Luke Stein
Stanford graduate economics core (2006–7)
Reference and review (last updated April 11, 2012)

Contents

---

1 Econometrics

1.1 Probability foundations

Basic set theory (C&B 1.1.4) All sets A, B, C satisfy:

1. Commutativity: $A \cup B = B \cup A$ and $A \cap B = B \cap A$;

2. Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$;

3. Distributive laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;

4. DeMorgan's Laws: $(A \cup B)^{\mathsf{C}} = A^{\mathsf{C}} \cap B^{\mathsf{C}}$ and $(A \cap B)^{\mathsf{C}} = A^{\mathsf{C}} \cup B^{\mathsf{C}}$.

Disjointness (C&B 1.1.5) Two events are disjoint (a.k.a. mutually exclusive) iff $A \cap B = \varnothing$. Events $\{A_i\}$ are pairwise disjoint or mutually exclusive iff $A_i \cap A_j = \varnothing$ for all $i \neq j$. Two events with nonzero probability cannot be both mutually exclusive and independent (see ex. 1.39).

Partition (C&B 1.1.6) $A_1, A_2, \ldots$ is a partition of S iff:

1. $S = \bigcup_i A_i$ (i.e., covering);

2. $A_1, A_2, \ldots$ are pairwise disjoint (i.e., non-overlapping).

Sigma algebra (C&B 1.2.1) Collection of subsets of S (i.e., a subset of the power set of S) is a sigma algebra, denoted $\mathcal{B}$ iff:

1. $\varnothing \in \mathcal{B}$;

2. If $A \in \mathcal{B}$, then $A^{\mathsf{C}} \in \mathcal{B}$ (closed under complementation—along with first axiom gives $S \in \mathcal{B}$);

3. If $\{A_i\} \subseteq \mathcal{B}$, then $\bigcup_i A_i \in \mathcal{B}$ (closed under countable unions).

A cdf completely determines the probability distribution of a random variable if its probability function is defined only for events in the Borel field $\mathcal{B}^1$, the smallest sigma algebra containing all the intervals of real numbers of the form (a, b), $[a, b)$, $(a, b]$, $[a, b]$. If probabilities are defined for a larger class of events, two random variables *may* have the same cdf but not the same probability for every event (see C&B p. 33).

Probability function (C&B 1.2.4, 8–9, 11) Given a sample space S and associated sigma algebra $\mathcal{B}$, a function $P \colon \mathcal{B} \to \mathbb{R}$ is a probability function iff it satisfies the Kolmogorov Axioms or Axioms of Probability:

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$;

2. $P(S) = 1$;

3. If $\{A_i\} \subseteq \mathcal{B}$ are pairwise disjoint, then $P(\bigcup_i A_i) = \sum_i P(A_i)$ (countable additivity for pairwise disjoint sets).

For any probability function P and $A, B \in \mathcal{B}$,

1. $P(\varnothing) = 0$;

2. $P(A) \leq 1$;

3. $P(A^{\mathsf{C}}) = 1 - P(A)$;

4. $P(B \cap A^{\mathsf{C}}) = P(B) - P(A \cap B)$ (B but not A is B minus both A and B);

5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;

6. If $A \subseteq B$, then $P(A) \leq P(B)$.

If $\{C_i\}$ partitions A, then $P(A) = \sum_i P(A \cap C_i)$.

Probability space (Hansen 1-31, 1-28) $(\Omega, \mathcal{F}, P)$ where:

1. Ω is the universe (e.g., S, the sample space);

2. $\mathcal{F}$ is the σ-field (e.g., $\mathcal{B}^1$);

3. P a probability measure (e.g., $\mathcal{P}$, the probability measure that governs all random variables).

A random variable X induces a probability measure P_X defined by $P_X(B) \equiv P(X \in B) = P(F)$. This gives the probability space $(\mathbb{R}, \mathcal{B}, P_X)$.

Counting (C&B sec. 1.2.3) The number of possible arrangement of size r from n objects is

	No replacement	With replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

where $\binom{n}{r} \equiv \frac{n!}{r!(n-r)!}$. (Unordered with replacement, a.k.a. "stars and bars.")

Conditional probability (C&B 1.3.2, ex. 1.38) For $A, B \in S$ with $P(B) > 0$, the conditional probability of A given B is $P(A|B) \equiv P(A \cap B)/P(B)$.

1. If A and B are disjoint ($A \cap B = \varnothing$), then $P(A|B) = P(B|A) = 0$.

2. If $P(B) = 1$, then $\forall A, P(A|B) = P(A)$.

3. If $A \subseteq B$, then $P(B|A) = 1$ and $P(A|B) = P(A)/P(B)$.

4. If A and B are mutually exclusive, $P(A|A \cup B) = P(A)/[P(A) + P(B)]$.

5. $P(A \cap B \cap C) = P(A|B \cap C) \cdot P(B|C) \cdot P(C)$.

Bayes' Rule (C&B 1.3.5) A formula for "turning around" conditional probabilities: $P(A|B) = P(B|A) \cdot P(A)/P(B)$. More generally, if $\{A_i\}$ partition the sample space and B is any set, then $\forall i$,

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_j P(B|A_j) \cdot P(A_j)}.$$

Independence of events (C&B 1.3.7, 9, 4.2.10) A, B statistically independent iff $P(A \cap B) = P(A) \cdot P(B)$ or identically iff $P(A|B) = P(A)$ (this happens iff $P(B|A) = P(B)$).

1. Iff A and B are independent, then the following pairs are also independent: A and B^{C}, A^{C} and B, A^{C} and B^{C}.

2. Two events with nonzero probability cannot be both mutually exclusive and independent (see ex. 1.39).

3. If X, Y independent r.v.s, then for any A, $B \subseteq \mathbb{R}$, events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.

Mutual independence of events (C&B 1.3.12) A collection of events $\{A_i\}$ are mutually independent iff for any subcollection $A_{i_1}, \ldots, A_{i_k}$ we have $P(\bigcap_j A_{i_j}) = \prod_j P(A_{i_j})$. Note that pairwise independence does *not* imply mutual independence.

4

1.2 Random variables

Random variable (1.4.1, 1.5.7–8, 10, sec. 3.2) A function $X\colon S \to \mathbb{R}$ where S is a sample space.

1. Continuous iff its cdf is a continuous function and discrete iff its cdf is a step function (i.e., if sample space is countable).

2. Identically distributed iff $\forall A \in \mathcal{B}^1$, $P(X \in A) = P(Y \in A)$, or identically iff $\forall x$, $F_X(x) = F_Y(x)$.

3. Note identical distribution says nothing about (in)dependence.

Random vector (C&B 4.1.1) n-dimensional random vector is a function $\mathbf{X}\colon S \to \mathbb{R}^n$ where S is a sample space.

Measurability (Hansen 1-28, 4-11) A r.v. $X\colon (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ is $\mathcal{F}$-measurable iff $\forall B \in \mathcal{B}$, $\{\omega \in \Omega\colon X(\omega) \in B\} \in \mathcal{F}$ (i.e., the preimage of every element of $\mathcal{B}$ is an element of $\mathcal{F}$).

If $\mathcal{F}$ and $\mathcal{G}$ are both σ-fields with $\mathcal{G} \subseteq \mathcal{F}$, then X is $\mathcal{G}$-measurable $\implies$ X is $\mathcal{F}$-measurable (i.e., if the preimage of every B is in $\mathcal{G}$, it is also in its superset $\mathcal{F}$).

Smallest σ-field (Hansen 4-12) The smallest σ-field that makes a r.v. $Z\colon (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ measurable is $\sigma(Z) \equiv \{G \subseteq \Omega\colon \exists B \in \mathcal{B}, G = Z^{-1}(B)\}$ (i.e., the set of preimages of elements of $\mathcal{B}$).

Independence of r.v.s (C&B 4.2.5, 7, p. 154, 4.3.5) X and Y independent r.v.s (written $X \perp\!\!\!\perp Y$) iff any of the following equivalent conditions hold:

1. $\forall A, B$, $P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$.

2. $F_{XY}(x, y) \equiv P(X \leq x, Y \leq y) = F_X(x) F_Y(y)$.

3. $f(x, y) = f_X(x) f_Y(y)$ (i.e., joint pdf/pmf is the product of marginal pdfs/pmfs).

4. $f(y|x) = f_Y(y)$ (i.e., conditional pdf/pmf equals marginal pdf/pmf).

5. $\exists g(x), h(y), \forall x, y, f(x, y) = g(x) h(y)$ (i.e., joint pdf/pmf is separable). Note that functional forms may appear separable, but limits may still depend on the other variable; if the support set of (X, Y) is not a cross product, then X and Y are *not* independent.

For any functions $g(t)$ and $h(t)$, $X \perp\!\!\!\perp Y \implies g(X) \perp\!\!\!\perp h(Y)$.

Independence of random vectors (C&B 4.6.5) $\mathbf{X}_1, \ldots, \mathbf{X}_n$ mutually independent iff for every $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the joint pdf/pmf is the product of the marginal pdfs/pmfs; i.e., $f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_i f_{\mathbf{X}_i}(\mathbf{x}_i)$.

1. Knowledge about the values of some coordinates gives us no information about the values of the other coordinates.

2. The conditional distribution of any subset of the coordinates, given the values of the rest of the coordinates, is the same as the marginal distribution of the subset.

3. Mutual independence implies pairwise independence, but pairwise independence does *not* imply mutual independence.

Mean independence (Metrics P.S. 3-4c, Metrics section) X is mean independent of Y (written $X \perp\!\!\!\perp_{\mathrm{m}} Y$) iff $\mathrm{E}(X|Y) = \mathrm{E}(X)$.

1. Mean independence is not transitive (i.e., $X \perp\!\!\!\perp_{\mathrm{m}} Y$ does *not* imply that $Y \perp\!\!\!\perp_{\mathrm{m}} X$).

2. Independence implies mean independence (i.e., $X \perp\!\!\!\perp Y \implies X \perp\!\!\!\perp_{\mathrm{m}} Y \wedge Y \perp\!\!\!\perp_{\mathrm{m}} X$).

3. $X \perp\!\!\!\perp_{\mathrm{m}} Y \implies \mathrm{E}[(X|g(Y)] = \mathrm{E}[X]$, for any function $g(\cdot)$.

4. $X \perp\!\!\!\perp_{\mathrm{m}} Y \implies \mathrm{Cov}(X, g(Y)) = 0$ for any function $g(\cdot)$.

Cumulative distribution function (C&B 1.5.1, 3, p. 147) $F_X(x) \equiv P(X \leq x)$. By the Fundamental Theorem of Calculus, $\frac{d}{dx} F_X(x) = f_X(x)$ for a continuous r.v. at continuity points of f_X. A function F is a cdf iff:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$;

2. $F(\cdot)$ nondecreasing;

3. $F(\cdot)$ right-continuous; i.e., $\forall x_0$, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

A random vector X has joint cdf $F_X(x_1, \ldots, x_n) \equiv P(X_1 \leq x_1, \ldots, X_n \leq x_n)$. By the Fundamental Theorem of Calculus, $\frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_X(\vec{x}) = f_X(\vec{x})$ for a continuous (in all dimensions) random vector at continuity points of f_X.

Probability mass function (C&B 1.6.1, 5, 4.1.3) For a discrete r.v., $f_X(x) \equiv P(X = x)$. A function f_X is a pmf iff:

1. $\forall x$, $f_X(x) \geq 0$;

2. $\sum_x f_X(x) = 1$.

f_X gives the probability of any event: $P(X \in B) = \sum_k 1_{(x_k \in B)} f_X(x_k)$.

A discrete random vector X has joint pmf $f_X(\vec{v}) \equiv P(X = \vec{v})$.

Marginal pmf (C&B 4.1.6, p. 178) For a discrete random vector,

$$f_{X_i}(x_i) \equiv P(X_i = x_i) = \sum_{x_{-i} \in \mathbb{R}^{n-1}} f_X(x);$$

i.e., hold $X_i = x_i$, and sum f_X over all remaining possible values of X.

We can also take the marginal pmf for multiple i by holding these and summing f_X over all remaining possible values of X.

Conditional pmf (C&B 4.2.1) For (X, Y) a discrete random vector, $f(y|x) \equiv P(Y = y|X = x) = f(x, y)/f_X(x)$, where $f(x, y)$ is joint pmf, and $f_X(x)$ is marginal pmf.

Probability density function (C&B 1.6.3, 5, 4.1.10) For a continuous r.v., $f_X(x)$ defined as the function which satisfies $F_X(x) = \int_{-\infty}^x f_X(t)\, dt$ for all x. A function f_X is a pdf iff:

1. $\forall x$, $f_X(x) \geq 0$;

2. $\int_{\mathbb{R}} f_X(x)\, dx = 1$.

f_X gives the probability of any event: $P(X \in B) = \int_{\mathbb{R}} 1_{(x \in B)} f_X(x)\, dx$.

A continuous (in all dimensions) random vector X has joint pdf $f_X(x_1, \ldots, x_n)$ iff $\forall A \subseteq \mathbb{R}^n$, $P(X \in A) = \int \cdots \int_A f_X(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$.

Marginal pdf (C&B p. 145, 178) For a continuous (in all dimensions) random vector,

$$f_{X_i}(x_i) \equiv \int \cdots \int_{\mathbb{R}^{n-1}} f_X(x)\, dx_1 \cdots dx_{i-1}\, dx_{i+1} \cdots dx_n,$$

i.e., hold $X_i = x_i$, and integrate f_X over $\mathbb{R}$ in all X_j for $i \neq j$.

We can also take the marginal pdf for multiple i by holding these and integrating f_X over $\mathbb{R}$ in all X_j that aren't being held.

Conditional pdf (C&B 4.2.3, p. 178) For (X, Y) a continuous random vector, $f(y|x) \equiv f(x, y)/f_X(x)$ as long as $f_X(x) \neq 0$, where $f(x, y)$ is joint pdf, and $f_X(x)$ is marginal pdf.

