Econometrics Notes

Matteo Courthoud

Last updated: 2020-04-27

Contents

1 Matrix Algebra 13

1.1 Basics 13
1.2 Spectral Decomposition 15
1.3 Quadratic Forms and Definite Matrices 16
1.4 Matrix Calculus 16
1.5 References 17

2 Probability Theory 19

2.1 Probability 19
2.2 Random Variables 20
2.3 Moments 21
2.4 Inequalities 22
2.5 Theorems 23
2.6 Statistical Models 23
2.7 References 24

3 Asymptotic Theory 25

3.1 Convergence 25
3.2 Theorems 26
3.3 Ergodic Theory 28
3.4 Asymptotic Properties of Estimators 31
3.5 Asymptotic Properties of Test Statistics 31
3.6 References 32

4 OLS Algebra 33

4.1 The Gauss Markov Model 33
  4.1.1 `Matlab` 34
4.2 The OLS estimator 34
  4.2.1 `Matlab` 36
4.3 OLS residual properties 37
  4.3.1 `Matlab` 38
4.4 Finite Sample Properties of the OLS estimator 39
  4.4.1 `Matlab` 41
4.5 References 41

5 OLS Inference **43**
5.1 Asymptotic Theory of the OLS Estimator 43
 5.1.1 Gaussian Error Term 43
 5.1.2 Homoskedastic Error Term 44
 5.1.3 Heteroskedastic Error Term 44
 5.1.4 Heteroskedastic and Autocorrelated Error Term 46
 5.1.5 Fixed b asymptotics 49
 5.1.6 Fixed G asymptotics 49
 5.1.7 `Matlab` 51
5.2 Inference 52
 5.2.1 Hypothesis Testing 52
 5.2.2 Confidence Intervals 54
 5.2.3 `Matlab` 55
5.3 References 55

6 Endogeneity **57**
6.1 Instrumental Variables 57
 6.1.1 `Matlab` 60
6.2 GMM 61
 6.2.1 1-step GMM 61
 6.2.2 2-step GMM 62
 6.2.3 `Matlab` 62
6.3 Testing Overidentifying Restrictions 63
 6.3.1 Naive Test 63
 6.3.2 Hansen's Test 64
 6.3.3 Special Case: Conditional Homoskedasticity 64
6.4 Small-Sample Properties of 2SLS 65
6.5 Many Instrument Robust Estimation 65
 6.5.1 LIML 65
 6.5.2 JIVE 67
6.6 Hausman Test 68
6.7 References 69

7 Non-Parametric Estimation **71**
7.1 Introduction 71
7.2 Discrete x - Cell Estimator 71
7.3 Local Non-Parametric Estimation - Kernels 72
 7.3.1 Estimator examples: 72
 7.3.2 Kernel examples: 75
 7.3.3 Choice of the optimal bandwidth 77
 7.3.4 Inference 77
 7.3.5 Bias-variance trade-off 79
7.4 Global Non-Parametric Estimation - Series 80
 7.4.1 Examples 81
 7.4.2 Estimation 81

7.4.3 Choice of the optimal K . . . 83
7.4.4 Inference . . . 83
7.4.5 Kernel vs Series . . . 86
7.5 References . . . 87

8 Variable Selection 89
8.1 Lasso . . . 89
8.1.1 Choosing the optimal lambda . . . 91
8.2 Pre-Testing . . . 93
8.2.1 Omitted Variable Bias . . . 93
8.2.2 Pre-test bias . . . 93
8.2.3 Partioned Regression . . . 98
8.3 Post Double Selection . . . 99
8.4 References . . . 101

9 Matlab Code 103
9.1 Lecture 1 . . . 103
9.2 Lecture 2 . . . 105
9.3 Lecture 3 . . . 106

List of Tables

List of Figures

5.1 Autocorrelation Function . . . 48
5.2 Fixed-b . . . 49

6.1 RJIVE . . . 66
6.2 IV to OLS . . . 66

7.1 NW regression . . . 73
7.2 LL regression . . . 74
7.3 Kernelsh . . . 76
7.4 Optimal Bandwidth . . . 78
7.5 Hermite Polynomials . . . 82
7.6 Smoothing . . . 84

8.1 Constraints . . . 90
8.2 Lasso Path . . . 92
8.3 Pre-test Bias . . . 95
8.4 Pre-test Bias . . . 96
8.5 Pre-test Bias . . . 97
8.6 Pre-test Bias . . . 100

Preface

Welcome to my lecture notes for PhD Econometrics! You can download the full PDF using the button at the top-left corner of the page.

These notes were initially born as my personal summary for the PhD Econometrics course of professor Damian Kozbur in Zurich. The first draft was the result of an intense collaborative effort together with Chiara Aina and Paolo Mengano. They helped a lot in both drafting and revising the original notes. During the years I have expanded the first draft in order to make it more comprehensive and include Matlab code examples. I will try to be as transparent as possible on the sources of the material but all errors are mine.

On the left, you can find the ordered table of contents. The first three lectures cover basics in matrix algebra, probability theory and large sample theory. The fourth lecture introduces the Gauss-Markov Model and the OLS estimator. The other lectures follow. At the end of each section, I will post some sample code in Matlab to replicate the main results. In a separate page I have collected all the Matlab code present in the notes.

All comments are very welcome. If you want to suggest edits or signal typos (I know there are many), please let me know at matteo.courthoud@econ.uzh.ch. This webpage was created using Bookdown.

References

- Kozbur (2019), PhD Econometrics - Lecture Notes.
- Hansen (2019), "*Econometrics*".
- Wooldridge (2010), "*Econometric Analysis of Cross Section and Panel Data*".
- Greene (2006), "*Econometric Analysis*".
- Hayashi (2000), "*Econometrics*".

Chapter 1

Matrix Algebra

1.1 Basics

A real $n \times m$ matrix A is an array

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nm} \end{bmatrix}$$

We write $[A]_{ij} = a_{ij}$ to indicate the (i, j)-element of A.

> We will usually take the convention that a real vector $x \in \mathbb{R}^n$ is identified with an $n \times 1$ matrix.

The $n \times n$ **identity matrix** I_n is given by

$$[I_n]_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Fundamental operations on matrices:

1. Two $n \times m$ matrices, A, B, are added element-wise so that $[A + B]_{ij} = [A]_{ij} + [B]_{ij}$.
2. A matrix A can be multiplied by a scalar $c \in \mathbb{R}$ in which case we set $[cA]_{ij} = c[A]_{ij}$.
3. An $n \times m$ matrix A can be multiplied with an $m \times p$ matrix B.
4. The product AB is defined according to the rule $[AB]_{ij} = \sum_{k=1}^{m} [A]_{ik}[B]_{kj}$.
5. An $n \times n$ matrix is invertible if there exists a matrix B such that $AB = I$. In this case, we use the notational convention of writing $B = A^{-1}$.
6. Matrix transposition is defined by $[A']_{ij} = [A]_{ji}$.

The **trace** of a square matrix A with dimension $n \times n$ is $\text{tr}(A) = \sum_{i=1}^{n} a_{ii}$.

The **determinant** of a square $n \times n$ matrix A is defined according to one of the following three (equivalent) definitions.

1. Recursively as $det(A) = \sum_{i=1}^{n} a_{ij}(-1)^{i+j} det([A]_{-i,-j})$ where $[A]_{-i,-j}$ is the matrix obtained by deleting the ith row and the jth column.
2. $A \mapsto det(A)$ under the unique alternating multilinear map on $n \times n$ matrices such that $I \mapsto 1$.

Vectors $x_1, ..., x_k$ are **linearly independent** if the only solution to the equation $b_1 x_1 + ... + b_k x_k = 0,\ b_j \in \mathbb{R}$, is $b_1 = b_2 = ... = b_k = 0$.

Useful matrix identities:

- $(A + B)' = A' + B'$
- $(AB)C = A(BC)$
- $A(B + C) = AB + AC$
- $(AB') = B'A'$
- $(A^{-1})' = (A')^{-1}$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $\text{tr}(cA) = c\text{tr}(A)$
- $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$
- $\text{tr}(AB) = \text{tr}(BA)$
- $det(I) = 1$
- $det(cA) = c^n det(A)$ if A is $n \times n$ and $c \in \mathbb{R}$
- $det(A) = det(A')$
- $det(AB) = det(A)det(B)$
- $det(A^{-1}) = (det(A))^{-1}$
- A^{-1} exists iff $det(A) \neq 0$
- $rank(A) = rank(A') = rank(A'A) = rank(AA')$
- A^{-1} exists iff $rank(A) = n$ for A $n \times n$
- $rank(AB) \leq \min\{rank(A), rank(B)\}$

The **rank** of a matrix, $rank(A)$ is equal to the maximal number of linearly independent rows for A.

Let A be an $n \times n$ matrix. The $n \times 1$ vector $x \neq 0$ is an **eigenvector** of A with corresponding **eigenvalue** λ is $Ax = \lambda x$.

The following types of matrices are defined:

1. A matrix A is diagonal if $[A]_{ij} \neq 0$ only if $i = j$.
2. An $n \times n$ matrix A is orthogonal if $A'A = I$
3. A matrix A is symmetric if $[A]_{ij} = [A]_{ji}$.
4. An $n \times n$ matrix A is idempotent if $A^2 = A$.
5. The matrix of zeros ($[A]_{ij} = 0$ for each i, j) is simply denoted 0.
6. An $n \times n$ matrix A is nilpotent if $A^k = 0$ for some integer $k > 0$.

1.2 Spectral Decomposition

Theorem: Let A be an $n \times n$ symmetric matrix. Then A can be factored as $A = C\Lambda C'$ where C is orthogonal and Λ is diagonal.

If we postmultiply A by C, we get

- $AC = C\Lambda C'C$ and
- $AC = C\Lambda$.

  This is a matrix equation which can be split into columns. The ith column of the equation reads $Ac_i = \lambda_i c_i$ which corresponds to the definition of eigenvalues and eigenvectors. So if the decomposition exists, then C is the eigenvector matrix and Λ contains the eigenvalues.

Theorem: The trace of a symmetric matrix equals the sum of its eigenvalues. The determinant of a symmetric matrix equals the product of its eigenvalues.

Theorem: The rank of a symmetric matrix equals the number of non zero eigenvalues.

Proof: $rank(A) = rank(C\Lambda C') = rank(\Lambda) = |\{i : \lambda_i \neq 0\}|$.

■

Theorem: The nonzero eigenvalues of AA' and $A'A$ are identical.

Theorem: The trace of a symmetric matrix equals the sum of its eignevalues.

Proof: $tr(A) = tr(C\Lambda C') = tr((C\Lambda)C') = tr(C'C\Lambda) = tr(\Lambda) = \sum_{i=1}^{n} \lambda_i$.

■

Theorem: The determinant of a symmetric matrix equals the product of its eignevalues.

Proof: $det(A) = det(C\Lambda C') = det(C)det(\Lambda)det(C') = det(C)det(C')det(\Lambda) = det(CC')det(\Lambda) = det(I)det(\Lambda) = det(\Lambda) = \prod_{i=1}^{n} \lambda_i$.

■

Theorem: For any symmetric matrix A, the eigenvalues of A^2 are the square of the eignevalues of A, and the eigenvectors are the same.

Proof: $A = C\Lambda C' \implies A^2 = C\Lambda C'C\Lambda C' = C\Lambda I\Lambda C' = C\Lambda^2 C'$

■

Theorem: For any symmetric matrix A, and any integer $k > 0$, the eigenvalues of A^k are the kth power of the eignevalues of A, and the eigenvectors are the same.

Theorem: Any square symmetric matrix A with positive eigenvalues can be written as the product of a lower triangular matrix L and its (upper triangular) transpose $L' = U$. That is $A = LU = LL'$

Note that

$$A = LL' = LU = U'U = (L')^{-1}L^{-1} = U^{-1}(U')^{-1}$$

where L^{-1} is lower triangular and U^{-1} is upper trianguar. You can check this for the 2×2 case. Also note that the validity of the theorem can be extended to symmetric matrices with non- negative eigenvalues by a limiting argument. However, then the proof is not constructive anymore.

1.3 Quadratic Forms and Definite Matrices

A **quadratic form** in the $n \times n$ matrix A and $n \times 1$ vector x is defined by the scalar $x'Ax$.

1. A is negative definite (ND) if for each $x \neq 0$, $x'Ax < 0$
2. A is negative semidefinite (NSD) if for each $x \neq 0$, $x'Ax \leq 0$
3. A is positive definite (PD) if for each $x \neq 0$, $x'Ax > 0$
4. A is positive semidefinite (PSD) if for each $x \neq 0$, $x'Ax \geq 0$

Theorem: Let A be a symmetric matrix. Then A is PD(ND) $\iff$ all of its eigenvalues are positive (negative).

Some more results:

1. If a symmetric matrix A is PD (PSD, ND, NSD), then $\det(A) > (\geq, <, \leq) 0$.
2. If symmetric matrix A is PD (ND) then A^{-1} is symmetric PD (ND).
3. The identity matrix is PD (since all eigenvalues are equal to 1).
4. Every symmetric idempotent matrix is PSD (since the eigenvalues are only 0 or 1).

Theorem: If A is $n \times k$ with $n > k$ and $rank(A) = k$, then $A'A$ is PD and AA' is PSD.

The **semidefinite partial order** is defined by $A \geq B$ iff $A - B$ is PSD.

Theorem: Let A, B be symmetric,square , PD, conformable. Then $A-B$ is PD iff $A^{-1}-B^{-1}$ is PD.

1.4 Matrix Calculus

We first define matrices blockwise when they are conformable. In particular, we assume that if A_1, A_2, A_3, A_4 are matrices with appropriate dimensions then the matrix

$$A = \begin{bmatrix} A_1 & A_1 \\ A_3 & A_4 \end{bmatrix}$$

is defined in the obvious way.

Let $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^p \times \mathbb{R}^q$ be a matrix valued function. More precisely, given a real $m \times n$ matrix X, $F(X)$ returns the $p \times q$ matrix

$$\begin{bmatrix} f_{11}(X) & ... & f_{1q}(X) \\ \vdots & \ddots & \vdots \\ f_{p1}(X) & ... & f_{pq}(X) \end{bmatrix}$$

The derivative of F with respect to the matrix X is the $mp \times nq$ matrix

$$\frac{\partial F(X)}{\partial X} = \begin{bmatrix} \frac{\partial F(X)}{\partial x_{11}} & \cdots & \frac{\partial F(X)}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F(X)}{\partial x_{m1}} & \cdots & \frac{\partial F(X)}{\partial x_{mn}} \end{bmatrix}$$

where each $\frac{\partial F(X)}{\partial x_{ij}}$ is a $p \times q$ matrix given by

$$\frac{\partial F(X)}{\partial x_{ij}} = \begin{bmatrix} \frac{\partial f_{11}(X)}{\partial x_{ij}} & \cdots & \frac{\partial f_{1q}(X)}{\partial x_{ij}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{p1}(X)}{\partial x_{ij}} & \cdots & \frac{\partial f_{pq}(X)}{\partial x_{ij}} \end{bmatrix}$$

The most important case is when $F : \mathbb{R}^n \to \mathbb{R}$ since this simplifies the derivation of the least squares estimator. Also, the trickiest thing is to make sure that dimensions are correct.

Useful results in matrix calculus:

1. $\frac{\partial b'x}{\partial x} = b$ for $dim(b) = dim(x)$
2. $\frac{\partial B'x}{\partial x} = B$ for arbitrary, conformable B
3. $\frac{\partial B'x}{\partial x'} = B'$ for arbitrary, conformable B
4. $\frac{\partial x'Ax}{\partial x} = (A + A')x$
5. $\frac{\partial x'Ax}{\partial A} = xx'$
6. $\frac{\partial x'Ax}{\partial x} = det(A)(A^{-1})'$
7. $\frac{\partial \ln det(A)}{\partial A} = (A^{-1})'$

1.5 References

- Kozbur (2019). PhD Econometrics - Lecture Notes.
- Greene (2006). "*Econometric Analysis*". Appendix A: Matrix Algebra.

Chapter 2

Probability Theory

2.1 Probability

A **probability space** is a triple $(\Omega, \mathcal{A}, P)$ where

- Ω is the sample space.
- $\mathcal{A}$ is the σ-algebra on Ω.
- P is a probability measure.

The **sample space** Ω is the space of all possible events.

A nonempty set (of subsets of Ω) $\mathcal{A} \in 2^{\Omega}$ is a **sigma algebra** (σ-algebra) of Ω if the following conditions hold:

1. $\Omega \in \mathcal{A}$
2. If $A \in \mathcal{A}$, then $(\Omega - A) \in \mathcal{A}$
3. If $A_1, A_2, ... \in \mathcal{A}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

   The smallest σ-algebra is $\{\emptyset, \Omega\}$ and the largest one is 2^{Ω} (in cardinality terms).

Suppose $\Omega = \mathbb{R}$. Let $\mathcal{C} = \{(a, b], -\infty \leq a < b < \infty\}$. Then the **Borel σ- algebra** on $\mathbb{R}$ is defined by

$$\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{C})$$

A **probability measure** P is a set function with domain $\mathcal{A}$ and codomain $[0, 1]$ such that

1. $P(A) \geq 0 \ \forall A \in \mathcal{A}$
2. P is σ-additive: is $A_n \in \mathcal{A}$ are pairwise disjoint events ($A_j \cap A_k = \emptyset$ for $j \neq k$), then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P\left(A_n\right)$$

3. $P(\Omega) = 1$

   Properties

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$
- For $A, B \in \mathcal{A}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- For $A, B \in \mathcal{A}$, if $A \subset B$ then $P(A) \leq P(B)$
- For $A_n \in \mathcal{A}$, $P\left(\cup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n)$
- For $A_n \in \mathcal{A}$, if $A_n \uparrow A$ then $\lim_{n\to\infty} P(A_n) = P(A)$

Let $A, B \in \mathcal{A}$ and $P(B) > 0$, the **conditional probability** of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Two events A and B are **independent** if $P(A \cap B) = P(A)P(B)$.

Theorem (Law of Total Probability): Let $(E_n)_{n\geq 1}$ be a finite or countable partition of Ω. Then, if $A \in \mathcal{A}$,

$$P(A) = \sum_n P(A|E_n)P(E_n)$$

Theorem (Bayes Theorem): Let $(E_n)_{n\geq 1}$ be a finite or countable partition of Ω, and suppose $P(A) > 0$. Then,

$$P(E_n|A) = \frac{P(A|E_n)P(E_n)}{\sum_m P(A|E_m)P(E_m)}$$

2.2 Random Variables

A **random variable** X on a probability space $(\Omega, \mathcal{A}, P)$ is a (measurable) mapping $X : \Omega \to \mathbb{R}$ such that

$$\forall B \in \mathcal{B}(\mathbb{R}), \quad X^{-1}(B) \in \mathcal{A}$$

> The measurability condition states that the inverse image is a measurable set of Ω i.e. $X^{-1}(B) \in \mathcal{A}$. This is essential since probabilities are defined only on $\mathcal{A}$.

Let X be a real valued random variable. The **distribution function** (also called cumulative distribution function) of X, commonly denoted $F_X(x)$ is defined by

$$F_X(x) = \Pr(X \leq x)$$

Properties

- F is monotone non-decreasing
- F is right continuous
- $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$

The random variables $(X_1, .., X_n)$ are independent if and only if

$$F_{(X_1,\ldots,X_n)}(x) = \prod_{i=1}^{n} F_{X_i}(x_i) \quad \forall x \in \mathbb{R}^n$$

Let X be a real valued random variable. X has a **probability density function** if there exists $f_X(x)$ such that for all measurable $A \subset \mathbb{R}$,

$$P(X \in A) = \int_A f_X(x)\mathrm{d}x$$

2.3 Moments

The **expected value** of a random variable, when it exists, is given by

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathrm{d}P$$

When X has a density, then

$$\mathbb{E}[X] = \int_\mathbb{R} x f_X(x)\mathrm{d}x = \int_\mathbb{R} x\mathrm{d}F_X(x)$$

The **empirical expectation** (or **sample average**) is given by

$$\mathbb{E}_n[z_i] = \frac{1}{n}\sum_{i=1}^{N} z_i$$

The **covariance** of two random variables X, Y defined on Ω is

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]E[Y]$$

In vector notation, $Cov(X,Y) = \mathbb{E}[XY'] - \mathbb{E}[X]\mathbb{E}[Y']$.

The **variance** of a random variable X, when it exists, is given by

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

In vector notation, $Var(X) = \mathbb{E}[XX'] - \mathbb{E}[X]\mathbb{E}[X']$.

Properties

Let $X, Y, Z, T \in \mathcal{L}^2$ and $a, b, c, d \in \mathbb{R}$

- $Cov(X,X) = Var(X)$
- $Cov(X,Y) = Cov(Y,X)$
- $Cov(aX + b, Y) = a\ Cov(X,Y)$
- $Cov(X + Z, Y) = Cov(X,Y) + Cov(Z,Y)$
- $Cov(aX + bZ, cY + dT) = ac * Cov(X,Y) + ad * Cov(X,T) + bc * Cov(Z,Y) + bd * Cov(Z,T)$

Let $X, Y \in \mathcal{L}^1$ be independent. Then, $\mathbb{E}[XY] = \mathbb{E}[X]E[Y]$.

If X and Y are independent, then $Cov(X,Y) = 0$.

Note that the converse does not hold: $Cov(X,Y) = 0 \not\Rightarrow X \perp Y$.

The **sample variance** is given by

$$Var_n(z_i) = \frac{1}{n}\sum_{i=1}^{N}(z_i - \bar{z})^2$$

where $\bar{z}_i = \mathbb{E}_n[z_i] = \frac{1}{n}\sum_{i=1}^{N} z_i$.

Theorem: The expected sample variance $\mathbb{E}[\sigma_n^2] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{N}\left(y_i - \mathbb{E}_n[y]\right)^2\right]$ gives an estimate of the population variance that is biased by a factor of $\frac{1}{n}$ and is therefore referred to as **biased sample variance**.

Proof:

$$\begin{aligned}
&\mathbb{E}[\sigma_n^2] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbb{E}_n[y]\right)^2\right] = \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)^2\right] = \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[y_i^2 - \frac{2}{n}y_i\sum_{j=1}^{n} y_j + \frac{1}{n^2}\sum_{j=1}^{n} y_j\sum_{k=1}^{n} y_k\right] = \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}\mathbb{E}[y_i^2] - \frac{2}{n}\sum_{j\neq i}\mathbb{E}[y_i y_j] + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j}\mathbb{E}[y_j y_k] + \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}[y_j^2]\right] = \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n}(\mu^2+\sigma^2) - \frac{2}{n}(n-1)\mu^2 + \frac{1}{n^2}n(n-1)\mu^2 + \frac{1}{n^2}n(\mu^2+\sigma^2)\right] = \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

■

2.4 Inequalities

- **Triangle Inequality**: if $\mathbb{E}[X] < \infty$, then

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$$

- **Markov's Inequality**: if $\mathbb{E}[X] < \infty$, then

$$\Pr(|X| > t) \leq \frac{1}{t}\mathbb{E}[|X|]$$

- **Chebyshev's Inequality**: if $\mathbb{E}[X^2] < \infty$, then

$$\Pr(|X-\mu| > t\sigma) \leq \frac{1}{t^2} \Leftrightarrow \Pr(|X-\mu| > t) \leq \frac{\sigma^2}{t^2}$$

- **Cauchy-Schwarz's Inequality**:

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

- **Minkowski Inequality**:

$$\left(\sum_{k=1}^{n}|x_k+y_k|^p\right)^{\frac{1}{p}} \leq \left(\sum_{k=1}^{n}|x_k|^p\right)^{\frac{1}{p}} + \left(\sum_{k=1}^{n}|y_k|^p\right)^{\frac{1}{p}}$$

- **Jensen's Inequality**: if $g(\cdot)$ is concave (e.g. logarithmic function), then

$$\mathbb{E}[g(x)] \leq g(\mathbb{E}[x])$$

  Similarly, if $g(\cdot)$ is convex (e.g. exponential function), then

$$\mathbb{E}[g(x)] \geq g(\mathbb{E}[x])$$

2.5 Theorems

Theorem: Law of Iterated Expectations

$$\mathbb{E}(Y) = \mathbb{E}_X[\mathbb{E}(Y|X)]$$

Theorem: Law of Total Variance

$$Var(Y) = Var_X(\mathbb{E}[Y|X]) + \mathbb{E}_X[Var(Y|X)]$$

Useful distributional results:

1. $\chi^2_q \sim \sum_{i=1}^{q} Z_i^2$ where $Z_i \sim N(0,1)$
2. $F(n_1,n_2) \sim \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$
3. $t_n \sim \frac{Z}{\sqrt{\chi^2_n/n}}$

   The t distribution is approximately standard normal but has heavier tails. The approximation is good for $n \geq 30$: $t_{n\geq 30} \sim N(0,1)$

2.6 Statistical Models

A **statistical model** is a set of probability distributions. More precisely, a **statistical model over data** $D \in \mathcal{D}$ is a set of probability distribution over datasets D which takes values in $\mathcal{D}$.

Suppose you have regression data $\{\curvearrowleft_i, y_i\}_{i=1}^{N}$ with $\curvearrowleft_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. The statistical model is

$$\Big\{P : y_i = f(\curvearrowleft_i) + \varepsilon_i,\ x_i \sim F_x,\ \varepsilon_i \sim F_\varepsilon,\ \varepsilon_i \perp \curvearrowleft_i,\ f \in C^2(\mathbb{R}^p)\Big\}$$

> In words: the statistical model is the set of distributions P such that an additive decomposition of y_i as $f(\curvearrowleft_i)+\varepsilon_i$ exists for some $\curvearrowleft_i$; where f is twice continuously differentiable.

A statistical model parameterized by $\theta \in \Theta$ is **well specified** if the data generating process corresponds to some θ_0 and $\theta_0 \in \Theta$. Otherwise, the statistical model is **misspecified**.

A statistical model can be parametrized as $\mathcal{F} = \{P_\theta\}_{\{\theta \in \Theta\}}$.

Categories of statistical models:

- **Parametric**: the stochastic features of the model are completly specified up to a finite dimensional parameter: $\{P_\theta\}_{\{\theta \in \Theta\}}$ with $\Theta \subseteq \mathbb{R}^k, k < \infty$;
- **Semiparametric**: it is a partially specified model, e.g., $\{P_\theta\}_{\{\theta \in \Theta, \gamma \in \Gamma\}}$ with Θ of finite dimension and Γ of infinite dimension;
- **Non parametric**: there is no finite dimensional component of the model.

In a **linear model data** are given by $D_n = \{(y_i, x_{i1}, \dots, x_{ik})\}_{i=1}^n \in \mathcal{D}$ where:

- D_n are the observed data;
- y_i is the dependent variable;
- $x_{i1}, \dots, x_{ik}$ are the regressors including a constant.

Let $\mathcal{D}$ be the set of possible data realizations. Let $D \in \mathcal{D}$ be your data. Let $\mathcal{F}$ be a statistical model indexed by θ. Let ν be a functional $\mathcal{F} \to \mathbb{R}$. Let $\alpha > 0$ be a small tolerance. An **estimator** is a map

$$\mathcal{D} \to \mathcal{F} \quad , \quad D \mapsto \hat{\theta} \qquad \text{or} \qquad \mathcal{D} \to \mathbb{R} \quad , \quad D \mapsto \hat{\nu}$$

Statistical **inference** is a map into subsets of $\mathcal{F}$ given by

$$\mathcal{D} \to \mathcal{G} \subseteq \mathcal{F} : \min_\theta P_\theta(\mathcal{G} | \theta \in \mathcal{G}) \geq 1-\alpha \qquad \text{or} \qquad \mathcal{D} \to A \subseteq \mathbb{R} : \min_\theta P_\theta(A | \nu(\theta) \in A) \geq 1-\alpha$$

A **data generating process** (DGP) is a single statistical distribution over $\mathcal{D}$.

Suppose you have a statistical model parametrized by θ and an estimator $\hat{\theta}$. The **bias** of $\hat{\theta}$ relative to θ is given by

$$Bias_\theta(\hat{\theta}) = \mathbb{E}_{x|\theta}[\hat{\theta}] - \theta = \mathbb{E}_{x|\theta}[\hat{\theta} - \theta]$$

Let $\hat{\theta}$ be an estimator for θ_0. We say $\hat{\theta}$ is an **unbiased** estimator for θ if $\mathbb{E}[\hat{\theta}] = \theta_0$.

2.7 References

- Kozbur (2019). PhD Econometrics - Lecture Notes.
- Greene (2006). "*Econometric Analysis*". Appendix B: Probability and Distribution Theory.
- Greene (2006). "*Econometric Analysis*". Appendix C: Estimation and Inference.

