquad \mathbb{E}\left(xu\right) = 0 \qquad \beta = \left(\mathbb{E}\left(xx^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(xy\right)$$

- short regression

$$y = x_1^{\mathrm{T}}\gamma_1 + u_1 \qquad \mathbb{E}\left(x_1 u_1\right) = 0 \qquad \gamma_1 = \left(\mathbb{E}\left(x_1 x_1^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(x_1 y\right)$$

  - recall, both coefficient and error change

- the linear projection coefficient from the short regression is
  - $\gamma_1 = \beta_1 + \left(\mathbb{E}\left(x_1 x_1^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(x_1 x_2\right)\beta_2 + \left(\mathbb{E}\left(x_1 x_1^{\mathrm{T}}\right)\right)^{-1}\mathbb{E}\left(x_1 u\right)$
    - ★ $\mathbb{E}\left(xu\right) = 0 \Rightarrow \mathbb{E}\left(x_1 u\right) = 0$
  - $\gamma_1 \neq \beta_1$ unless $\mathbb{E}\left(x_1 x_2\right) = 0$ (or $\beta_2 = 0$)
    - ★ "omitted variable" bias

Short Regression Coefficient

in many cases, short regression coefficient is a bound

- Long regression: linear projection of log(wage) on $x = (ed, ab)$
  - ab intellectual ability, unobserved
- Short regression: linear projection of log(wage) on $x = ed$
  - ed and ab likely positively correlated
  - conditional on ed, ab likely increases wages
  - therefore $\gamma_1 = \beta_1 + c$, $c > 0$
    - an upper bound (not very useful here)

Review

- How do we express β in terms of covariances? (consider a single covariate)
- $\beta = Cov\left(x, y\right) / Var\left(x\right)$

How does this change if $Var\left(x\right) = Var\left(y\right)$?

- $\beta = Cov\left(x, y\right) / \left(sd\left(x\right) sd\left(y\right)\right) = Corr\left(x, y\right)$

How does this change, if we regress x on y?

- $\beta = Cov\left(x, y\right) / Var\left(y\right) = Corr\left(x, y\right)$ Identical!

How does the short (regression) coefficient differ from the long coefficient?

- $\gamma_{1(short)} = \beta_{1(long)} + \left(Var\left(x_1\right)\right)^{-1} Cov\left(x_1, x_2\right) \beta_{2(long)}$

When does the short (regression) coefficient provide a useful bound on the long coefficient?

- when the bias attenuated the measured response (sign of $\beta_{1(long)}$ differs from sign of $Cov\left(x_1, x_2\right) \beta_{2(long)}$)

Specification of Conditional Expectation Functions

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara

Heteroskedasticity

homoskedasticity

$$\mathbb{E}\left(e^2|x\right) = \sigma^2 \quad \text{(does not depend on } x\text{)}$$

heteroskedasticity

$$\mathbb{E}\left(e^2|x\right) = \sigma^2\left(x\right) \quad \text{(does depend on } x\text{)}$$

- unconditional variance $\mathbb{E}\left(\mathbb{E}\left(e^2|x\right)\right)$ constant by construction
  - formally, *conditional* heteroskedasticity
- heteroskedasticity is the leading case for empirical analysis

Review

- Why focus on $\mathbb{E}(y|x)$?
- "best" predictor

Suppose $\mathbb{E}(y|x) = x^{\mathrm{T}}\beta$. How to do you interpret β?

- $\nabla_x \mathbb{E}(y|x)$

What is required for causality?

- $\nabla_x e = 0$

When is $\mathbb{E}(y|x)$ known?

- x discrete

What is the leading empirical approach for dispersion?

- (conditional) heteroskedasticity $\mathbb{E}\left(e^2|x\right) = \sigma^2(x)$

241B Lecture
Stochastic Processes

Martingales

Let X_t be an element of Z_t. Then X_t is a martingale with respect to Z_t if

$$E\left[X_t | Z_{t-1}, Z_{t-2}, \ldots, Z_1\right] = X_{t-1} \text{ for all } t \geq 2.$$

- The collection $(Z_{t-1}, Z_{t-2}, \ldots)$ is called the information set at $t-1$
- If the conditioning information set is $(X_{t-1}, X_{t-2}, \ldots)$, then X_t is a martingale (it is implicit that X_t is a martingale with respect to X)
- If X_t is a martingale with respect to Z_t then X_t is a martingale (because Z_t contains X_t)
- The vector Z_t is a martingale if $E\left[Z_t | Z_{t-1}, Z_{t-2}, \ldots, Z_1\right] = Z_{t-1}$ for all $t \geq 2$
- If the process started in the infinite past, there is no need to include the qualifier "for all $t \geq 2$"
- For our large-sample results, it does not matter if the process started in the infinite past (simply that the process started before the first observation)