We can also condition for/on multiple coordinates: e.g., for (X_1, X_2, X_3, X_4) a continuous random vector, $f(x_3, x_4|x_1, x_2) \equiv f(x_1, x_2, x_3, x_4)/f_{X_1 X_2}(x_1, x_2)$, where f is a joint pdf, and $f_{X_1 X_2}$ is the marginal pdf in X_1 and X_2.

Borel Paradox (4.9.3) Be careful when we condition on events of probability zero: two events of probability zero may be equivalent, but the probabilities conditional on the two events is different!

Stochastic ordering (C&B ex. 1.49, ex. 3.41-2) cdf F_X stochastically greater than cdf F_Y iff $F_X(t) \leq F_Y(t)$ at all t, with strict inequality at some t. This implies $P(X > t) \geq P(Y > t)$ at all t, with strict inequality at some t.

A family of cdfs $\{F(x|\theta)\}$ is stochastically increasing in θ iff $\theta_1 > \theta_2 \implies F(x|\theta_1)$ stochastically greater than $F(x|\theta_2)$. A location family is stochastically increasing in its location parameter; if a scale family has sample space $[0, \infty)$, it is stochastically increasing in its scale parameter.

Support set (C&B eq. 2.1.7) Support set (a.k.a. support) of a r.v. X is $\mathcal{X} \equiv \{x\colon f_X(x) > 0\}$, where f_X a cdf or pdf (or in general, any nonnegative function).

1.3 Transformations of random variables

Transformation $\mathbb{R}^1 \to \mathbb{R}^1$ (C&B 2.1.3, 5, 8) A discrete r.v. can be transformed into a discrete r.v. A continuous r.v. can be transformed into either a continuous or a discrete r.v. (or mixed!). When $Y = g(X)$ and $\mathcal{Y} \equiv g(\mathcal{X})$ (where $\mathcal{X}$ is the support of X),

1. If g monotone increasing on $\mathcal{X}$, then $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$;

2. If g monotone decreasing on $\mathcal{X}$ *and X a continuous r.v.*, then $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.

If g monotone, f_X continuous on $\mathcal{X}$, and g^{-1} has continuous derivative on $\mathcal{Y}$, then:

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|, & y \in \mathcal{Y}; \\ 0, & \text{otherwise.} \end{cases}$$

If $\{A_i\}_{i=0}^{k}$ partitions $\mathcal{X}$, with $P(X \in A_0) = 0$; f_X continuous on each A_i; and $\exists \{A_i\}_{i=1}^{k}$ satisfying:

1. $g(x) = g_i(x)$ for $x \in A_i$,

2. g_i monotone on A_i,

3. $\exists \mathcal{Y}, \forall i, g_i(A_i) = \mathcal{Y}$ (i.e., all A_i have same image under their respective g_is) [Hansen note 2-15 suggests this need not hold],

4. $\forall i, g_i^{-1}$ has a continuous derivative on $\mathcal{Y}$; then:

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X\left(g_i^{-1}(y)\right)\left|\frac{d}{dy}g_i^{-1}(y)\right|, & y \in \mathcal{Y}; \\ 0, & \text{otherwise.} \end{cases}$$

Transformation $\mathbb{R}^2 \to \mathbb{R}^2$ (C&B p. 158, 185) Let $U = g_1(X,Y)$ and $V = g_2(X,Y)$ where:

1. (X,Y) has pdf f_{XY} and support $\mathcal{A}$;

2. g_1 and g_2 define a 1-to-1 transformation from $\mathcal{A}$ to $\mathcal{B} \equiv \{(u,v): u \in g_1(\mathcal{A}), v \in g_2(\mathcal{A})\}$ (i.e., the support of (U,V));

3. Inverse transform is $X = h_1(U,V)$ and $Y = h_2(U,V)$; then:

$$f_{UV}(u,v) = \begin{cases} f_{XY}(h_1(u,v), h_2(u,v))\,|J|, & (u,v) \in \mathcal{B}; \\ 0, & \text{otherwise;} \end{cases}$$

where J is the Jacobian,

$$J \equiv \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}.$$

If the transformation is not 1-to-1, we can partition $\mathcal{A}$ into $\{A_i\}$ such that 1-to-1 transformations exist from each $\mathcal{A}_i$ to $\mathcal{B}$ which map $(x,y) \mapsto (u,v)$ appropriately. Letting

$x = h_{1i}(u,v)$ and $y = h_{2i}(u,v)$ be the inverses, and J_i the Jacobian, on $\mathcal{A}_i$;

$$f_{UV}(u,v) = \begin{cases} \sum_i f_{XY}(h_{1i}(u,v), h_{2i}(u,v))\,|J_i|, & (u,v) \in \mathcal{B}; \\ 0, & \text{otherwise.} \end{cases}$$

For generalization to $\mathbb{R}^n \to \mathbb{R}^n$ case for $n > 2$, see C&B p. 185.

Convolution formulae (C&B 5.2.9, ex. 5.6) $X \perp\!\!\!\perp Y$ both continuous. Then:

1. $f_{X+Y}(z) = \int_{\mathbb{R}} f_X(w) f_Y(z - w)\,dw$.

2. $f_{X-Y}(z) = \int_{\mathbb{R}} f_X(w) f_Y(w - z)\,dw$.

3. $f_{XY}(z) = \int_{\mathbb{R}} |\frac{1}{w}| f_X(w) f_Y(z/w)\,dw$.

4. $f_{X/Y}(z) = \int_{\mathbb{R}} |w| f_X(wz) f_Y(w)\,dw$.

Probability integral transformation (C&B 2.1.10, ex. 2.10) If $Y = F_X(X)$ (for X continuous) then $Y \sim \text{Unif}(0,1)$. Can be used to generate random samples from a particular distribution: generate a uniform random and apply inverse of cdf of target distribution. If X is discrete, Y is stochastically greater than $\text{Unif}(0,1)$.

1.4 Properties of random variables

Expected value, mean (C&B 2.2.1, 5–6, 4.6.6)

$$\mathrm{E}\, g(X) \equiv \begin{cases} \int_{\mathbb{R}} g(x) f_X(x)\,dx, & \text{if } X \text{ continuous;} \\ \sum_{x \in \mathcal{X}} g(x) f_X(x)\,dx, & \text{if } X \text{ discrete;} \end{cases}$$

provided the integral or sum exists *and* that $\mathrm{E}\,|g(X)| \neq \infty$. For constants a, b, c and functions g_1, g_2 such that $\mathrm{E}(g_1(X))$, $\mathrm{E}(g_2(X))$ exist,

1. $\mathrm{E}[a g_1(X) + b g_2(X) + c] = a\,\mathrm{E}(g_1(X)) + b\,\mathrm{E}(g_2(X)) + c$ (i.e., expectation is a linear operator);

2. If $\forall x, g_1(x) \geq 0$, then $\mathrm{E}\, g_1(X) \geq 0$;

3. If $\forall x, g_1(x) \geq g_2(x)$, then $\mathrm{E}\, g_1(X) \geq \mathrm{E}\, g_2(X)$;

4. If $\forall x, a \leq g_1(x) \leq b$, then $a \leq \mathrm{E}\, g_1(X) \leq b$.

The mean is the MSE minimizing predictor for X; i.e., $\min_b \mathrm{E}(X - b)^2 = \mathrm{E}(X - \mathrm{E}\,X)^2$. If $X_1, \dots, X_n$ mutually independent, then $\mathrm{E}[g_1(X_1) \cdot \dots \cdot g_n(X_n)] = \mathrm{E}[g_1(X_1)] \cdot \dots \cdot \mathrm{E}[g_n(X_n)]$.

Conditional expectation (C&B p. 150; Hansen 4-14–6; Hayashi 138–9) a.k.a. regression of Y on X. $\mathrm{E}(Y|X)$ is a r.v. which is a function of X. For discrete (X,Y), $\mathrm{E}(g(Y)|x) \equiv \sum_Y g(y) f(y|x)$. For continuous (X,Y), $\mathrm{E}(g(Y)|x) \equiv \int_{\mathbb{R}} g(y) f(y|x)\,dy$. Conditional expected value has all the standard properties of expected value. Also:

1. $\mathrm{E}[g(X)|X] = g(X)$ for any function g.

2. $\mathrm{E}[g(X)h(Y)|X] = g(X)\,\mathrm{E}[h(Y)|X]$ for any functions g and h.

3. $X \perp\!\!\!\perp Y \implies \mathrm{E}(Y|X) = \mathrm{E}(Y)$ (i.e., knowing X gives us no additional information about $\mathrm{E}\,Y$).

4. $\mathrm{E}(Y|X) = \mathrm{E}(Y) \implies \mathrm{Cov}(X,Y) = 0$

5. $\mathrm{E}(Y|X)$ is the MSE minimizing predictor of Y based on knowledge of X, (i.e., $\min_{g(x)} \mathrm{E}[Y - g(X)]^2 = \mathrm{E}[Y - \mathrm{E}(Y|X)]^2$).

Let X be a r.v. that takes values in $(\mathbb{R}, \mathcal{B})$, let $\mathcal{G} = \sigma(X)$ (i.e., the smallest sigma field measuring X), and assume $\mathrm{E}\,|Y| < \infty$. Conditional expected value of Y given X, is defined implicitly (and non-uniquely) as satisfying:

1. $\mathrm{E}\,|\mathrm{E}(Y|X)| < \infty$;

2. $\mathrm{E}(Y|X)$ is $\mathcal{G}$-measurable (i.e., $Y|X$ cannot rely on more information than X does);

3. $\forall G \in \mathcal{G}, \int_G \mathrm{E}(Y|X)\,dP(\omega) = \int_G Y\,dP(\omega)$ (i.e., $\mathrm{E}[\mathrm{E}(Y|X)|X \in G] = \mathrm{E}[Y|X \in G]$);

where the notation $\int_B \cdot dP_X(x)$ means $\int_B \cdot f_X(x)\,dx$ if X is continuous, and means $\sum_{x \in B} \cdot f_X(x)$ if X is discrete.

Two-way rule for expectations (C&B p. 58, ex. 2.21) If $Y = g(X)$, then $\mathrm{E}\, g(X) = \mathrm{E}\,Y$; i.e., $\int_{\mathbb{R}} g(x) f_X(x)\,dx = \int_{\mathbb{R}} y f_Y(y)\,dy$.

Law of Iterated Expectations (C&B 4.4.3; Hansen 4-21) $\mathrm{E}\,X = \mathrm{E}[\mathrm{E}(X|Y)]$, provided the expectations exist. More generally, when $\mathcal{L} \subseteq \mathcal{M}$ (i.e., $\mathcal{L}$ contains less information, $\mathcal{M}$ contains more),

$$\mathrm{E}[X|\mathcal{L}] = \mathrm{E}[\mathrm{E}(X|\mathcal{M})|\mathcal{L}] = \mathrm{E}[\mathrm{E}(X|\mathcal{L})|\mathcal{M}].$$

Median (C&B ex. 2.17–18) m such that $P(X \leq m) \geq \frac{1}{2}$ and $P(X \geq m) \geq \frac{1}{2}$. If X continuous, the median minimizes absolute deviation; i.e., $\min_a \mathrm{E}\,|X - a| = \mathrm{E}\,|X - m|$.

Mode (C&B ex. 2.27) $f(x)$ is unimodal with mode equal to a iff $a \geq x \geq y \implies f(a) \geq f(x) \geq f(y)$ and $a \leq x \leq y \implies f(a) \geq f(x) \geq f(y)$.

1. Modes are not necessarily unique.

2. If f is symmetric and unimodal, then the point of symmetry is a mode.

Symmetric distribution (C&B ex. 2.25–26) If f_X is symmetric about a (i.e., $\forall \varepsilon, f_X(a + \varepsilon) = f_X(a - \varepsilon)$), then:

1. X and $2a - X$ are identically distributed;

2. If $a = 0$, then M_X is symmetric about 0;

3. a is the median;

4. If $\mathrm{E}\,X$ exists, then $\mathrm{E}\,X = a$.

5. For odd k, the kth central moment μ_k is zero (if it exists); if the distribution is symmetric about 0, then all odd moments are zero (if they exist).

Moment (C&B 2.3.1, 11; Hansen 2-37) For $n \in \mathbb{Z}$, the nth moment of X is $\mu'_n \equiv \mathrm{E}\, X^n$. Also denote $\mu'_1 = \mathrm{E}\, X$ as μ. The nth central moment is $\mu_n \equiv \mathrm{E}(X - \mu)^n$.

1. Two different distributions *can* have all the same moments, but only if the variables have unbounded support sets.

2. A distribution is uniquely determined by its moments if all moments are defined and $\lim_{n \to \infty} \sum_{k=1}^n \mu'_k r^k / k!$ exists for all r in a neighborhood of zero.

Variance (C&B 2.3.2, 4, p. 60, 4.5.6, ex. 4.58) $\mathrm{Var}\, X \equiv \mu_2 = \mathrm{E}(X - \mathrm{E}\, X)^2 = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2$. Often parametrized as σ^2.

1. For constants a, b, if $\mathrm{Var}\, X \neq \infty$, then $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}\, X$;

2. Assuming variances exist, $\mathrm{Var}(aX + bY) = a^2 \mathrm{Var}\, X + b^2 \mathrm{Var}\, Y + 2ab\, \mathrm{Cov}(X, Y)$;

3. $\mathrm{Var}[Y - \mathrm{E}(Y|X)] = \mathrm{E}[\mathrm{Var}(Y|X)]$.

Multivariate variance (?) $\mathrm{Var}\, \mathbf{X} \equiv \mathrm{E}[\mathbf{XX}'] - \mathrm{E}[\mathbf{X}]\, \mathrm{E}[\mathbf{X}]'$. Thus:

1. $\mathrm{Var}(\mathbf{X} + \mathbf{Y}) = \mathrm{Var}(\mathbf{X}) + \mathrm{Cov}(\mathbf{X}, \mathbf{Y}) + \mathrm{Cov}(\mathbf{X}, \mathbf{Y})' + \mathrm{Var}(\mathbf{Y})$;

2. $\mathrm{Var}(\mathbf{AX}) = \mathbf{A}\, \mathrm{Var}(\mathbf{X})\mathbf{A}'$.