Chapter 3

Asymptotic Theory

3.1 Convergence

A sequence of nonrandom numbers $\{a_n\}$ **converges** to a (has limit a) if for all $\varepsilon > 0$, there exists n_ε such that if $n > n_\varepsilon$, then $|a_n - a| < \varepsilon$. We write $a_n \to a$ as $n \to \infty$.

A sequence of nonrandom numbers $\{a_n\}$ is **bounded** if and only if there is some $B < \infty$ such that $|a_n| \leq B$ for all $n = 1, 2, ...$ Otherwise, we say that $\{a_n\}$ is unbounded.

A sequence of nonrandom numbers $\{a_n\}$ is $O(N^\delta)$ (at most of order N^δ) if $N^{-\delta}a_n$ is bounded. When $\delta = 0$, a_n is bounded, and we also write $a_n = O(1)$ (big oh one).

A sequenc of nonrandom numberse $\{a_n\}$ is $o(N^\delta)$ if $N^{-\delta}a_n \to 0$. When $\delta = 0$, a_n converges to zero, and we also write $a_n = o(1)$ (little oh one).

> From the definitions, it is clear that if $a_n = o(N^\delta)$, then $a_n = O(N^\delta)$; in particular, if $a_n = o(1)$, then $a_n = O(1)$. If each element of a sequence of vectors or matrices is $O(N^\delta)$, we say the sequence of vectors or matrices is $O(N^\delta)$, and similarly for $o(N^\delta)$.

A sequence of random variables $\{X_n\}$ **converges in probability** to a constant $c \in \mathbb{R}$ if for all $\varepsilon > 0$

$$\Pr\Big(|X_n - c| > \varepsilon\Big) \to 0 \qquad \text{as } n \to \infty$$

We write $X_n \xrightarrow{p} c$ and say that a is the probability limit (*plim*) of X_n: $\text{plim} X_n = c$. In the special case where $c = 0$, we also say that $\{X_n\}$ is $o_p(1)$ (little oh p one). We also write $X_n = o_p(1)$ or $X_n \xrightarrow{p} 0$.

A sequence of random variables $\{X_n\}$ is bounded in probability if for every $\varepsilon > 0$, there exists a $B_\varepsilon < \infty$ and an integer n_ε such that

$$\Pr\Big(|x_n| > B_\varepsilon\Big) < \varepsilon \qquad \text{for all } n > n_\varepsilon$$

We write $X_n = O_p(1)$ ($\{X_n\}$ is big oh p one).

A sequence of random variables $\{X_n\}$ is $o_p(a_n)$ where $\{a_n\}$ is a nonrandom positive sequence, if $X_n/a_n = o_p(1)$. We write $X_n = o_p(a_n)$.

A sequence of random variables $\{X_n\}$ is $O_p(a_n)$ where $\{a_n\}$ is a nonrandom positive sequence, if $X_n/a_n = O_p(1)$. We write $X_n = O_p(a_n)$.

A sequence of random variables $\{X_n\}$ **converges almost surely** to a constant $c \in \mathbb{R}$ if

$$\Pr\left(X_n \overset{p}{\to} c\right) = 1$$

We write $X_n \overset{as}{\to} c$.

A sequence of random variables $\{X_n\}$ **converges in mean square** to a constant $c \in \mathbb{R}$ if

$$\mathbb{E}[(X_n - c)^2] \to 0 \qquad \text{as } n \to \infty$$

We write $X_n \overset{ms}{\to} c$.

Let $\{X_n\}$ be a sequence of random variables and F_n be the cumulative distribution function (cdf) of X_n. We say that X_n **converges in distribution** to a random variable x with cdf F if the cdf F_n of X_n converges to the cdf F of x *at every continuity point* of F. We write $X_n \overset{d}{\to} x$ and we call F the **asymptotic distribution** of X_n.

Lemma: Let $\{X_n\}$ be a sequence of random variables and $c \in \mathbb{R}$

- $X_n \overset{ms}{\to} c \;\Rightarrow\; X_n \overset{p}{\to} c$
- $X_n \overset{as}{\to} c \;\Rightarrow\; X_n \overset{p}{\to} c$
- $X_n \overset{p}{\to} c \;\Rightarrow\; X_n \overset{d}{\to} c$

  Note that all the above definitions naturally extend to a sequence of random vectors by requiring element-by-element convergence. For example, a sequence of $K \times 1$ random vectors $\{X_n\}$ **converges in probability** to a constant $c \in \mathbb{R}^K$ if for all $\varepsilon > 0$

$$\Pr\left(|X_{nk} - c_k| > \varepsilon\right) \to 0 \qquad \text{as } n \to \infty \quad \forall k = 1...K$$

3.2 Theorems

Slutsky Theorem: Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables, x a random variable and c a constant such that $\{X_n\} \overset{d}{\to} X$ and $\{Y_n\} \overset{p}{\to} c$. Then

- $X_n + Y_n \overset{d}{\to} X + c$
- $X_n \cdot Y_n \overset{d}{\to} X \cdot c$

Continuous Mapping Theorem: Let $\{X_n\}$ be sequence of $K \times 1$ random vectors and $g : \mathbb{R}^K \to \mathbb{R}^J$ a continuous function that does not depend on n.Then

- $x_n \overset{as}{\to} x \;\Rightarrow\; g(X_n) \overset{as}{\to} g(x)$
- $x_n \overset{p}{\to} x \;\Rightarrow\; g(X_n) \overset{p}{\to} g(x)$

- $x_n \xrightarrow{d} x \;\Rightarrow\; g(X_n) \xrightarrow{d} g(x)$

Weak Law of Large Numbers: Let $\{x_i\}_{i=1}^n$ be a sequence of independent, identically distributed random variables such that $\mathbb{E}[|x_i|] < \infty$. Then the sequence satisfies the **weak law of large numbers (WLLN)**:

$$\mathbb{E}_n[x_i] = \frac{1}{n}\sum_{i=1}^n x_i \xrightarrow{p} \mu \qquad \text{where } \mu \equiv \mathbb{E}[x_i]$$

Proof: The independence of the random variables implies no correlation between them, and we have that

$$Var\left(\mathbb{E}_n[x_i]\right) = Var\left(\frac{1}{n}\sum_{i=1}^n x_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^n x_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Using Chebyshev's inequality on $\mathbb{E}_n[x_i]$ results in

$$\Pr\Big(|\mathbb{E}_n[x_i] - \mu| > \varepsilon\Big) \le \frac{\sigma^2}{n\varepsilon^2}$$

As n approaches infinity, the right hand side approaches 0. And by definition of convergence in probability, we have obtained $\mathbb{E}_n[x_i] \xrightarrow{p} \mu$ as $n \to \infty$.

■

Intuitions for the law of large numbers:

- Cancellation with high probability.
- Re-visiting regions of the sample space over and over again.

Lindberg-Levy Central Limit Theorem: Let $\{x_i\}_{i=1}^n$ be a sequence of independent, identically distributed random variables such that $\mathbb{E}[x_i^2] < \infty$, and $\mathbb{E}[x_i] = \mu$. Then $\{x_i\}$ satisfies the **central limit theorem (CLT)**; that is,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (x_i - \mu) \xrightarrow{d} N(0, \sigma^2)$$

where $\sigma^2 = Var(x_i) = \mathbb{E}[x_i x_i']$ is necessarily positive semidefinite.

Proof: Suppose $\{x_i\}$ are independent and identically distributed random variables, each with mean μ and finite variance σ^2. The sum $x_1 + ... + X_n$ has mean $n\mu$ and variance $n\sigma^2$. Consider the random variable

$$Z_n = \frac{x_1 + ... + X_n - n\mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^n \frac{x_i - \mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^n \frac{1}{\sqrt{n}}\tilde{x}_i$$

where in the last step we defined the new random variables $\tilde{x}_i = \frac{x_i - \mu}{\sigma}$ each with zero mean and unit variance. The characteristic function of Z_n is given by

$$\varphi_{Z_n}(t) = \varphi_{\sum_{i=1}^n \frac{1}{\sqrt{n}}\tilde{x}_i}(t) = \varphi_{\tilde{x}_1}\left(\frac{t}{\sqrt{n}}\right) \times ... \times \varphi_{Y_n}\left(\frac{t}{\sqrt{n}}\right) = \left[\varphi_{\tilde{x}_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

where in the last step we used the fact that all of the $\tilde{x}_i$ are identically distributed. The characteristic function of $\tilde{x}_1$ is, by Taylor's theorem,

$$\varphi_{\tilde{x}_1}\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) \qquad \text{for } n \to \infty$$

where $o(t^2)$ is "little o notation" for some function of t that goes to zero more rapidly than t^2. By the limit of the exponential function, the characteristic function of Z_n equals

$$\varphi_{Z_n}(t) = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \to e^{-\frac{1}{2}t^2} \qquad \text{for } n \to \infty$$

Note that all of the higher order terms vanish in the limit $n \to \infty$. The right hand side equals the characteristic function of a standard normal distribution $N(0,1)$, which implies through Lévy's continuity theorem that the distribution of Z_n will approach $N(0,1)$ as $n \to \infty$. Therefore, the sum $x_1 + ... + x_n$ will approach that of the normal distribution $N(n_\mu, n\sigma^2)$, and the sample average

$$\mathbb{E}_n[x_i] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

converges to the normal distribution $N(\mu, \sigma^2)$, from which the central limit theorem follows.

■

Delta Method: Let $\{X_n\}$ be a sequence of independent, identically distributed $K \times 1$ random vectors such that $\sqrt{n}(X_n - c) \xrightarrow{d} Z$ for some fixed $c \in \mathbb{R}^K$ and Σ a $K \times K$ positive definite matrix. Suppose $g : \mathbb{R}^K \to \mathbb{R}^J$ with $J \leq K$ is continuously differentiable and full rank at c, then

$$\sqrt{n}\Big[g(X_n) - g(c)\Big] \xrightarrow{d} GZ$$

where $G = \frac{\partial g(c)}{\partial x}$ is the $J \times K$ matrix of partial derivatives evaluated at c.

> Note that the most common utilization is with the random variable $\mathbb{E}_n[x_i]$. In fact, under the assumptions of the CLT, we have that
>
> $$\sqrt{n}\Big[g\big(\mathbb{E}_n[x_i]\big) - g(\mu)\Big] \xrightarrow{d} N(0, G\Sigma G')$$

3.3 Ergodic Theory

Let $(\Omega, \mathcal{B}, P)$ be a probability space and $T : \Omega \to \Omega$ a measurable map. T is a **probability preserving transformation** if the probability of the pre-image of every set is the same as the probability of the set itself, i.e. $\forall G, \Pr(T^{-1}(G)) = \Pr(G)$.

Let $(\Omega, \mathcal{B}, P)$ be a probability space and $T : \Omega \to \Omega$ a PPT. A set $G \in \mathcal{B}$ is **invariant** if $T^{-1}(G) = G$.

> Note that it does not have to work the other way around: $G \neq T(G)$.

Let $(\Omega, \mathcal{B}, P)$ be a probability space and $T : \Omega \to \Omega$ a PPT. T is **ergodic** if every invariant set $G \in \mathcal{B}$ has probability zero or one, i.e. $\Pr(G) = 0 \vee \Pr(G) = 1$.

Poincarè Recurrence Theorem: Let $(\Omega, \mathcal{B}, P)$ be a probability space and $T : \Omega \to \Omega$ a PPT. Suppose $A \in \mathcal{B}$ is measurable. Then, for almost every $\omega \in A$, $T^n(\omega) \in A$ for infinitely many n.

Proof: We follow 5 steps:

1. Let $G = \{\omega \in A : T^K(\omega) \notin A \quad \forall k > 0\}$: the set of all points of A that never "return" in A.
2. Note that $\forall j \geq 1$, $T^{-j}(G) \cap G = \emptyset$. In fact, suppose $\omega \in T^{-j}(G)$. Then $\omega \notin G$ since otherwise we would have $\omega \in G \subseteq A$ and $\omega \in T^J(G) \subseteq A$ which contradicts the definition of G.
3. It follows that $\forall l, n \geq 1$, $T^{-l}(G) \cap T^{-n}(G) = \emptyset$
4. Since T is a PPT, $\Pr(T^{-j}(G)) = \Pr(G)\ \forall j$
5. Then

$$\Pr(T^{-1}(G) \cup T^{-2}(G) \cup ... \cup T^{-l}(G)) = l \cdot \Pr(G) \leq 1 \Rightarrow \Pr(G) \leq \frac{1}{l} \quad \Rightarrow \quad \lim_{l \to \infty} \Pr(G) = 0$$

■

$$\frac{1}{n} \sum_{i=1}^{n} f(T^i x) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ergodic Theorem: Let T be an ergodic PPT on Ω. Let x be a random variable on Ω with $\mathbb{E}[x] < \infty$. Let $x_i = x \circ T^i$. Then,

$$\frac{1}{n} \sum_{i=1}^{n} x_i \overset{as}{\to} \mathbb{E}[x]$$

> To figure out whether a PPT is ergodic, it's useful to draw a graph with $T^{-1}(G)$ on the y-axis and G on the x-axis.

From the ergodic theorem, we have that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f(T^i x) g(x) = f^*(x) g(x) \quad \Rightarrow \quad \lim_{n \to \infty} \Pr(T^{-n} G \cap H) = \Pr(G) \Pr(H)$$

where $f^*(x) = \int f(x) dx = \mathbb{E}[f]$.

Halmos: *We have seen that if a transformation* T *is ergodic, then* $\Pr(T^{-n} G \cap H)$ *converges in the sense of Cesaro to* $\Pr(G) \Pr(H)$. *The validity of this condition for all* G *and* H *is, in fact, equivalent to ergodicity. To prove this, suppose that* A *is a measurable invariant set, and take both* G *and* H *equal to* A. *It follows that* $\Pr(A) = (\Pr(A))^2$, *and hence that* $\Pr(A)$ *is either 0 or 1.*

The Cesaro convergence condition has a natural intuitive interpretation. We may visualize the transformation T *as a particular way of stirring the contents of a vessel (of total volume*

1) full of an incompressible fluid, which may be thought of as 90 per cent gin (G) and 10 per cent vermouth (H). If H is the region originally occupied by the vermouth, then, for any part G of the vessel, the relative amount of vermouth in G, after n repetitions of the act of stirring, is given by $\Pr(T^{-n}G \cap H)/\Pr(H)$. *The ergodicity of T implies therefore that on the average this relative amount is exactly equal to 10 per cent. In general, in physical situations like this one, one expects to be justified in making a much stronger statement, namely that, after the liquid has been stirred sufficiently often* ($n \to \infty$), *every part G of the container will contain approximately 10 per cent vermouth. In mathematical language this expectation amounts to replacing Cesaro convergence by ordinary convergence, i.e., to the condition* $\lim_{n\to\infty} \Pr(T^{-n}G \cap H) = \Pr(G)\Pr(H)$. *If a transformation T satisfies this condition for every pair G and H of measurable sets, it is called mixing, or, in distinction from a related but slightly weaker concept, strongly mixing."*

Let $\{\Omega, \mathcal{B}, P\}$ be a probability space. Let T be a probability preserving transform. Then T is **strongly mixing** if for every invariant sets G,$H \in \mathcal{B}$

$$P(G \cap T^{-k}H) \to P(G)P(H) \quad \text{as } k \to \infty$$

where $T^{-k}H$ is defined as $T^{-k}H = T^{-1}(...T^{-1}(T^{-1}H)...)$ repeated k times.

Let $\{X_i\}_{i=-\infty}^{\infty}$ be a two sided sequence of random variables. Let $\mathcal{B}_{-\infty}^{n}$ be the sigma algebra generated by $\{X_i\}_{i=-\infty}^{n}$ and $\mathcal{B}_{n+k}^{\infty}$ the sigma algebra generated by $\{X_i\}_{i=n+k}^{\infty}$. Define the mixing coefficient

$$\alpha(k) = \sup_{n\in\mathbb{Z}} \sup_{G\in\mathcal{B}_{-\infty}^{n}} \sup_{H\in\mathcal{B}_{n+k}^{\infty}} |\Pr(G \cap H) - \Pr(G)\Pr(H)|$$

$\{X_i\}$ is **α-mixing** if $\alpha(k) \to 0$ if $k \to \infty$.

Note that mixing implies ergodicity.

Let $X_i : \Omega \to \mathbb{R}$ be a (two sided) sequence of random variables with $i \in \mathbb{Z}$. X_i is **strongly stationary** or simply stationary if

$$\Pr(X_{i_1} \le a_1, ..., X_{i_k} \le a_k) = \Pr(X_{i_{1-s}} \le a_1, ..., X_{i_{k-s}} \le a_k) \quad \text{for every } i_1, ..., i_k, a_1, ..., a_k, s \in \mathbb{R}.$$

Let $X_i : \Omega \to \mathbb{R}$ be a (two sided) sequence of random variables with $i \in \mathbb{Z}$. X_i is **covariance stationary** if $\mathbb{E}[X_i] = \mathbb{E}[X_j]$ for every i, j and $\mathbb{E}[X_iX_j] = \mathbb{E}[X_{i+k}X_{j+k}]$ for all i, j, k. All of the second moments above are assumed to exist.

Let $X_t : \Omega \to \mathbb{R}$ be a sequence of random variables indexed by $t \in \mathbb{Z}$ such that $\mathbb{E}[|X_t|] < 1$ for each t. X_t is a **martingale** if $\mathbb{E}[X_t|X_{t-1}, X_{t-2}, ...] = X_t$. X_t is a **martingale difference** if $\mathbb{E}[X_t|X_{t-1}, X_{t-2}, ...] = 0$.

Gordin's Central Limit Theorem: Let $\{z_i\}$ be a stationary, α-mixing sequence of random variables. If moreover

- $\sum_{m=1}^{\infty} \alpha(m)^{\frac{\delta}{2+\delta}} < \infty$
- $\mathbb{E}[z_i] = 0$

- $\mathbb{E}\Big[||z_i||^{2+\delta}\Big] < \infty$

Then

$$\sqrt{n}\mathbb{E}_n[z_i] \xrightarrow{d} N(0,\Omega) \quad \text{where} \quad \Omega = \lim_{n\to\infty} Var(\sqrt{n}\mathbb{E}_n[z_i])$$

Let $\Omega_k = \mathbb{E}[z_i z_{i+k}']$. Then a necessary condition for Gordin's CLT is covariance summability: $\sum_{k=1}^{\infty} \Omega_k < \infty$.

Ergodic Central Limit Theorem: Let $\{z_i\}$ be a stationary, ergodic, martingale difference sequence. Then

$$\sqrt{n}\mathbb{E}_n[z_i] \xrightarrow{d} N(0,\Omega) \quad \text{where} \quad \Omega = \lim_{n\to\infty} Var(\sqrt{n}\mathbb{E}_n[z_i])$$

3.4 Asymptotic Properties of Estimators

Let $\{\theta_n\}$ be a sequence of estimators, if

$$\hat{\theta} \xrightarrow{p} \theta_0$$

then we say $\hat{\theta}$ is a **consistent** estimator of θ_0.

Let $\{\theta_n\}$ be a sequence of estimators, if

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$$

then we say that $\hat{\theta}$ is $\sqrt{\ltimes}$**-asymptotically distributed** and V is the **asymptotic variance** of $\sqrt{n}(\hat{\theta} - \theta_0)$, denoted as $AVar\Big(\sqrt{n}(\hat{\theta} - \theta_0)\Big)$.

Let $\hat{\theta}$ and $\tilde{\theta}$ be estimators of θ_0 each satisfying asymptotic normality, with asymptotic variances $V = AVar\mathcal{B}ig(\sqrt{n}(\hat{\theta} - \theta_0)\mathcal{B}ig)$ and $D = AVar\mathcal{B}ig(\sqrt{n}(\tilde{\theta} - \theta_0)\mathcal{B}ig)$ (these generally depend on the value of θ_0, but we suppress that consideration here). Then

1. $\hat{\theta}$ is **asymptotically efficient** relative to $\tilde{\theta}$ if $D - V$ is positive semidefinite for all θ_0,
2. $\hat{\theta}$ and $\tilde{\theta}$ are $\sqrt{\ltimes}$**-equivalent** if $\sqrt{n}(\hat{\theta} - \tilde{\theta}) = o_p(1)$.

3.5 Asymptotic Properties of Test Statistics

The **asymptotic size** of a testing procedure is defined as the limiting probability of rejecting H_0 when H_0 is true. Mathematically, we can write this as $\lim_{n\to\infty} \Pr_n(\text{reject } H_0|H_0)$, where the n subscript indexes the sample size.

A test is said to be **consistent** against the alternative H_1 if the null hypothesis is rejected with probability approaching 1 when H_1 is true: $\lim_{N\to\infty} \Pr_N(\text{reject } H_0|H_1) \xrightarrow{p} 1$.

Theorem: Suppose that $\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{d}N(0,V)$, where V is positive definite. Then for any non-stochastic $Q\times P$ matrix R, $Q\leq P$, with $\text{rank}(R)=Q$

$$\sqrt{n}R(\hat{\theta}-\theta_0)\sim N(0,RVR')$$

and

$$[\sqrt{n}R(\hat{\theta}-\theta_0)]'[RVR']^{-1}[\sqrt{n}R(\hat{\theta}-\theta_0)]\xrightarrow{d}\chi^2_Q$$

In addition, if $\text{plim}\hat{V}_n=V$, then

$$[\sqrt{n}R(\hat{\theta}-\theta_0)]'[RVR']^{-1}[\sqrt{n}R(\hat{\theta}-\theta_0)]=(\hat{\theta}-\theta_0)'R'[R(\hat{V}_n/n)R']^{-1}R(\hat{\theta}-\theta_0)\xrightarrow{d}\chi^2_Q$$

For testing the null hypothesis $H_0: R\theta_0=r$, where r is a $Q\times 1$ random vector, define the **Wald statistic** for testing H_0 against $H_1: R\theta_0\neq r$ as

$$W_n=(R\hat{\theta}-r)'[R(\hat{V}_n/n)R']^{-1}(R\hat{\theta}-r)$$

Under H_0, $W_n\xrightarrow{d}\chi^2_Q$. If we abuse the asymptotics and we treat $\hat{\theta}$ as being distributed as Normal we get the equation exactly.

3.6 References

- Kozbur (2019). PhD Econometrics - Lecture Notes.
- Wooldridge (2010). "*Econometric Analysis of Cross Section and Panel Data*". Chapter 3: Basic Asymptotic Theory.
- Halmos (2006). "*Lectures on Ergodic Theory*".
- Greene (2006). "*Econometric Analysis*". Appendix D: Large Sample Distribution Theory.
- Hayashi (2000). "*Econometrics*". Chapter 2: Large-Sample Theory.

Chapter 4

OLS Algebra

4.1 The Gauss Markov Model

A statistical model for regression data is the **Gauss Markov Model** if each of its distributions satisfies the conditions (1)-(4): linearity, strict exogeneity, no multicollinearity, and spherical error variance. The **Extended Gauss Markov Model** also satisfies assumption (5).

1. **Linearity**: a statistical model $\mathcal{F}$ over data $\mathcal{D}$ satisfies linearity if for each element of $\mathcal{F}$, the data can be decomposed in

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i = x_i'\beta + \varepsilon_i \\ \underset{n\times 1}{y} &= \underset{n\times k}{X} \cdot \underset{k\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \end{aligned}$$

2. **Strict Exogeneity**: $\mathbb{E}[\varepsilon_i | x_1, \dots, x_n] = 0, \forall i$.
3. **No Multicollinerity**: $\mathbb{E}_n[x_i x_i']$ is strictly positive definite almost surely. Equivalent to require $rank(X) = k$ with probability $p \to 1$. Intuition: no regressor is a linear combination of other regressors.
4. **Spherical Error Variance**: -$\mathbb{E}[\varepsilon_i^2 | x] = \sigma^2 > 0, \ \forall i$ -$\mathbb{E}[\varepsilon_i \varepsilon_j | x] = 0, \ \forall \ 1 \leq i < j \leq n$
5. (Extended GM model) **Normal error term**: $\varepsilon | X \sim N(0, \sigma^2 I_n)$ and $\varepsilon \perp X$.

Note that by (2) and (4) you get **homoskedasticity**:

$$Var(\varepsilon_i | x) = \mathbb{E}[\varepsilon_i^2 | x] - \mathbb{E}[\varepsilon_i | x]^2 = \sigma^2 I \qquad \forall i$$

Implications:

- Strict exogeneity is not restrictive since it is sufficient to include a constant in the regression to enforce it

$$y_i = \alpha + x_i'\beta + (\varepsilon_i - \alpha) \quad \Rightarrow \quad \mathbb{E}[\varepsilon_i] = \mathbb{E}_x[\mathbb{E}[\varepsilon_i | x]] = 0$$

- This implies $\mathbb{E}[x_{jk}\varepsilon_i] = 0$ by the LIE.
- These two conditions together imply $Cov(x_{jk}\varepsilon_i) = 0$.

A map $\Pi : V \to V$ is a **projection** if $\Pi \circ \Pi = \Pi$.

The Gauss Markov Model assumes that the **conditional expectation function (CEF)** $f(X) = \mathbb{E}[Y|X]$ and the **linear projection** $g(X) = X\beta$ coincide.

4.1.1 Matlab

This code draws 100 observations from the model $y = 2x_1 - x_2 + \varepsilon$ where $x_1, x_2 \sim U[0,1]$ and $\varepsilon \sim N(0,1)$.

```
% Set seed
rng(123)

% Set the number of observations
n = 100;

% Set the dimension of X
k = 2;

% Draw a sample of explanatory variables
X = rand(n, k);

% Draw the error term
sigma = 1;
e = randn(n,1)*sqrt(sigma);

% Set the parameters
b = [2; -1];

% Calculate the dependent variable
y = X*b + e;
```

4.2 The OLS estimator

The **sum of squared residuals (SSR)** is given by

$$Q_n(\beta) \equiv \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \frac{1}{n}(y - X\beta)'(y - X\beta)$$

Consider a dataset $\mathcal{D}$ and define $Q_n(\beta) = \mathbb{E}_n[(y_i - x_i'\beta)^2]$. Then the **ordinary least squares (OLS)** estimator $\hat{\beta}_{OLS}$ is the value of β that minimizes $Q_n(\beta)$.

When we can write $D = (y, X)$ in matrix form, then

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \frac{1}{n}(y - X\beta)'(y - X\beta)$$

Theorem: Under the assumption that X has full rank, the OLS estimator is unique and it is determined by the normal equations. More explicitly, $\hat{\beta}$ is the OLS estimate precisely when $X'X\hat{\beta} = X'y$.

Proof: Taking the FOC:

$$\frac{\partial Q_n(\beta)}{\partial \beta} = -\frac{2}{n}X'y + \frac{2}{n}X'X\beta = 0 \quad \Leftrightarrow \quad X'X\beta = X'y$$

Since $(X'X)^{-1}$ exists by assumption,

Finally, $\frac{\partial^2 Q_n(\beta)}{\partial\beta\partial\beta'} = X'X/n$ is positive definite since $X'X$ is positive semi-definite and $(X'X)^{-1}$ exists because X is full rank. Therefore, $Q_n(\beta)$ minimized at $\hat{\beta}_n$.

■

The k equations $X'X\hat{\beta} = X'y$ are called **normal equations**.

We can now define the following objects:

- Fitted coefficient: $\hat{\beta}_{OLS} = (X'X)^{-1}X'y = \mathbb{E}_n[x_i x_i']\mathbb{E}_n[x_i y_i]$
- Fitted residual: $\hat{\varepsilon}_i = y_i - x_i'\hat{\beta}$
- Fitted value: $\hat{y}_i = x_i'\hat{\beta}$
- Predicted coefficient: $\hat{\beta}_{-i} = \mathbb{E}_n[x_{-i}x_{-i}']\mathbb{E}_n[x_{-i}y_{-i}]$
- Prediction error: $\hat{\varepsilon}_{-i} = y_i - x_i'\hat{\beta}_{-i}$
- Predicted value: $\hat{y}_i = x_i'\hat{\beta}_{-i}$

  Notes on the orthogonality conditions:

  - The normal equations are equivalent to the moment condition $\mathbb{E}_n[x_i\varepsilon_i] = 0$.
  - The algebraic result $\mathbb{E}_n[x_i\hat{\varepsilon}_i] = 0$ is called **ortogonality property** of the OLS residual $\hat{\varepsilon}_i$.
  - If we have included a constant in the regression, $\mathbb{E}_n[\hat{\varepsilon}_i] = 0$.
  - $\mathbb{E}\Big[\mathbb{E}_n[x_i\varepsilon_i]\Big] = 0$ by strict exogeneity (assumed in GM), but $\mathbb{E}_n[x_i\varepsilon_i] \neq \mathbb{E}[x_i\varepsilon_i] = 0$. This is why $\hat{\beta}_{OLS}$ is just an estimate of β_0.
  - Calculating OLS is like replacing the j equations $\mathbb{E}[x_{ij}\varepsilon_i] = 0 \ \forall j$ with $\mathbb{E}_n[x_{ij}\varepsilon_i] = 0 \ \forall j$ and forcing them to hold (remindful of GMM).