Random Walks

A leading example of a martingale process is a random walk. Let $\{U_t\}$ be vector independent white noise, so $EU_t = 0$ and the covariance matrix of U_t is finite. A random walk is a sequence of cumulative sums

$$Z_1 = U_1, Z_2 = U_1 + U_2, \ldots.$$

As the underlying white noise can be deduced from $\{Z_t\}$ via

$$U_1 = Z_1, U_2 = Z_2 - Z_1, \ldots$$

the two processes contain the same information (and the first difference of a random walk is independent white noise).

The following CLT extends the Lindberg-Levy CLT to stationary and ergodic m.d.s.

Ergodic Stationary Martingale Difference CLT: *Let $\{U_t\}$ be a vector martingale difference sequence that is stationary and ergodic with $E(U_t U_t') = \Omega$ and $\bar{U} = \frac{1}{n}\sum_{t=1}^{n} U_t$. Then*

$$\sqrt{n}\bar{U} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} U_t \underset{d}{\rightarrow} N(0, \Omega).$$

This CLT is applicable not just to i.i.d. sequences, but also to stationary martingale difference sequences, such as ARCH processes (although we have not yet allowed for serially correlated processes).

1. Because $\{U_t\}$ is an m.d.s. with mean zero, there is no need to subtract a mean from $\bar{U}$.

2. Because $\{U_t\}$ is stationary, the covariance matrix does not depend on t. It is implicit that the moments exist and are finite.

Laws of Large Numbers and Central Limit Theorems

To verify convergence in probability from basic principles, we would need to know the joint distribution of the first n elements in the sequence. To verify convergence almost surely, we would need to know the joint distribution of the entire (infinite dimensional) sequence. To verify convergence in quadratic mean, we would need to know the first two moments of the finite sample distribution of the estimator. In general, we do not know any of these distributions. Instead, we turn to laws of large numbers. Laws of large numbers are one method of establishing convergence in probability (or almost surely) to a constant. They are applied to sequences of random variables, for which each member of the sequence is an (increasing) sum of random variables, such as $\left\{\bar{Y}_n\right\}$. An LLN is strong if the convergence is almost sure and weak if the convergence is in probability.

Chebychev's Weak LLN:

$$" \lim_{n\to\infty} E\left(\bar{Y}_n\right) = \mu, \quad \lim_{n\to\infty} Var\left(\bar{Y}_n\right) = 0" \Rightarrow \bar{Y}_n \xrightarrow{p} \mu.$$

The following strong LLN assumes that $\{Y_t\}$ is i.i.d. but the variance does not need to be finite

Kolmogorov's Second Strong LLN: *Let $\{Y_t\}$ be i.i.d. with $EY_t = \mu$. Then $\bar{Y}_n \xrightarrow{a.s.} \mu$.*

Central limit theorems are a principal method of establishing convergence in distribution, as they govern the limit behavior of the difference between $\bar{Y}_n$ and $E\bar{Y}_n$ (which equals EY_t if $\{Y_t\}$ is i.i.d.) blown up by $\sqrt{n}$. For i.i.d. sequences, the only CLT we need is

Lindberg-Levy CLT: Let $\{Y_t\}$ be i.i.d. with $EY_t = \mu$ and $Var\left(Y_t\right) = \Omega$. Then

$$\sqrt{n}\left(\bar{Y}_n - \mu\right) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(Y_t - \mu\right) \xrightarrow{d} N\left(0, \Omega\right).$$

We read the above CLT as: a sequence of random vectors $\sqrt{n}\left(\bar{Y}_n - \mu\right)$ converges in distribution to a random vector whose distribution is $N\left(0, \Omega\right)$. To understand how to relate the above CLT for random vectors, with the CLT for random scalars, let λ be any vector of real numbers with the same dimension as Y_t. Now $\{\lambda' Y_t\}$ is a sequence of random variables with $E\left(\lambda' Y_t\right) = \lambda'\mu$ and $Var\left(\lambda' Y_t\right) = \lambda'\Omega\lambda$. The scalar version of Lindberg-Levy implies

$$\sqrt{n}\left(\lambda'\bar{Y}_n - \lambda'\mu\right) = \lambda'\sqrt{n}\left(\bar{Y}_n - \mu\right) \xrightarrow{d} N\left(0, \lambda'\Omega\lambda\right).$$