Conditional variance (C&B p. 151, 4.4.7; Greene 81–4) a.k.a. scedastic function. $\mathrm{Var}(Y|X) \equiv E[(Y - E[Y|X])^2|X] = \mathrm{E}[Y^2|X] - (E[Y|X])^2$.

1. $X \perp\!\!\!\perp Y \implies \mathrm{Var}(Y|X) = \mathrm{Var}(Y)$.

2. Conditional variance identity: provided the expectations exist,
$$\mathrm{Var}(Y) = \underbrace{\mathrm{E}[\mathrm{Var}(Y|X)]}_{\text{residual variance}} + \underbrace{\mathrm{Var}[\mathrm{E}(Y|X)]}_{\text{regression variance}}.$$

Implies that on average, conditioning reduces the variance of the variable subject to conditioning ($\mathrm{Var}(Y) \geq \mathrm{E}[\mathrm{Var}(Y|X)]$).

Standard deviation (C&B 2.3.2) $\sigma \equiv \sqrt{\mathrm{Var}\, X}$.

Covariance (C&B 4.5.1, 3, ex. 4.58–9; Greene 77) $\mathrm{Cov}(X, Y) \equiv \mathrm{E}[(X - \mathrm{E}\, X)(Y - \mathrm{E}\, Y)] = \mathrm{E}[(X - \mathrm{E}\, X)Y] = \mathrm{E}[X(Y - \mathrm{E}\, Y)] = \mathrm{E}(XY) - (\mathrm{E}\, X)(\mathrm{E}\, Y)$. If X, Y, Z all have finite variances, then:

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}[X, \mathrm{E}(Y|X)]$;

2. $\mathrm{Cov}[X, Y - \mathrm{E}(Y|X)] = 0$;

3. $\mathrm{Cov}(X, Y) = \mathrm{E}[\mathrm{Cov}(X, Y|Z)] + \mathrm{Cov}[\mathrm{E}(X|Z), \mathrm{E}(Y|Z)]$.

4. $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$.

5. $\mathrm{Cov}(aX + bY, cX + dY) = ac\, \mathrm{Var}(X) + bd\, \mathrm{Var}(Y) + (ad + bc)\, \mathrm{Cov}(X, Y)$.

Multivariate covariance (Hansen 5-28; Hayashi 75–6) $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) \equiv \mathrm{E}[(\mathbf{X} - \mathrm{E}\, \mathbf{X})(\mathbf{Y} - \mathrm{E}\, \mathbf{Y})'] = \mathrm{E}(\mathbf{XY}') - (\mathrm{E}\, \mathbf{X})(\mathrm{E}\, \mathbf{Y}')$. Thus:

1. $\mathrm{Cov}(\mathbf{AX}, \mathbf{BY}) = \mathbf{A}\, \mathrm{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}'$;

2. $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathrm{Cov}(\mathbf{Y}, \mathbf{X})'$.

Correlation (C&B 4.5.2, 5, 7) $\mathrm{Corr}(X, Y) \equiv \rho_{XY} \equiv \mathrm{Cov}(X, Y)/(\sigma_X \sigma_Y)$.

1. $\mathrm{Corr}(a_1 X + b_1, a_2 Y + b_2) = \mathrm{Corr}(X, Y)$.

2. Correlation is in the range $[-1, 1]$, with ± 1 indicating a perfectly linear relationship ($+1$ for positive slope, -1 for negative slope), by the Cauchy-Schwarz Inequality.

3. $X \perp\!\!\!\perp Y \implies \mathrm{Cov}(X, Y) = \rho_{XY} = 0$ (assuming finite moments); note however that zero covariance need *not* imply independence.

Skewness (C&B ex. 2.28; Greene 66) $\alpha_3 \equiv \mu_3 \cdot (\mu_2)^{-3/2}$, where μ_i is the ith central moment. Measures the lack of symmetry in the pdf. 0 for any normal, t, or uniform; 2 for exponential, $2\sqrt{2/r}$ for χ_r^2, $2\sqrt{a}/a$ for gamma.

Kurtosis (C&B ex. 2.28) $\alpha_4 \equiv \mu_4 \cdot \mu_2^{-2}$, where μ_i is the ith central moment. Measures the "peakedness" of the pdf. $\alpha_4 = 3$ for any normal. (Sometimes normalized by subtracting 3.)

Moment generating function (C&B 2.3.6–7, 11–12, 15, 4.2.12, 4.6.7, 9) $M_X(t) \equiv \mathrm{E}\, e^{tX}$ as long as the expectation exists for t in a neighborhood 0. If M_X exists, then $\forall n \in \mathbb{Z}$, $n \geq 0$,
$$\mu'_n \equiv \mathrm{E}\, X^n = \frac{d^n}{dt^n} M_X(t) \bigg|_{t=0}.$$

1. It is possible for all moments to exist, but not the mgf.

2. If r.v.s have equal mgfs in some neighborhood of 0, then the variables are identically distributed (i.e., an extant mgf characterizes a distribution).

3. If the mgfs of a sequence of r.v.s converge to M_X in some neighborhood of zero, then the cdfs of the sequence converge to F_X at all points where F_X is continuous.

4. For constants a, b, if M_X exists, then $M_{aX+b}(t) = e^{bt} M_X(at)$.

5. For $X \perp\!\!\!\perp Y$, $M_{X+Y}(t) = M_X(t)M_Y(t)$. For $X_1, \dots, X_n$ mutually independent, $M_{\sum X_i} = \prod_i M_{X_i}$.

6. For $X_1, \dots, X_n$ mutually independent, $Z \equiv (a_1 X_1 + b_1) + \cdots + (a_n X_n + b_n)$, then $M_Z(t) = \left(e^{t(\sum b_i)}\right) \prod_i M_{X_i}(a_i t)$.

Characteristic function (C&B sec. 2.6.2) $\phi_X(t) \equiv \mathrm{E}\, e^{itX}$, where $i = \sqrt{-1}$.

1. The cf always exists.

2. A cf completely determines a distribution: if the cfs of a sequence of r.v.s converge to ϕ_X in some neighborhood of zero, then the cdfs of the sequence converge to F_X at all points where F_X is continuous.

3. For $X \sim \mathrm{N}(0, 1)$, $\phi_X(t) = e^{-t^2/2}$.

4. We can recover probability from a cf: for all a, b such that $P(X = a) = P(X = b) = 0$,
$$P(X \in [a, b]) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t)\, dt.$$

Other generating functions (C&B sec. 2.6.2) Cumulant generating function $\equiv \log[M_X(t)]$, if the mgf exists.

Factorial moment generating function (a.k.a. probability-generating function when X is discrete) $\equiv \mathrm{E}\, t^X$, if the expectation exists.

1.5 Distributions

Normal distribution (C&B p. 102–4, 2.1.9, 3.6.5, 4.2.14, 4.3.4, 6, 5.3.3; Wikipedia) Normal (a.k.a. Gaussian) particularly important because it is analytically tractable, has a familiar symmetric bell shape, and CLT shows that it can approximate many distributions in large samples. If X is normal with mean (and median) μ and variance σ^2, then $X \sim \mathrm{N}(\mu, \sigma^2)$ with pdf
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right).$$

f_X has maximum at μ and inflection points at $\mu \pm \sigma$. Moments are $\mathrm{E}\, X = \mu$, $\mathrm{E}\, X^2 = \mu^2 + \sigma^2$, $\mathrm{E}\, X^3 = \mu^3 + 3\mu\sigma^2$, $\mathrm{E}\, X^4 = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$.

Stein's Lemma: If $g(\cdot)$ is differentiable with $\mathrm{E}\, |g'(X)| < \infty$, then $\mathrm{E}[g(X)(X - \mu)] = \sigma^2 \mathrm{E}\, g'(X)$.

$Z \equiv (X - \mu)/\sigma$ is distributed $\mathrm{N}(0, 1)$ (i.e., "standard normal"). $\mathrm{E}[Z^k] = 0$ if k odd, $\mathrm{E}[Z^k] = 1 \cdot 3 \cdot 5 \cdots (n - 1)$ if k even. CDF denoted $\Phi(\cdot)$; pdf is
$$\phi(z) \equiv f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

1. $P(|X - \mu| \leq \sigma) = P(|Z| \leq 1) \approx 68.26\%$;

2. $P(|X - \mu| \leq 2\sigma) = P(|Z| \leq 2) \approx 95.44\%$;

3. $P(|X - \mu| \leq 3\sigma) = P(|Z| \leq 3) \approx 99.74\%$.

Independence and zero-covariance are equivalent for linear functions of normally distributed r.v.s. If normally distributed random vectors are pairwise independent, they are mutually independent.

Given iid sample $X_i \sim \mathrm{N}(\mu, \sigma^2)$, log-likelihood is
$$L(\mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2.$$

Many distributions can be generated with manipulations/combinations of normals:

1. Square of standard normal is χ_1^2.

2. If $X \sim \mathrm{N}(\mu, \sigma^2)$, $Y \sim \mathrm{N}(\gamma, \tau^2)$, and $X \perp\!\!\!\perp Y$, then $X + Y \sim \mathrm{N}(\mu + \gamma, \sigma^2 + \tau^2)$ (i.e., independent normals are additive in mean and variance).

3. The sum and difference of independent normal r.v.s are independent normal as long as the variances are equal.

4. Ratio of independent standard normals is Cauchy ($\sigma = 1$, $\theta = 0$); look for the kernel of the exponential distribution when dividing normals.

Bivariate normal distribution (C&B 4.5.10, ex. 4.45) Parameters $\mu_X, \mu_Y \in \mathbb{R}$; $\sigma_X, \sigma_Y > 0$; $\rho \in [-1, 1]$; and pdf (on $\mathbb{R}^2$):

$$f(x, y) = \left(2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}\right)^{-1}$$

$$\times \exp\left(\frac{-1}{2(1 - \rho^2)}\left(\left(\frac{x - \mu_X}{\sigma_X}\right)^2\right.\right.$$

$$\left.\left. -2\rho\left(\frac{x - \mu_X}{\sigma_X}\right)\left(\frac{y - \mu_Y}{\sigma_Y}\right) + \left(\frac{y - \mu_Y}{\sigma_Y}\right)^2\right)\right).$$

1. The marginal distributions of X and Y are $\mathrm{N}(\mu_X, \sigma_X^2)$ and $\mathrm{N}(\mu_Y, \sigma_Y^2)$ (note marginal normality does *not* imply joint normality).

2. The correlation between X and Y is ρ.

3. For any constants a and b, the distribution of $aX + bY$ is $\mathrm{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$.

4. All conditional distributions of Y given $X = x$ are normal: $Y|X = x \sim$

$$\mathrm{N}(\mu_Y + \underbrace{\rho(\sigma_Y/\sigma_X)}_{\mathrm{Cov}(X,Y)/\sigma_X^2}(x - \mu_X), \sigma_Y^2(1 - \rho^2)).$$

Multivariate normal distribution (Hansen 5-17–35; MaCurdy p. 6; Greene 94) p-dimensional normal, $\mathrm{N}_p(\mu, \boldsymbol{\Sigma})$ has pdf

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}(\mathbf{x} - \mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu)\right],$$

where $\mu = \mathrm{E}[\mathbf{X}]$ and $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j)$.

A linear transformation of a normal is normal: if $\mathbf{X} \sim \mathrm{N}_p(\mu, \boldsymbol{\Sigma})$, then for any $\mathbf{A} \in \mathbb{R}^{q \times p}$ with full row rank (which implies $q \leq p$), and any $\mathbf{b} \in \mathbb{R}^q$, we have $\mathbf{AX} + \mathbf{b} \sim \mathrm{N}_q(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. In particular, $\boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \mu) \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, where $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1} = \mathbf{H}\boldsymbol{\Lambda}^{-1/2}\mathbf{H}'$.

The following transformations of $\mathbf{X} \sim \mathrm{N}_p(\mu, \boldsymbol{\Sigma})$ are independent iff $\mathbf{A\Sigma B}' = \mathrm{Cov}(\mathbf{AX}, \mathbf{BX}) = \mathbf{0}$:

1. $\mathbf{AX} \sim \mathrm{N}(\mathbf{A}\mu, \mathbf{A\Sigma A}')$ and $\mathbf{BX} \sim \mathrm{N}(\mathbf{B}\mu, \mathbf{B\Sigma B}')$,

2. $\mathbf{AX} \sim \mathrm{N}(\mathbf{A}\mu, \mathbf{A\Sigma A}')$ and $\mathbf{X}'\mathbf{BX} \sim \chi^2_{\mathrm{rank}(\mathbf{B\Sigma})}$ (where $\mathbf{B\Sigma}$ is an idempotent matrix),

3. $\mathbf{X}'\mathbf{AX} \sim \chi^2_{\mathrm{rank}(\mathbf{A\Sigma})}$ and $\mathbf{X}'\mathbf{BX} \sim \chi^2_{\mathrm{rank}(\mathbf{B\Sigma})}$ (where $\mathbf{A\Sigma}$ and $\mathbf{B\Sigma}$ are idempotent matrices).

Chi squared distribution (C&B 5.3.2; Hansen 5-29–32; MaCurdy p. 6; Greene 92) χ_n^2 (Chi squared with n degrees of freedom) has mean n and variance $2n$. Can be generated from normal:

1. If $Z \sim \mathrm{N}(0, 1)$, then $Z^2 \sim \chi_1^2$ (i.e., the square of standard normal is a chi squared with 1 degree of freedom);

2. If $X_1, \ldots, X_n$ are independent with $X_i \sim \chi_{p_i}^2$, then $\sum X_i \sim \chi^2_{\sum p_i}$ (i.e., independent chi squared variables add to a chi squared, and the degrees of freedom add).

3. If $\mathbf{X} \sim \mathrm{N}_n(\mu, \boldsymbol{\Sigma})$, then $(\mathbf{X} - \mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mu) \sim \chi_n^2$.