The **projection matrix** is given by $P = X(X'X)^{-1}X'$. It has the following properties:

- $PX = X$
- $P\hat{\varepsilon} = 0$ $\quad$ (P, ε orthogonal)
- $Py = X(X'X)^{-1}X'y = X\hat{\beta} = \hat{y}$
- Symmetric: $P = P'$, Idempotent: $PP = P$
- $tr(P) = tr(X(X'X)^{-1}X') = tr(X'X(X'X)^{-1}) = tr(I_k) = k$
- Its diagonal elements are $h_{ii} = x_i(X'X)^{-1}x_i'$ and are called **leverage**.

$h_{ii} \in [0,1]$ is a normalized length of the observed regressor vector x_i. In the OLS regression framework it captures the relative influence of observation i on the estimated coefficient. Note that $\sum_n h_{ii} = k$.

The **annihilator matrix** is given by $M = I_n - P$. It has the following properties:

- $MX = 0 \quad (M,\, X \text{ orthogonal})$
- $M\hat{\varepsilon} = \hat{\varepsilon}$
- $My = \hat{\varepsilon}$
- Symmetric: $M = M'$, idempotent: $MM = M$
- $tr(M) = n - k$
- Its diagonal elements are $1 - h_{ii} \in [0,1]$

Then we can equivalently write $\hat{y}$ (defined by stacking $\hat{y}_i$ into a vector) as $\hat{y} = Py$.

4.2.1 Matlab

```
% Estimate beta
b_hat = inv(X'*X)*(X'*y) % = 1.9020, -0.9305

% Equivalent but faster formulation
b_hat = (X'*X)\(X'*y);

% Even faster (but less intuitive) formulation
b_hat = X\y;

% Note that is generally not equivalent to Var(X)^-1 * Cov(X,y)...
Var_X = cov(X);
Cov_Xy = n/(n-1) * (mean(X .* y) - mean(X).*mean(y));
b_alternative = inv(Var_X) * Cov_Xy' % = 2.1525, -0.7384

% ...unless you include a constant
a = 3;
y = a + X*b + e;
b_hat_1 = [ones(n,1), X]\y % = 2.1525, -0.7384
Var_X = cov(X);
Cov_Xy = n/(n-1) * (mean(X .* y) - mean(X).*mean(y));
b_alternative = inv(Var_X) * Cov_Xy' % = 2.1525, -0.7384

% Predicted y
y_hat = X*b_hat;

% Residuals
e_hat = y - X*b_hat;

% Projection matrix
```

```
P = X*inv(X'*X)*X';

% Annihilator matrix
M = eye(n) - P;

% Leverage
h = diag(P);
```

4.3 OLS residual properties

The error is **homoskedastic** if $\mathbb{E}[\varepsilon^2|x] = \sigma^2$ does not depend on x.

$$Var(\varepsilon) = I\sigma^2 = \begin{bmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix}$$

The error is **heteroskedastic** if $\mathbb{E}[\varepsilon^2|x] = \sigma^2(x)$ does depend on x.

$$Var(\varepsilon) = I\sigma_i^2 = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n^2 \end{bmatrix}$$

The OLS **residual variance** can be an object of interest even in a heteroskedastic regression. Its method of moments estimator is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Note that $\hat{\sigma}^2$ can be rewritten as

$$\hat{\sigma}^2 = \frac{1}{n}\varepsilon' M' M \varepsilon = \frac{1}{n} tr(\varepsilon' M \varepsilon) = \frac{1}{n} tr(M \varepsilon' \varepsilon)$$

However, the method of moments estimator is a biesed estimator. In fact

$$\mathbb{E}[\hat{\sigma}^2|X] = \frac{1}{n}\mathbb{E}[tr(M\varepsilon'\varepsilon)|X] = \frac{1}{n} tr(M\mathbb{E}[\varepsilon'\varepsilon|X]) = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})\sigma_i^2$$

Under conditional homoskedasticity, the above expression simplifies to

$$\mathbb{E}[\hat{\sigma}^2|X] = \frac{1}{n} tr(M)\sigma^2 = \frac{n-k}{n}\sigma^2$$

The OLS **residual sample variance** is denoted by s^2 and is given by

$$s^2 = \frac{SSR}{n-k} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Furthermore, the square root of s^2, denoted s, is called the standard error of the regression (SER) or the standard error of the equation (SEE). Not to be confused with other notions of standard error to be defined later in the course.

The sum of squared residuals can be rewritten as: $SSR = \hat{\varepsilon}'\hat{\varepsilon} = \varepsilon' M \varepsilon$.

The OLS residual sample variance is an unbiased estimator of the error variance σ^2.

Another unbiased estimator of σ^2 is given by

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-1}\hat{\varepsilon}_i^2$$

One measure of the variability of the dependent variable y_i is the sum of squares $\sum_{i=1}^{n} y_i^2 = y'y$. There is a decomposition:

$$\begin{aligned} y'y &= (\hat{y}+e)'(\hat{y}+\hat{\varepsilon}) \\ &= \hat{y}'\hat{y} + 2\hat{y}'\hat{\varepsilon} + \hat{\varepsilon}'\hat{\varepsilon}e \\ &= \hat{y}'\hat{y} + 2b'X'\hat{\varepsilon} + \hat{\varepsilon}'\hat{\varepsilon} \quad (\text{since } \hat{y} = Xb) \\ &= \hat{y}'\hat{y} + \hat{\varepsilon}'\hat{\varepsilon} \quad (\text{since } X'\hat{\varepsilon} = 0) \end{aligned}$$

The **uncentered $\mathbf{R^2}$** is defined as:

$$R^2_{uc} \equiv 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y} = 1 - \frac{\mathbb{E}_n[\hat{\varepsilon}_i^2]}{\mathbb{E}_n[y_i^2]} = \frac{\mathbb{E}[\hat{y}_i^2]}{\mathbb{E}[y_i^2]}$$

A more natural measure of variability is the sum of centered squares $\sum_{i=1}^{n}(y_i - \bar{y})^2$, where $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$. If the regressors include a constant, it can be decomposed as

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

The **coefficient of determination**, $\mathbf{R^2}$, is defined as

$$R^2 \equiv 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\mathbb{E}_n[(\hat{y}_i-\bar{y})^2]}{\mathbb{E}_n[(y_i-\bar{y})^2]}$$

Always use the centered R^2 unless you really know what you are doing.

4.3.1 `Matlab`

```
% Biased variance estimator
sigma_hat = e_hat'*e_hat / n;

% Unbiased estimator 1
sigma_hat_2 = e_hat'*e_hat / (n-k);

% Unbiased estimator 2
sigma_hat_3 = mean( e_hat.^2 ./ (1-h) );

% R squared - uncentered
R2_uc = (y_hat'*y_hat)/ (y'*y);

% R squared
y_bar = mean(y);
R2 = ((y_hat-y_bar)'*(y_hat-y_bar))/ ((y-y_bar)'*(y-y_bar));
```

4.4 Finite Sample Properties of the OLS estimator

Theorem: Under the GM assumptions (1)-(3), the OLS estimator is **conditionally unbiased**, i.e. the distribution of $\hat{\beta}_{OLS}$ is centered at β_0: $\mathbb{E}[\hat{\beta}|X] = \beta_0$.

Proof:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|X] &= \mathbb{E}[(X'X)^{-1}X'y|X] = \\
&= (X'X)^{-1}X'\mathbb{E}[y|X] = \\
&= (X'X)^{-1}X'\mathbb{E}[X\beta + \varepsilon|X] = \\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\mathbb{E}[\varepsilon|X] = \\
&= \beta
\end{aligned}$$

■

Theorem: Under the GM assumptions (1)-(3), $Var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$.

Proof:

$$\begin{aligned}
Var(\hat{\beta}|X) &= Var((X'X)^{-1}X'y|X) = \\
&= ((X'X)^{-1}X')Var(y|X)((X'X)^{-1}X')' = \\
&= ((X'X)^{-1}X')Var(X\beta + \varepsilon|X)((X'X)^{-1}X')' = \\
&= ((X'X)^{-1}X')Var(\varepsilon|X)((X'X)^{-1}X')' = \\
&= ((X'X)^{-1}X')\sigma^2 I((X'X)^{-1}X')' = \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

■

Higher correlation of the X implies higher variance of the OLS estimator.

Intuition: individual observations carry less information. You are exploring a smaller region of the X space.

Theorem: Under the GM assumptions (1)-(3), $Cov(\hat{\beta}, \hat{\varepsilon}) = 0$.

Theorem: Under the GM assumptions (1)-(3), $\hat{\beta}_{OLS}$ is the best (most efficient) linear, unbiased estimator (**BLUE**), i.e., for any unbiased linear estimator b: $Var(b|X) \geq Var(\hat{\beta}|X)$.

Proof:
Consider four steps:

1. Define three objects: (i) $b = Cy$, (ii) $A = (X'X)^{-1}X'$ such that $\hat{\beta} = Ay$, and (iii) $D = C - A$.
2. Decompose b as
$$\begin{aligned} b &= (D + A)y = \\ &= Dy + Ay = \\ &= D(X\beta + \varepsilon) + \hat{\beta} = \\ &= DX\beta + D\varepsilon + \hat{\beta} \end{aligned}$$
3. By assumption, b must be unbiased:
$$\begin{aligned} \mathbb{E}[b|X] &= \mathbb{E}[D(X\beta + \varepsilon) + Ay|X] = \\ &= \mathbb{E}[DX\beta|X] + \mathbb{E}[D\varepsilon|X] + \mathbb{E}[\hat{\beta}|X] = \\ &= DX\beta + D\mathbb{E}[\varepsilon|X] + \beta \\ &= DX\beta + \beta \end{aligned}$$
Hence, it must be that $DX = 0$
4. We know by (2)-(3) that $b = D\varepsilon + \hat{\beta}$. We can now calculate its variance.
$$\begin{aligned} Var(b|X) &= Var(\hat{\beta} + D\varepsilon|X) = \\ &= Var(Ay + D\varepsilon|X) = \\ &= Var(AX\beta + (D + A)\varepsilon|X) = \\ &= Var((D + A)\varepsilon|X) = \\ &= (D + A)\sigma^2 I(D + A)' = \\ &= \sigma^2 I(DD' + AA' + DA' + AD') = \\ &= \sigma^2 I(DD' + AA') \geq \\ &\geq \sigma^2 AA' = \\ &= \sigma^2 (X'X)^{-1} = \\ &= Var(\hat{\beta}|X) \end{aligned}$$
since $DA' = AD' = 0$, $DX = 0$ and $AA' = (X'X)^{-1}$.

■

$Var(b|X) \geq Var(\hat{\beta}|X)$ is meant in a positive definite sense.

4.4.1 Matlab

```
% Ideal variance of the OLS estimator
var_b = sigma*inv(X'*X);

% Standard errors
std_b = sqrt(diag(var_b));
```

4.5 References

- Kozbur (2019). PhD Econometrics - Lecture Notes.
- Hansen (2019). “*Econometrics*”.
- Wooldridge (2010). “*Econometric Analysis of Cross Section and Panel Data*”.
- Greene (2006). “*Econometric Analysis*”.
- Hayiashi (2000). “*Econometrics*”.

Chapter 5

OLS Inference

5.1 Asymptotic Theory of the OLS Estimator

Theorem: Assume that $(x_i, y_i)_{i=1}^n$ i.i.d. , $\mathbb{E}[x_i x_i'] = Q$ positive definite, $\mathbb{E}[x_i x_i'] < \infty$ and $\mathbb{E}[y_i^2] < \infty$, then $\hat{\beta}_{OLS}$ is a **consistent** estimator of β_0, i.e. $\hat{\beta} = \mathbb{E}_n[x_i x_i']\mathbb{E}_n[x_i y_i] \xrightarrow{p} \beta_0$.

Proof:
We consider 4 steps:

1. $\mathbb{E}_n[x_i x_i'] \xrightarrow{p} \mathbb{E}[x_i x_i']$ by WLLN since $x_i x_i'$ iid and $\mathbb{E}[x_i x_i'] < \infty$.
2. $\mathbb{E}_n[x_i y_i] \xrightarrow{p} \mathbb{E}[x_i y_i]$ by WLLN, due to $x_i y_i$ iid, Cauchy-Schwarz and finite second moments of x_i and y_i
$$\mathbb{E}\left[x_i y_i\right] \leq \sqrt{\mathbb{E}[x_i^2]\mathbb{E}[y_i^2]} < \infty$$
3. $\mathbb{E}_n[x_i x_i']^{-1} \xrightarrow{p} \mathbb{E}[x_i x_i']^{-1}$ by CMT.
4. $\mathbb{E}_n[x_i x_i']^{-1}\mathbb{E}_n[x_i y_i] \xrightarrow{p} \mathbb{E}[x_i x_i']^{-1}\mathbb{E}[x_i y_i] = \beta$ by CMT.

■

Now we are going to investigate the variance of $\hat{\beta}_{OLS}$ progressively relaxing the underlying assumptions.

- Gaussian error term.
- Homoskedastic error term.
- Heteroskedastic error term.
- Heteroskedastic and autocorrelated error term.

5.1.1 Gaussian Error Term

Theorem: Under the GM assumption (1)-(5), $\hat{\beta} - \beta | X \sim N(0, \sigma^2 (X'X)^{-1})$

Proof:
We follow 2 steps:

1. We can rewrite $\hat{\beta}$ as
$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta+\varepsilon)\\ &= \beta + (X'X)^{-1}X'\varepsilon = \\ &= \beta + \mathbb{E}_n[x_i x_i']^{-1}\mathbb{E}_n[x_i\varepsilon_i]\end{aligned}$$
2. Therefore: $\hat{\beta}-\beta = \mathbb{E}_n[x_i x_i']^{-1}\mathbb{E}_n[x_i\varepsilon_i]$.
$$\begin{aligned}\hat{\beta}-\beta|X &\sim (X'X)^{-1}X'N(0,\sigma^2 I_n) = \\ &= N(0,\sigma^2(X'X)^{-1}X'X(X'X)^{-1}) = \\ &= N(0,\sigma^2(X'X)^{-1})\end{aligned}$$

■

Does it make sense to assume that ε is gaussian? Not much. But does it make sense to assume that $\hat{\beta}$ is gaussian? Yes, because it's an average.

5.1.2 Homoskedastic Error Term

Theorem: Under the assumptions of the previous theorem, plus $\mathbb{E}[x^4] < \infty$, the OLS estimate has an asymptotic normal distribution: $\hat{\beta}|X \xrightarrow{d} N(\beta, \sigma^2(X'X)^{-1})$.

Proof:
$$\sqrt{n}(\hat{\beta}-\beta) = \underbrace{\mathbb{E}_n[x_i x_i']^{-1}}_{\xrightarrow{p} Q^{-1}}\underbrace{\sqrt{n}\mathbb{E}_n[x_i\varepsilon_i]}_{\xrightarrow{d} N(0,\Omega)} \to N(0,\Sigma)$$
where in general $\Omega = Var(x_i\varepsilon_i) = \mathbb{E}[(x_i\varepsilon_i)^2]$ and $\Sigma = Q^{-1}\Omega Q^{-1}$. \

■

Given that $Q = \mathbb{E}[x_i x_i']$ is unobserved, we estimate it with $\hat{Q} = \mathbb{E}_n[x_i x_i']$. Since we have assumed homoskedastic error term, we have $\Omega = \sigma^2(X'X)^{-1}$. Since we do not observe σ^2 we estimate it as $\hat{\sigma}^2 = \mathbb{E}_n[\hat{\varepsilon}_i^2]$.

The terms $x_i\varepsilon_i$ are called **scores** and we can already see their central importance for inference.

5.1.3 Heteroskedastic Error Term

Assumption: $\mathbb{E}[\varepsilon_i x_i \varepsilon_j' x_j'] = 0$, for all $j \neq i$ and $\mathbb{E}[\varepsilon_i^4] \leq \infty$, $\mathbb{E}[||x_i||^4] \leq C < \infty$ a.s.

Theorem: Under GM assumptions (1)-(4) plus heteroskedastic error term, the following estimators are consistent, i.e. $\hat{\Sigma} \xrightarrow{p} \Sigma$.

Note that we are only specifying Ω of the $\Sigma = Q^{-1}\Omega Q^{-1}$ matrix.

- **HC0**: use the observed residual $\hat{\varepsilon}_i$
$$\Omega_{HC0} = \mathbb{E}_n[x_i x_i' \hat{\varepsilon}_i^2]$$
When k is too big relative to n – i.e., $k/n \to c > 0$ – $\hat{\varepsilon}_i^2$ are too small (Ω_{HC0} biased towards zero). Ω_{HC1}, Ω_{HC2} and Ω_{HC3} try to correct this small sample bias. \

- **HC1**: degree of freedom correction (default `robust` in Stata)
$$\Omega_{HC1} = \frac{1}{n-k}\mathbb{E}_n[x_i x_i' \hat{\varepsilon}_i^2]$$
- **HC2**: use standardized residuals
$$\Omega_{HC2} = \mathbb{E}_n[x_i x_i' \hat{\varepsilon}_i^2 (1-h_{ii})^{-1}]$$
where $h_{ii} = [X(X'X)^{-1}X']_{ii}$ is the **leverage** of the i^{th} observation. A large h_{ii} means that observation i is unusual in the sense that the regressor x_i is far from its sample mean.
- **HC3**: use prediction error, equivalent to Jack-knife estimator, i.e., $\mathbb{E}_n[x_i x_i' \hat{\varepsilon}_{(-i)}^2]$
$$\Omega_{HC3} = \mathbb{E}_n[x_i x_i' \hat{\varepsilon}_i^2 (1-h_{ii})^{-2}]$$
This estimator does not overfit when k is relatively big with respect to n. Idea: you exclude the corresponding observation when estimating a particular ε_i: $\hat{\varepsilon}_i = y_i - x_i'\hat{\beta}_{-i}$.

Theorem: Under regularity conditions HC0 is consistent, i.e. $\hat{\Sigma}_{HC0} \xrightarrow{p} \Sigma$.

$$\hat{\Sigma} = \hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1} \xrightarrow{p} \Sigma \qquad \text{with } \hat{\Omega} = \mathbb{E}_n[x_i x_i' \hat{\varepsilon}_i^2] \quad \text{and } \hat{Q} = \mathbb{E}_n[x_i x_i']^{-1}$$

Why is the proof relevant? You cannot directly apply the WLLN to $\hat{\Sigma}$.

Proof:
For the case $\dim(x_i) = 1$.

1. $\hat{Q}^{-1} \xrightarrow{p} Q^{-1}$ by WLLN since x_i is iid, $\mathbb{E}[x_i^4] < \infty$
2. $\bar{\Omega} = \mathbb{E}_n[\varepsilon_i^2 x_i x_i'] \xrightarrow{p} \Omega$ by WLLN since $\mathbb{E}_n[\varepsilon_i^4] < c$ and x_i bounded.
3. By the triangle inequality,
$$|\hat{\Omega} - \hat{\Omega}| \leq \underbrace{|\Omega - \bar{\Omega}|}_{\xrightarrow{p} 0} + \underbrace{|\bar{\Omega} - \hat{\Omega}|}_{\text{WTS:} \xrightarrow{p} 0}$$
4. We want to show $|\bar{\Omega} - \hat{\Omega}| \xrightarrow{p} 0$
$$\begin{aligned} |\bar{\Omega} - \hat{\Omega}| &= \mathbb{E}_n[\varepsilon_i^2 x_i^2] - \mathbb{E}_n[\hat{\varepsilon}_i^2 x_i^2] = \\ &= \mathbb{E}_n[\left(\varepsilon_i^2 - \hat{\varepsilon}_i^2\right) x_i^2] \leq \\ &\leq \mathbb{E}_n\left[\left(\varepsilon_i^2 - \hat{\varepsilon}_i^2\right)^2\right]^{\frac{1}{2}} \mathbb{E}_n[x_i^4]^{\frac{1}{2}} \end{aligned}$$
where $\mathbb{E}_n[x_i^4]^{\frac{1}{2}} \xrightarrow{p} \mathbb{E}[x_i^4]^{\frac{1}{2}}$ by x_i bounded, iid and CMT.
5. We want to show that $\mathbb{E}_n\left[(\varepsilon_i^2 - \hat{\varepsilon}_i^2)^2\right] \leq \eta$ with $\eta \to 0$. Let $L = \max_i |\hat{\varepsilon}_i - \varepsilon_i|$ (RV depending on n), with $L \xrightarrow{p} 0$ since
$$|\hat{\varepsilon}_i - \varepsilon_i| = |x_i\hat{\beta} - x_i\beta| \leq |x_i||\hat{\beta} - \beta| \xrightarrow{p} c \cdot 0$$

We can depompose

$$\begin{aligned}\left(\varepsilon_i^2 - \hat{\varepsilon}_i^2\right)^2 &= (\varepsilon_i - \hat{\varepsilon}_i)^2 (\varepsilon_i + \hat{\varepsilon}_i)^2 \leq \\ &\leq (\varepsilon_i + \hat{\varepsilon}_i)^2 L^2 = \\ &= (2\varepsilon_i - \varepsilon_i + \hat{\varepsilon}_i)^2 L^2 \leq \\ &\leq \left(2(2\varepsilon_i)^2 + 2(\hat{\varepsilon}_i - \varepsilon_i)^2\right)^2 L^2 \leq \\ &\leq (8\varepsilon_i^2 + 2L^2)L^2\end{aligned}$$

Hence

$$\mathbb{E}\left[\left(\varepsilon_i^2 - \hat{\varepsilon}_i^2\right)^2\right] \leq L^2 \left(8\mathbb{E}_n[\varepsilon_i^2] + 2\mathbb{E}_n[L^2]\right) \xrightarrow{p} 0$$

■

5.1.4 Heteroskedastic and Autocorrelated Error Term

Assumption: There esists a $\bar{d}$ such that:

- $\mathbb{E}[\varepsilon_i x_i \varepsilon'_{i-d} x'_{i-d}] \neq 0 \quad$ for $d \leq \bar{d}$
- $\mathbb{E}[\varepsilon_i x_i \varepsilon'_{i-d} x'_{i-d}] = 0 \quad$ for $d > \bar{d}$

Intuition: observations far enough from each other are not correlated.\

We can express the variance of the score as

$$\begin{aligned}\Omega_n &= Var(\sqrt{n}\mathbb{E}_n[x_i\varepsilon_i]) = \\ &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n x_i\varepsilon_i\right)\left(\frac{1}{n}\sum_{j=1}^n x_j\varepsilon_j\right)\right] = \\ &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[x_i\varepsilon_i x'_j\varepsilon'_j] = \\ &= \frac{1}{n}\sum_{i=1}^n\sum_{j:|i-j|\leq\bar{d}} \mathbb{E}[x_i\varepsilon_i x'_j\varepsilon'_j] = \\ &= \frac{1}{n}\sum_{d=0}^{\bar{d}}\sum_{i=d}^n \mathbb{E}[x_i\varepsilon_i x'_{i-d}\varepsilon'_{i-d}]\end{aligned}$$

We estimate Ω_n by

$$\hat{\Omega}_n = \frac{1}{n}\sum_{d=0}^{\bar{d}}\sum_{i=d}^n x_i\hat{\varepsilon}_i x'_{i-d}\hat{\varepsilon}'_{i-d}$$

Theorem: If $\bar{d}$ is a fixed integer, then

$$\hat{\Omega}_n - \Omega_n \xrightarrow{p} 0$$

What if $\bar{d}$ does not exist (all x_i, x_j are correlated)?

$$\hat{\Omega}_n = \frac{1}{n} \sum_{d=0}^{n} \sum_{i=d}^{n} x_i \hat{\varepsilon}_i x'_{i-d} \hat{\varepsilon}'_{i-d} = n \mathbb{E}_n [x_i \hat{\varepsilon}_i]^2 = 0$$

By the orthogonality property of the OLS residual.

HAC with Uniform Kernel

$$\hat{\Omega}_h = \frac{1}{n} \sum_{i,j} x_i \hat{\varepsilon}_i x'_j \hat{\varepsilon}'_j \mathbb{I}|i-j| \leq h$$

where h is the **bandwidth** of the kernel. The bandwidth is chosen such that $\mathbb{E}[x_i \varepsilon_i x'_{i-d} \varepsilon'_{i-d}]$ is small for $d > h$. How small? Small enough for the estimates to be consistent.

HAC with General Kernel

$$\hat{\Omega}^{HAC}_{k,h} = \frac{1}{n} \sum_{i,j} x_i \hat{\varepsilon}_i x'_j \hat{\varepsilon}'_j k \left(\frac{|i-j|}{n} \right)$$

Theorem: If the joint distribution is stationary and α-mixing with $\sum_{k=1}^{\infty} k^2 \alpha(k) < \infty$ and

- $\mathbb{E}[|x_{ij} \varepsilon_i|^\nu] < \infty \ \forall \nu$
- $\hat{\varepsilon}_i = y_i - x'_i \hat{\beta}$ for some $\hat{\beta} \xrightarrow{p} \beta_0$
- k smooth, symmetric, $k(0) \to \infty$ as $z \to \infty$, $\int k^2 < \infty$
- $\frac{h}{n} \to 0$
- $h \to \infty$

Then the HAC estimator is **consistent**.

$$\hat{\Omega}^{HAC}_{k,h} - \Omega_n \xrightarrow{p} 0$$

We want to choose h small relative to n in order to avoid estimation problems. But we also want to choose h large so that the remainder is small:

$$\begin{aligned} \Omega_n &= Var(\sqrt{n} \mathbb{E}_n [x_i \varepsilon_i]) = \\ &= \underbrace{\frac{1}{n} \sum_{i,j : |i-j| \leq h} \mathbb{E}[x_i \varepsilon_i x'_j \varepsilon'_j]}_{\Omega_n^h} + \underbrace{\frac{1}{n} \sum_{i,j : |i-j| > h} \mathbb{E}[x_i \varepsilon_i x'_j \varepsilon'_j]}_{\text{remainder: } R_n} = \\ &= \Omega_n^h + R_n \end{aligned}$$

In particular, HAC theory requires:

$$\hat{\Omega}^{HAC} \xrightarrow{p} \Omega \quad \text{if} \quad \begin{cases} \frac{h}{n} \to 0 \\ h \to \infty \end{cases}$$

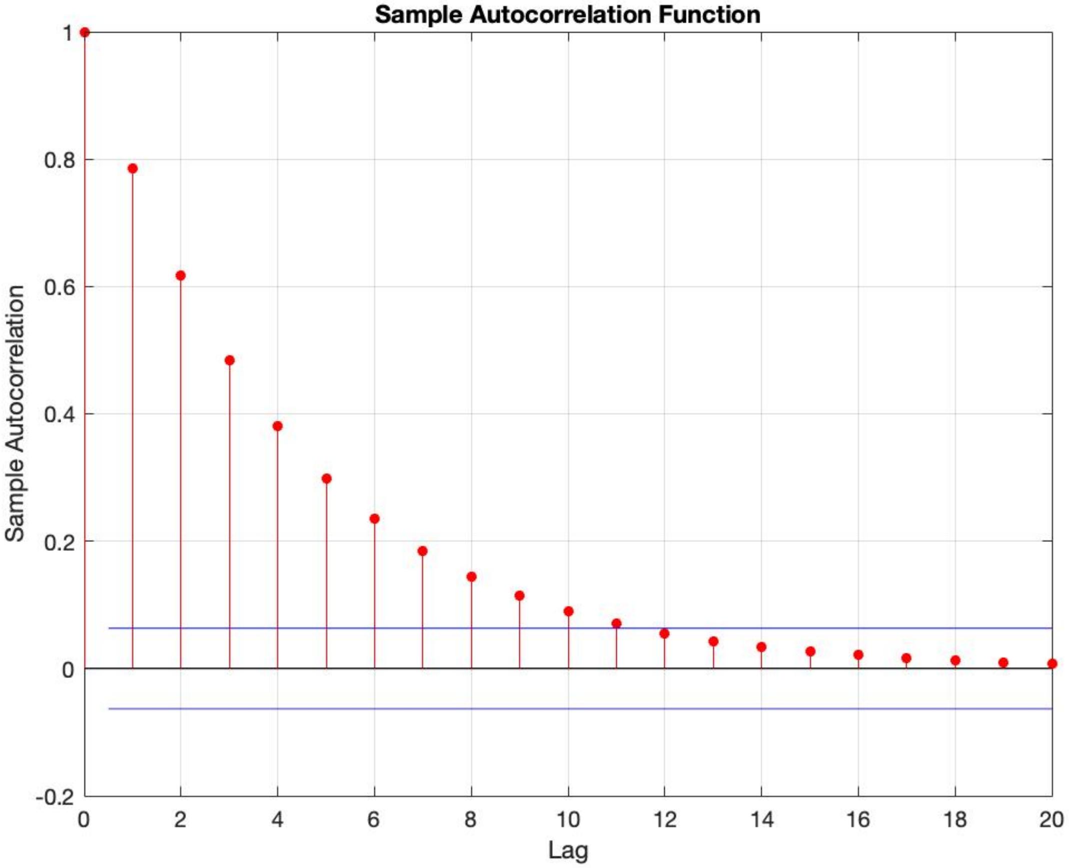


Figure 5.1: Autocorrelation Function

But in practice, long-run estimation implies $\frac{h}{n} \simeq 0$ which is not “safe" in the sense that it does not imply $R_n \simeq 0$. On the other hand, if $h \simeq n$, $\hat{\Omega}^{HAC}$ does not converge in probability because it's too noisy.