4. If $\mathbf{X} \sim \mathrm{N}_n(\mathbf{0}, \mathbf{I})$ and $\mathbf{P}_{n \times n}$ is an idempotent matrix, then $\mathbf{X}'\mathbf{PX} \sim \chi^2_{\mathrm{rank}(\mathbf{P})} = \chi^2_{\mathrm{tr}(\mathbf{P})}$.

5. If $\mathbf{X} \sim \mathrm{N}_n(\mathbf{0}, \mathbf{I})$ then the sum of the squared deviations from the sample mean $\mathbf{X}'\mathbf{M}_\iota\mathbf{X} \sim \chi^2_{n-1}$.

6. If $\mathbf{X} \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$ and $\mathbf{B}_{n \times n}\boldsymbol{\Sigma}$ is an idempotent matrix, then $\mathbf{X}'\mathbf{BX} \sim \chi^2_{\mathrm{rank}(\mathbf{B\Sigma})} = \chi^2_{\mathrm{tr}(\mathbf{B\Sigma})}$.

Student's t distribution (C&B 5.3.4; Greene 69–70) If $X_1, \ldots, X_n$ are iid $\mathrm{N}(\mu, \sigma^2)$, then $\sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathrm{N}(0, 1)$. However, we will generally not know σ. Using the sample variance rather than the true variance gives $\sqrt{n}(\bar{X} - \mu)/s \sim t_{n-1}$.

Generally, $\mathrm{N}(0,1)/\sqrt{\chi^2_{n-1}/(n-1)} \sim t_{n-1}$. If a t distribution has p degrees of freedom, there are only $p - 1$ defined moments. t has thicker tails than normal.

t_1 is Cauchy distribution (the ratio of two independent standard normals). t_∞ is standard normal.

Snedecor's F distribution (C&B 5.3.6–8) $(\chi_p^2/p)/(\chi_q^2/q) \sim F_{p,q}$. The F distribution is also related by transformation with several other distributions:

1. $1/F_{p,q} \sim F_{q,p}$ (i.e., the reciprocal of an F r.v. is another F with the degrees of freedom switched);

2. $(t_q)^2 \sim F_{1,q}$;

3. $(p/q)F_{p,q}/(1 + (p/q)F_{p,q}) \sim \mathrm{beta}(p/2, q/2)$.

Lognormal distribution (C&B p. 625) If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $Y \equiv e^X$ is lognormally distributed. (Note: a lognormal is *not* the log of a normally distributed r.v.).

$$\mathrm{E}\, Y = e^{\mu + (\sigma^2/2)};$$

$$\mathrm{Var}\, Y = e^{2(\mu + \sigma^2)} - e^{2\mu + \sigma^2}.$$

Exponential families (C&B 3.4; Mahajan 1-5–6, 11) Any family of pdfs or pmfs that can be expressed as

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right),$$

where $h(x) \geq 0$, $\{t_i(x)\}$ are real-valued functions, $c(\theta) \geq 0$, $\{w_i(\theta)\}$ are real-valued functions, and the support does not depend on θ.

Includes normal, gamma, beta, χ^2, binomial, Poisson, and negative binomial. C&B Theorem 3.4.2 gives results that may help calculate mean and variance using differentiation, rather than summation/integration.

Can be re-parametrized as:

$$f(x|\eta) = h(x)c^*(\eta)\exp\left(\sum_{i=1}^{k}\eta_i t_i(x)\right),$$

over "natural parameter space" $\mathcal{H} \equiv \{\eta = (\eta_1, \ldots, \eta_k): \int_\mathbb{R} h(x)\exp(\sum_{i=1}^{k}\eta_i t_i(x))\, dx < \infty\}$, where for all $\eta \in \mathcal{H}$, we have $c^*(\eta) \equiv [\int_\mathbb{R} h(x)\exp(\sum_{i=1}^{k}\eta_i t_i(x))\, dx]^{-1}$ to ensure the pdf integrates to 1.

The joint distribution of an iid sample from an exponential family will also be an exponential family (closure under random sampling).

Location and Scale families (C&B 3.5.1–7, p. 121) If $f(x)$ a pdf, μ, σ constants with $\sigma > 0$, then $g(x)$ also a pdf:

$$g(x) \equiv \frac{1}{\sigma}f\left(\frac{x - \mu}{\sigma}\right).$$

$X \sim g(x)$ iff $\exists Z \sim f$, $X = \sigma Z + \mu$. Assume X and Z exist; $P(X \leq x) = P(Z \leq (x - \mu)/\sigma)$, and if $\mathrm{E}\, Z$ and $\mathrm{Var}\, Z$ exist, then $\mathrm{E}\, X = \sigma\, \mathrm{E}\, Z + \mu$ and $\mathrm{Var}(X) = \sigma^2\, \mathrm{Var}\, Z$.

1. Family of pdfs $f(x - \mu)$ indexed by μ is the "location family" with standard pdf $f(x)$ and location parameter μ.

2. Family of pdfs $\frac{1}{\sigma}f(x/\sigma)$ indexed by $\sigma > 0$ is the "scale family" with standard pdf $f(x)$ and scale parameter σ. (e.g., exponential.)

3. Family of pdfs $\frac{1}{\sigma}f((x - \mu)/\sigma)$ indexed by μ, and $\sigma > 0$ is the "location-scale family" with standard pdf $f(x)$, location parameter μ, and scale parameter σ. (e.g., uniform, normal, double exponential, Cauchy.)

Stable distribution (Hansen 5-15) Let X_1, X_2 be iid F, and define $Y = aX_1 + bX_2 + c$. Then F is a stable distribution iff $\forall a$, b, c, $\exists d$, e such that $dY + e \sim F$.

1.6 Random samples

Random sample, iid (5.1.1) R.v.s $X_1, \ldots, X_n$ are a random sample of size n from population $f(x)$ (a.k.a. n iid r.v.s with pdf/pmf $f(x)$) if they are mutually independent, each with marginal pdf/pmf $f(x)$. By independence, $f(x_1, \ldots, x_n) = \prod_i f(x_i)$.

Statistic (C&B 5.2.1; Mahajan 1-7) A r.v. $Y \equiv T(X_1, \ldots, X_n)$, where T is a real or vector-valued function $T(x_1, \ldots, x_n)$ whose domain includes the support of $(X_1, \ldots, X_n)$. The distribution of Y is called its sampling distribution.

Alternately, any measurable function *of the data* (as distinct from a parameter—a function of the distribution).

Unbiased estimator (Hansen 5-14; C&B 7.3.2) An statistic $\hat{\theta}$ is unbiased for θ iff $\mathrm{E}_\theta(\hat{\theta}) = \theta$ for all θ. That is, if $\mathrm{Bias}[\hat{\theta}] \equiv \mathrm{E}_\theta \hat{\theta} - \theta = 0$ for all θ.

Sample mean (C&B 5.2.2, 4, 6–8, 10, p. 216, 5.5.2) $\bar{X} \equiv \frac{1}{n} \sum_i X_i$ (i.e., the arithmetic average of the values in a random sample). For any real numbers, the arithmetic average minimizes SSR (i.e., $\bar{x} \in \mathrm{argmin}_a \sum_i (x_i - a)^2$).

1. If $\mathrm{E} X_i = \mu < \infty$, then $\mathrm{E} \bar{X} = \mu$ (i.e., $\bar{X}$ is an unbiased estimator of μ)

2. If $\mathrm{Var} X_i = \sigma^2 < \infty$, then $\mathrm{Var} \bar{X} = \sigma^2/n$.

3. If X_i have mgf $M_X(t)$, then $M_{\bar{X}}(t) = [M_X(t/n)]^n$.

4. (Law of Large Numbers) If $\{X_i\}$ iid, $\mathrm{E} X_i = \mu < \infty$ and $\mathrm{Var} X_i = \sigma^2 < \infty$, then the series $\bar{X}_n \xrightarrow{\text{as}} \mu$ (this also implies convergence in probability, the "Weak" LLN).

The distribution of the X_i, together with n, characterize the distribution of $\bar{X}$:

1. If $X_i \sim \mathrm{N}(\mu, \sigma^2)$, then $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$.

2. If $X_i \sim \mathrm{gamma}(\alpha, \beta)$, then $\bar{X} \sim \mathrm{gamma}(n\alpha, \beta/n)$.

3. If $X_i \sim \mathrm{Cauchy}(\theta, \sigma)$, then $\bar{X} \sim \mathrm{Cauchy}(\theta, \sigma)$.

4. If $X_i \sim (1/\sigma) f((x - \mu)/\sigma)$ are members of a location-scale family, then $\bar{X} = \sigma \bar{Z} + \mu$, where $\{Z_i\}_{i=1}^n$ is a random sample with $Z_i \sim f(z)$.

Sample variance (C&B 5.2.3–4, 6; Greene 102–4)

$$s^2 \equiv \frac{1}{n-1} \sum_i (X_i - \bar{X})^2 = \frac{1}{n-1} \Big[\sum_i X_i^2 - n\bar{X}^2 \Big].$$

Sample standard deviation is $s \equiv \sqrt{s^2}$. If $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then $\mathrm{E} s^2 = \sigma^2$ (i.e., s^2 is an unbiased estimator of σ^2). $s_{aX}^2 = a^2 s_X^2$.

For any real numbers $\{x_i\}_{i=1}^n$, we have $\sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2$.

Sample covariance (Greene 102–4)

$$s_{XY} \equiv \frac{1}{n-1} \sum_i (X_i - \bar{X})(Y_i - \bar{Y})$$

$$= \frac{1}{n-1} \Big[\sum_i X_i Y_i - n\bar{X}\bar{Y} \Big].$$

If $\mathrm{Cov}(X_i, Y_i) = \sigma_{XY} < \infty$, then $\mathrm{E} s_{XY} = \sigma_{XY}$ (i.e., s_{XY} is an unbiased estimator of σ_{XY}). $s_{aX,bY} = |ab| s_{XY}$.

For any real numbers $\{x_i, y_i\}_{i=1}^n$, we have $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y}$.

Sample correlation (Greene 102–4) $r_{XY} \equiv s_{XY}/(s_X s_Y)$.

$$r_{aX,bY} = (ab/|ab|) r_{XY}.$$

Order statistic (C&B 5.4.1–4) The order statistics of a sample $X_1, \ldots, X_n$ are the ordered values from $X_{(1)}$ (the sample minimum) to $X_{(n)}$ (the sample maximum). Thus the sample median is

$$M \equiv \begin{cases} X_{((n+1)/2)}, & n \text{ is odd}; \\ \frac{1}{2}(X_{(n/2)} + X_{(n/2+1)}), & n \text{ is even}. \end{cases}$$

If $\{X_i\}_{i=1}^n$ iid continuous r.v.s, then

$$F_{X_{(j)}}(x) = \sum_{k=j}^n \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k};$$

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j}.$$

See C&B 5.4.3 for discrete r.v.

Samples from the normal distribution (C&B 5.3.1, 6) $\{X_i\}_{i=1}^n$ iid $\mathrm{N}(\mu, \sigma^2)$ gives:

1. $\bar{X} \perp\!\!\!\perp s^2$ (can also be shown with Basu's Theorem);

2. $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$;

3. $\mathrm{Var}(s^2) = 2\sigma^4/(n-1)$;

4. $(n-1)s^2/\sigma^2 \sim \chi_{n-1}^2$.

If $\{X_i\}_{i=1}^n$ iid $\mathrm{N}(\mu_X, \sigma_X^2)$ and $\{Y_i\}_{i=1}^m$ iid $\mathrm{N}(\mu_Y, \sigma_Y^2)$,

$$\frac{s_X^2/s_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}.$$

1.7 Convergence of random variables
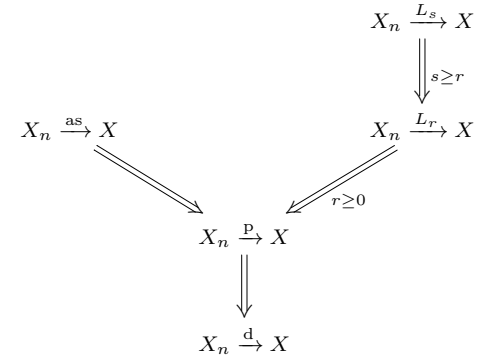
See more LLNs and CLTs in "Time-series concepts."

Convergence in probability (C&B 5.5.1–4, 12; Hansen 5-41; Hayashi 89; D&M 103; MaCurdy p. 9) $\{X_i\}_{i=1}^\infty$ converges in probability to X iff, $\forall \varepsilon > 0$, $\lim_{n \to \infty} P(|X_n - X| \geq \varepsilon) = 0$, or equivalently $\lim_{n \to \infty} P(|X_n - X| < \varepsilon) = 1$. Written as $X_n \xrightarrow{\mathrm{p}} X$ or $X_n - X = o_p(1)$ or $\mathrm{plim}_{n \to \infty} X_n = X$.

1. Convergence in probability is implied by almost sure convergence or convergence in L_p (for $p > 0$).

2. Convergence in probability implies convergence in distribution (but *not* conversely).

3. (Weak Law of Large Numbers) If $\{X_i\}$ iid with $\mathrm{E} X_i = \mu < \infty$ and $\mathrm{Var} X_i = \sigma^2 < \infty$, then the series $\bar{X}_n \xrightarrow{\mathrm{p}} \mu$ (stronger result gives convergence almost surely).

4. (Continuous Mapping Theorem) If $X_n \xrightarrow{\mathrm{p}} X$ and h is a continuous function, then $h(X_n) \xrightarrow{\mathrm{p}} h(X)$.

5. If $\mathrm{E} X_n \to \mu$ and $\mathrm{Var} X_n \to 0$, then $X_n \xrightarrow{\mathrm{p}} \mu$.

Uniform convergence in probability (Hayashi 456–7; MaCurdy p. 14; D&M 137) $\{Q_i(\theta)\}_{i=1}^\infty$ converges in probability to $Q_0(\theta)$ iff, $\sup_{\theta \in \Theta} \|Q_n(\theta) - Q_0(\theta)\| \xrightarrow{\mathrm{p}} 0$.

That is $\forall \varepsilon > 0$, $\lim_{n \to \infty} P(\sup_{\theta \in \Theta} \|Q_n(\theta) - Q_0(\theta)\| \geq \varepsilon) = 0$, or equivalently $\lim_{n \to \infty} P(\sup_{\theta \in \Theta} \|Q_n(\theta) - Q_0(\theta)\| < \varepsilon) = 1$.

This is stronger than pointwise convergence. Uniform convergence in probability is the regularity condition required to pass plims through functions or to reverse the order of differentiation and integration.

Little o error notation (D&M 108–13; Hansen 5-42; MathWorld) Roughly speaking, a function is $o(z)$ iff is of lower asymptotic order than z.

$f(n) = o(g(n))$ iff $\lim_{n \to \infty} f(n)/g(n) = 0$. If $\{f(n)\}$ is a sequence of random variables, then $f(n) = o_p(g(n))$ iff $\mathrm{plim}_{n \to \infty} f(n)/g(n) = 0$.

We write $X_n - X = o_p(n^{-\gamma})$ iff $n^\gamma (X_n - X) \xrightarrow{\mathrm{p}} 0$.

Big O error notation (D&M 108–13; MathWorld) Roughly speaking, a function is $O(z)$ iff is of the same asymptotic order as z.

$f(n) = O(g(n))$ iff $|f(n)/g(n)| < K$ for all $n > N$ and some positive integer N and some constant $K > 0$. If $\{f(n)\}$ is a sequence of random variables, then $f(n) = o_p(g(n))$ iff $\text{plim}_{n\to\infty} f(n)/g(n) = 0$.

Order symbols (D&M 111-2)

$$O(n^p) \pm O(n^q) = O(n^{\max\{p,q\}}).$$
$$o(n^p) \pm o(n^q) = o(n^{\max\{p,q\}}).$$
$$O(n^p) \pm o(n^q) = \begin{cases} O(n^p), & p \geq q; \\ o(n^q), & p < q. \end{cases}$$
$$O(n^p) \cdot O(n^q) = O(n^{p+q}).$$
$$o(n^p) \cdot o(n^q) = o(n^{p+q}).$$
$$O(n^p) \cdot o(n^q) = o(n^{p+q}).$$

These identities cover sums, *as long as the number of terms summed is independent of n.*

Asymptotic equivalence (D&M 110–1) $f(n) \overset{a}{=} g(n)$ iff $\lim_{n\to\infty} f(n)/g(n) = 1$.

Convergence in L_p (Hansen 5-43–5; Hayashi 90; MaCurdy p. 11) $\{X_i\}_{i=1}^\infty$ converges in L_p to X iff, $\lim_{n\to\infty} \mathrm{E}(|X_n - X|^p) = 0$. Note this requires the existence of a pth moment. Written $X_n \xrightarrow{L_p} X$. Convergence in L_2 a.k.a. convergence in mean square/quadratic mean.

1. Convergence in L_p is implied by convergence in L_q for $q \geq p$.

2. Convergence in L_p implies convergence in L_j for $j \leq p$.

3. Convergence in L_p (for $p > 0$) implies convergence in probability and in distribution (but *not* conversely).

4. (Continuous Mapping Theorem extension) If $X_n \xrightarrow{L_p} X$ and h is a continuous function, then $h(X_n) \xrightarrow{L_p} h(X)$.

Convergence in L_2 to a constant requires $\lim_{n\to\infty} \mathrm{E}[(X_n - X)'(X_n - X)] = \lim_{n\to\infty} \text{Bias}^2 + \text{Var}^2 = 0$. Thus it is necessary and sufficient that squared bias and squared variance go to zero.

Almost sure convergence (C&B 5.5.6, 9; Hayashi 89; D&M 106, 131; MaCurdy p. 10) $\{X_i\}_{i=1}^\infty$ converges in probability to X iff, $\forall \varepsilon > 0$, $P(\lim_{n\to\infty} |X_n - X| < \varepsilon) = 1$. Written $X_n \xrightarrow{as} X$. Also known as strong convergence, or convergence almost everywhere.

1. Almost sure convergence implies convergence in probability and in distribution (but *not* conversely).

2. (Strong Law of Large Numbers) If $\{X_i\}$ iid with $\mathrm{E}\, X_i = \mu < \infty$ and $\text{Var}\, X_i = \sigma^2 < \infty$, then the series $\bar{X}_n \xrightarrow{as} \mu$.

3. (Strong Law of Large Numbers, niid) If $\{X_i\}$ niid with $\mathrm{E}\, X_i = 0$ and $\lim_{n\to\infty} n^{-2} \sum_i \text{Var}\, X_i = \infty$, then the series $\bar{X}_n \xrightarrow{as} 0$.

4. (Continuous Mapping Theorem extension) If $X_n \xrightarrow{as} X$ and h is a continuous function, then $h(X_n) \xrightarrow{as} h(X)$.

Convergence in distribution (C&B 5.5.10–13; Hayashi 90–1; Greene 119–20; D&M 107) $\{X_i\}_{i=1}^\infty$ converges in distribution to X iff, $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all points where F_X is continuous. Written as $X_n \xrightarrow{d} X$ or $X_n - X = O_p(1)$ or as "X is the limiting distribution of X_n."

1. Convergence in probability is implied by almost sure convergence, convergence in L_p (for $p > 0$), or convergence in probability.

2. Convergence in distribution implies convergence in probability *if the series converges to a constant.*

Central Limit Theorem for iid samples (C&B 5.5.14–15; Hansen 5-60–65; Hayashi 96) Lindeberg-Levy CLT: $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} \mathrm{N}(0, 1)$ as long as the iid X_is each have finite mean, and finite, nonzero variance. Note a weaker form requires mgfs of X_i to exist in a neighborhood of zero.

In multivariate case, iid $\mathbf{X}_i \sim (\mu, \boldsymbol{\Sigma})$ satisfy $\sqrt{n}(\bar{\mathbf{X}}_n - \mu) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Proved using Cramér-Wold Device.

We also have CLT for niid samples (Lyapounov's Theorem), Ergodic stationary mds CLT, and CLT for MA(∞) processes.

Central Limit Theorem for niid samples (Hansen 5-62; D&M 126; MaCurdy p. 21–2) [Lyapounov's Theorem] If $X_i \sim \text{niid}(\mu, \sigma_i^2)$ and a $(2 + \delta)$th moment exists for each X_i, then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathrm{N}(0, \bar{\sigma}^2),$$

where $\bar{\sigma}^2 \equiv \lim_{n\to\infty} \frac{1}{n} \sum_i \sigma_i^2$, as long as the iid X_is each have finite mean, and finite, nonzero variance. Note a weaker form requires mgfs of X_i to exist in a neighborhood of zero.

Implies that if $\varepsilon_i \sim \text{niid}(0, \sigma^2)$ (with extant $(2 + \delta)$th moment), and $\{z_i\}$ a series of (non-stochastic) constants, then $n^{-1/2} Z'\varepsilon \xrightarrow{d} \mathrm{N}(0, \sigma^2 S_{zz})$ where $S_{zz} \equiv \lim_{n\to\infty} \frac{1}{n} \sum_i z_i^2 = \lim_{n\to\infty} \frac{1}{n} Z'Z$.

Slutsky's Theorem (C&B 5.5.17; Hayashi 92–3) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$, where a is a constant, then:

1. $Y_n X_n \xrightarrow{d} aX$;

2. $X_n + Y_n \xrightarrow{d} X + a$.

Delta Method (C&B 5.5.24, 26, 28; Wikipedia; Hayashi 93–4) Let $\{X_i\}_{i=1}^\infty$ be a sequence of r.v.s satisfying $\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$. For a given function g and specific θ, suppose $g'(\theta)$ exists and $g'(\theta) \neq 0$. Then:

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{d} \mathrm{N}(0, \sigma^2 [g'(\theta)]^2).$$

If $g'(\theta) = 0$, but $g''(\theta)$ exists and $g''(\theta) \neq 0$, we can apply the second-order Delta Method and get $n[g(X_n) - g(\theta)] \xrightarrow{d} \frac{1}{2}\sigma^2 g''(\theta)\chi_1^2$.

Alternate formulation: If B is an estimator for β then the variance of a function $h(B) \in \mathbb{R}$ is $\text{Var}(h(B)) \approx \nabla h(\beta)' \text{Var}(B)\nabla h(\beta)$. If $h(B)$ is vector-valued, the variance is $H \text{Var}(B) H'$, where $H = \frac{\partial h}{\partial \beta'}$ (i.e., $H_{ij} \equiv \partial_j h_i(\beta)$).

1.8 Parametric models

Parametric model (Mahajan 1-1–2) Describe an (unknown) probability distribution P that is assumed to be a member of a family of distributions $\mathcal{P}$. We describe $\mathcal{P}$ with a parametrization: a map from a (simpler) space Θ to $\mathcal{P}$ such that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. If Θ is a "nice" subset of Euclidean space, and the mapping P_θ is "smooth," then $\mathcal{P}$ is called a parametric model. A regular parametric model if either all P_θ are continuous, or all are discrete.

Parameter (Mahajan 1-2) Mapping from the family of distributions $\mathcal{P}$ to another space (typically a subset of Euclidean Space). Can be explicitly or implicitly defined.

A function *of the distribution* (as distinct from a statistic—a function of the data).

Identification (Mahajan 1-3–4; C&B 11.2.2) A parameter is (point) identified if it is uniquely determined for every distribution; i.e., if the mapping is one-to-one. When the parameter is implicitly defined as the solution to an optimization problem (e.g., $\theta(P) = \text{argmax}_{b\in\Theta} Q_0(b, P)$), identification corresponds to existence of a unique solution to the optimization.

Two elements $\theta_1, \theta_2 \in \Theta$ are observationally equivalent iff they imply $P_{\theta_1} = P_{\theta_2}$. Identification of θ means there is no other element of Θ observationally equivalent to θ; i.e., $P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2$.

Identification in exponential families (Mahajan 1-5–6, Metrics P.S. 5-1) For iid sampling from an exponential family where $\eta_i(\theta) = \theta_i$, if the $k \times k$ (Fisher Information) matrix

$$I(\theta^*) \equiv \mathrm{E}\left[\left(\frac{d \log p(x, \theta^*)}{d\theta} \right) \left(\frac{d \log p(x, \theta^*)}{d\theta} \right)' \right]$$

is nonsingular for every $\theta* \in \Theta$, then θ is identified.

Conditional homoscedasticity (Mahajan 2-17) The assumption that $\text{Var}(Y|Z) = \sigma^2$; i.e., variance of Y does not depend on Z.

Regression model (Mahajan 1-6–7, 3-9 Hayashi 465–6, Metrics P.S. 7-7) $\{Y_i, X_i\}_{i=1}^n$, where $Y_i \in \mathbb{R}$ the dependent variable (a.k.a. response) and $X_i \in \mathbb{R}^d$ the covariates or explanatory variables. Possible specifications include:

1. $E(Y_i|X_i) = g(X_i)$ for some unknown function g (non-parametric regression);

2. $E(Y_i|X_i) = g(X_i'\theta_0)$ for some unknown $\theta_0 \in \mathbb{R}^d$ and unknown function g (single-index model; semi-parametric);

3. $E(Y_i|X_i) = X_i'\theta_0$ for some unknown $\theta_0 \in \mathbb{R}^d$;

4. $(Y_i|X_i) \sim N(X_i'\theta_0, \sigma^2)$ for some unknown $\theta_0 \in \mathbb{R}^d$ and $\sigma^2 \in \mathbb{R}_+$ (Gaussian regression model; θ_0 is identified and conditional MLE $\hat{\theta} = (X'X)^{-1}X'Y$ is consistent if $E(X_iX_i')$ is nonsingular, conditional MLE $\hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\theta})'(Y - X\hat{\theta})$).

Linear regression model with non-stochastic covariates (Mahajan 1-10–1, 18–9, Metrics P.S. 5-3) For two-dimensional Gaussian regression model with $X_i = (1, x_i)'$ known. The parameter (θ_0, σ^2) is identified as long as the x_i are not all identical, with complete sufficient statistic $(\sum Y_i, \sum Y_i^2, \sum x_i Y_i)$. MLE computed in problem set.

Seemingly Unrelated Regressions (Mahajan 2-11, 21) $\{Y_i, X_i\}_{i=1}^n$ where $Y_i \in \mathbb{R}^m$ and $X_i = (X_{1i}', \ldots, X_{mi}')' \in \mathbb{R}^{(m^2)}$. We assume Y_i are (multivariate) normal distributed with means $E[Y_{si}] = x_{si}'\beta_s$ where $\beta_s \in \mathbb{R}^m$ and the parameter of interest is $\beta = (\beta_1', \ldots, \beta_m')' \in \mathbb{R}^{(m^2)}$.

Probit model (Mahajan 2-12–3, 3-9–10; Hayashi 466) iid $\{W_i\}_{i=1}^n = \{Y_i, Z_i\}_{i=1}^n$ where $Y_i \in \{0,1\}$ and Y_i have conditional distribution $P(Y_i = 1|Z_i) = \Phi(\theta'Z_i)$ where $\Phi(\cdot)$ is the cdf of the standard normal. θ is identified and MLE is consistent if $E(Z_iZ_i')$ is nonsingular.

Alternate motivation is threshold crossing model: $Y_i^* = \theta'Z_i - \varepsilon_i$ where $\varepsilon_i \perp Z_i$ and standard normal, and $Y_i = I\{Y_i^* > 0\}$.

Nonlinear least squares (Mahajan 2-16–7, 3-6–7, 17–9) iid $\{Y_i, Z_i\}_{i=1}^n$ where Y_i have conditional expectation $E[Y|Z] = \psi(Z, \theta)$. The parameter θ can also be defined implicitly as $\theta = \arg\min_{b \in \Theta} E[Y - \psi(Z, b)]^2$. Identification condition is that for all $b \neq \theta$, we have $P(\psi(Z, b) \neq \psi(Z, \theta)) > 0$.