Example: How to choose h? Look at the score autocorrelation function (ACF).

```
% Set seed
seed(123)

% Autocorrelated process
X = rand(1000,1);
for t=11:length(X)
    for p=1:10
        X(t,:) = X(t,:) + X(t-p,:)*0.3;
    end
end
autocorr(X)
```

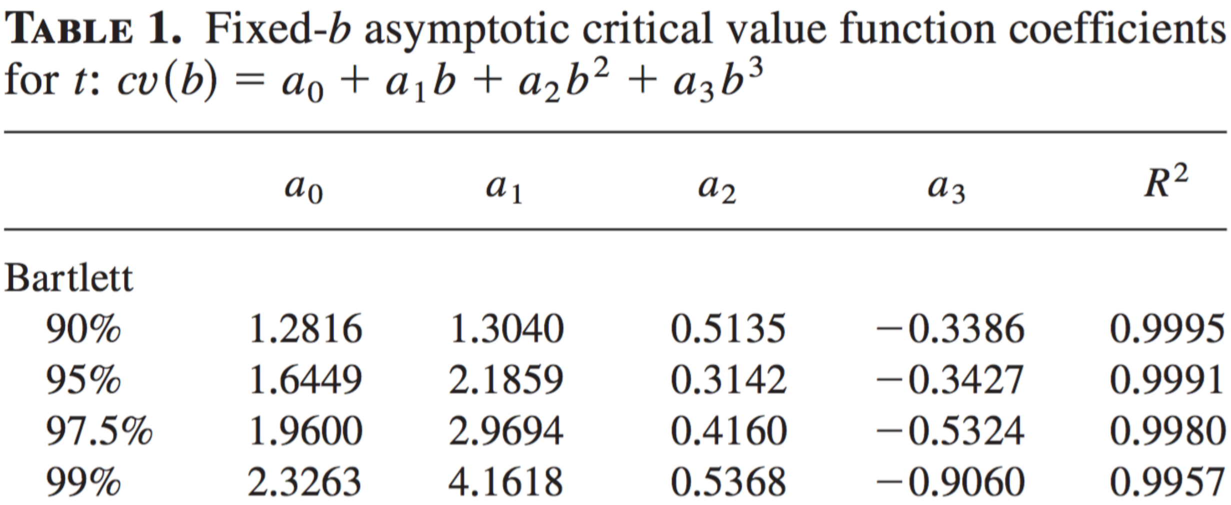


TABLE 1. Fixed-b asymptotic critical value function coefficients for t: $cv(b) = a_0 + a_1 b + a_2 b^2 + a_3 b^3$

	a_0	a_1	a_2	a_3	R^2
Bartlett					
90%	1.2816	1.3040	0.5135	−0.3386	0.9995
95%	1.6449	2.1859	0.3142	−0.3427	0.9991
97.5%	1.9600	2.9694	0.4160	−0.5324	0.9980
99%	2.3263	4.1618	0.5368	−0.9060	0.9957

Figure 5.2: Fixed-b

It looks like after 10 periods the empirical autocorrelation is quite small but still not zero.

5.1.5 Fixed b asymptotics

[Neave, 1970]: “*When proving results on the asymptotic behavior of estimates of the spectrum of a stationary time series, it is invariably assumed that as the sample size n tends to infinity, so does the truncation point h, but at a slower rate, so that* $\frac{h}{n}$ *tends to zero. This is a convenient assumption mathematically in that, in particular, it ensures consistency of the estimates, but it is unrealistic when such results are used as approximations to the finite case where the value of* $\frac{h}{n}$ *cannot be zero.*”"

Theorem: Under regularity conditions,

$$\sqrt{n}\Big(V^{HAC}_{k,h}\Big)(\hat{\beta} - \beta_0) \xrightarrow{d} F$$

The asymptotic critical values of the F statistic depend on the choice of the kernel. In order to do hypothesis testing, Kiefer and Vogelsang(2005) provide critical value functions for the t-statistic for each kernel-confidence level combination using a cubic equation:

$$cv(b) = a_0 + a_1 b + a_2 b^2 + a_3 b^3$$

Example for the Bartlett kernel:

5.1.6 Fixed G asymptotics

[Bester, 2013]: “*Cluster covariance estimators are routinely used with data that has a group structure with independence assumed across groups. Typically, inference is conducted in such settings under the assumption that there are a large number of these independent groups.*”"

“*However, with enough weakly dependent data, we show that groups can be chosen by the researcher so that group-level averages are approximately independent. Intuitively, if groups are large enough and well shaped (e.g. do not have gaps), the majority of points in a group will be far from other groups, and hence approximately independent of observations from other groups provided the data are weakly dependent. The key prerequisite for our methods is the researcher's ability to construct groups whose averages are approximately independent. As we show later, this often requires that the number of groups be kept relatively small, which is why our main results explicitly consider a fixed (small) number of groups.*”"

Assumption: Suppose you have data $D = (y_{it}, x_{it})_{i=1,t=1}^{N,T}$ where $y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_{it}$ where i indexes the observational unit and t indexes time (could also be space).

Let

$$\tilde{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

$$\tilde{x}_{it} = x_{it} - \frac{1}{T}\sum_{t=1}^{T} x_{it}$$

$$\tilde{\varepsilon}_{it} = \varepsilon_{it} - \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it}$$

Then

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{\varepsilon}_{it}$$

The $\tilde{\varepsilon}_{it}$ are by construction correlated between each other even if the original ε was iid. The **cluster score variance estimator** is given by:

$$\hat{\Omega}^{CL} = \frac{1}{T-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T} \tilde{x}_{it}\hat{\tilde{\varepsilon}}_{it}\tilde{x}_{is}\hat{\tilde{\varepsilon}}_{is}$$

> It's very similar too the HAC estimator since we have *dependent cross-products* here as well. However, here we do not consider the $i \times j$ cross-products. We only have time-dependency (state).

On T and n:

- If T is fixed and $n \to \infty$, then the number of cross-products considered is much smaller than the total number of cross-products.
- If $T >> n$ issues arise since the number of cross products considered is close to the total number of cross products. As in HAC estimation, this is a problem because it implies that the algebraic estimate of the cluster score variance gets close to zero because of the orthogonality property of the residuals.
- The panel assumption is that observations across individuals are not correlated.

  Strategy: as in HAC, we want to limit the correlation across clusters (individuals). We hope that observations are **negligibly dependent** between cluster sufficiently distant from each other.

Classical cluster robust estimator:

$$\hat{\Omega}^{CL} = \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i x_j' \varepsilon_j' \mathbb{I} i, j \text{ in the same cluster}$$

On clusters:

- If the number of observations near a boundary is small relative to the sample size, ignoring the dependence should not affect inference too adversely.
- The higher the dimension of the data, the easier it is to have observations near boundaries (*curse of dimensionality*).
- We would like to have few clusters in order to make less independence assumptions. However, few clusters means bigger blocks and hence a larger number of cross-products to estimate. If the number of cross-products is too large (relative to the sample size), $\hat{\Omega}^{CL}$ does not converge

Theorem: Under regularity conditions:

$$\hat{t} \xrightarrow{d} \sqrt{\frac{G}{G-1}} t_{G-1}$$

5.1.7 Matlab

```
% Homoskedastic standard errors
std_h = var(e_hat) * inv(X'*X);

% HC0 variance and standard errors
omega_hc0 = X' * diag(e_hat.^2) * X;
std_hc0 = sqrt(diag(inv(X'*X) * omega_hc0 * inv(X'*X))) % = 0.9195, 0.8631

% HC1 variance and standard errors
omega_hc1 = n/(n-k) * X' *  diag(e_hat.^2) * X;
std_hc1 = sqrt(diag(inv(X'*X) * omega_hc1 * inv(X'*X))) % = 0.9289, 0.8719

% HC2 variance and standard errors
omega_hc2 = X' * diag(e_hat.^2./(1-h)) * X;
std_hc2 = sqrt(diag(inv(X'*X) * omega_hc2 * inv(X'*X))) % = 0.9348, 0.8768

% HC3 variance and standard errors
omega_hc3 = X' * diag(e_hat.^2./(1-h).^2) * X;
std_hc3 = sqrt(diag(inv(X'*X) * omega_hc3 * inv(X'*X))) % = 0.9504, 0.8907

% Note what happens if you allow for full autocorrelation
omega_full = X'*e_hat*e_hat'*X;
```

5.2 Inference

In order to do inference on $\hat{\beta}$ we need to know its distribution. We have two options: (i) assume gaussian error term (extended GM) or (ii) rely on asymptotic approximations (CLT).

5.2.1 Hypothesis Testing

A statistical hypothesis is a subset of a statistical model, $\mathcal{K} \subset \mathcal{F}$. A hypothesis test is a map $\mathcal{D} \to \{0,1\}$, $D \mapsto T$. If $\mathcal{F}$ is the statistical model and $\mathcal{K}$ is the statistical hypothesis, we use the notation $H_0 : \Pr \in \mathcal{K}$.

> Generally, we are interested in understanding whether it is likely that data D are drawn from $\mathcal{K}$ or not.

A hypothesis test, T is our tool for deciding whether the hypothesis is consistent with the data. $T(D) = 0$ implies fail to reject H_0 and test inconclusive $T(D) = 1 \implies$ reject H_0 and D is inconsistent with any $\Pr \in \mathcal{K}$.

Let $\mathcal{K} \subseteq \mathcal{F}$ be a statistical hypothesis and T a hypothesis test.

1. Suppose $\Pr \in \mathcal{K}$. A Type I error (relative to Pr) is an event $T(D) = 1$ under Pr.

2. Suppose $\Pr \in \mathcal{K}^c$. A Type II error (relative to Pr) is an event $T(D) = 0$ under Pr.

The corresponding probability of a type I error is called **size**. The corresponding probability of a type II error is called **power** (against the alternative Pr).

In this section, we are interested in testing three hypotheses, under the assumptions of linearity, strict exogeneity, no multicollinearity, normality on the error term. They are:

1. $H_0 : \beta_{0k} = \bar{\beta}_{0k}$ (single coefficient, $\bar{\beta}_{0k} \in \mathbb{R}$, $k \leq K$)
2. $a'\beta_0 = c$ (linear combination, $a \in \mathbb{R}^K, c \in \mathbb{R}$)
3. $R\beta_0 = r$ (linear restrictions, $R \in \mathbb{R}^{p \times K}$, full rank, $r \in \mathbb{R}^p$) \

Consider the testing problem $H_0 : \beta_{0k} = \bar{\beta}_{0k}$ where $\bar{\beta}_{0k}$ is a pre-specified value under the null. The t-statistic for this problem is defined by

$$t_k := \frac{b_k - \bar{\beta}_{0k}}{SE(b_k)}, \quad SE(b_k) := \sqrt{s^2[(X'X)^{-1}]_{kk}}$$

Theorem: In the testing procedure above, the sampling distribution under the null H_0 is given by

$$t_k|X \sim t_{n-k} \ \text{ and so } \ t_k \sim t_{n-k}$$

$t_{(n-K)}$ denotes the t-distribution with $(n-k)$ degress of freedom. The test can be one sided or two sided. The above sampling distribution can be used to construct a confidence interval.

Example: We want to asses whether or not the “true" coefficient β_0 equals a specific value $\hat{\beta}$. Specifically, we are interested in testing H_0 against H_1, where:

- *Null Hypothesis*: $H_0 : \beta_0 = \hat{\beta}$
- *Alternative Hypothesis*: $H_1 : \beta_0 \neq \hat{\beta}$.

Hence, we are interested in a statistic informative about H_1, which is the Wald test statistic

$$|T^*| = \left| \frac{\hat{\beta} - \beta_0}{\sigma(\hat{\beta})} \right| \sim N(0,1)$$

However, the true variance $\sigma^2(\hat{\beta})$ is not known and has to be estimated. Therefore we plug in the sample variance $\hat{\sigma}^2(\hat{\beta}) = \frac{n}{n-1}\mathbb{E}_n[\hat{e}_i^2]$ and we use

$$|T| = \left| \frac{\hat{\beta} - \beta_0}{\hat{\sigma}(\hat{\beta})} \right| \sim t_{(n-k)}$$

Hypothesis testing is like proof by contradiction. Imagine the sampling distribution was generated by β. If it is highly improbable to observe $\hat{\beta}$ given $\beta_0 = \beta$ then we reject the hypothesis that the sampling distribution was generated by β.

Then, given a realized value of the statistic $|T|$, we take the following decision:

- *Do not reject* H_0: it is consistent with random variation under true H_0—i.e., $|T|$ small as it has an exact student t distribution with $(n-k)$ degree of freedom in the normal regression model.
- *Reject* H_0 *in favor of* H_1: $|T| > c$, with c being the critical values selected to control for false rejections: $\Pr(|t_{n-k}| \geq c) = \alpha$. Moreover, you can also reject H_0 if the p-value p is such that: $p < \alpha$.

The probability of false rejection is decreasing in c, i.e. the critical value for a given significant level.

$$\begin{aligned}
\Pr(\text{Reject } H_0|H_0) &= \Pr(|T| > c|H_0) = \\
&= \Pr(T > c|H_0) + \Pr(T < -c|H_0) = \\
&= 1 - F(c) + F(-c) = 2(1 - F(c))
\end{aligned}$$

Example: Consider the testing problem $H_0 : a'\beta_0 = c$ where a is a pre-specified linear combination under study. The t-statistic for this problem is defined by:

$$t_k := \frac{a'b - c}{SE(a'b)}, \quad SE(a'b) := \sqrt{s^2 a'(X'X)^{-1}a}$$

Theorem: In the testing procedure above, the sampling distribution under the null H_0 is given by

$$t_a|X \sim t_{n-K} \quad \text{and so} \quad t_a \sim t_{n-K}$$

Like in the previous test, $t_{(n-K)}$ denotes the t-distribution with $(n-K)$ degress of freedom. The test can again be one sided or two sided. The above sampling distribution can be used to construct a confidence interval.

Example: Consider the testing problem

$$H_0 : R\beta_0 = r$$

where $R \in \mathbb{R}^{p\times k}$ is a presepecified set of linear combinations and $r \in \mathbb{R}^p$ is a restriction vector.

The F-statistic for this problem is given by

$$F := \frac{(Rb - r)'[R(X'X)R']^{-1}(Rb - r)/p}{s^2}$$

Theorem:
For the problem, the sampling distribution of the F-statistic under the null H_0 :

$$F|X \sim F_{p,n-K} \text{ and so } F \sim F_{p,n-K}$$

The test is intrinsically two-sided. The above sampling distribution can be used to construct a confidence interval.

Theorem:
Consider the testing problem $H_0 : R\beta_0 = r$ where $R \in \mathbb{R}^{p\times K}$ is a presepecified set of linear combinations and $r \in \mathbb{R}^p$ is a restriction vector.

Consider the restricted least squares estimator, denoted $\hat{\beta}_R$: $\hat{\beta}_R := \arg\min_{\beta:R\beta=r} Q(\beta)$. Let $SSR_U = Q(b)$, $SSR_R = Q(\hat{\beta}_R)$. Then the F statistic is numerically equivalent to the following expression: $F = \frac{(SSR_R - SSR_U)/p}{SSR_U/(n-K)}$.

5.2.2 Confidence Intervals

A **confidence interval at** $(1 - \alpha)$ is a random set C such that

$$\Pr(\beta_0 \in C) \geq 1 - \alpha$$

i.e. the probability that C covers the true value β is fixed at $(1 - \alpha)$.

Since C is not known, it has to be estimated ($\hat{C}$). We construct confidence intervals such that:

- they are symmetric around $\hat{\beta}$;
- their length is proportional to $\sigma(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$.

A CI is equivalent to the set of parameter values such that the t-statistic is less than c, i.e.,

$$\hat{C} = \left\{\beta : |T(\beta)| \leq c\right\} = \left\{\beta : -c \leq \frac{\beta - \hat{\beta}}{\sigma(\hat{\beta})} \leq c\right\}$$

In practice, to construct a 95% confidence interval for a single coefficient estimate $\hat{\beta}_j$, we use the fact that

$$\Pr\left(\frac{|\hat{\beta}_j - \beta_{0,j}|}{\sqrt{\sigma^2[(X'X)^{-1}]_{jj}}} > 1.96\right) = 0.05$$

5.2.3 Matlab

```
% t-test for beta=0
t = abs(b_hat./(std_hc1));

% p-value
p_val = 1 - normcdf(t);

% F statistic of joint significance
SSR_u = e_hat'*e_hat;
SSR_r = y'*y;
F = (SSR_r - SSR_u)/k / (SSR_u/(n-k));

% 95% confidente intervals
conf_int = [b_hat - 1.96*std_hc1, b_hat + 1.96*std_hc1];
```

5.3 References

- Kozbur (2019). PhD Econometrics - Lecture Notes.
- Hansen (2019). "*Econometrics*".
- Kiefer and Vogelsang (2005). "*A new asymptotic theory for heteroskedasticity-autocorrelation robust tests*".
- Wooldridge (2010). "*Econometric Analysis of Cross Section and Panel Data*".
- Greene (2006). "*Econometric Analysis*".
- Hayiashi (2000). "*Econometrics*".

Chapter 6

Endogeneity

6.1 Instrumental Variables

We say that there is **endogeneity** in the linear regression model if $\mathbb{E}[x_i \varepsilon_i] \neq 0$.

The random vector z_i is an **instrumental variable** in the linear regression model if the following conditions are met.

- **Exclusion restriction**: the instruments are uncorrelated with the regression error
$$\mathbb{E}_n[z_i \varepsilon_i] = 0$$
almost surely, i.e. with probability $p \to 1$.
- **Rank condition**: no linearly redundant instruments
$$\mathbb{E}_n[z_i z_i'] \neq 0$$
almost surely, i.e. with probability $p \to 1$.
- **Relevance condition** (need $L > K$):
$$rank\ (\mathbb{E}_n[z_i x_i']) = K$$
almost surely, i.e. with probability $p \to 1$.

Let $K = dim(x_i)$ and $L = dim(z_i)$. We say that the model is **just-identified** if $L = K$ (method: IV) and **over-identified** if $L > K$ (method: 2SLS).

Assume z_i satisfies the instrumental variable assumptions above and $dim(z_i) = dim(x_i)$, then the **instrumental variables (IV)** estimator $\hat{\beta}_{IV}$ is given by

$$\begin{aligned}
\hat{\beta}_{IV} &= \mathbb{E}_n[z_i x_i']^{-1} \mathbb{E}_n[z_i y_i] = \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} z_i x_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} z_i y_i \right) = \\
&= (Z'X)^{-1}(Z'y)
\end{aligned}$$

Assume z_i satisfies the instrumental variable assumptions above and $dim(z_i) > dim(x_i)$, then the **two-stage-least squares (2SLS)** estimator $\hat{\beta}_{2SLS}$ is given by

$$\hat{\beta}_{2SLS} = \Big(X'Z(Z'Z)^{-1}Z'X\Big)^{-1}\Big(X'Z(Z'Z)^{-1}Z'y\Big)$$

Where $\hat{x}_i$ is the predicted x_i from the **first stage** regression of x_i on z_i. This is equivalent of the IV estimator using $\hat{x}_i$ as an instrument for x_i.

On the algebra of 2SLS:

- The estimator is called *two-stage-least squares* since it can be rewritten as an IV estimator that uses $\hat{X}$ as instrument:

$$\begin{aligned}\hat{\beta}_{\text{2SLS}} &= \Big(X'Z(Z'Z)^{-1}Z'X\Big)^{-1}\Big(X'Z(Z'Z)^{-1}Z'y\Big) = \\ &= (\hat{X}'X)^{-1}\hat{X}'y = \\ &= \mathbb{E}_n[\hat{x}_i x_i']^{-1}\mathbb{E}_n[\hat{x}_i y_i]\end{aligned}$$

- Moreover it can be rewritten as

$$\begin{aligned}\hat{\beta}_{\text{2SLS}} &= (\hat{X}'X)^{-1}\hat{X}'y = \\ &= (X'P_Z X)^{-1}X'P_Z y = \\ &= (X'P_Z P_Z X)^{-1}X'P_Z y = \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'y = \\ &= \mathbb{E}_n[\hat{x}_i\hat{x}_i]^{-1}\mathbb{E}_n[\hat{x}_i y_i]\end{aligned}$$

- How to the test the relevance condition? Rule of thumb: F-test in the first stage > 10 (joint test on z_i). **Problem**: as $n \to \infty$, with finite L, $F \to \infty$ (bad rule of thumb).

Theorem: If $K = L$, $\hat{\beta}_{\text{2SLS}} = \hat{\beta}_{\text{IV}}$.

Proof: If $K = L$, $X'Z$ and $Z'X$ are squared matrices and, by the relevance condition, non-singular (invertible).

$$\begin{aligned}\hat{\beta}_{\text{2SLS}} &= \Big(X'Z(Z'Z)^{-1}Z'X\Big)^{-1}\Big(X'Z(Z'Z)^{-1}Z'y\Big) = \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y = \\ &= (Z'X)^{-1}(Z'Z)(Z'Z)^{-1}Z'y = \\ &= (Z'X)^{-1}(Z'y) = \\ &= \hat{\beta}_{\text{IV}}\end{aligned}$$

■

Example from Hayiashi (2000) page 187: demand and supply simultaneous equations.

$$\begin{aligned}q_i^D(p_i) &= \alpha_0 + \alpha_1 p_i + u_i \\ q_i^S(p_i) &= \beta_0 + \beta_1 p_i + v_i\end{aligned}$$

We have an endogeneity problem. To see why, we solve the system of equations for (p_i, q_i):

$$\begin{aligned} p_i &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1} \\ q_i &= \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1} \end{aligned}$$

Then the price variable is not independent from the error term in neither equation:

$$\begin{aligned} Cov(p_i, u_i) &= -\frac{Var(u_i)}{\alpha_1 - \beta_1} \\ Cov(p_i, v_i) &= \frac{Var(v_i)}{\alpha_1 - \beta_1} \end{aligned}$$

As a consequence, the two coefficient estimates are not consistent:

$$\begin{aligned} \hat{\alpha}_{1,OLS} &\xrightarrow{p} \alpha_1 + \frac{Cov(p_i, u_i)}{Var(p_i)} \\ \hat{\beta}_{1,OLS} &\xrightarrow{p} \beta_1 + \frac{Cov(p_i, v_i)}{Var(p_i)} \end{aligned}$$

In general, running OLS on $q_i = \gamma p_i + \varepsilon_i$ you estimate

$$\begin{aligned} \hat{\gamma}_{OLS} &\xrightarrow{p} \frac{Cov(p_i, q_i)}{Var(p_i)} = \\ &= \frac{\alpha_1 Var(v_i) + \beta_1 Var(u_i)}{(\alpha_1 - \beta_1)^2} \left(\frac{Var(v_i) + Var(u_i)}{(\alpha_1 - \beta_1)^2} \right)^{-1} = \\ &= \frac{\alpha_1 Var(v_i) + \beta_1 Var(u_i)}{Var(v_i) + Var(u_i)} \end{aligned}$$

Which is neither α_1 nor β_1 but a variance weighted average of the two.

Suppose we have a supply shifter z_i such that $\mathbb{E}[z_i v_i] \neq 0$ and $\mathbb{E}[z_i u_i] = 0$. We combine the second condition and $\mathbb{E}[u_i] = 0$ to get a system of 2 equations in 2 unknowns: α_0 and α_1.

$$\begin{aligned} \mathbb{E}[z_i u_i] &= \mathbb{E}[z_i(q_i^D(p_i) - \alpha_0 - \alpha_1 p_i)] = 0 \\ \mathbb{E}[u_i] &= \mathbb{E}[q_i^D(p_i) - \alpha_0 - \alpha_1 p_i] = 0 \end{aligned}$$

We could try to solve for the vector α that solves

$$\begin{aligned} \mathbb{E}_n[z_i(q_i^D - x_i\alpha)] &= 0 \\ \mathbb{E}_n[z_i q_i^D] - \mathbb{E}_n[z_i x_i \alpha] &= 0 \end{aligned}$$

If $\mathbb{E}_n[z_i x_i]$ is invertible, we get $\hat{\alpha} = \mathbb{E}_n[z_i x_i]^{-1}\mathbb{E}_n[z_i q_i^D]$ which is indeed the IV estimator of α using z_i as an instrument for the endogenous variable p_i.

6.1.1 Matlab

```
% Set the dimension of Z
l = 3;

% Draw instruments
Z = randn(n,l);

% Correlation matrix for error terms
S = eye(2,2); S(1,2)=.8; S(2,1)=.8;

% Endogenous X
gamma = [2, 0; 0, -1; -1, 3];
e = randn(n,2)*chol(S);
X = Z*gamma + e(:,1);

% Calculate Y
Y = X*b + e(:,2);

% Estimate beta OLS
beta_OLS = (X'*X)\(X'*Y) % = 2.1957, -0.9022

% IV: l=k=2 instruments
Z_IV = Z(:,1:k);
beta_IV = (Z_IV'*X)\(Z_IV'*Y) % = 2.1207, -1.3617

% Calculate standard errors
ehat = Y - X*beta_IV;
V_NHC_IV = var(ehat) * inv(Z_IV'*X)*Z_IV'*Z_IV*inv(Z_IV'*X);
V_HC0_IV = inv(Z_IV'*X)*Z_IV' * diag(ehat.^2) * Z_IV*inv(Z_IV'*X);

% 2SLS: l=3 instruments
Pz = Z*inv(Z'*Z)*Z';
beta_2SLS = (X'*Pz*X)\(X'*Pz*Y) % = 2.0723, -0.9628

% Calculate standard errors
ehat = Y - X*beta_2SLS;
V_NCH_2SLS = var(ehat) * inv(X'*Pz*X);
V_HC0_2SLS = inv(X'*Pz*X)*X'*Pz * diag(ehat.^2) *Pz*X*inv(X'*Pz*X);
```

6.2 GMM

Setting: we have a system of L moment conditions

$$\begin{aligned} &\mathbb{E}[g_1(\omega_i, \delta_0)] = 0 \\ &\vdots \\ &\mathbb{E}[g_L(\omega_i, \delta_0)] = 0 \end{aligned}$$

If $L = \dim(\delta_0)$, no problem. If $L > \dim(\delta_0)$, there may be no solution to the system of equations. There are two possibilities.

1. **First Solution**: add moment conditions until the system is identified

   $$\mathbb{E}[a' g(\omega_i, \delta_0)] = 0$$

   Solve $\mathbb{E}[A g(\omega_i, \delta)] = 0$ for $\hat{\delta}$. How to choose A? Such that it minimizes $Var(\hat{\delta})$.

2. **Second Solution**: generalized method of moments (GMM)

   $$\begin{aligned} \hat{\delta}_{GMM} &= \arg\min_{\delta} \quad \Big|\Big| \mathbb{E}_n[g(\omega_i, \delta)] \Big|\Big| = \\ &= \arg\min_{\delta} \quad n\mathbb{E}_n[g(\omega_i, \delta)]' W \mathbb{E}_n[g(\omega_i, \delta)] \end{aligned}$$

   The choice of A and W are closely related to each other.