See Mahajan 3-17–9 for asymptotic properties, including heteroscedasticity robust asymptotic variance matrix

Linear instrumental variables (Mahajan 2-22–3, 3-11–2) $Y_i = X_i\theta + \varepsilon_i$ with moment conditions $E[(Y_i - X_i\theta)Z_i] = 0$ for random vector Z_i. The Z_i are "endogenous instruments" for regressors X_i, endogenous because $E \varepsilon_i X_i \neq 0$. Identification condition for θ is $E(Z_iX_i')$ has full column rank and that dimension of Z_i be at least dimension of X_i.

1.9 Statistics

Sufficient statistic (Mahajan 1-8–10; C&B 6.2.1) $T(\mathbf{X})$ is sufficient for $\{P_\theta: \theta \in \Theta\}$ (or more compactly for θ) iff the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ does not depend on θ; i.e., $p(\mathbf{x}|T(\mathbf{X})) = p(\mathbf{x}|T(\mathbf{x}), \theta)$. Once the value of a sufficient statistic is known, the sample does not carry any further information about θ. Useful for:

1. Decision theory: base decision rules on sufficient statistics (for any decision rule, we can always come up with rule based only on a sufficient statistic that has the same risk);

2. Dealing with nuisance parameters in hypothesis testing: find sufficient statistics for the nuisance parameters and condition decision rules on them;

3. Unbiased estimation: look for unbiased estimators that are functions of sufficient statistics.

Any one-to-one function of a sufficient statistic is also sufficient. Outside exponential families, it is rare to have a sufficient statistic of smaller dimension than the data.

Factorization theorem (Mahajan 1-9; C&B 6.2.6) In a regular parametric model $\{P_\theta: \theta \in \Theta\}$, a statistic $T(\mathbf{X})$ (with range $\mathcal{T}$) is sufficient for θ iff there exists a function $g: \mathcal{T} \times \Theta \to \mathbb{R}$ and a function h such that $f(\mathbf{x}, \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$ for all $\mathbf{x}$ and θ.

Minimal sufficient statistic (Mahajan 1-12, 19; C&B 6.2.11) $T(\mathbf{X})$ is minimal sufficient if it is sufficient, and for any other sufficient statistic $S(\mathbf{X})$ we can find a function r such that $T(\mathbf{X}) = r(S(\mathbf{X}))$. This means that a minimal sufficient statistic induces the coarsest possible partition on the data; i.e., it has achieved the maximal amount of data reduction possible while still retaining all information about the parameter.

Any one-to-one function of a minimal sufficient statistic is minimal sufficient. If a minimal sufficient statistic exists, then any complete sufficient statistic is also a minimal sufficient statistic.

Likelihood function (Mahajan 1-13–4) $L(\mathbf{x}, \theta) \equiv p(\mathbf{x}, \theta)$. This is the same as the pdf/pmf, but considered as a function of θ instead of $\mathbf{x}$.

The likelihood ratio $\Lambda(\mathbf{x}, \cdot) \equiv L(\mathbf{x}, \cdot)/L(\mathbf{x}, \theta_0)$, where $\theta_0 \in \Theta$ is fixed and known, with the support of P_θ a subset of the support of P_{θ_0} for all $\theta \in \Theta$. The likelihood ratio is minimal sufficient for θ.

Ancillary statistic (Mahajan 1-14; C&B 6.2.16) $S(\mathbf{X})$ is ancillary for θ iff the distribution of $S(\mathbf{X})$ does not depend on θ.

It is first-order ancillary iff $E\, S(\mathbf{X})$ does not depend on θ.

Complete statistic (Mahajan 1-16; C&B 6.2.21, 28) $T: \mathcal{X} \to \mathcal{T}$ is complete iff for every measurable real function $g: \mathcal{T} \to \mathbb{R}$ such that $\forall \theta \in \Theta$, $E_\theta[g(T)] = 0$ implies that $g(T) = 0$ almost everywhere. Equivalently, T is complete if no non-constant function of T is first-order ancillary.

If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

Basu's theorem (Mahajan 1-19; C&B 6.2.24, 28) If $T(\mathbf{X})$ is a complete minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic. Note the minimal wasn't really necessary: if a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

Statistics in exponential families (Mahajan 1-11, 17; C&B 6.2.10, 25) By the factorization theorem, $T(\mathbf{X}) \equiv (T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))$ is sufficient for θ. (N.B. The statistic must contain *all* the T_i.)

If $\mathbf{X}$ is an iid sample, then $T(\mathbf{X}) \equiv (\sum_i T_1(X_1), \ldots, \sum_i T_k(X_i))$ is a complete statistic if the set $\{(\eta_1(\theta), \ldots, \eta_k(\theta)): \theta \in \Theta\}$ contains an open set in $\mathbb{R}^k$. (Usually, all we'll check is dimensionality.)

1.10 Point estimation

Estimator (Mahajan 2-2) Any measurable function of the data. Note this must *not* be a function of any parameters of the distribution.

Extremum estimator (Hayashi 446) An estimator $\hat{\theta}$ such that there is a scalar ("objective") function $Q_n(\theta)$ such that $\hat{\theta}$ maximizes $Q_n(\theta)$ subject to $\theta \in \Theta \subseteq \mathbb{R}^p$. The objective function depends not only on θ, but also on the data (a sample of size n).

Analogy principle (Mahajan 2-2–3) Consider finding an estimator that satisfies the same properties in the sample that the parameter satisfies in the population; i.e., seek to estimate $\theta(P)$ with $\theta(P_n)$ where P_n is the empirical distribution which puts mass $\frac{1}{n}$ at each sample point. Note this distribution converges uniformly to P.

Consistent estimator (Mahajan 3-2; Hansen 5-41–2; C&B 10.1.1, 3) The sequence of estimators $\{\hat{\theta}_n\}_{n=1}^\infty$ is consistent for θ iff $\hat{\theta}_n \xrightarrow{P} \theta(P)$. The sequence is superconsistent iff $\hat{\theta}_n - \theta = o_p(n^{-1/2})$. Superconsistency implies consistency.

If $\lim_{n\to\infty} \text{Var}_\theta \hat{\theta}_n = 0$ (variance goes to zero) and $\lim_{n\to\infty} E_\theta \hat{\theta}_n = \theta$ (bias goes to zero) for every $\theta \in \Theta$, then $\{\hat{\theta}_n\}$ is consistent (sufficient, not necessary, by Chebychev's Inequality).

Consistency with compact parameter space (Mahajan 3-4–5, 11; Hayashi 457) Let $\hat{\theta}_n \equiv \arg\max_{b \in \Theta} Q_n(\mathbf{W}, b) \equiv \arg\max_{b \in \Theta} Q_n(b)$. (In the last equivalence we have suppressed dependence on the data $\mathbf{W}$.) This covers M-estimators, MLE, and GMM estimators. Suppose that:

1. Θ is a *compact* subset of $\mathbb{R}^d$ [generally not satisfied];

2. $Q_n(b)$ is continuous in b for any realization of the data $\mathbf{W}$ ["usually easily checked"];

3. $Q_n(b)$ is a measurable function of the data for all $b \in \Theta$ [generally assumed].

These conditions ensure that $\hat{\theta}_n$ is well-defined. Suppose there exists a function $Q_0(b)$ such that:

1. *Identification*: $Q_0(\cdot)$ is uniquely (globally) maximized on Θ at $\theta \in \Theta$;

2. *Uniform convergence*: $Q_n(\cdot)$ converges uniformly in probability to $Q_0(\cdot)$ [can be verified by checking more primitive conditions; in particular, for M-estimators a Uniform Weak LLN will suffice].

Then $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.

Consistency without compact parameter space (Mahajan 3-5, 6; Hayashi 458) Let $\hat{\theta}_n \equiv \mathrm{argmax}_{b \in \Theta} Q_n(\mathbf{W}, b) \equiv \mathrm{argmax}_{b \in \Theta} Q_n(b)$ as above. Suppose that:

1. True parameter $\theta \in \mathrm{interior}(\Theta)$;

2. Θ is a convex set;

3. $Q_n(b)$ is concave in b for any realization of the data $\mathbf{W}$ [will be true for MLE, since log-likelihood is concave, if only after a re-parametrization];

4. $Q_n(b)$ is a measurable function of the data for all $b \in \Theta$.

These conditions ensure that $\hat{\theta}_n$ is well-defined. Suppose there exists a function $Q_0(b)$ such that:

1. *Identification*: $Q_0(\cdot)$ is uniquely (globally) maximized on Θ at $\theta \in \Theta$;

2. *Pointwise convergence*: $Q_n(b) \xrightarrow{\mathrm{P}} Q_0(b)$ for all $b \in \Theta$.

Then $\hat{\theta}_n$ exists with probability approaching 1 and $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.

See Mahajan 3-6 for M-estimators.

Uniform (Weak) Law of Large Numbers (Mahajan 3-6; Hayashi 459) Suppose $\{W_i\}_i$ is ergodic stationary and that:

1. Θ is compact;

2. $q(W_i, b)$ is continuous in b for all W_i;

3. $q(W_i, b)$ is measurable in W_i for all b;

4. $\mathrm{E}_P[\sup_{b \in \Theta} |q(W_i, b)|] < \infty$.

Then $\frac{1}{n} \sum_i q(W_i, b)$ converges uniformly to $\mathrm{E}[q(W_i, b)]$, and $\mathrm{E}[q(W_i, b)]$ is a continuous function of b.

Asymptotically normal estimator (Mahajan 3-2, 13–4) The sequence of estimators $\{\hat{\theta}_n\}_i$ is $(\sqrt{n})$ asymptotically normal iff $\sqrt{n}(\hat{\theta}_n - \theta(P)) \xrightarrow{\mathrm{d}} \mathrm{N}(0, V(P))$ for some symmetric positive definite matrix $V(P)$ (somewhat inaccurately referred to as the asymptotic variance of $\hat{\theta}_n$).

Suppose that

1. $\hat{\theta}_n$ is consistent for θ;

2. $\theta \in \mathrm{interior}(\Theta)$;

3. $Q_n(b)$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ;

4. $\sqrt{n}\frac{\partial Q_n(\theta)}{\partial \theta} \xrightarrow{\mathrm{d}} \mathrm{N}(0, \boldsymbol{\Sigma})$;

5. *Uniform convergence of the Hessian:* There exists a matrix $H(b)$ that is continuous and nonsingular at θ such that

$$\sup_{b \in \mathcal{N}} \left\| \frac{\partial^2 Q_n(b)}{\partial \theta \, \partial \theta'} - H(b) \right\| \xrightarrow{\mathrm{P}} 0.$$

Then $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, H(\theta)^{-1} \boldsymbol{\Sigma} H(\theta)^{-1})$.

Asymptotic variance (C&B 10.1.9) If $k_n[\hat{\theta}_n - \theta] \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2)$ for some sequence of constants $\{k_n\}$, then σ^2 is the asymptotic variance.

Asymptotically efficient estimator (C&B 10.1.11–2) $\{\hat{\sigma}_n\}$ is asymptotically efficient if $\sqrt{n}[\hat{\theta}_n - \theta] \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2)$ and σ^2 is the CRLB.

Under regularity conditions, the MLE is consistent and asymptotically efficient.

Maximum Likelihood estimator (Mahajan 2-4–10, 36; Hayashi 448–9, 463–5) $\hat{\theta} \equiv \mathrm{argmax}_{b \in \Theta} L(\mathbf{X}, b)$. Equivalently, for iid data, $\hat{\theta} = \mathrm{argmax}_{b \in \Theta} \frac{1}{n} \sum \log p(X_i, b)$.* Estimating $\theta = \mathrm{argmax}_{b \in \Theta} \mathrm{E}_{P_\theta} \log p(\mathbf{X}, b)$. An M-Estimator with $q(X_i, b) \equiv -\log p(X_i, b)$. Note:

1. The identification condition is that the parameter being estimated is identified.

2. MLE need not always exist, and if they do, need not be unique.

3. We may not be able to get a closed-form expression for $\hat{\theta}$, and therefore have to characterize it as the solution to a maximization problem (or its FOCs).

4. The expected value of (log) likelihood is uniquely maximized at the true parameter as long as θ is identified; i.e., the Kullback-Liebler Divergence

$$K(b, \theta) \equiv \mathrm{E}_\theta \left[\log \left(\frac{p(\mathbf{X}, \theta)}{p(\mathbf{X}, b)} \right) \right] > 0 \text{ for all } b \neq \theta,$$

or equivalently, if $p(\mathbf{x}, b) = p(\mathbf{x}, \theta)$ for all $\mathbf{x}$ implies that $b = \theta$.

5. (Invariance property) If $\hat{\theta}$ is an MLE of θ, then $h(\hat{\theta})$ is an MLE of $h(\theta)$.

Consistency for for MLE (Hayashi 463–465; Mahajan 3-8–10) Let $\{y_t, \mathbf{x}_t\}$ be ergodic stationary, and let $\hat{\theta}$ be the conditional MLE that maximizes the average log conditional likelihood (derived under the assumption that $\{y_t, \mathbf{x}_t\}$ is iid).

Suppose conditions (specified on Hayashi 464) allow us to apply a general consistency theorem. Then $\hat{\theta} \xrightarrow{\mathrm{P}} \theta_0$ *despite* the fact that the MLE was derived under the iid assumption.

Asymptotic normality for MLE (Mahajan 3-20–2; C&B 10.1.12; Hayashi 474–6; D&M 258, 260–3, 270–4) Suppose $\{W_i\} \equiv \{Y_i, Z_i\}$ is an iid sample, that Z is ancillary for θ, that $\hat{\theta}_n \equiv \mathrm{argmax}_{b \in \Theta} \frac{1}{n} \sum \log p(Y_i | Z_i, b) \equiv \mathrm{argmax}_{b \in \Theta} Q_n(b)$. Define score $s(W_i, b) \equiv \frac{\partial \log p(Y_i | Z_i, b)}{\partial b}$ and Hessian $H(W_i, b) \equiv \frac{\partial s(W_i, b)}{\partial b} = \frac{\partial^2 \log p(Y_i | Z_i, b)}{\partial b \, \partial b'}$. Suppose:

1. $\hat{\theta}_n$ is consistent for θ—generally fails either because number of parameters increases with sample size, or model not asymptotically identified (even if it is identified by any finite sample);

2. $\theta \in \mathrm{interior}(\Theta)$;

3. $p(Y|Z, b)$ is twice continuously differentiable in b for any (Y, Z);

4. $\mathrm{E}[s(W, \theta)] = 0$ and $-\mathrm{E}[H(W, \theta)] = \mathrm{E}[s(W, \theta)s(W, \theta)']$ (this is stated as an assumption, but is the Information Equality and hence holds if its requirements do);

5. $\frac{1}{\sqrt{n}} \sum s(W_i, \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \boldsymbol{\Sigma})$ for some $\boldsymbol{\Sigma} > 0$;

6. $\mathrm{E}(\sup_{b \in \mathcal{N}} \|H(W, b)\|) < \infty$, which implies via ULLN that

$$\sup_{b \in \mathcal{N}} \left\| \frac{1}{n} \sum_{i=1}^n H(W_i, b) - \mathrm{E}[H(W, b)] \right\| \xrightarrow{\mathrm{P}} 0;$$

7. $\mathrm{E}[H(W, \theta)]$ is nonsingular (only required at true parameter).

Then $\hat{\theta}_n$ is asymptotically normal with variance $\mathcal{I}(\theta)^{-1}$ (note this is the Fisher information for *one* observation, not the joint distribution). The MLE is not necessarily unbiased, but in the limit the variance attains the CRLB hence MLE is *asymptotically efficient*. No GMM estimator can achieve lower variance.

Estimation of asymptotic variance can either be done by estimating the Hessian or the score (either gives a consistent expression using the Fisher Information Equality).

M-estimator (Mahajan 2-15–6; Hayashi 447) $\hat{\theta} \equiv \mathrm{argmin}_{b \in \Theta} \frac{1}{n} \sum q(X_i, b)$ (assuming iid data). Estimating $\theta = \mathrm{argmin}_{b \in \Theta} \mathrm{E}_P q(\mathbf{X}, b)$.

1. MLE is an M-Estimator with $q(X_i, b) \equiv -\log L(X_i, b)$.

*This is "quasi-ML" if used for non-iid data. It *can* be consistent even for (non-iid) ergodic stationary processes—see Hayashi 464–5.

2. Sample mean is an M-Estimator with $q(X_i, b) \equiv (X_i - b)^2$.

Asymptotic normality for M-estimator (Mahajan 3-14–7; Hayashi 470–4; D&M 593) Suppose $\{W_i\}$ is an iid sample, that $\hat{\theta}_n \equiv \text{argmax}_{b \in \Theta} \frac{1}{n} \sum q(W_i, b) \equiv \text{argmax}_{b \in \Theta} Q_n(b)$. Define "score" $s(W_i, b) \equiv \frac{\partial q(W_i, b)}{\partial b}$ and Hessian $H(W_i, b) \equiv \frac{\partial s(W_i, b)}{\partial b} = \frac{\partial^2 q(W_i, b)}{\partial b \partial b'}$. Suppose:

1. $\hat{\theta}_n$ is consistent for θ;
2. $\theta \in \text{interior}(\Theta)$;
3. $q(W, b)$ is twice continuously differentiable in b for any W;
4. $\frac{1}{\sqrt{n}} \sum s(W_i, \theta) \xrightarrow{d} N(0, \boldsymbol{\Sigma})$ for some $\boldsymbol{\Sigma} > 0$;[*]
5. $E(\sup_{b \in \mathcal{N}} \|H(W, b)\|) < \infty$, which implies via ULLN that

$$\sup_{b \in \mathcal{N}} \left\| \frac{1}{n} \sum_{i=1}^n H(W_i, b) - E[H(W, b)] \right\| \xrightarrow{p} 0;$$

6. $E[H(W, \theta)]$ is nonsingular (only required at true parameter).

Then $\hat{\theta}_n$ is asymptotically normal with variance $(E[H(W, \theta)])^{-1} \boldsymbol{\Sigma} (E[H(W, \theta)])^{-1}$.

"Under appropriate conditions for a ULLN to apply," this variance is estimated by:

$$\hat{V} = \left[\frac{1}{n} \sum H(W_i, \theta_n) \right]^{-1} \underbrace{\left[\frac{1}{n} \sum s(W_i, \theta_n) s(W_i, \theta_n)' \right]}_{\hat{\boldsymbol{\Sigma}}} \left[\frac{1}{n} \sum H(W_i, \theta_n) \right]^{-1}.$$

Method of Moments estimator (Mahajan 2-19–20) Suppose iid data from a parametric model with $\theta \in \mathbb{R}^d$ identified and the first d moments of P_θ exist: $\{m_j(\theta)\}_{j=1}^d \equiv \{E_\theta X^j\}_{j=1}^d$. Method of Moments estimator gives moments equal to sample moments:

$$m_j(\hat{\theta}) = \hat{m}_j \equiv \frac{1}{n} \sum_{i=1}^n X_i^j \text{ for all } j \in \{1, \ldots, d\}.$$

Generalized Method of Moments estimator (Mahajan 2-20–1; Hayashi 447, 468) GMM estimates parameter $\theta \in \mathbb{R}^d$ satisfying $E[\mathbf{m}(\mathbf{X}, \theta)] = 0$ where $\mathbf{m} \colon \mathcal{X} \times \Theta \to \mathbb{R}^m$ a vector of m moment conditions. If θ is identified, it is the unique solution to the moment conditions.

When $m \geq d$ we typically can't find a $\hat{\theta}$ satisfying all moment conditions, so instead we seek $\hat{\theta} = \text{argmin}_{b \in \Theta} Q_n(b)$ where

$$Q_n(b) \equiv \left(\frac{1}{n} \sum m(X_i, \theta) \right)' S \left(\frac{1}{n} \sum m(X_i, \theta) \right)$$

for some specified weight matrix $S_{m \times m}$ symmetric and positive definite. The quadratic form in S defines a norm. Correct specification requires orthogonality condition $E[m(X_i, \theta)] = 0$.

Extremum estimators can typically be thought of as GMM estimators when we characterize them by FOCs. Includes MLE and M-estimators that can be characterized by FOCs.

Consistency for GMM estimators (Mahajan 3-10–1; Hayashi 467–8) Let $\hat{\theta}_n \equiv \text{argmax}_{b \in \Theta} Q_n(b)$ (as above), where

$$Q_n(b) = -\frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^n m(X_i, b) \right)' S_n \left(\frac{1}{n} \sum_{i=1}^n m(X_i, b) \right).$$

The true parameter satisfies $E_P[m(W, \theta)] = 0$ and hence uniquely maximizes limit function $Q_0(b) = -\frac{1}{2} E(m(W, b))' S E(m(W, b))$. Suppose that:

1. Θ is a *compact* subset of $\mathbb{R}^d$;
2. $m(b)$ is continuous in b for any realization of the data;
3. $m(b)$ is a measurable function of the data for all $b \in \Theta$ (this ensures that $\hat{\theta}$ is measurable);
4. The weight matrices S_n converge in probability to some symmetric positive definite matrix S.

Suppose further that:

1. *Identification*: $E[m(W, b)] \neq 0$ for any $b \neq \theta$;
2. *Dominance to ensure ULLN applies*: $E[\sup_{b \in \Theta} \|m(W, b)\|] < \infty$.

Then $\hat{\theta}_n \xrightarrow{p} \theta$.

Showing identification and dominance for nonlinear GMM is quite difficult and usually just assumed. If objective function is concave, we can replace compact Θ; continuous, measurable m; and dominance by requirement that $E[m(W, b)]$ exist and be finite for all $b \in \Theta$.

Asymptotic normality for GMM estimator (Mahajan 3-24–6; Hayashi 478–81) Let $\hat{\theta}_n \equiv \text{argmax}_{b \in \Theta} -\frac{1}{2}[m_n(b)]' W_n [m_n(b)] \equiv \text{argmax}_{b \in \Theta} Q_n(b)$, where $m_n(b) \equiv \frac{1}{n} \sum m(X_i, b_{d \times 1})_{m \times 1}$. Jacobian $M_n(b)_{d \times m} \equiv \frac{1}{n} \sum \frac{\partial m(b)}{\partial b} = \frac{\partial m_n(b)}{\partial b}$. Suppose:

1. The matrix $M_n(\theta_n) W_n M_n(\bar{b}_n)$ is invertible;
2. $\sqrt{n} m_n(\theta) \xrightarrow{d} N(0, E[m(X, \theta) m(X, \theta)']) \equiv N(0, S(\theta))$ (by a CLT);
3. $M_n(b_n) \xrightarrow{p} E[\frac{\partial m(X, \theta)}{\partial b}] \equiv M(\theta)$ (by a ULLN);
4. $W_n \xrightarrow{p} W$.

Then $\hat{\theta}_n$ is asymptotically normal with variance

$$[M(\theta) W M(\theta)']^{-1} M(\theta) W S(\theta) W M(\theta)' [M(\theta) W M(\theta)']^{-1}.$$

If we choose $W = S(\theta)^{-1}$ (the efficient choice), the asymptotic variance reduces to $[M(\theta) S(\theta)^{-1} M(\theta)']^{-1}$.

Assuming conditions for ULLN apply, we can estimate terms using consistent estimators: $\hat{S} \equiv \frac{1}{n} \sum m(X_i, \hat{\theta}_n) m(X_i, \hat{\theta}_n)'$ and $\hat{M} \equiv \frac{1}{n} \sum \frac{\partial m(X_i, \hat{\theta}_n)}{\partial b}$.

Efficient GMM (Mahajan 3-26–7; Hayashi 212–3) Given above GMM estimator, if $W^* = S(\theta)^{-1}$ (the inverse of the variance of the moment conditions), then the asymptotic variance reduces to $V_e = [M(\theta) S(\theta)^{-1} M(\theta)']^{-1}$.

This is the lowest variance (in the matrix sense) that can be achieved. Therefore the optimal choice of weighting matrices is any sequence of random matrices that converge to S^{-1}. A natural choice is to develop preliminary estimates $\tilde{\theta}_n \xrightarrow{p} \theta$ (often using the identity as weighting matrix) and generating

$$W_n = \left[\frac{1}{n} \sum m(X_i, \tilde{\theta}_n) \sum m(X_i, \tilde{\theta}_n)' \right]^{-1}.$$

Under conditions necessary to implement ULLN, $W_n \xrightarrow{p} S^{-1}$.

Minimum Distance estimator (Mahajan 2-23–4) Estimator $\hat{\theta}_n \equiv \text{argmin}_{b \in \Theta} g_n(b)' S_n g_n(b)$ for some square weight matrix S_n. GMM is a special case where g_n are sample averages of some function.

Asymptotic normality for MD estimator (Mahajan 3-22–4) Let $\hat{\theta}_n \equiv \text{argmax}_{b \in \Theta} -\frac{1}{2}[g_n(b)]' W_n [g_n(b)] \equiv \text{argmax}_{b \in \Theta} Q_n(b)$. Define Jacobian [transpose of way we usually write derivatives?] $G_n(b) \equiv \frac{\partial g_n(b)}{\partial b}$. Suppose:

1. The matrix $G_n(\theta_n) W_n G_n(\bar{b}_n)$ is invertible (where $\bar{b}_n$ is a point between θ and $\hat{\theta}_n$ for which a mean value expansion holds);
2. $\sqrt{n} g_n(\theta) \xrightarrow{d} N(0, S(\theta))$ (by a CLT);
3. $G_n(b_n) \xrightarrow{p} G(\theta)$ (by a ULLN);
4. $W_n \xrightarrow{p} W$.

Then $\hat{\theta}_n$ is asymptotically normal with variance

$$[G(\theta) W G(\theta)']^{-1} G(\theta) W S(\theta) W G(\theta)' [G(\theta) W G(\theta)']^{-1}.$$

If we choose $W = S(\theta)^{-1}$ (the efficient choice), the asymptotic variance reduces to $[G(\theta) S(\theta)^{-1} G(\theta)']^{-1}$.

Uniformly minimum variance unbiased estimator (Mahajan 2-27–9, Metrics P.S. 6-1; C&B 7.3.7, 17, 19–20, 23, 7.5.1) An unbiased estimator $\phi(X)$ of a quantity $g(\theta)$ is a UMVUE (a.k.a. best unbiased estimator) iff ϕ has finite variance and for every unbiased estimator $\delta(X)$ of $g(\theta)$, we have $\text{Var } \phi(X) \leq \text{Var } \delta(X)$ for all θ. Note:

---

[*]If $\{W_i\}$ is non-iid ergodic stationary, then $\boldsymbol{\Sigma}$ is the long-run variance of $\{s(W_i, \theta)\}$; Gordin's conditions are sufficient for this convergence.

1. Unbiased estimators may not exist;

2. Not every unbiased estimator is a UMVUE;

3. If a UMVUE exists, it is unique;

4. (Rao-Blackwell) If $h(X)$ is an unbiased estimator of $g(\theta)$, and $T(X)$ is a sufficient statistic for θ, then $\phi(T) \equiv \mathrm{E}[h(X)|T]$ is unbiased for $g(\theta)$, and has variance (weakly) lower than $h(X)$ for all θ—means we only need consider statistics that are functions of the data only through sufficient statistics;

5. If $\phi(T)$ is an unbiased estimator of $g(\theta)$ and is a function of a complete statistic $T(X)$, then all other unbiased estimators that are functions of T are equal to $\phi(T)$ almost everywhere;

6. (Lehmann-Scheffé) If $\phi(T)$ is a function (only) of a complete statistic $T(X)$, then $\phi(T)$ is the unique UMVUE of $\mathrm{E}\,\phi(T)$;

7. (Hausman Principle) W is a UMVUE for $\mathrm{E}\,W$ iff it is uncorrelated with every unbiased estimator of 0 (practically, impossible to prove, except for the case where W is a function only of a complete statistic).