6.2.1 1-step GMM

Since $J(\delta, W)$ is a quadratic form, a closed form solution exists:

$$\hat{\delta}(W) = \Big(\mathbb{E}_n[z_i x_i'] W \mathbb{E}_n[z_i x_i'] \Big)^{-1} \mathbb{E}_n[z_i x_i'] W \mathbb{E}_n[z_i y_i]$$

Assumptions for consistency of the GMM estimator given data $\mathcal{D} = \{y_i, x_i, z_i\}_{i=1}^n$:

- **Linearity**: $y_i = x_i \gamma_0 + \varepsilon_i$
- **IID**: (y_i, x_i, z_i) iid
- **Orthogonality**: $\mathbb{E}[z_i(y_i - x_i\gamma_0)] = \mathbb{E}[z_i \varepsilon_i] = 0$
- **Rank identification**: $\Sigma_{xz} = \mathbb{E}[z_i x_i']$ has full rank

Theorem: Under linearity, independence, orthogonality and rank conditions, if $\hat{W} \xrightarrow{p} W$ positive definite, then

$$\hat{\delta}(\hat{W}) \to \delta(W)$$

If in addition to the above assumption, $\sqrt{n}\mathbb{E}_n[g(\omega_i, \delta_0)] \xrightarrow{d} N(0, S)$ for a fixed positive definite S, then

$$\sqrt{n}(\hat{\delta}(\hat{W}) - \delta(W)) \xrightarrow{d} N(0, V)$$

where $V = (\Sigma_{xz}' W \Sigma_{xz})^{-1} \Sigma_{xz} W S W \Sigma_{xz} (\Sigma_{xz}' W \Sigma_{xz})^{-1}$.

Finally, if a consistent estimator $\hat{S}$ of S is available, then using sample analogues $\hat{\Sigma}_{xz}$ it follows that

$$\hat{V} \xrightarrow{p} V$$

> If $W = S^{-1}$ then V reduces to $V = (\Sigma'_{xz} W \Sigma_{xz})^{-1}$. Moreover, $(\Sigma'_{xz} W \Sigma_{xz})^{-1}$ is the smallest possible form of V, in a positive definite sense.

Therefore, to have an efficient estimator, you want to construct $\hat{W}$ such that $\hat{W} \xrightarrow{p} S^{-1}$.

6.2.2 2-step GMM

Estimation steps:

- Choose an arbitrary weighting matrix $\hat{W}_{init}$ (usually the identity matrix I_K)
- Estimate $\hat{\delta}_{init}(\hat{W}_{init})$
- Estimate $\hat{S}$ (asymptotic variance of the moment condition)
- Estimate $\hat{\delta}(\hat{S}^{-1})$

  On the procedure:

  - This estimator achieves the semiparametric efficiency bound.
  - This strategy works only if $\hat{S} \xrightarrow{p} S$ exists.
  - For iid cases: we can use $\hat{\delta} = \mathbb{E}_n[(\hat{\varepsilon}_i z_i)(\hat{\varepsilon}_i z_i)']$ where $\hat{\varepsilon}_i = y_i - x_i \hat{\delta}(\hat{W}_{init})$.

6.2.3 `Matlab`

```
% GMM 1-step: inefficient weighting matrix
W_1 = eye(l);

% Objective function
gmm_1 = @(b) ( Y - X*b )' * Z * W_1 *  Z' * ( Y - X*b );

% Estimate GMM
beta_gmm_1 = fminsearch(gmm_1, beta_OLS) % = 2.0763, -0.9548
ehat = Y - X*beta_gmm_1;

% Standard errors GMM
S_hat = Z'*diag(ehat.^2)*Z;
d_hat = -X'*Z;
V_gmm_1 = inv(d_hat * inv(S_hat) * d_hat');

% GMM 2-step: efficient weighting matrix
W_2 = inv(S_hat);
gmm_2 = @(b) ( Y - X*b )' * Z * W_2 *  Z' * ( Y - X*b );
beta_gmm_2 = fminsearch(gmm_2, beta_OLS) % = 2.0595, -0.9666

% Standard errors GMM
```

```
ehat = Y - X*beta_gmm_2;
S_hat = Z'*diag(ehat.^2)*Z;
d_hat = -X'*Z;
V_gmm_2 = inv(d_hat * inv(S_hat) * d_hat');
```

6.3 Testing Overidentifying Restrictions

If the equations are **exactly identified**, then it is possible to choose δ so that all the elements of the sample moments $\mathbb{E}_n[g(\omega_i;\delta)]$ are zero and thus that the distance

$$J(\delta,\hat{W}) = n\mathbb{E}_n[g(\omega_i,\delta)]'\hat{W}\mathbb{E}_n[g(\omega_i,\delta)]$$

is zero. (The δ that does it is the IV estimator.)

If the equations are **overidentified**, i.e. L (number of instruments) $> K$ (number of equations), then the distance cannot be zero exactly in general, but we would expect the minimized distance to be *close* to zero.

6.3.1 Naive Test

Suppose your model is overidentified ($L > K$) and you use the following naive testing procedure:

1. Estimate $\hat{\delta}$ using a subset of dimension K of instruments $\{z_1,..,z_K\}$ for $\{x_1,...,x_K\}$
2. Set $\hat{\varepsilon}_i = y_i - x_i\hat{\delta}_{\text{GMM}}$
3. Infer the size of the remaining $L-K$ moment conditions $\mathbb{E}[z_{i,K+1}\varepsilon_i],...,\mathbb{E}[z_{i,L}\varepsilon_i]$ looking at their empirical counterparts $\mathbb{E}_n[z_{i,K+1}\hat{\varepsilon}_i],...,\mathbb{E}_n[z_{i,L}\hat{\varepsilon}_i]$
4. Reject exogeneity if the empirical expectations are high. How high? Calculate p-values.

Example If you have two invalid instruments and you use one to test the validity of the other, it might happen by chance that you don't reject it.

- Model: $y_i = x_i + \varepsilon_i$ and $x_i = \frac{1}{2}z_{i1} - \frac{1}{2}z_{i2} + u_i$
- Have

$$Cov(z_{i1}, z_{i2}, \varepsilon_i, u_i) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{bmatrix}$$

- You want to test whether the second instrument is valid (is not since $\mathbb{E}[z_2\varepsilon] \neq 0$). You use z_1 and estimate $\hat{\beta} \to$ the estimator is consistent.
- You obtain $\mathbb{E}_n[z_{i2}\hat{\varepsilon}_i] \simeq 0$ even if z_2 is invalid
- Problem: you are using an invalid instrument in the first place.

6.3.2 Hansen's Test

Theorem: We are interested in testing $H_0 : \mathbb{E}[z_i \varepsilon_i] = 0$ against $H_1 : \mathbb{E}[z_i \varepsilon_i] \neq 0$. Suppose $\hat{S} \xrightarrow{p} S$. Then

$$J(\hat{\delta}(\hat{S}^{-1}), \hat{S}^{-1}) \xrightarrow{d} \chi^2_{L-K}$$

For c satisfying $\alpha = 1 - G_{L-K}(c)$, $\Pr(J > c | H_0) \to \alpha$ so the test *reject* H_0 *if* $J > c$ has asymptotic size α.

On Hansen's test:

- The degrees of freedom of the asymptotic distribution are the number of overidentifying restrictions.
- This is a specification test, testing whether all model assumptions are true jointly. Only when we are confident that about the other assumptions, can we interpret a large J statistic as evidence for the endogeneity of some of the L instruments included in x.
- Unlike the tests we have encountered so far, the test is not consistent against some failures of the orthogonality conditions (that is, it is not consistent against some fixed elements of the alternative).
- Several papers in the July 1996 issue of JBES report that the finite-sample null rejection probability of the test can far exceed the nominal significance level α.

6.3.3 Special Case: Conditional Homoskedasticity

The main implication of conditional homoskedasticity is that efficient GMM becomes 2SLS. With efficient GMM estimation, the weighting matrix is $\hat{S}^{-1} = \mathbb{E}n[z_i z_i' \varepsilon_i^2]^{-1}$. With conditional homoskedasticity, the efficient weighting matrix is $\mathbb{E}_n[z_i z_i']^{-1} \sigma^{-2}$, or equivalently $\mathbb{E}_n[z_i z_i']^{-1}$. Then, the GMM estimator becomes

$$\hat{\delta}(\hat{S}^{-1}) = \left(\mathbb{E}_n[z_i x_i']' \underbrace{\mathbb{E}_n[z_i z_i']^{-1} \mathbb{E}[z_i x_i']}_{\text{ols of } x_i \text{ on } z_i} \right)^{-1} \mathbb{E}_n[z_i x_i']' \underbrace{\mathbb{E}_n[z_i z_i']^{-1} \mathbb{E}[z_i y_i']}_{\text{ols of } y_i \text{ on } z_i} = \hat{\delta}_{2SLS}$$

Proof: Consider the matrix notation.

$$\begin{aligned}
\hat{\delta}\left(\frac{Z'Z}{n}\right) &= \left(\frac{X'Z}{n} \left(\frac{Z'Z}{n}\right)^{-1} \frac{Z'X}{n} \right)^{-1} \frac{X'Z}{n} \left(\frac{Z'Z}{n}\right)^{-1} \frac{Z'Y}{n} = \\
&= \left(X'Z(Z'Z)^{-1}Z'X \right)^{-1} X'Z(Z'Z)^{-1}Z'Y = \\
&= \left(X'P_Z X \right)^{-1} X'P_Z Y = \\
&= \left(X'P_Z P_Z X \right)^{-1} X'P_Z Y = \\
&= \left(\hat{X}_Z' \hat{X}_Z \right)^{-1} \hat{X}_Z' Y = \\
&= \hat{\delta}_{2SLS}
\end{aligned}$$

■

6.4 Small-Sample Properties of 2SLS

Theorem: When the number of instruments is equal to the sample size ($L = n$), then $\hat{\delta}_{2SLS} = \hat{\delta}_{OLS}$

Proof: We have a perfect prediction problem. The first stage estimated coefficient $\hat{\gamma}$ is such that it solves the normal equations: $\hat{\gamma} = z_i^{-1} x_i$. Then

$$\begin{aligned}
\hat{\delta}_{2SLS} &= \mathbb{E}_n[\hat{x}_i x_i']^{-1} \mathbb{E}_n[\hat{x}_i y_i] = \\
&= \mathbb{E}_n[z_i z_i^{-1} x_i x_i']^{-1} \mathbb{E}_n[z_i z_i^{-1} x_i y_i] = \\
&= \mathbb{E}_n[x_i x_i']^{-1} \mathbb{E}_n[x_i y_i] = \\
&= \hat{\delta}_{OLS}
\end{aligned}$$

■

You have this overfitting problem in general when the number of instruments is large relative to the sample size. This problem arises even if the instruments are valid.

Example from Angrist (1992)

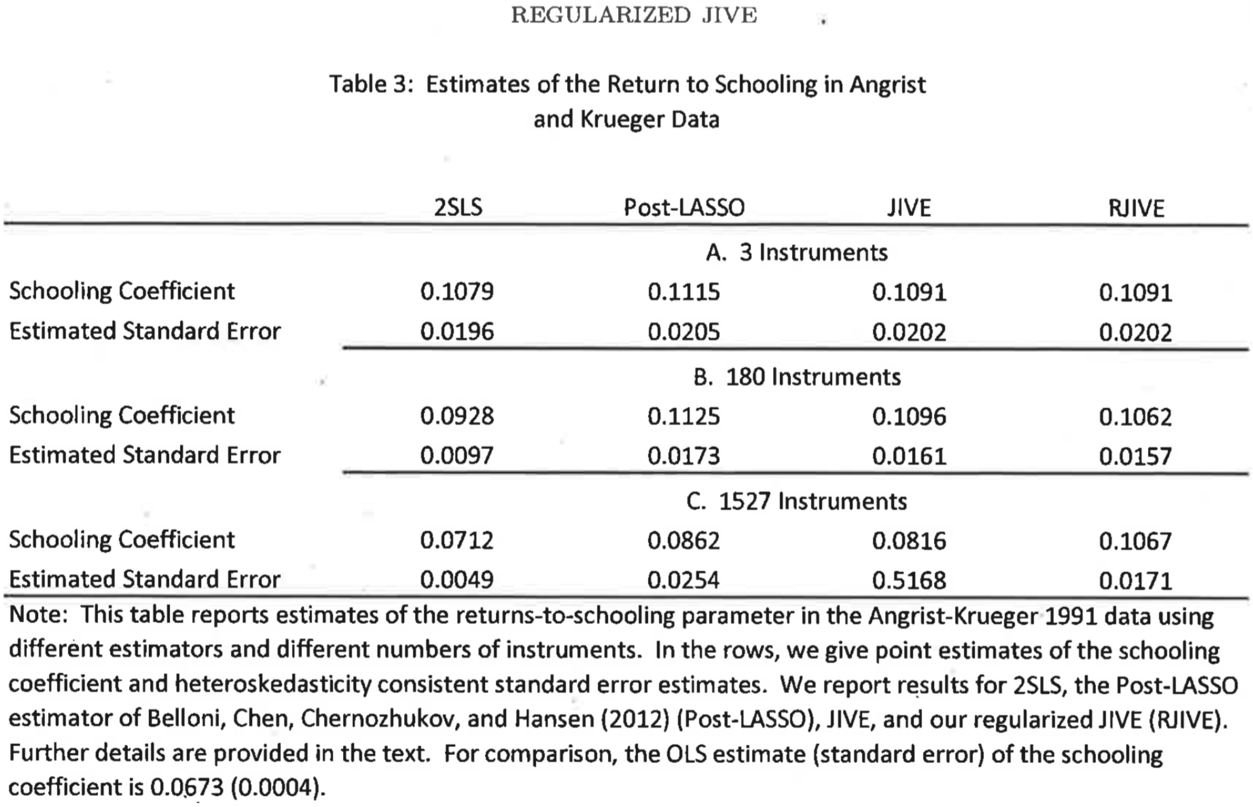
- They regress wages on years of schooling.
- Problem: endogeneity: both variables are correlated with skills which are unobserved.
- Solution: instrument years of schooling with the quarter of birth. Idea: if born in the first three quarters, can attend school from the year of your sixth birthday. Otherwise, you have to wait one more year.
- Problem: quarters of birth are three dummies. In order to “improve the first stage fit" they interact them with year of birth (180 effective instruments) and also with the state (1527 effective instruments). This mechanically increases the R^2 but also increases the bias of the 2SLS estimator.
- Solutions: LIML, JIVE, RJIVE (Hansen et al., 2014), Post-Lasso (Belloni et al., 2012).

6.5 Many Instrument Robust Estimation

Why having too many instruments is problematic? As the number of instruments increases, the estimated coefficient gets closer to OLS which is biased. As seen in the theorem above, for $L = n$, the two estimators coincide.

6.5.1 LIML

An alternative method to estimate the parameters of the structural equation is by maximum likelihood. Anderson and Rubin (1949) derived the maximum likelihood estimator for the joint distribution of (y_i, x_i). The estimator is known as **limited information maximum likelihood**, or **LIML**. This estimator is called “limited information" because it is based



REGULARIZED JIVE

Table 3: Estimates of the Return to Schooling in Angrist and Krueger Data

	2SLS	Post-LASSO	JIVE	RJIVE
		A. 3 Instruments		
Schooling Coefficient	0.1079	0.1115	0.1091	0.1091
Estimated Standard Error	0.0196	0.0205	0.0202	0.0202
		B. 180 Instruments		
Schooling Coefficient	0.0928	0.1125	0.1096	0.1062
Estimated Standard Error	0.0097	0.0173	0.0161	0.0157
		C. 1527 Instruments		
Schooling Coefficient	0.0712	0.0862	0.0816	0.1067
Estimated Standard Error	0.0049	0.0254	0.5168	0.0171

Note: This table reports estimates of the returns-to-schooling parameter in the Angrist-Krueger 1991 data using different estimators and different numbers of instruments. In the rows, we give point estimates of the schooling coefficient and heteroskedasticity consistent standard error estimates. We report results for 2SLS, the Post-LASSO estimator of Belloni, Chen, Chernozhukov, and Hansen (2012) (Post-LASSO), JIVE, and our regularized JIVE (RJIVE). Further details are provided in the text. For comparison, the OLS estimate (standard error) of the schooling coefficient is 0.0673 (0.0004).

Figure 6.1: RJIVE

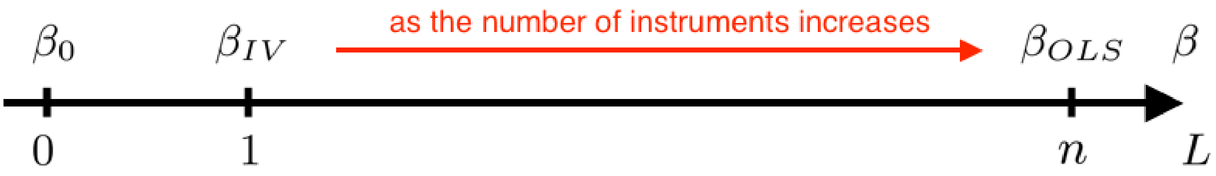


Figure 6.2: IV to OLS

on the structural equation for (y_i, x_i) combined with the reduced form equation for x_i. If maximum likelihood is derived based on a structural equation for x_i as well, then this leads to what is known as **full information maximum likelihood (FIML)**. The advantage of the LIML approach relative to FIML is that the former does not require a structural model for x_i, and thus allows the researcher to focus on the structural equation of interest - that for y_i.

The **k-class** estimators have the form

$$\hat{\delta}(\alpha) = (X'P_Z X - \alpha X'X)^{-1}(X'P_Z Y - \alpha X'Y)$$

The limited information maximum likelihood estimator **LIML** is the k-class estimator $\hat{\delta}(\alpha)$ where

$$\alpha = \lambda_{min}\Big(([X',Y]^{-1}[X',Y])^{-1}[X',Y]^{-1}P_Z[X',Y]\Big)$$

If $\alpha = 0$ then $\hat{\delta}_{\text{LIML}} = \hat{\delta}_{\text{2SLS}}$ while for $\alpha \to \infty$, $\hat{\delta}_{\text{LIML}} \to \hat{\delta}_{\text{OLS}}$.

Comments on LIML:

- The particular choice of α gives a many instruments robust estimate
- The LIML estimator has no finite sample moments. $\mathbb{E}[\delta(\alpha_{LIML})]$ does not exist in general
- In simulation studies performs well
- Has good asymptotic properties

Asymptotically the LIML estimator has the same distribution as 2SLS. However, they can have quite different behaviors in finite samples. There is considerable evidence that the LIML estimator has superior finite sample performance to 2SLS when there are many instruments or the reduced form is weak. However, on the other hand there is worry that since the LIML estimator is derived under normality it may not be robust in non-normal settings.

6.5.2 JIVE

The **Jacknife IV** procedure is the following

- Regress $\{x_j\}_{j\neq i}$ on $\{z_j\}_{j\neq i}$ and estimate π_{-i} (leave the i^{th} observation out).
- Form $\hat{x}_i = \hat{\pi}_{-i} z_i$.
- Run IV using $\hat{x}_i$ as instruments.

$$\hat{\delta}_{JIVE} = \mathbb{E}_n[\hat{x}_i x_i']^{-1}\mathbb{E}_n[\hat{x}_i y_i']$$

Comments on JIVE:

- Prevents overfitting.
- With many instruments you get bad out of sample prediction which implies low correlation between $\hat{x}_i$ and x_i: $\mathbb{E}_n[\hat{x}_i x_i'] \simeq 0$.
- Use lasso/ridge regression in the first stage in case of too many instruments.

6.6 Hausman Test

Here we consider testing the validity of OLS. OLS is generally preferred to IV in terms of precision. Many researchers only doubt the (joint) validity of the regressor z_i instead of being certain that it is invalid (in the sense of not being predetermined). So then they wish to choose between OLS and 2SLS, assuming that they have an instrument vector x_i whose validity is not in question. Further, assume for simplicity that $L = K$ so that the efficient GMM estimator is the IV estimator.

The **Hausman test statistic**

$$H \equiv n(\hat{\delta}_{IV} - \hat{\delta}_{OLS})'[A\hat{v}ar(\hat{\delta}_{IV} - \hat{\delta}_{OLS})]^{-1}(\hat{\delta}_{IV} - \hat{\delta}_{OLS})$$

is asymptotically distributed as a χ^2_{L-s} under the null where $s = \# z_i \cup x_i$: the number of regressors that are retained as instruments in x_i.

> In general, the idea of the Hausman test is the following. If you have two estimators, one which is efficient under H_0 but inconsistent under H_1 (in this case, OLS), and another which is consistent under H_1 (in this case, IV), then construct a test as a quadratic form in the differences of the estimators. Another classic example arises in panel data with the hypothesis H_0 of unconditional strict exogeneity. In that case, under H_0 Random Effects estimators are efficient but under H_1 they are inconsistent. Fixed Effects estimators instead are consistent under H_1.
>
> The Hausman test statistic can be used as a pretest procedure: select either OLS or IV according to the outcome of the test. Although widely used, this pretest procedure is not advisable. When the null is false, it is still possible that the test *accepts* the null (committing a Type 2 error). In particular, this can happen with a high probability when the sample size is *small* and/or when the regressor z_i is *almost valid.* In such an instance, estimation and also inference will be based on incorrect methods. Therefore, the overall properties of the Hausman pretest procedure are undesirable.
>
> The Hausman test is an example of a specification test. There are many other specification tests. One could for example test for conditional homoskedasticity. Unlike for the OLS case, there does not exist a convenient test for conditional homoskedasticity for the GMM case. A test statistic that is asymptotically chi-squared under the null is available but is extremely cumbersome; see White (1982, note 2). If in doubt, it is better to use the more generally valid inference methods that allow for conditional heteroskedasticity. Similarly, there does not exist a convenient test for serial correlation for the GMM case. If in doubt, it is better to use the more generally valid inference methods that allow for serial correlation; for example, when data are collected over time (that is, time-series data).

6.7 References

- Belloni, A., Chen, H., Chernozhukov, V., & Hansen, C. B. (2012). *Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain*. Econometrica, 80(6), 2369–2429.
- Hansen (2019). "*Econometrics*". Chapters 12 and 13.
- Hayiashi (2000). "*Econometrics*".
- Kozbur (2019). PhD Econometrics - Lecture Notes.

Chapter 7

Non-Parametric Estimation

7.1 Introduction

Non-parametric regression is a flexible estimation procedure for (i) regression functions $\mathbb{E}[y|x] = g(x)$ and (ii) density functions $f(x)$. You want to let your data to tell you how flexible you can afford to be in terms of estimation procedures. Non-parametric regression is naturally introduced in terms of fitting a curve.

Consider the problem of estimating the Conditional Expectation Function, defined as $\mathbb{E}[y_i|x_i] = g(x_i)$ given data $D = (x_i, y_i)_{i=1}^n$ under minimal assumption of $g(\cdot)$, e.g. smoothness. There are two main methods:

1. Local methods: Kernel-based estimation
2. Global methods: Series-based estimation

Another way of looking at non-parametrics is to do estimation/inference without specifying functional forms. With no assumptions, informative inference is impossible. Non parametrics tries to work with functional restrictions—continuity, differentiability, etc.—rather than pre-specifying functional form.

7.2 Discrete x - Cell Estimator

Suppose that x can take R distinct values, e.g. gender $R = 2$, years of schooling $R = 20$, gender×years of schooling $R = 2 \times 20$.

A simple way for estimating $\mathbb{E}\left[y|x\right] = g(x)$ is to split the sample to include observations with $x_i = x$ and calculate the sample mean of $\bar{y}$ for these observations. Note that this requires no assumptions about how $\mathbb{E}[y_i|x_i]$ varies with x since we fit a different value for each value x.

$$\hat{g}(x) = \frac{1}{\#\{i : x_i = x\}} \sum_{i:x_i=x} y_i$$

Issues:

- **Curse of dimensionality**: if R is big compared to n, there will be only a small number of observations per x values. If x_i is continuous, $R = n$ with probability 1. Solution: we can borrow information about $g_0(x)$ using neighboring observations of x.
- Averaging for each separate x_r value is only feasible in cases where x_i is coarsely discrete.

7.3 Local Non-Parametric Estimation - Kernels

Suppose we believe that $\mathbb{E}[y_i|x_i]$ is a smooth function of x_i – e.g. continuous, differentiable, etc. Then it should not change too much across values of x that are close to each other: we can estimate the conditional expectation at $x = \bar{x}$ by averaging y's over the values of x that are "close"" to $\bar{x}$. This procedure relies on two (three) arbitrary choices:

- Choice of the **kernel function** $K(\cdot)$; it is used to weight "far out"" observations, such that
  - $K : \mathbb{R} \to \mathbb{R}$
  - K is symmetric: $K(\bar{x} + x_i) = K(\bar{x} - x_i)$
  - $\lim_{x_i \to \infty} K(x_i - \bar{x}) = 0$
- Choice of the **bandwidth} h: it measures the size of a "small" window around $\bar{x}$, e.g. $(\bar{x} - h, \bar{x} + h)$.
- Choice of the local estimation procedure. Examples are locally constant, a.k.a. Nadaraya-Watson (**NW**), and locally linear (**LL**).

  Generally, the choice of h is more important than $K(\cdot)$ in low dimensional settings.

We need to define what is an "optimal" h, depending on the smoothness level of g_0, typically unknown. The choice of h relates to the bias-variance trade-off:

- large h: small variance, higher bias;
- small h: high variance, smaller bias.

  Note that $K_h(\cdot) = K(\cdot/h)$.

7.3.1 Estimator examples:

- **Nadaraya-Watson** estimator, or locally constant estimator. It assumes the CEF locally takes the form $g(x) = \beta_0(x)$. The local parameter is estimated as:

$$\hat{\beta}_0(\bar{x}) = \arg\min_{\beta_0} \quad \mathbb{E}_n\Big[K_h(x_i - \bar{x}) \cdot \Big(y_i - \beta_0\Big)^2\Big]$$

The Nadaraya-Watson estimate of the CEF takes the form:

$$\mathbb{E}_n\left[y|x = \bar{x}\right] = \hat{g}(\bar{x}) = \frac{\sum_{i=1}^n y_i K_h(x_i - \bar{x})}{\sum_{i=1}^n K_h(x_i - \bar{x})}$$

- **Local Linear** estimator. It assumes the CEF locally takes the form $g(x) = \beta_0(x) + \beta_1(x)x$. The local parameters are estimated as:

$$\Big(\hat{\beta}_0(\bar{x}), \hat{\beta}_1(\bar{x})\Big) = \arg\min_{\beta_0, \beta_1} \quad \mathbb{E}_n\Big[K_h(x_i - \bar{x}) \cdot \Big(y_i - \beta_0 - (x_i - \bar{x})\beta_1\Big)^2\Big]$$

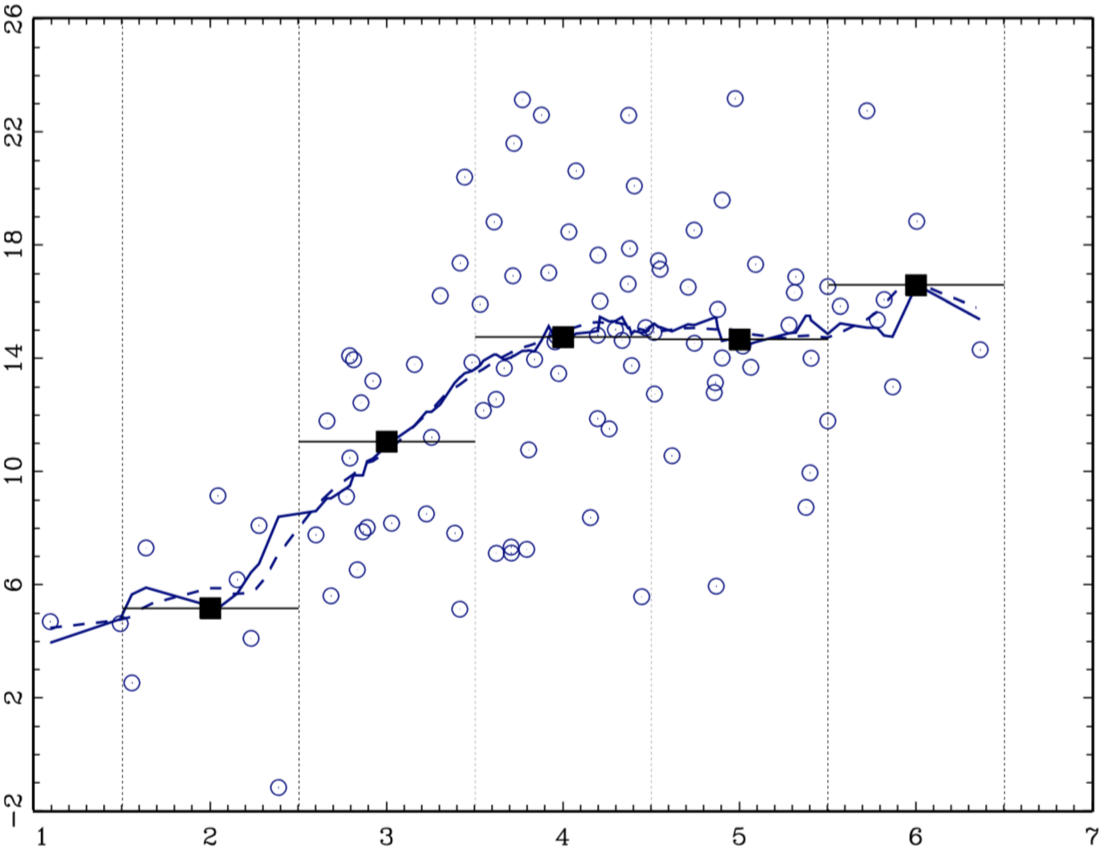


Figure 17.1: Scatter of (y_i, x_i) and Nadaraya-Watson regression

Figure 7.1: NW regression

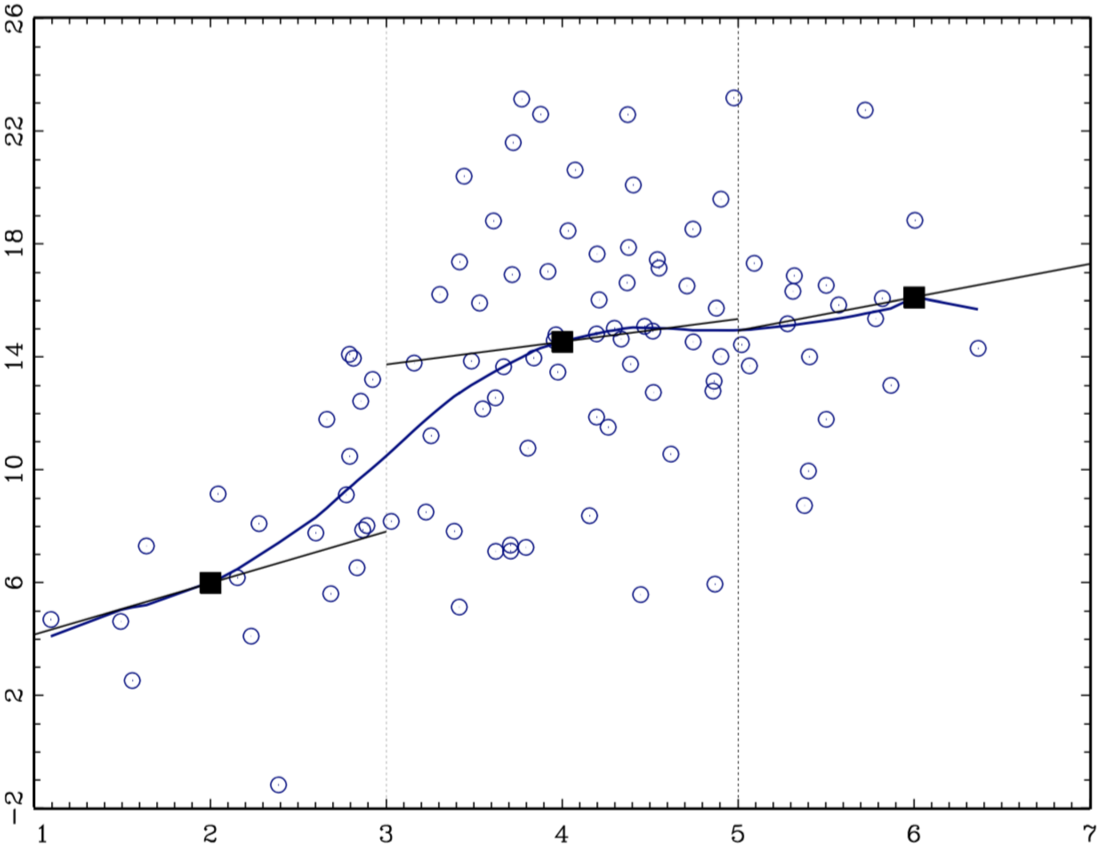


Figure 17.2: Scatter of (y_i, x_i) and Local Linear fitted regression

Figure 7.2: LL regression

In this case, we do LS estimate with i's contribution of residual weighted by the kernel $K_h(x_i - \bar{x})$. The final estimate at $\bar{x}$ is given by:

$$\hat{g}(\bar{x}) = \hat{\beta}_0(\bar{x}) + (\bar{x} - \bar{x})\hat{\beta}_1(\bar{x}) = \hat{\beta}_0(\bar{x})$$

since we have centered the x_s at $\bar{x}$ in the kernel. - It is possible to add linearly higher order polynomials, e.g. do locally quadratic least squares using loss function:

$$\mathbb{E}_n \left[K_h(x_i - \bar{x}) \Big(y_i - \beta_0 - (x_i - \bar{x})\beta_1 - (x_i - \bar{x})^2 \beta_2 \Big)^2 \right]$$

7.3.2 Kernel examples:

- **Uniform kernel**. LS restricted to sample i such that x_i within h of $\bar{x}$.

  $$K(\cdot) = \mathbb{I}\{\cdot \in [-1, 1]\}$$
  $$K_h(\cdot) = \mathbb{I}\{\cdot / h \in [-1, 1]\} = \mathbb{I}\{\cdot \in [-h, h]\}$$
  $$K_h(x_i - \bar{x}) = \mathbb{I}\{x_i - \bar{x} \in [-h, h]\} = \mathbb{I}\{x_i \in [\bar{x} - h, \bar{x} + h]\}$$

  Employed together with the locally linear estimator, the estimation procedure reduces to **local least squares}. The loss function is:

  $$\mathbb{E}_n \Big[K_n(x_i - \bar{x}) \Big(y_i - \beta_0 - \beta_1(x_i - \bar{x}) \Big)^2 \Big] = \frac{1}{n} \sum_{i: x_i \in [\bar{x} - h, \bar{x} + h]} \Big(y_i - \beta_0 - \beta_1(x_i - \bar{x}) \Big)^2$$

The more local is the estimation, the more appropriate the linear regression: if g_0 is smooth, $g_0(\bar{x}) + g_0'(\bar{x})(x_i - \bar{x})$ is a better approximation for $g_0(x_i)$.

However, the uniform density is not a good kernel choice as it produces discontinuous CEF estimates. The following are two popular alternative choices that produce continuous CEF estimates.

- **Epanechnikov kernel**

  $$K_h(x_i - \bar{x}) = \frac{3}{4} \Big(1 - (x_i - \bar{x})^2 \Big) \mathbb{I}\{x_i \in [\bar{x} - h, \bar{x} + h]\}$$

- **Normal or Gaussian kernel**

  $$K_\phi(x_i - \bar{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(- \frac{(x_i - \bar{x})^2}{2} \right)$$

- **K-Nearest Neighbors (KNN)**: choose bandwidth so that there is a fixed number of observations in each kernel. This kernel is different from the others since it takes a nonparamentric form.

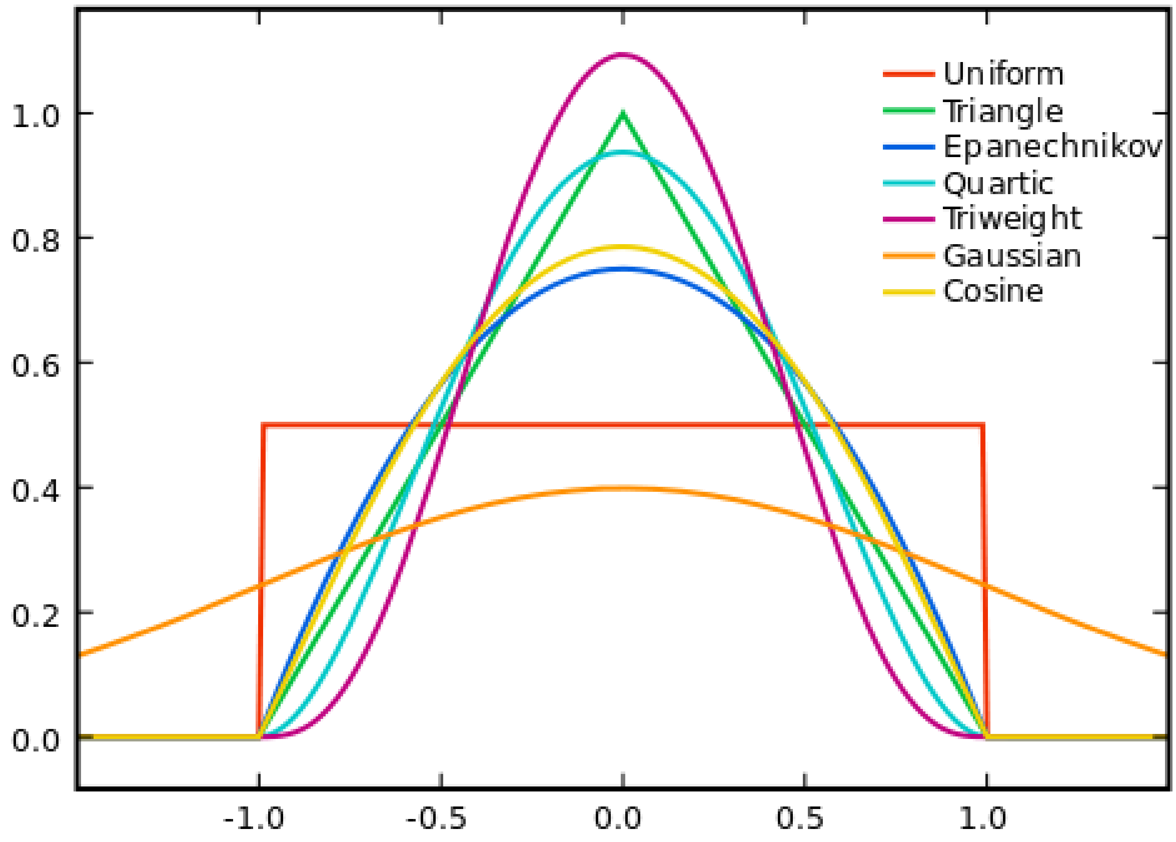


Figure 7.3: Kernelsh

7.3.3 Choice of the optimal bandwidth

Practical methods:

- **Eyeball Method.** (i) Choose a bandwidth (ii) Estimate the regression function (iii) Look at the result: if it looks more wiggly than you would like, increase the bandwidth: if it looks more smooth than you would like, decrease the bandwidth. Con: It only works for $\dim(x_i) = 1$ or 2.

- **Rule of Thumb.** For example, Silverman's rule of thumb: $h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}}$. Con: It requires too much knowledge about g_0 (i.e. normality) which you don't have.

- **Cross Validation.** Under some assumptions, CV will approximately gives the MSE optimal bandwidth. The basic idea is to evaluate quality of the bandwidth by looking at how well the resulting estimator forecasts in the given sample.

Leave-one-out CV. For each $h > 0$ and each i, $\hat{g}_{-i}(x_i)$ is the estimate of the conditional expectation at x_i using bandwidth h and all observations expect observation i. The CV bandwidth is defined as

$$\hat{h} = \arg\min_h CV(h) = \arg\min_h \sum_{i=1}^n \Big(y_i - \hat{g}_{-i}(x_i)\Big)^2$$

How to choose the optimal bandwidth:

- Select a value for h.
- For each observation i, calculate

$$\hat{g}_{-i}(x_i) = \frac{\sum_{j \neq i} y_j K_h(x_j - x_i)}{\sum_{i=1}^n K_h(x_j - x_i)}, \qquad e_{i,h}^2 = (y_i - \hat{g}_{-i}(x_i))^2$$

- Calculate $\text{CV}(h) = \sum_{i=1}^n e_{i,h}^2$.
- Repeat for each h and choose the one that minimizes $\text{CV}(h)$.

7.3.4 Inference

Theorem: Consider data $\{y_i, x_i\}_{i=1}^n$, iid and suppose that $y_i = g(x_i) + \varepsilon_i$ where $\mathbb{E}[\varepsilon_i | x_i] = 0$. Assume that $x_i \in Interior(X)$ where $X \subseteq \mathbb{R}$, $g(x)$ and $f(x)$ are three times continuously differentiable, and $f(x) > 0$ on X. $f(x)$ is the probability density of $x \in X$, and $g(x)$ is the function of interest. Suppose that $K(\cdot)$ is a kernel function. Suppose $n \to \infty$, $h \to 0$, $nh \to \infty$, and $nh^7 \to 0$. Then for any fixed $x \in X$,

$$AMSE = \sqrt{nh}\Big(\hat{g}(x) - g(x) - h^2 B(x)\Big) \xrightarrow{d} N\left(0, \frac{\kappa \sigma^2(x)}{f(x)}\right)$$

for $\sigma^2(x) = Var(y_i | x_i = x)$, $\kappa = \int K^2(v) dv$, and $B(x) = \frac{\kappa_2}{2} \frac{f'(x)g'(x) + f(x)g''(x)}{f(x)}$ where $\kappa_2 = \int v^2 K(v) dv$.

Remark on the theorem:

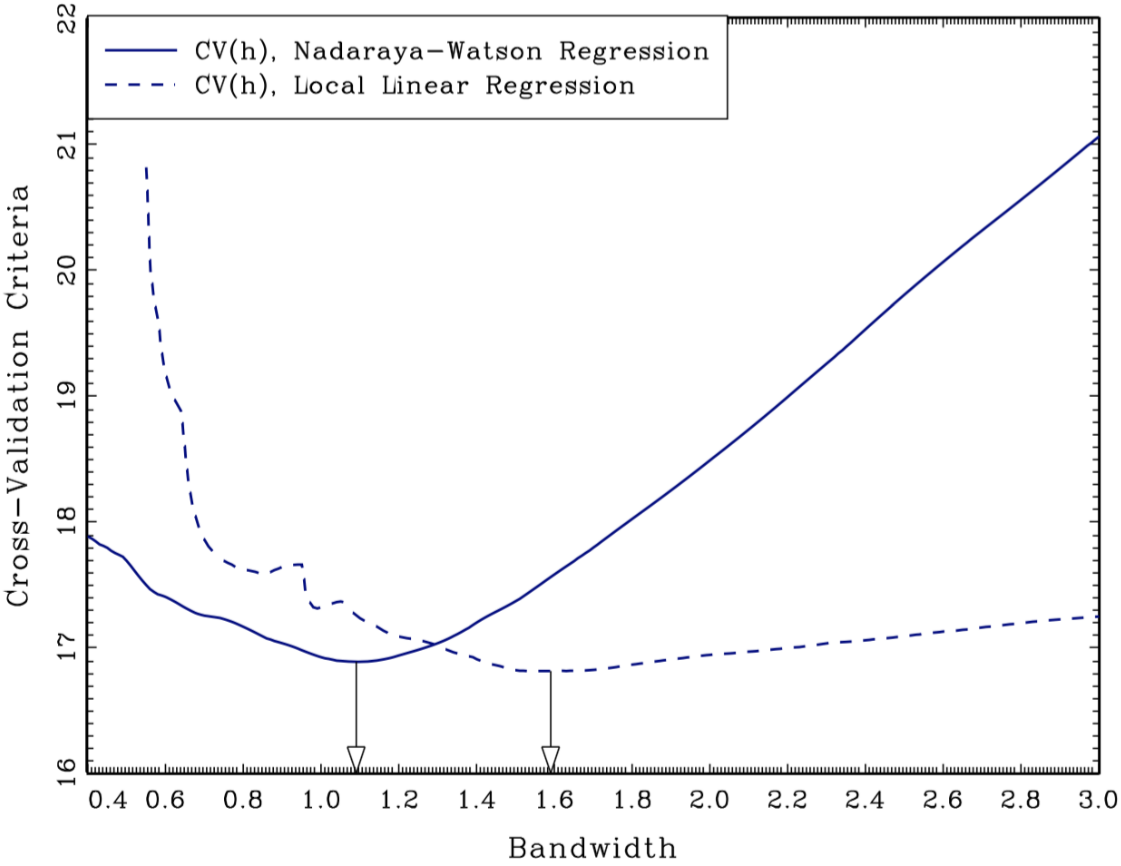


Figure 7.4: Optimal Bandwidth

- If the function is smooth enough and the bandwidth small enough, you can ignore the bias relative to sampling variation. To make this plausible, use a smaller bandwidth than would be the "optimal".
- All kernel regression estimators can be written as a weighted average

  $$\hat{g}(x) = \frac{1}{n}\sum_{i=1}^{n} w_i(x)y_i, \qquad \text{with} \qquad w_i(x) = \frac{nK_h(x_i - x)}{\sum_{i=1}^{n} K_h(x_i - x)}$$

  Do inference as if you were estimating a mean $\mathbb{E}[z_i]$ with sample mean $\frac{1}{n}\sum_{i=1}^{n} z_i$ using $z_i = w_i(x)y_i$.
- If you are doing inference at more than one value of x, do inference as in the previous point, treating each value of x as a different sample mean and note that even with independent data, these means will be correlated in general because there will generally be some common observations in to each of the averages. If you have a time series, make sure you account for correlation between the observations going in the different averages even if they don't overlap.

Issue when doing inference: the estimation of the bandwidth from the data is generally not accounted for in the distributional approximation (when doing inference). In large-samples, this is unlikely to lead to large changes, but uncertainty is understated in small samples.

7.3.5 Bias-variance trade-off

Theorem: For any estimator mean-square error MSE is decomposable into variance and bias-squared:

$$\text{MSE}(\bar{x}, \hat{g}) = \mathbb{E}\left[(\hat{g}(\bar{x}) - g_0(\bar{x}))^2\right] = \mathbb{E}\Big[\underbrace{\hat{g}(\bar{x}) - g_0(\bar{x})}_{\text{Bias}}\Big]^2 + Var(\hat{g}(\bar{x})).$$

The theorem follows from the following corollary.

Corollary: Let A be a random variable and θ_0 a fixed parameter. Then,

$$\mathbb{E}[(A - \theta_0)^2] = Var(A) + \mathbb{E}[A - \theta_0]^2$$

Proof:

$$\begin{aligned}
\mathbb{E}[(A - \theta_0)^2] &= \mathbb{E}[A^2] - 2\mathbb{E}[A\theta_0] + \mathbb{E}[\theta_0] \\
&= \mathbb{E}[A^2] \underbrace{-\mathbb{E}[A]^2 + E[A]^2}_{\text{add and subtract}} -2\mathbb{E}[A\theta_0] + \mathbb{E}[\theta_0] \\
&= Var(A) + \mathbb{E}[A]^2 - 2\theta_0\mathbb{E}[A] + \mathbb{E}[\theta_0] \\
&= Var(A) + \mathbb{E}[A - \theta_0]^2
\end{aligned}$$

Note that $\mathbb{E}[(A - \theta_0)^2] = \mathbb{E}[A - \theta_0]^2$.

■

Criteria to solve the bias-variance trade-off:

- **Mean squared error (MSE)}:**

  $$\mathbf{MSE}(\bar{x})(\hat{g}) = \mathbb{E}\left[(\hat{g}(\bar{x}) - g_0(\bar{x}))^2\right]$$

  NB!** This is the criterium we are going to use.

- **Integrated mean squared error (IMSE)**:

  $$\text{IMSE}(\hat{g}) = \mathbb{E}\left[\int |\hat{g}(x) - g_0(x)|^2 \mathrm{d}F(x)\right]$$

- Type I - Type II error.

Hansen (2019): the theorem above implies that we can asymptotically approximate the MSE as

$$\text{AMSE} = \left(h^2\sigma_k^2 B(x)\right)^2 + \frac{\kappa\sigma^2(x)}{nhf(x)} \approx \text{const} \cdot \left(h^4 + \frac{1}{nh}\right)$$

Where

- $Var \propto \frac{1}{hn}$, where you can think of nh as the **effective sample size**.
- Bias $\propto h^2$, derived if g_0 is twice continuously differentiable using Taylor expansion.

Trade-off: the asymptotic MSE is dominated by the larger of h^4 and $\frac{1}{hn}$. Notice that the bias is increasing in h and the variance is decreasing in h (more smoothing means more observations are used for local estimation: this increases the bias but decreases estimation variance). To select h to minimize the asymptotic MSE, these two components should balance each other:

$$\frac{1}{hn} \propto h^4 \quad \Rightarrow \quad h \propto n^{-1/5}$$

This result means that the bandwidth should take the form $h = c \cdot n^{-1/5}$. The optimal constant c depends on the kernel k the bias function $B(x)$ and the marginal density $f_x(x)$. A common misinterpretation is to set $h = n^{-1/5}$ which is equivalent to setting $c = 1$ and is completely arbitrary. Instead, an empirical bandwidth selection rule such as cross-validation should be used in practice.

7.4 Global Non-Parametric Estimation - Series

The goal is to try to globally approximate the CEF with a function $g(x)$. Series methods are based on the **Stone-Weierstrass theorem}: a real-valued continuous function $g(x)$ defined in a compact set can be approximated with polynomials for any degree of accuracy

$$g_0(x) = p_1(x)\beta_1 + \cdots + p_K(x)\beta_K + r(x)$$

where $p_1(x), \ldots, p_K(x)$ are called “a dictionary of approximating series” and $r(x)$ is a remainder function. If $p_1(x), \ldots, p_K(x)$ are sufficiently rich, $r(x)$ will be small. If $K \to \infty$, then $r \to 0$.

Example - Taylor series: if $g(x)$ is infinitely differentiable, then

$$g(x) = \sum_{k=0}^{\infty} a_k x^k$$

where $a_k = \frac{1}{k!}\frac{\partial^k g_0}{\partial x^k}$.

The basic idea is to approximate the infinite sum by chopping it off after K terms and then estimate the coefficients by OLS.

Series estimation:

- Choose K, i.e. the number of series terms, and an approximating dictionary $p_1(x), \ldots, p_K(x)$
- Expand data to $D = (y_i, p_1(x_i), \ldots, p_K(x_i))_{i=1}^n$
- Estimate OLS to get $\hat{\beta}_1, \ldots, \hat{\beta}_K$
- Set $\hat{g}(x) = p_1(x)\hat{\beta}_1 + \cdots + p_K(x)\hat{\beta}_K$

7.4.1 Examples

- **Monomials**: $p_1(x) = 1, p_2(x) = x, p_3(x) = x^2, \ldots$
- **Hermite Polynomials**: $p_1(x) = 1$, $p_2(x) = x$, $p_3(x) = x^2 - 1$, $p_4(x) = x^3 - 3x, \ldots$. Con: **edge effects**. The estimated function is particularly volatile at the edges of the sample space (Gibbs effect)
- **Trig Polynomials**: $p_1(x) = 1$, $p_2(x) = \cos 2\pi x$, $p_3(x) = \sin 2\pi x$, $p_4(x) = \cos 2\pi x \cdot 2x \ldots$. Pro: cyclical therefore good for series. Con: edge effects
- **B-splines**: recursively constructed using knot points

$$B_{i,0} = \begin{cases} 1 & \text{if } t_i \leq x < t_{i+1} \\ 0 & \text{otherwise} \end{cases} \qquad B_{i_k}(x) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(x)$$

where $t_0, \ldots, t_i, \ldots$ are knot points and k is the order of the spline. Pro: faster rate of convergence and lower asymptotic bias.

7.4.2 Estimation

Given K, inference proceeds exactly as if one had run an OLS of y on $(p_k)_{k=1}^K$. The idea is that you ignore that you are doing non-parametric regression as long as you believe you have put enough terms (high K). Then the function is smooth enough so that the bias of the approximation is small relative to the variance (see Newey, 1997). Note that his approximation does not account for data-dependent estimation of the bandwidth.

Newey (1997): results about consistency of $\hat{g}$ and asymptotic normality of $\hat{g}$.

- OLS: $\hat{\beta} \xrightarrow{p} \beta_0$

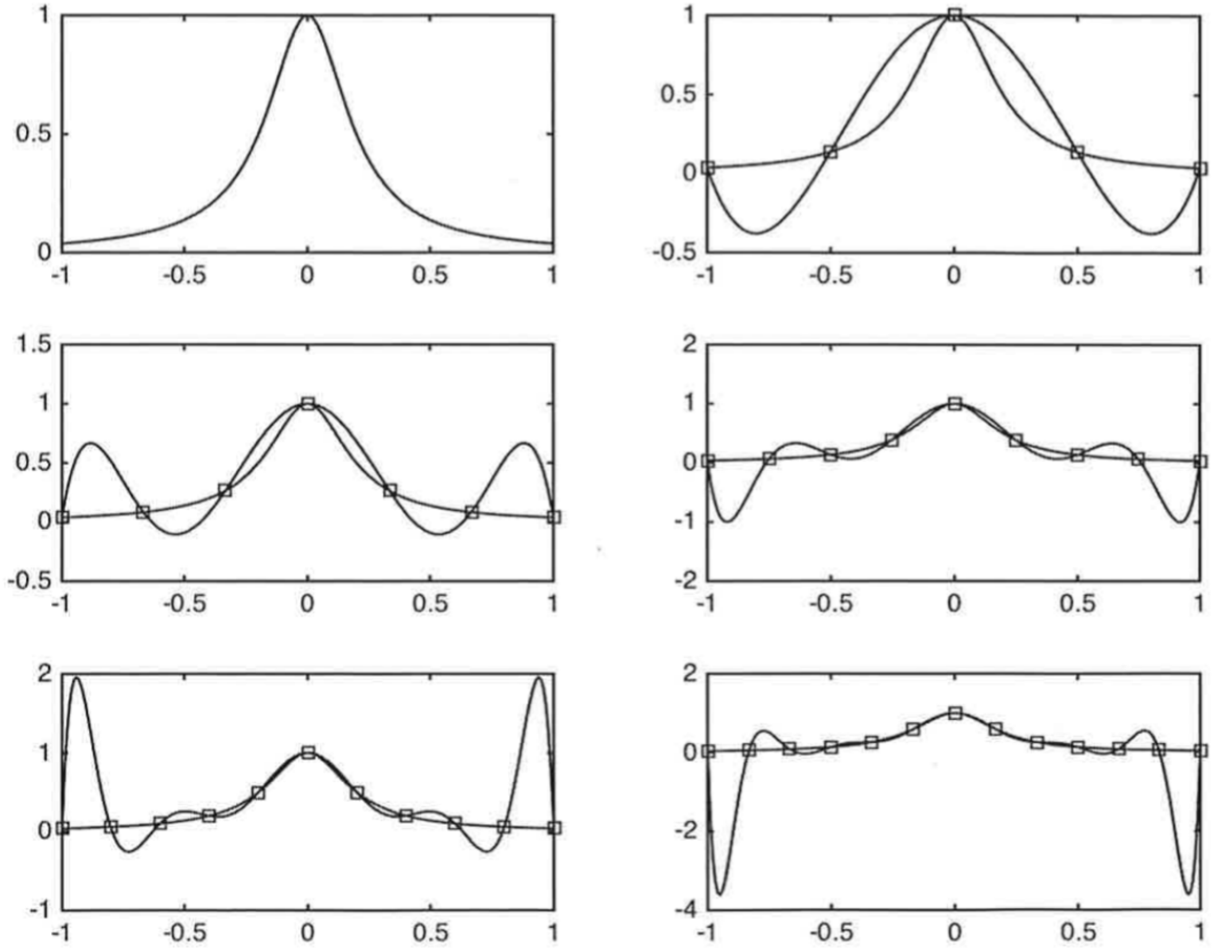


Figure 7.5: Hermite Polynomials

- Non-parametric: you have a sequence $\{\beta_k\}_{k=1}^K$ with $\hat\beta_k \xrightarrow{p} \beta_k$ as $n \to \infty$ (as $k \to \infty$). However, this does not make sense because $\{\beta_k\}$ is not constant. Moreover, β_k is not the quantity of interest. We want to make inference on $\hat g(x)$.

Theorem: Under regularity conditions, including $||\hat\beta - \beta_0|| \xrightarrow{p} 0$,

- Uniform Consistency: $\sup_x |\hat g(x) - g_0(x)| \xrightarrow{p} 0$
- Mean-square Consistency: $\int |\hat g(x) - g_0(x)|^2 \mathrm{d}F(x) \xrightarrow{p} 0$

Theorem: Under the following assumptions:

- (x_i, y_i) are iid and $Var(y_i|x_i)$ is bounded;
- For all K, there exists a non-singular matrix B such that $A = [(Bp(x))(Bp(x))']$ where $p(x) = (p_1(x), \dots, p_K(x))$ has the properties that $\lambda_{\min}(A)^{-1} = O(1)$. In addition, $\sup_x ||Bp(x)|| = o(\sqrt{K/n})$.
- There exists α and β_K for all K such that

$$\sup_x |g_0(x) - p(x)\beta_K| = O_p(K^{-\alpha})$$

Then, it holds that

$$\text{IMSE} = \int (g_0(x) - \hat g(x))^2 \, \mathrm{d}F(x) = O_p\left(\frac{K}{n} + K^{-2\alpha}\right)$$

7.4.3 Choice of the optimal K

The bias-variance trade-off for series comes in through the choice of K:

- Higher K: smaller bias, since we are leaving out less terms form the infinite sum.
- Smaller K: smaller variance, since we are estimating less regression coefficients from the same amount of data.

Cross-validation for series: For each $K \geq 0$ and for each $i = 1, \dots, n$, consider

$$D_{-i} = \{(x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots (x_n, y_n)\}$$

and calculate $\hat g_{-i}^{(K)}(x)$ using series estimate with $p_1(x), \dots, p_K(x)$ in order to get $e_i^{(K)} = y_i - \hat g_{-i}^{(K)}(x_i)$. Choose $\hat K$ such that

$$\hat K = \arg\min_K \mathbb{E}_n \left[{e_i^{(K)}}^2 \right]$$

7.4.4 Inference

Consider the data $D = \{(x_i, y_i)\}_{i=1}^n$ such that $y_i = g_0(x_i) + \varepsilon_i$. You may want to form confidence intervals for quantities that depends on g_0.

Example: θ_0 functional forms of interests:

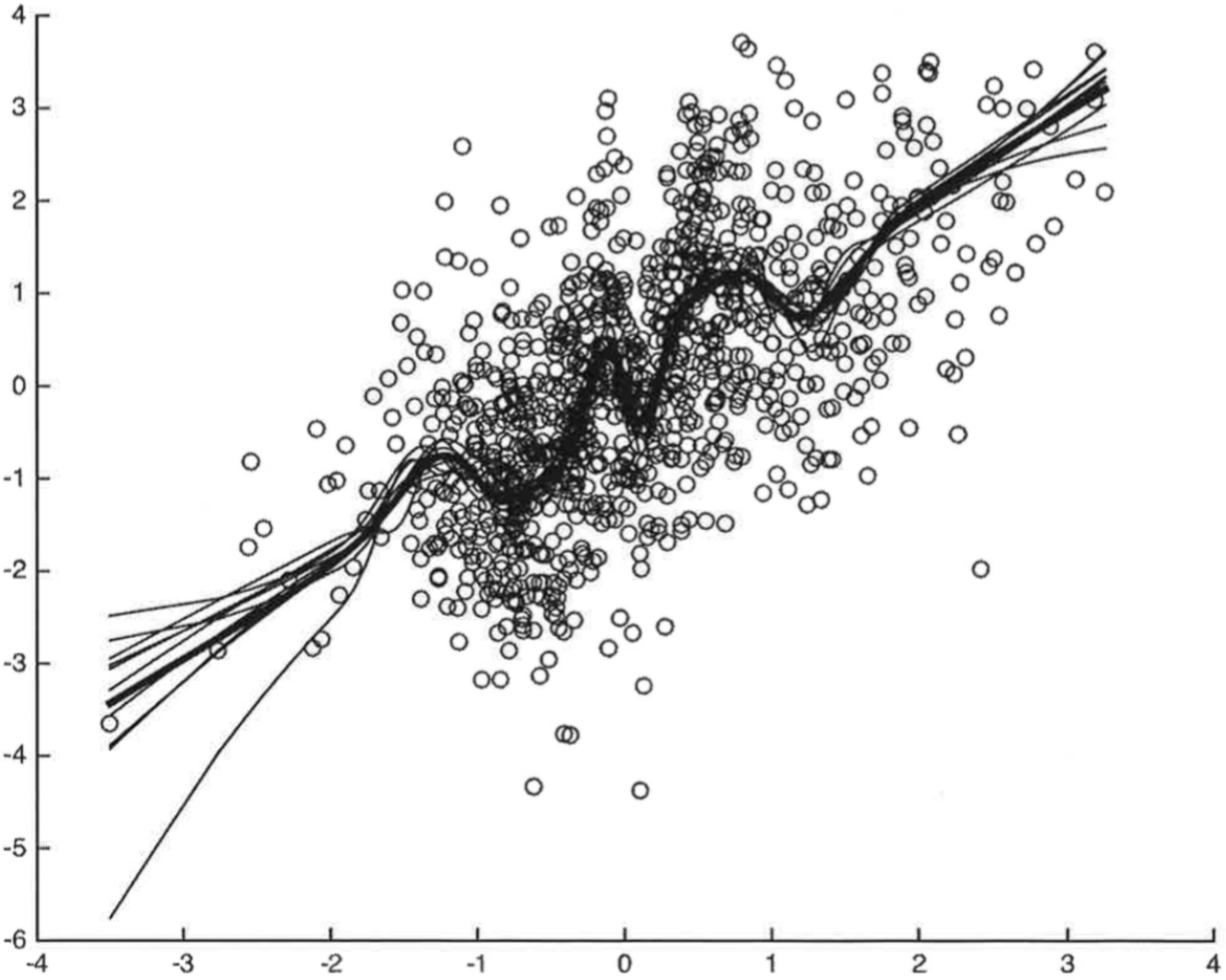


Figure 7.6: Smoothing

- Point estimate: $\theta_0 = g_0(\bar{x})$ for fixed $\bar{x}$
- Interval estimate: $\theta_0 = g_0(\bar{x}_2) - g_0(\bar{x}_1)$
- Point derivative estimate: $\theta_0 = g_0'(\bar{x})$ at $\bar{x}$
- Average derivative $\theta_0 = \mathbb{E}[g_0'(x)]$
- Consumer surplus: $\theta_0 = \int_a^b g_0(x)dx \quad$ when g_0 is a demand function.

Those estimates are functionals: maps from a function to a real number. We are doing inference on a function now, not on a point estimate.

In order to form a confidence interval for θ_0, with series you can

- **Undersmooth**: in order to apply a *central limit theorem*, you need deviations around the function to be approximately gaussian. Undersmoothing makes the function oscillate much more than the curve you are estimating in order to obtain such guassian deviations.

  Example: if on the contrary you oversmooth (e.g. g_0 linear), errors are going to constantly be on either one or the other side of the curve → not gaussian!

- Use the **delta method**. It would usually require more series terms than a criterion like cross-validation would suggest.

Theorem: Under the assumptions of the consistency theorem

$$\frac{\sqrt{n}\Big(\hat{\theta}-\theta_0+B(r_K)\Big)}{\sqrt{v_K}} \xrightarrow{d} N(0,1)$$

Theorem: Under the assumptions of the consistency theorem and $\sqrt{n}K^{-\alpha}=o(1)$ (or equivalently $nK^{-2\alpha}=O(1)$ in Hansen),

$$\frac{\sqrt{n}\Big(\hat{\theta}-\theta_0\Big)}{\sqrt{v_K}} \xrightarrow{d} N(0,1)$$

Remark on convergence rate:

- The rate of convergence of splines is faster than for power series (Newey 1997).
- We have **undersmoothing} if $\sqrt{n}K^{\alpha}=o(1)$ (see comment below)
- Usually, in order to prove asymptotic normality, we first prove unbiasedness. However here we have a **biased** estimator but we make the bias converge to zero faster than the variance.

Hansen (2019): The critical condition is the assumption that $\sqrt{n}K^{\alpha}=o(1)$ This requires that $K\to\infty$ at a rate faster than $n^{\frac{1}{2\alpha}}$ This is a troubling condition. The optimal rate for estimation of $g(x)$ is $K=O(n^{\frac{1}{1+2\alpha}})$. If we set $K=n^{\frac{1}{1+2\alpha}}$ by this rule then $nK^{-2\alpha}=n^{\frac{1}{1+2\alpha}}\to\infty$ not zero. Thus this assumption is equivalent to assuming that K is much larger than optimal. The reason why this trick works (that is, why the bias is negligible) is that by increasing K the asymptotic bias decreases and the asymptotic variance increases and thus the variance dominates. Because K is larger than optimal, we typically say that $\hat{g}(x)$ is **undersmoothed** relative to the optimal series estimator.

> Many authors like to focus their asymptotic theory on the assumptions in the theorem, as the distribution of θ appears cleaner. However, it is a poor use of asymptotic theory. There are three problems with the assumption $\sqrt{n}K^{-\alpha}=o(1)$ and the approximation of the theorem.
>
> - First, it says that if we intentionally pick K to be larger than optimal, we can increase the estimation variance relative to the bias so the variance will dominate the bias. But why would we want to intentionally use an estimator which is sub-optimal?
> - Second, the assumption $\sqrt{n}K^{-\alpha}=o(1)$ does not eliminate the asymptotic bias, it only makes it of lower order than the variance. So the approximation of the theorem is technically valid, but the missing asymptotic bias term is just slightly smaller in asymptotic order, and thus still relevant in finite samples.

- Third, the condition $\sqrt{n}K^{\alpha} = o(1)$ is just an assumption, it has nothing to do with actual empirical practice. Thus the difference between the two theorems is in the assumptions, not in the actual reality or in the actual empirical practice. Eliminating a nuisance (the asymptotic bias) through an assumption is a trick, not a substantive use of theory. My strong view is that the result (1) is more informative than (2). It shows that the asymptotic distribution is normal but has a non-trivial finite sample bias.

7.4.5 Kernel vs Series

Hansen (2019): in this and the previous chapter we have presented two distinct methods of nonparametric regression based on kernel methods and series methods. Which should be used in practice? Both methods have advantages and disadvantages and there is no clear overall winner.

First, while the asymptotic theory of the two estimators appear quite different, they are actually rather closely related. When the regression function $g(x)$ is twice differentiable ($s = 2$) then the rate of convergence of both the MSE of the kernel regression estimator with optimal bandwidth h and the series estimator with optimal K is $n^{-\frac{2}{k+4}}$ (where $k = \dim(x)$). There is no difference. If the regression function is smoother than twice differentiable ($s > 2$) then the rate of the convergence of the series estimator improves. This may appear to be an advantage for series methods, but kernel regression can also take advantage of the higher smoothness by using so-called higher-order kernels or local polynomial regression, so perhaps this advantage is not too large.

Both estimators are asymptotically normal and have straightforward asymptotic standard error formulae. The series estimators are a bit more convenient for this purpose, as classic parametric standard error formula work without amendment.

An advantage of kernel methods is that their distributional theory is easier to derive. The theory is all based on local averages which is relatively straightforward. In contrast, series theory is more challenging, dealing with increasing parameter spaces. An important difference in the theory is that for kernel estimators we have explicit representations for the bias while we only have rates for series methods. This means that plug-in methods can be used for bandwidth selection in kernel regression. However, typically we rely on cross-validation, which is equally applicable in both kernel and series regression.

Kernel methods are also relatively easy to implement when the dimension of x, k, is large. There is not a major change in the methodology as k increases. In contrast, series methods become quite cumbersome as k increases as the number of cross-terms increases exponentially. E.g ($K = 2$) with $k = 1$ you have only $\{x_1, x_1^2\}$; with $k = 2$ you have to add $\{x_2, x_2^2, x_1 x_2\}$; with $k = 3$ you have to add $\{x_3, x_3^2, x_1 x_3, x_2 x_3\}$, etc..

A major advantage of series methods is that it has inherently a high degree of flexibility, and the user is able to implement shape restrictions quite easily. For example, in series estimation it is relatively simple to implement a partial linear CEF, an additively separable CEF, monotonicity, concavity or convexity. These restrictions are harder to implement in kernel regression.

7.5 References

- Newey, W. K. (1997). *Convergence rates and asymptotic normality for series estimators.* Journal of Econometrics, 79(1), 147–168.
- Hansen (2019). "*Econometrics*". Chapters 19, 20 and 21.
- Kozbur (2019). PhD Econometrics - Lecture Notes.

Chapter 8

Variable Selection

8.1 Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) is a particularly popular method for high dimensional regression. It does variable selection and estimation simultaneously. It is a non-parametric (series) estimation technique part of a general class of estimators called *penalized estimators*. It allows number of series terms higher than n.

Consider data $D = \{x_i, y_i\}_{i=1}^n$ with $\dim(x_i) = p$. Assume that p is large relative to n. Two possible reasons:

- we have an intrinsic problem of high dimensionality
- p indicates the number of expansion terms of small number of underlying important variables (e.g. series estimation)

Assumption: $y_i = x_i'\beta_0 + r_i + \varepsilon_i$ where β_0 depends on p, r_i is a remainder term.

Note that in classic non-parametrics, we have $x_i'\beta_0$ as $p_1(x_i)\beta_{1,K} + \cdots + p_K(x_i)\beta_{K,K}$. For simplicity, we assume $r_i = 0$, as if we had extreme undersmoothing. Hence the model becomes:

$$y_i = x_i'\beta_0 + \varepsilon_i, \qquad p \geq n$$

We cannot run OLS because $p \geq n$, thus the rank condition is violated.

We define the **Lasso estimator** as

$$\hat{\beta}_L = \arg\min \quad \underbrace{\mathbb{E}_n\Big[(y_i - x_i'\beta)^2\Big]}_{\text{SSR term}} + \underbrace{\frac{\lambda}{n}\sum_{j=1}^{P}|\beta_j|}_{\text{Penalty term}}$$

where λ is called **penalty parameter**.

The **penalty term** discourages large values of $|\beta_j|$. The choice of λ is analogous to the choice of K in series estimation and h in the kernel estimation.

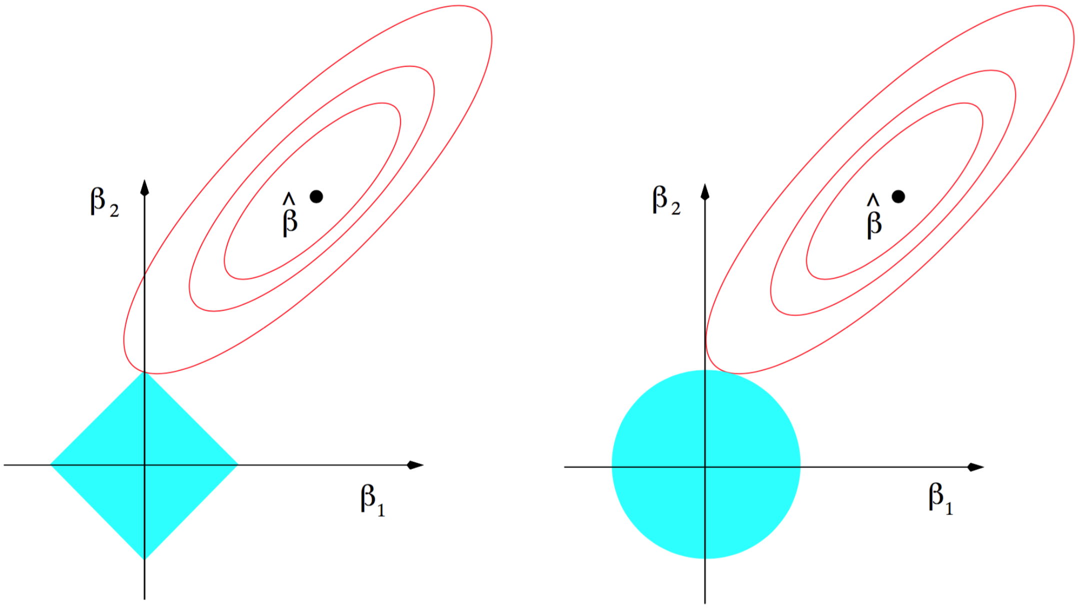


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Figure 8.1: Constraints

The shrinkage to zero of the coefficients directly follows from the $|| \cdot ||_1$ norm. On the contrary, another famous penalized estimator, *ridge regression*, uses the $|| \cdot ||_2$ norm and does not have this property.

Minimizing SSR + penalty is equivalent to minimize SSR $s.t.$ pen $\leq c$ (clear from the picture).

Let $S_0 = \{j : \beta_{0,j} \neq 0\}$, we define $s_0 = \#S_0$ as the **sparsity** of β_0. If $s_0/n \to 0$, we are dealing with a **sparse regression** (analogous of smooth regression).

Remark on sparsity:

- In words, sparsity means that even if we have a lot of variables, only a small number of them (relative to n) have an effect on the dependent variable.
- *Approximate sparsity imposes a restriction that only* s_0 *variables among all of* x_{ij}*, where* s_0 *is much smaller than* n*, have associated coefficients* β_{0j} *that are different from zero, while permitting a nonzero approximation error. Thus, estimators for this kind of model attempt to learn the identities of the variables with large nonzero coefficients, while simultaneously estimating these coefficients.* (Belloni et al., 2004)
- Sparsity is an assumption. β_0 is said to be s_0-sparse with $s_0 < n$ if

$$|\{j : \beta_{0j} \neq 0\}| \leq s_0$$

Theorem: Suppose that for data $D_n = (y_i, x_i)_{i=1}^N$ with $y_i = x_i'\beta + \varepsilon_i$. Let $\hat{\beta}_L$ be the Lasso estimator. Let $\mathcal{S} = 2\max_j |\mathbb{E}[x_{ij}\varepsilon_i]|$. Suppose $|support(\beta_0) \leq s_0$ (sparsity assumption). Let $c_0 = (\mathcal{S} + \lambda/n)/(-\mathcal{S} + \lambda/n)$. Let

$$\kappa_{c_0,s_0} = \min_{d\in\mathbb{R}^p, A\subseteq\{1,\ldots,p\}:|A|\leq s_0, ||d_{A^c}||\leq c_0||d_A||_1} \sqrt{\frac{s_0 d'\mathbb{E}_n[x_i x_i']d}{||d_A||_1^2}}$$

Then

$$\mathbb{I}_{\left\{\frac{\lambda}{n}>\mathcal{S}\right\}}\mathbb{E}_n[(x_i\beta_0 - x_i\beta_L)^2]^{\frac{1}{2}} \leq 2\frac{\lambda}{n}\frac{\sqrt{s_0}}{\kappa_{c_0,s_0}}$$

Intuition: for a sufficiently high lambda the root mean squared error of Lasso is approximately zero.

$$\text{RMSE } : \mathbb{E}_n[(x_i\beta_0 - x_i\beta_L)^2]^{\frac{1}{2}} \simeq 0 \quad \Leftrightarrow \quad \frac{\lambda}{n} > \mathcal{S}$$

Note on the theorem:

- The minimization region is the set of “essentially sparse" vectors $d \in \mathbb{R}^p$ where essentially sparse is defined by $\mathcal{C}, \mathcal{S}$. In particular the condition $k_{\mathcal{C},\mathcal{S}} > 0$ means that no essentially sparse vector d has $\mathbb{E}[x_i x_i']d = 0$, i.e. regressors were not added multiple times.
- Need to dominate the score with the penalty term λ.
- Need no collinearity on a small ($\leq s_0$) subset of regressors ($\to k_{c_0,s_0} > 0$).

When Lasso? For prediction problems in high dimensional environments. **NB!** Lasso is not good for inference, only for prediction.

In particular, in econometrics it’s used for selecting either

- instruments (predicting $\hat{x}$ in the first stage)
- control variables (next section: double prediction problem, in the first stage and in the reduced form)

8.1.1 Choosing the optimal lambda

The choice of λ determines the bias-variance tradeoff:

- if λ is too big: $\lambda \approx \infty \mathbb{R}ightarrow\hat{\beta} \approx 0$;
- if λ is too small: $\lambda \approx 0\mathbb{R}ightarrow$ overfitting.

Possible solutions: Bonferroni correction, bootstrapping or $\frac{\lambda}{n} \asymp \sqrt{\frac{\log(p)}{n}}$ (asymptotically equal to), $\mathcal{S}$ behaves like the maximum of gaussians.

The Lasso Path: how the estimated $\hat{\beta}$ depends on the penalty parameter λ?

Post Lasso: fit OLS without the penalty with all the nonzero coeficients selected by Lasso in the first step.

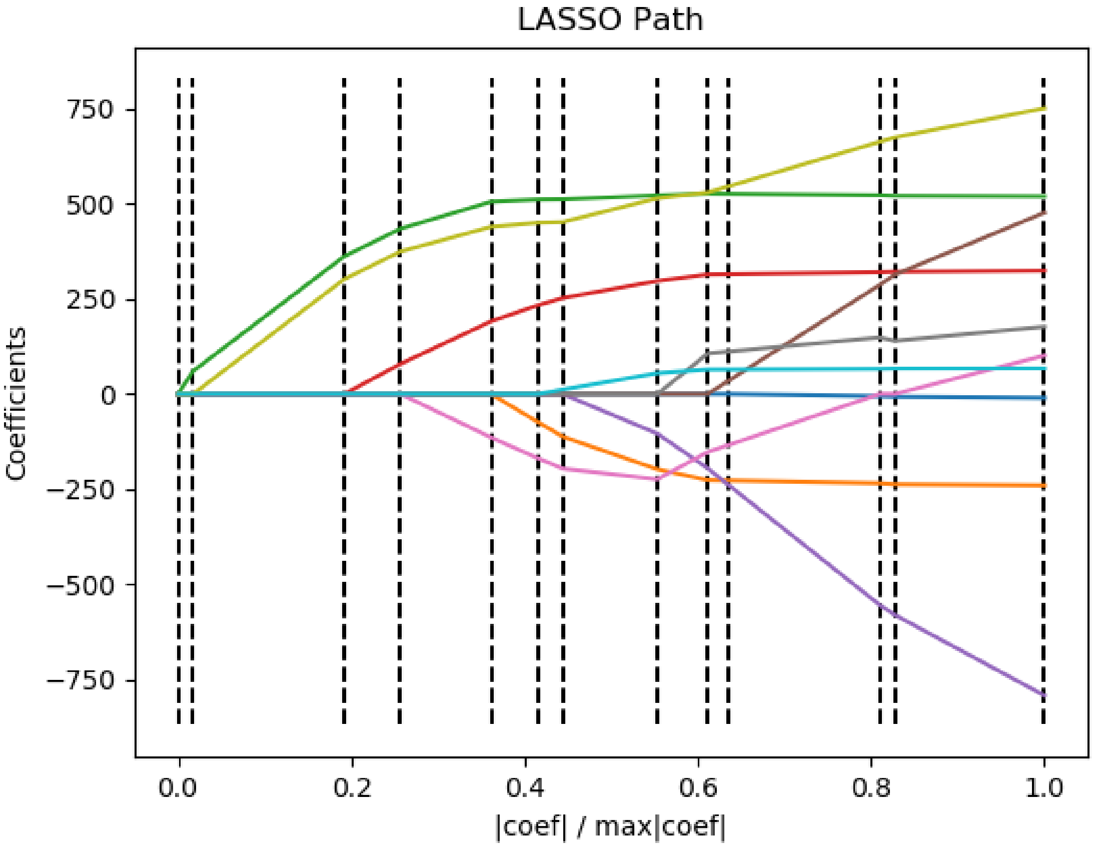


Figure 8.2: Lasso Path

On Lasso:

- Do not do inference with post-Lasso because standard errors are not uniformely valid.
- As $n \to \infty$ the CV and the **score domination** bounds converge to a unique bound.
- What is the problem of cross-validation? In high dimensional settings you can overfit in so many ways that CV doesn't work and still overfits.
- Using λ with $\frac{\lambda}{n} > \mathcal{S}$ small coefficients get shrunk to zero with high probability. In this case with small we mean $\propto \frac{1}{\sqrt{n}}$ or $2 \max_j |\mathbb{E}_n[\varepsilon_i x_{ij}]|$.
- If $|\beta_{0j}| \leq \frac{c}{\sqrt{n}}$ for a sufficiently small constant c, then $\hat{\beta}_{LASSO} \xrightarrow{p} 0$.
- In standard t-tests $c = 1.96$.
- $\sqrt{n}$ factor is important since it is the demarcation line for reliable statistical detection.

8.2 Pre-Testing

Main reference for this section: Belloni, Chernozhukov and Hansen *Inference for Treatment Effects with High Dimensional Controls* in the *Review of Economic Studies* (2014).

8.2.1 Omitted Variable Bias

Consider two separate statistical models. Assume the following **long regression** of interest:

$$y_i = x_i'\alpha_0 + z_i'\beta_0 + \varepsilon_i$$

Define the corresponding **short regression**

$$y_i = x_i'\alpha_0 + v_i \quad \text{ with } v_i = z_i'\beta_0 + \varepsilon_i$$

Theorem: Suppose that the DGP for the long regression corresponds to α_0, β_0. Suppose further that $\mathbb{E}[x_i] = 0$, $\mathbb{E}[z_i] = 0$, $\mathbb{E}[\varepsilon_i | x_i, z_i] = 0$. Then, unless $\beta_0 = 0$ or z_i is orthogonal to x_i, the (sole) stochastic regressor x_i is correlated with the error term in the short regression which implies that the OLS estimator of the short regression is inconsistent for α_0 due to the omitted variable bias. In particular, one can show that the plim of the OLS estimator of $\hat{\alpha}_{SHORT}$ from the short regression is

$$\hat{\alpha}_{SHORT} \xrightarrow{p} \frac{Cov(y_i, x_i)}{Var(x_i)} = \alpha_0 + \beta_0 \frac{Cov(z_i, x_i)}{Var(x_i)}$$

8.2.2 Pre-test bias

Consider data $D = (y_i, x_i, z_i)_{i=1}^n$, where the true model is:

$$\begin{aligned} y_i &= x_i'\alpha_0 + z_i'\beta_0 + \varepsilon_i \\ x_i &= z_i'\gamma_0 + u_i \end{aligned}$$

Where x_i is the variable of interest (we want to make inference on α_0) and z_i is a high dimensional set of control variables. \

From now on, we will work under the following assumptions:

- $\dim(x_i) = 1$ for all n
- β_0 uniformely bounded in n
- Strict exogeneity: $\mathbb{E}[\varepsilon_i|x_i, z_i] = 0$ and $\mathbb{E}[u_i|z_i] = 0$
- β_0 and γ_0 have dimension (and hence value) that depend on n

Pre Testing procedure:

1. Regress y_i on x_i and z_i
2. For each $j = 1, ..., p = \dim(z_i)$ calculate a test statistic t_j
3. Let $\hat{T} = \{j : |t_j| > C > 0\}$ for some constant C (set of statistically significant coefficients).
4. Re-run the new “model" using $(x_i, z_{\hat{T},i})$ (i.e. using the selected covariates).
5. Perform statistical inference (i.e. confidence intervals and hypothesis tests) as if no model selection had been done.

This procedure is unfortunately not desirable in most practical settings since it introduces **omitted variable bias**.

Concept of **uniformity**: the DGP varies with n. Instead of having a fixed “true” parameter β_0, you have a sequence $\beta_0(n)$. Then, we could plot a sequence of ordered pairs (n, DGP) as below. In order to study the behavior of $\hat{\beta}$ at your particular n, you study its behavior along the sequence.

> Remark: pre-test bias is problematic because the post-selection estimator is not asymptotically normal. Moreover, for particular data generating processes, it even fails to be consistent at the rate of $\sqrt{n}$ (belloni et al., 2014). An approximate depiction of the DGPs for which consistency fails is depicted below.

The intuition for the three different regions (from below to above) is the following.

1. When $\beta_0 = o(1/\sqrt{n})$, z_i is excluded with probability $p \to 1$. But, given that β_0 is small enough, failing to control for z_i does not introduce large omitted variables bias [BCH 14].
2. If however the coefficient on the control is “moderately close to zero" ($\beta_0 = O(1/\sqrt{n})$), the t-test set-up above still cannot distinguish this coefficient from 0, and the control z_i is dropped with probability $p \to 1$. However now the omitted variable bias created by dropping z_i scaled by $\sqrt{n}$, diverges to infinity. That is, the standard post-selection estimator is not asymptotically normal and even fails to be consistent at the rate of $\sqrt{n}$ [BCH 14].
3. Lastly, when β_0 is large enough, the null pre-testing hypothesis $H_0 : \beta_0 = 0$ will be rejected sufficiently often so that the bias is negligible.

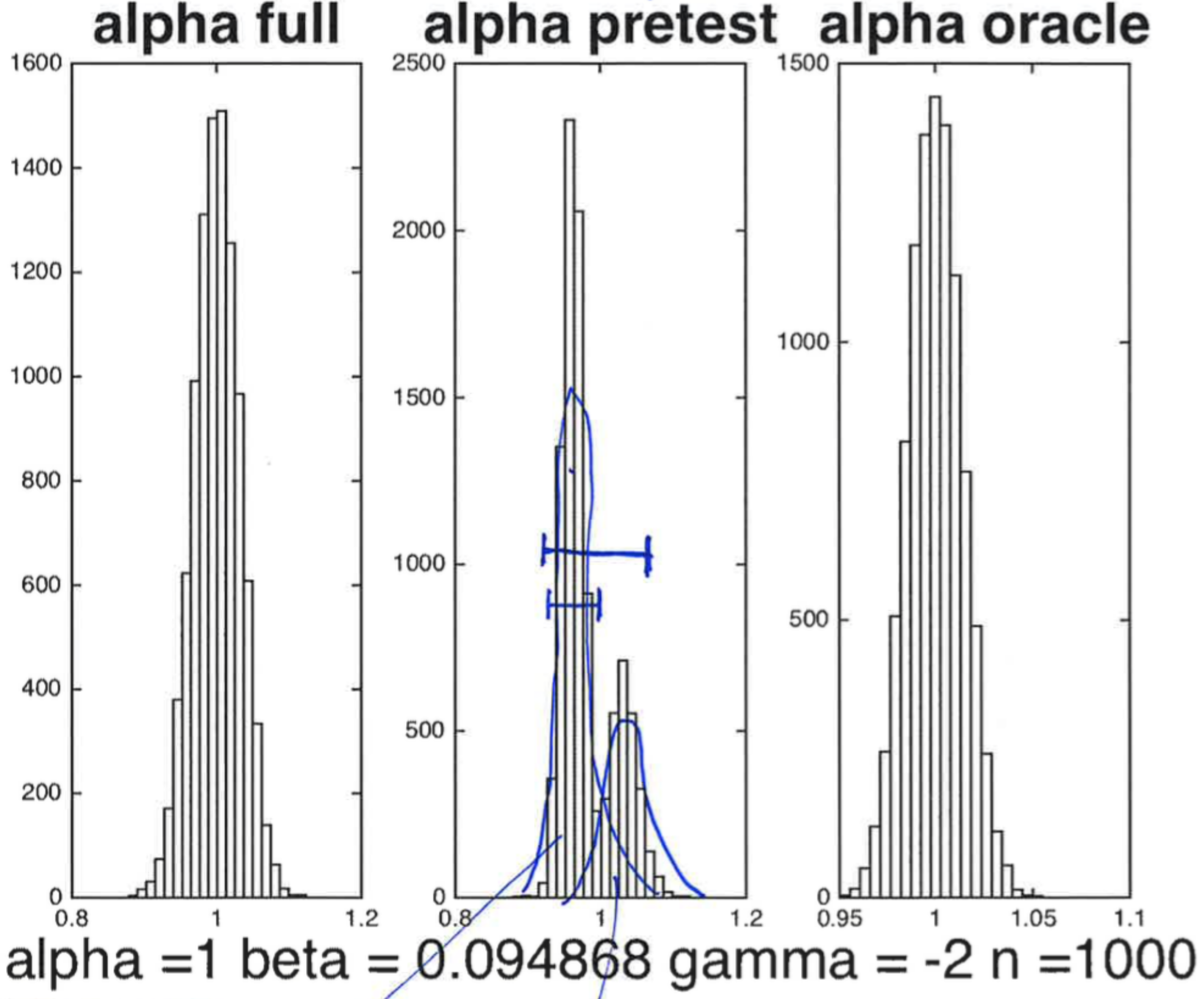
The post-double-selection estimator, $\hat{\alpha}_{PDS}$ solves this problem by doing variable selection via standard t-tests or Lasso-type selectors with the two “true model" equations (**first stage**

$\alpha_0 = 1$ (True Value)

$\underbrace{y = \alpha_0 x + \beta_0 z + \varepsilon} \to \hat{\alpha}$

$\underbrace{\text{pretest}} \to \hat{\alpha}$

reg$(y - \beta_0 z)$ on x $\to \hat{\alpha}$



times when rejected.

times when failed to reject

Figure 8.3: Pre-test Bias

$\alpha_0 = 1$ (True Value)

$\underbrace{y = \alpha_0 x + \beta_0 z + \varepsilon}_{\to \hat{\alpha}}$ $\underbrace{\text{pretest}}_{\to \hat{\alpha}}$ reg$(y - \beta_0 z)$ on x $\to \hat{\alpha}$

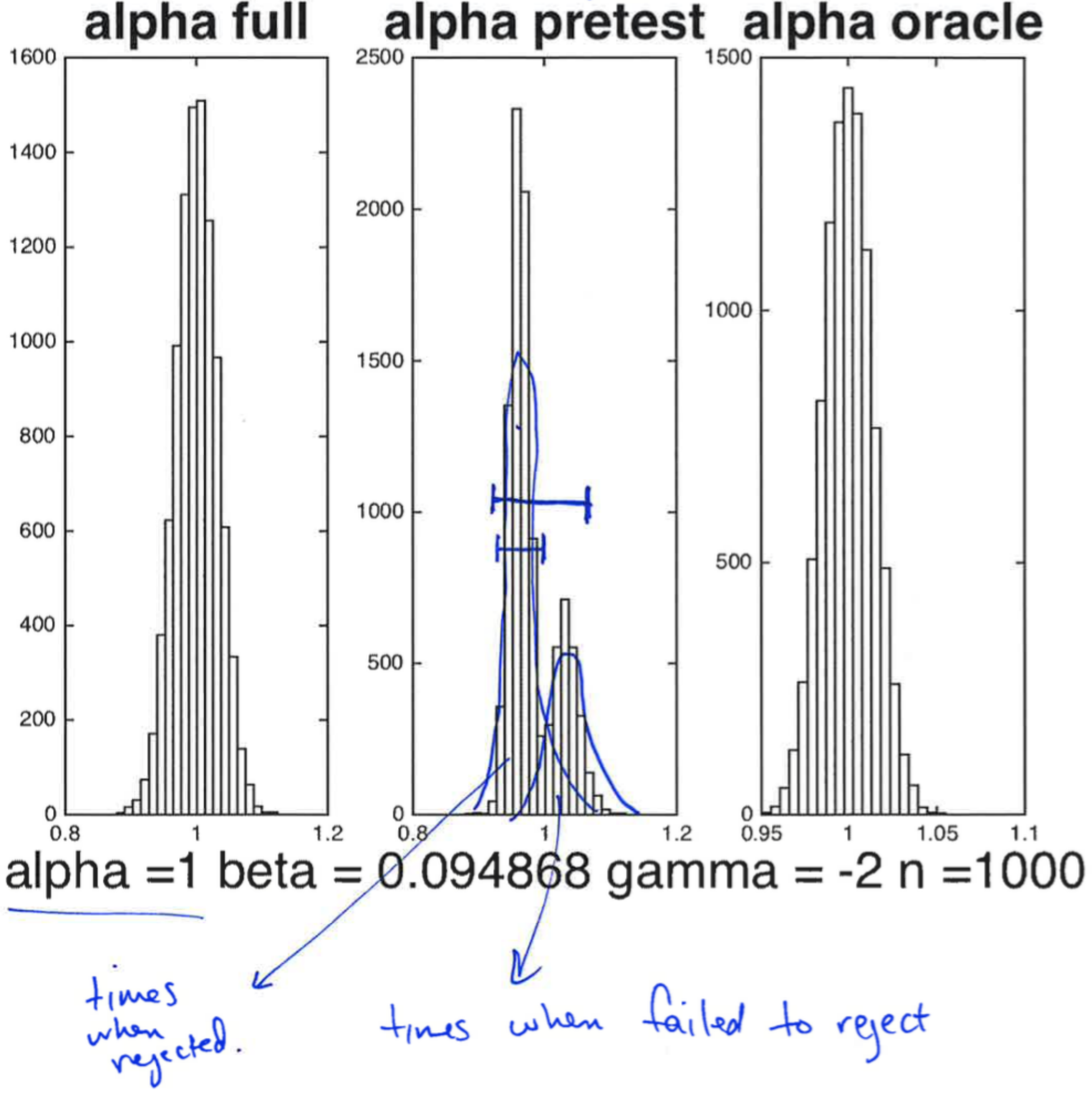


Figure 8.4: Pre-test Bias

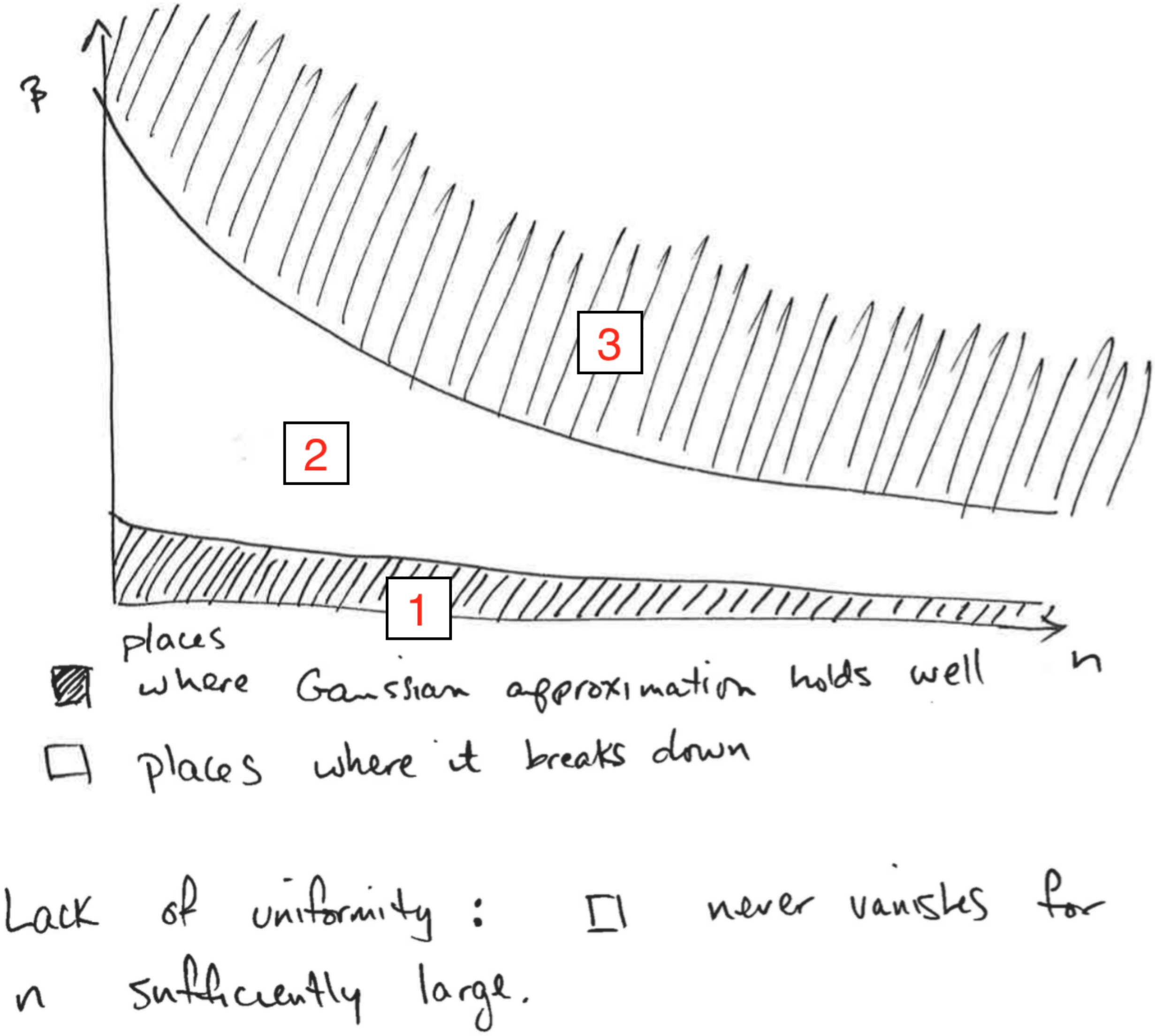

β
3
2
1
n
places
▨ where Gaussian approximation holds well
☐ places where it breaks down
Lack of uniformity: ☐ never vanishes for n sufficiently large.

Figure 8.5: Pre-test Bias

and **reduced form**) that contain the information from the model and then estimating α_0 by regressing y_i on x_i and the union of the selected controls. By doing so, z_i is omitted only if its coefficient in both equations is small which greatly limits the potential for omitted variables bias (Beloni et al., 2014).

8.2.3 Partioned Regression

Frisch-Waugh Theorem: Consider the data $D = \{x_i, y_i, z_i\}_{i=1}^{\infty}$ with DGP: $Y = X\alpha + Z\beta + \varepsilon$. The following estimators of α are numerically equivalent (if $[X, Z]$ has full rank):

- $\hat{\alpha}$ from regressing Y on X, Z
- $\tilde{\alpha}$ from regressing Y on $\tilde{X}$
- $\bar{\alpha}$ from regressing $\tilde{Y}$ on $\tilde{X}$

where the operation of passing to Y, X to $\tilde{Y}, \tilde{X}$ is called *projection out* Z, e.g.$\tilde{X}$ are the residuals from regressing X on Z.

Proof: We want to show that $\hat{\alpha} = \tilde{\alpha}$.

Claim: $\hat{\alpha} = \tilde{\alpha} \Leftrightarrow \tilde{X}' \left[(X - \tilde{X})\hat{\alpha} + Z\hat{\beta} + \hat{\varepsilon} \right] = 0$.

Proof of the claim: if $\hat{\alpha} = \tilde{\alpha}$, we can write Y as

$$Y = X\hat{\alpha} + Z\hat{\beta} + \hat{\varepsilon} = \tilde{X}\hat{\alpha} + \underbrace{(X - \tilde{X})\hat{\alpha} + Z\hat{\beta} + \hat{\varepsilon}}_{\text{residual of } Y \text{ on } \tilde{X}} = \tilde{X}\tilde{\alpha} + \nu_i$$

Therefore, by the orthogonality property of the OLS residual, it must be that $\tilde{X}'\nu_i = 0$.

■

Having established the claim, we want to show that the normal equation $\tilde{X}' \left[(X - \tilde{X})\hat{\alpha} + Z\hat{\beta} + \hat{\varepsilon} \right] = 0$ is satisfied. We follow 3 steps:

1. First we have that $\tilde{X}'(X - \tilde{X})\hat{\alpha} = 0$. This follows from the fact that $\tilde{X}' = X'M_Z$ and hence:
$$\begin{aligned} \tilde{X}'(X - \tilde{X}) &= X'M_Z(X - M_Z) = X'M_Z X - X' \overbrace{M_Z M_Z}^{M_Z} X \\ &= X'M_Z X - X'M_Z X = 0 \end{aligned}$$
2. $\tilde{X}'Z\hat{\beta} = 0$ since $\tilde{X}$ is the residual from the regression of X on Z, by normal equation it holds that $\tilde{X}'Z = 0$.
3. $\tilde{X}'\hat{\varepsilon} = 0$. This follows from (i) $M_Z' M_{X,Z} = M_{X,Z}$ and (ii) $X'M_{X,Z} = 0$:
$$\tilde{X}'\hat{\varepsilon} = (M_Z X)'(M_{X,Z}\varepsilon) = X'M_Z' M_{X,Z}\varepsilon = \underbrace{X'M_{X,Z}}_{0}\varepsilon = 0.$$

■

The coefficient $\hat{\alpha}$ is a *partial regression* coefficient identified from the variation in X that is orthogonal to Z. This is often known as **residual variation**.

8.3 Post Double Selection

We would like to guard against pretest bias if possible, in order to handle high dimensional models. A good pathway towards motivating procedures which guard against pretest bias is a discussion of classical partitioned regression.

Consider a regression y_i on x_i and z_i. x_i is the 1-dimensional variable of interest, z_i is a high-dimensional set of control variables. We have the following procedure:

1. **First Stage** selection: lasso x_i on z_i. Let the selected variables be collected in the set $S_{FS} \subseteq z_i$
2. **Reduced Form** selection: lasso y_i on z_i. Let the selected variables be collected in the set $S_{RF} \subseteq z_i$
3. Regress y_i on x_i and $S_{FS} \cup S_{RF}$

Theorem: Let $\{P^n\}$ be a sequence of data-generating processes for $D_n = (y_i, x_i, z_i)_{i=1}^n \in (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^p)^n$ where p depends on n. For each n, the data are iid with $yi = x_i'\alpha_0^{(n)} + z_i'\beta_0^{(n)} + \varepsilon_i$ and $x_i = z_i'\gamma_0^{(n)} + u_i$ where $\mathbb{E}[\varepsilon_i|x_i, z_i] = 0$ and $\mathbb{E}[u_i|z_i] = 0$. The sparsity of the vectors $\beta_0^{(n)}$, $\gamma_0^{(n)}$ is controlled by $||\beta_0^{(n)}||_0 \leq s$ with $s^2(\log p)^2/n \to 0$. Suppose that additional regularity conditions on the model selection procedures and moments of the random variables y_i , x_i , z_i as documented in Belloni et al. (2014). Then the confidence intervals, CI, from the post double selection procedure are uniformly valid. That is, for any confidence level $\xi \in (0, 1)$

$$\Pr(\alpha_0 \in CI) \to 1 - \xi$$

In order to have valid confidence intervals you want their bias to be negligibly. Since

$$CI = \left[\hat{\beta} \pm \frac{1.96 \cdot \hat{\sigma}}{\sqrt{n}}\right]$$

If the bias is $o(n^{-1/2})$ then there is no problem since it is asymptotically negligible w.r.t. the magnitude of the confidence interval. If however the the bias is $O(n^{-1/2})$ then it has the same magnitude of the confidence interval and it does not asymptotically vanish.

The idea of the proof is to use partitioned regression. An alternative way to think about the argument is: bound the omitted variables bias. Omitted variable bias comes from the product of 2 quantities related to the omitted variable:

1. Its partial correlation with the outcome, and
2. Its partial correlation with the variable of interest.

If both those partial correlations are $O(\sqrt{\log p/n})$, then the omitted variables bias is $(s \times O(\sqrt{\log p/n})^2 = o(n^{-1/2})$, provided $s^2(\log p)^2/n \to 0$. Relative to the $n^{-1/2}$ convergence rate, the omitted variables bias is negligible.

In our omitted variable bias case, we want $|\gamma_0\delta_0| = o\left(\frac{1}{\sqrt{n}}\right)$. Post-double selection guarantees that

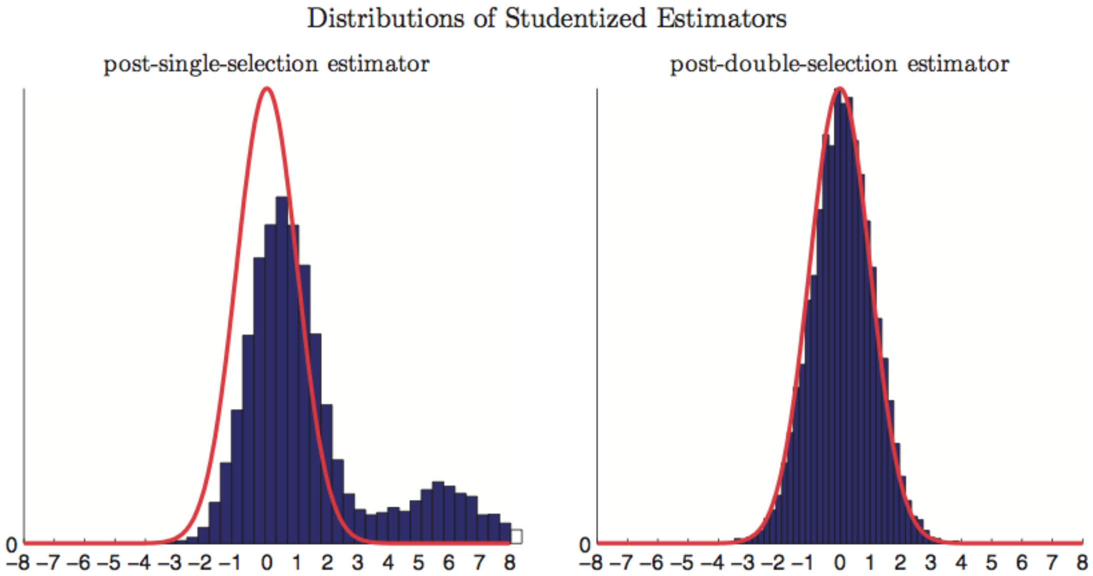


Figure 8.6: Pre-test Bias

- *First stage* selection: any "missing" variable has $|\gamma_{0j}| \leq \frac{c}{\sqrt{n}}$
- *Reduced form* selection: any "missing" variable has $|\delta_{0j}| \leq \frac{c}{\sqrt{n}}$

As a consequence, as long as the number of omitted variables is finite, the omitted variable bias is

$$OVB = |\gamma_{0j}| \cdot |\delta_{0j}| \leq \frac{c}{\sqrt{n}} \cdot \frac{c}{\sqrt{n}} = \frac{c^2}{n} = o\left(\frac{1}{\sqrt{n}}\right)$$

Under homoskedasticity, the above estimator achieves the semiparametric efficiency bound.

Example: $y_i = \alpha_0 x_i + \delta_0 z_{i1} + \gamma_0 z_{i2} + \varepsilon_i$

- $|\delta_0| = \frac{1}{2}$ $|\gamma_0| = \frac{1}{2}$ then the bias is $= \frac{1}{4}$
- $|\delta_0| = \frac{1}{2}$ $|\gamma_0| = \frac{1}{\sqrt{n}}$ then the bias is $= \frac{1}{2\sqrt{n}}$
- $|\delta_0| = \frac{1}{\sqrt{n}}$ $|\gamma_0| = \frac{1}{2}$ then the bias is $= \frac{1}{2\sqrt{n}}$
- $|\delta_0| = \frac{1}{\sqrt{n}}$ $|\gamma_0| = \frac{1}{\sqrt{n}}$ then the bias is $= \frac{1}{n}$

Only in the last case the bias is $o\left(\frac{1}{\sqrt{n}}\right)$.

What is the criterium that should guide the selection of λ?

$$\frac{\lambda}{n} \geq 2\mathbb{E}_n[x_{ij}\varepsilon_i] \qquad \forall j \qquad \text{if } Var(x_{ij}\varepsilon_i) = 1$$

How to choose the optimal λ:

- Decide the coverage of the confidence intervals $(1-\alpha)$:

$$\Pr\left(\sqrt{n}\Big|\mathbb{E}_n[x_{ij}\varepsilon_i]\Big| > t\right) = 1 - \alpha$$

- Solve for t
- Get λ such that all scores are dominated by $\frac{\lambda}{n}$ with $\alpha\%$ probability.

  It turns out that the optimal $t \propto \sqrt{\log(p)}$

8.4 References

- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). *Inference on Treatment Effects after Selection among High-Dimensional Controls.* The Review of Economic Studies, 81(2), 608–650.
- Hastie, Tibshirani, Friedman (2001). "*The Elements of Statistical Learning*".
- Hansen (2019). "*Econometrics*". Chapter 24.
- Kozbur (2019). PhD Econometrics - Lecture Notes.

Chapter 9

Matlab Code

In this appendix I collect all the matlab code from the lecture notes.

9.1 Lecture 1

```
% Set seed
rng(123)

% Set the number of observations
n = 100;

% Set the dimension of X
k = 2;

% Draw a sample of explanatory variables
X = rand(n, k);

% Draw the error term
sigma = 1;
e = randn(n,1)*sqrt(sigma);

% Set the parameters
b = [2; -1];

% Calculate the dependent variable
y = X*b + e;

% Estimate beta
b_hat = inv(X'*X)*(X'*y) % = 1.9020, -0.9305
```

```
% Equivalent but faster formulation
b_hat = (X'*X)\(X'*y);

% Even faster (but less intuitive) formulation
b_hat = X\y;

% Note that is generally not equivalent to Var(X)^-1 * Cov(X,y)...
Var_X = cov(X);
Cov_Xy = n/(n-1) * (mean(X .* y) - mean(X).*mean(y));
b_alternative = inv(Var_X) * Cov_Xy' % = 2.1525, -0.7384

% ...unless you include a constant
a = 3;
y = a + X*b + e;
b_hat_1 = [ones(n,1), X]\y % = 2.1525, -0.7384
Var_X = cov(X);
Cov_Xy = n/(n-1) * (mean(X .* y) - mean(X).*mean(y));
b_alternative = inv(Var_X) * Cov_Xy' % = 2.1525, -0.7384

% Predicted y
y_hat = X*b_hat;

% Residuals
e_hat = y - X*b_hat;

% Projection matrix
P = X*inv(X'*X)*X';

% Annihilator matrix
M = eye(n) - P;

% Leverage
h = diag(P);

% Biased variance estimator
sigma_hat = e_hat'*e_hat / n;

% Unbiased estimator 1
sigma_hat_2 = e_hat'*e_hat / (n-k);

% Unbiased estimator 2
sigma_hat_3 = mean( e_hat.^2 ./ (1-h) );

% R squared - uncentered
```

```
R2_uc = (y_hat'*y_hat)/ (y'*y);

% R squared
y_bar = mean(y);
R2 = ((y_hat-y_bar)'*(y_hat-y_bar))/ ((y-y_bar)'*(y-y_bar));

% Ideal variance of the OLS estimator
var_b = sigma*inv(X'*X);

% Standard errors
std_b = sqrt(diag(var_b));
```

9.2 Lecture 2

```
% Set seed
rng(123)

% Homoskedastic standard errors
std_h = var(e_hat) * inv(X'*X);

% HC0 variance and standard errors
omega_hc0 = X' * diag(e_hat.^2) * X;
std_hc0 = sqrt(diag(inv(X'*X) * omega_hc0 * inv(X'*X))) % = 0.9195, 0.8631

% HC1 variance and standard errors
omega_hc1 = n/(n-k) * X' *  diag(e_hat.^2) * X;
std_hc1 = sqrt(diag(inv(X'*X) * omega_hc1 * inv(X'*X))) % = 0.9289, 0.8719

% HC2 variance and standard errors
omega_hc2 = X' * diag(e_hat.^2./(1-h)) * X;
std_hc2 = sqrt(diag(inv(X'*X) * omega_hc2 * inv(X'*X))) % = 0.9348, 0.8768

% HC3 variance and standard errors
omega_hc3 = X' * diag(e_hat.^2./(1-h).^2) * X;
std_hc3 = sqrt(diag(inv(X'*X) * omega_hc3 * inv(X'*X))) % = 0.9504, 0.8907

% Note what happens if you allow for full autocorrelation
omega_full = X'*e_hat*e_hat'*X;

% t-test for beta=0
t = abs(b_hat./(std_hc1));

% p-value
```

```
p_val = 1 - normcdf(t);

% F statistic of joint significance
SSR_u = e_hat'*e_hat;
SSR_r = y'*y;
F = (SSR_r - SSR_u)/k / (SSR_u/(n-k));

% 95% confidente intervals
conf_int = [b_hat - 1.96*std_hc1, b_hat + 1.96*std_hc1];
```

9.3 Lecture 3

```
% Set seed
rng(123)

% Set the dimension of Z
l = 3;

% Draw instruments
Z = randn(n,l);

% Correlation matrix for error terms
S = eye(2,2); S(1,2)=.8; S(2,1)=.8;

% Endogenous X
gamma = [2, 0; 0, -1; -1, 3];
e = randn(n,2)*chol(S);
X = Z*gamma + e(:,1);

% Calculate Y
Y = X*b + e(:,2);

% Estimate beta OLS
beta_OLS = (X'*X)\(X'*Y) % = 2.1957, -0.9022

% IV: l=k=2 instruments
Z_IV = Z(:,1:k);
beta_IV = (Z_IV'*X)\(Z_IV'*Y) % = 2.1207, -1.3617

% Calculate standard errors
ehat = Y - X*beta_IV;
V_NHC_IV = var(ehat) * inv(Z_IV'*X)*Z_IV'*Z_IV*inv(Z_IV'*X);
V_HC0_IV = inv(Z_IV'*X)*Z_IV' * diag(ehat.^2) * Z_IV*inv(Z_IV'*X);
```

```
% 2SLS: l=3 instruments
Pz = Z*inv(Z'*Z)*Z';
beta_2SLS = (X'*Pz*X)\(X'*Pz*Y) % = 2.0723, -0.9628

% Calculate standard errors
ehat = Y - X*beta_2SLS;
V_NCH_2SLS = var(ehat) * inv(X'*Pz*X);
V_HC0_2SLS = inv(X'*Pz*X)*X'*Pz * diag(ehat.^2) *Pz*X*inv(X'*Pz*X);

% GMM 1-step: inefficient weighting matrix
W_1 = eye(l);

% Objective function
gmm_1 = @(b) ( Y - X*b )' * Z * W_1 *  Z' * ( Y - X*b );

% Estimate GMM
beta_gmm_1 = fminsearch(gmm_1, beta_OLS) % = 2.0763, -0.9548
ehat = Y - X*beta_gmm_1;

% Standard errors GMM
S_hat = Z'*diag(ehat.^2)*Z;
d_hat = -X'*Z;
V_gmm_1 = inv(d_hat * inv(S_hat) * d_hat');

% GMM 2-step: efficient weighting matrix
W_2 = inv(S_hat);
gmm_2 = @(b) ( Y - X*b )' * Z * W_2 *  Z' * ( Y - X*b );
beta_gmm_2 = fminsearch(gmm_2, beta_OLS) % = 2.0595, -0.9666

% Standard errors GMM
ehat = Y - X*beta_gmm_2;
S_hat = Z'*diag(ehat.^2)*Z;
d_hat = -X'*Z;
V_gmm_2 = inv(d_hat * inv(S_hat) * d_hat');
```