Fisher Information (Mahajan 2-30–1, Metrics P.S. 6-3)

$$\mathcal{I}(\theta) \equiv \mathrm{E}_\theta \left[\underbrace{\left(\tfrac{\partial}{\partial\theta} \log f(x,\theta)\right)}_{\text{Score}} \left(\tfrac{\partial}{\partial\theta} \log f(x,\theta)\right)' \right].$$

Fisher information for an iid sample is n times information for each individual observation. For (univariate) normal,

$$\mathcal{I}(\mu,\sigma^2) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \tfrac{1}{2}\sigma^{-4} \end{bmatrix}; \quad \mathcal{I}(\mu,\sigma^2)^{-1} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}.$$

Cramér-Rao Inequality (Mahajan 2-30–1, 5; C&B 7.3.9–11, 15) Given a sample $\mathbf{X} \sim f(\mathbf{x}|\theta)$ and an estimator $\phi(\mathbf{X})$ with $\mathrm{E}_\theta\,\phi(\mathbf{X}) = g(\theta)$, suppose that

1. The support does not depend on θ;

2. pdf $f(\mathbf{x}|\theta)$ is differentiable in θ almost everywhere;

3. $\mathrm{E}_\theta\,|\phi| < \theta$ (or per C&B, $\mathrm{Var}_\theta\,\phi < \infty$);

4. The operations of differentiation and integration can be interchanged in $\frac{d}{d\theta}\int \phi(\mathbf{x})f(\mathbf{x},\theta)\,d\mathbf{x}$;

5. Fisher Information $\mathcal{I}(\theta)$ is nonsingular. Note under previous conditions,

   (a) $\mathcal{I}(\theta) = \mathrm{Var}_\theta[\frac{\partial}{\partial\theta}\log f(\mathbf{x},\theta)]$;

   (b) If $f(\cdot,\theta)$ is twice differentiable and double integration and differentiation under the integral sign can be interchanged, then (Fisher Information Equality):

   $$\mathcal{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta\,\partial\theta'} \log f(\mathbf{x},\theta)\right].$$

Then

$$\mathrm{Var}_\theta\,\phi(\mathbf{X}) \geq \left(\frac{dg(\theta)}{d\theta}\right)' \mathcal{I}(\theta)^{-1}\left(\frac{dg(\theta)}{d\theta}\right).$$

Attained iff there exists a function $a(\theta)$ such that

$$a(\theta)[\phi(\mathbf{x}) - g(\theta)] = \tfrac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta).$$

1.11 Decision theory

Loss function (Mahajan 2-24–5) We observe $\mathbf{x} \in \mathcal{X}$ with unknown distribution $P_\theta \in \mathcal{P}$. Loss function $l\colon \mathcal{P} \times \mathcal{A} \to \mathbb{R}^+$ (where $\mathcal{A}$ is the action space), $l(P_b, a)$ gives the loss that occurs if the statistician chooses action a and the true distribution is P_b. Common examples when estimating parameter $v(P)$ include

1. Quadratic loss: $l(P, a) \equiv [v(P) - a]^2$;

2. Absolute value loss: $l(P, a) \equiv |v(P) - a|$;

3. 0-1 loss: action space is $\{0, 1\}$ and $l(P_\theta, a)$ is zero if $\theta \in \Theta_a$, one otherwise.

Decision rule (Mahajan 2-25) A mapping from the sample space $\mathcal{X}$ to the action space $\mathcal{A}$.

Risk function (Mahajan 2-25) Expected loss (expectation is taken using the *true* distribution): $R(P_\theta, \phi) = \mathrm{E}_\theta[l(P_\theta, \phi)]$.

Mean squared error risk function (Mahajan 2-25–6; Greene 109–11; C&B 7.3.1) If loss function $l(\cdot, \cdot)$ is quadratic loss, risk function is MSE risk function: $R(P_\theta, \phi) \equiv \mathrm{E}_\theta[\phi(\mathbf{X}) - g(\theta)]^2$, where $g(\theta)$ is the parameter being estimated. Note $R(P_\theta, \phi) = \mathrm{Bias}_\theta[\phi(\mathbf{X})]^2 + \mathrm{Var}_\theta[\phi(\mathbf{X})]$.

Typically it is infeasible to minimize MSE, since it depends on the (presumably unknown) parameter $g(\theta)$. So we often use UMVUE instead.

1.12 Hypothesis testing

Hypothesis testing (Mahajan 4-1–2) We observe data $\mathbf{x}$ from distribution $P \in \mathcal{P}$ (usually $\{P_\theta\colon \theta \in \Theta\}$), and must decide whether $P \in \mathcal{P}_H \subseteq \mathcal{P}$. The null hypothesis H is that $P \in \mathcal{P}_H \subseteq \mathcal{P}$, or equivalently $\theta \in \Theta_H$. If $\mathcal{P}_H$ is a singleton we call H simple, otherwise composite (identically for K). The alternate hypothesis K is $P \in \mathcal{P}_K \subseteq \mathcal{P}$ (or $\theta \in \Theta_K$), where $\mathcal{P}_H \cap \mathcal{P}_K = \varnothing$ and we assume the maintained hypothesis $P \in \mathcal{P}_H \cup \mathcal{P}_K$.

Test statistic (Mahajan 4-2) Test statistic (a.k.a. test function) is a decision rule $\delta\colon \mathcal{X} \to \{0, 1\}$, where 1 corresponds to rejecting H, and 0 to accepting it. Typically we evaluate a test over this action space using the 0-1 loss function.

Equivalently defined by a critical region in the sample space $C \equiv \{\mathbf{x} \in \mathcal{X}\colon \delta(\mathbf{x}) = 1\}$.

Type I/Type II error (Mahajan 4-2) Type I error: rejecting H when in fact $\theta \in \Theta_H$. Type II error: accepting H when $\theta \in \Theta_K$.

Power (Mahajan 4-2–3) $\beta_\delta(\theta) \equiv P_\theta(\delta(\mathbf{X}) = 1)$, for $\theta \in \Theta_K$. The chance of (correctly) rejecting H given that the true parameter is $\theta \in \Theta_K$. One minus the probability of Type II error at a given θ.

Size (Mahajan 4-3; C&B 8.3.5) For $\theta \in \Theta_H$, the power function $\beta_\delta(\theta) \equiv P_\theta(\delta(\mathbf{X}) = 1)$ gives the probability of a Type I error (rejecting H incorrectly). Size is the largest probability of this occurring: $\sup_{\theta \in \Theta_H} \beta_\delta(\theta)$.

Level (Mahajan 4-3; C&B 8.3.6) A test is called level α for some $\alpha \in [0, 1]$ if its size is at most α.

Distance function principle (Mahajan 4-7–8) When the hypothesis is framed in terms of a parameter θ for which we have a good estimate $\hat\theta$, it may be reasonable to reject if the distance between estimate and $\theta \in \Theta_H$ is large; i.e., reject if $T(\mathbf{x}) \equiv \inf_{b \in \Theta_H} d(\hat\theta, b) > k$, where $d(\cdot, \cdot)$ is some distance function.

p-value (Mahajan 4-8–9; C&B 8.3.26) The smallest level of significance at which a researcher using a given test would reject H on the basis of observed data $\mathbf{x}$. The magnitude of the P-value can be interpreted as a measure of the strength of the evidence for H.

C&B offers a different conception: a test statistic with $p(\mathbf{x}) \in [0, 1]$ for all $\mathbf{x} \in \mathcal{X}$, and $P_\theta(p(\mathbf{x}) \leq \alpha) \leq \alpha$ for every $\theta \in \Theta_H$ and every $\alpha \in [0, 1]$.

If we observe an iid sample from normal with known variance and use $T(\mathbf{x}) = \sqrt{n}\bar{X}/\sigma$, the P-value is $\Phi(-\sqrt{n}\bar{X}/\sigma)$.

Uniformly most powerful test (Mahajan 4-10, 19; C&B 8.3.11) Let $\mathcal{C}$ be a set of tests. A test $c^* \in \mathcal{C}$ with power function $\beta_*(\theta)$ is a UMP class $\mathcal{C}$ test of H against $K\colon \theta \in \Theta_K$ iff for every $c \in \mathcal{C}$ and every $\theta \in \Theta_K$, we have $\beta_*(\theta) \geq \beta_c(\theta)$. That is, for every $\theta \in \Theta_K$, the test c^* is the most powerful class $\mathcal{C}$ test.

In many cases, UMP level α tests do not exist.

Simple likelihood ratio statistic (Mahajan 4-10) For simple null and alternate hypotheses, $L(\mathbf{x}, \theta_H, \theta_K) \equiv p(\mathbf{x}|\theta_K)/p(\mathbf{x}|\theta_H)$ (where p is either a pdf or pmf). By convention, $L(\mathbf{x}, \theta_H, \theta_K) = 0$ when both numerator and denominator are zero.

Neyman-Pearson Lemma (Mahajan 4-12; C&B 8.3.12) Consider testing $H\colon \theta = \theta_0$ against $K\colon \theta = \theta_1$ where pdf/pmf is $p(\cdot|\theta)$. Consider (randomized) likelihood ratio test function:

$$\phi_k(\mathbf{x}) = \begin{cases} 1, & L(\mathbf{x}, \theta_0, \theta_1) > k; \\ 0, & L(\mathbf{x}, \theta_0, \theta_1) < k; \\ \gamma \in (0, 1), & L(\mathbf{x}, \theta_0, \theta_1) = k. \end{cases}$$

1. If ϕ_k is a size α (> 0) test (i.e., $E_{\theta_0}[\phi_k(\mathbf{X})] = \alpha$), then ϕ_k is the most powerful in the class of level α tests.

2. For each $\alpha \in [0,1]$ there exists a MP size α likelihood ratio test of the form ϕ_k.

3. If a test $\tilde{\phi}$ is a MP level α test, then it must be a level α likelihood ratio test. That is, there exists k and γ such that for any $\theta \in \{\theta_0, \theta_1\}$, we have $P_\theta(\tilde{\phi}(\mathbf{X}) \neq \phi(\mathbf{X})) = 0$.

Corollary: Power of a MP test $\geq$ its size. Consider the test that rejects for all data with probability α; it has power α, so MP test can't do worse.

Monotone likelihood ratio models (Mahajan 4-16–8; C&B 8.3.16) $\{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}$ is a MLR family if $L(\mathbf{x}, \theta_1, \theta_2)$ is a monotone function of some function of the data $T(\mathbf{x})$ in the same direction for all pairs $\theta_2 > \theta_1$. If $L(\mathbf{x}, \theta_1, \theta_2)$ is increasing (decreasing) in $T(\mathbf{x})$, the family is increasing (decreasing) MLR in $T(\mathbf{x})$.

In the one parameter exponential family $f(\mathbf{x}|\theta) = h(\mathbf{x}) \exp[\eta(\theta)T(\mathbf{x}) - B(\theta)]$,

$L(\mathbf{x}, \theta_1, \theta_2) = \exp\left[\left(\eta(\theta_2) - \eta(\theta_1)\right)T(\mathbf{x}) - \left(B(\theta_2) - B(\theta_1)\right)\right]$;

and if $\eta(\theta)$ is strictly increasing in $\theta \in \Theta$, the family is MLR increasing in $T(\mathbf{x})$.

Suppose $\{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}$ is MLR increasing in $T(\mathbf{x})$. Let $\delta_t(\cdot)$ be the test that rejects if $T(\mathbf{x}) > t$ (and possibly with some random probability if $T(\mathbf{x}) = t$). Then,

1. The power function $\beta_{\delta_t}(\theta)$ is increasing in θ for all $\theta \in \Theta$ and all t;

2. If $E_0[\delta_t(\mathbf{X})] = \alpha > 0$, then $\delta_t(\cdot)$ is UMP level α for testing $H : \theta \leq \theta_0$ against $K : \theta > \theta_0$.

Unbiased test (Mahajan 4-20; C&B 8.3.9; Greene 157–8) A test ϕ for $H : \theta \in \Theta_H$ against $K : \theta \in \Theta_K$ is unbiased of level α iff

1. $\beta_\phi(\theta) \leq \alpha$ for all $\theta \in \Theta_H$ (i.e., level α); and

2. $\beta_\phi(\theta) \geq \alpha$ for all $\theta \in \Theta_K$ (i.e., power $\geq \alpha$ everywhere).

If ϕ is biased, then there is some $\theta_H \in \Theta_H$ and $\theta_K \in \Theta_K$ such that $\beta_\phi(\theta_H) > \beta_\phi(\theta_K)$; i.e., we are more likely to reject under some true value than some false value.

Uniformly most powerful unbiased test (Mahajan 4-20–3) A test ϕ^* for $H : \theta \in \Theta_H$ against $K : \theta \in \Theta_K$ is UMPU of level α iff it is unbiased at level α and it is more powerful (for all $\theta \in \Theta_K$) than any other unbiased test of level α.

The first condition is not strictly necessary, since the constant randomized level α test is unbiased. If ϕ is UMP level α, it will also be UMPU level α.

See Mahajan 4-23 for UMPU tests in one parameter exponential families.

Likelihood ratio test (Mahajan 4-24–5, 34–6; Greene 159–67; C&B 10.3.1, 3; D&M 275) Test that rejects for large values of

$$L(\mathbf{X}) \equiv \frac{\sup_{\theta \in \Theta_K} p(\mathbf{x}, \theta)}{\sup_{\theta \in \Theta_H} p(\mathbf{x}, \theta)}.$$

Because distribution properties are often complex (and because Θ_H is often of smaller dimension than $\Theta = \Theta_H \cup \Theta_K$, which means sup over Θ_K equals sup over Θ), we often use

$$\lambda(\mathbf{x}) \equiv \frac{\sup_{\theta \in \Theta} p(\mathbf{x}, \theta)}{\sup_{\theta \in \Theta_H} p(\mathbf{x}, \theta)} = \frac{p(\mathbf{x}, \hat{\theta})}{p(\mathbf{x}, \hat{\theta}_C)}$$

where $\hat{\theta}$ is the MLE and $\hat{\theta}_C$ is the constrained MLE.

We will generally base our test on some monotonic function of $\lambda(\mathbf{x})$. Often $2\log\lambda(\mathbf{x}) = 2[\log p(\mathbf{x}, \hat{\theta}) - \log p(\mathbf{x}, \hat{\theta}_C)]$. Written more generally, we often have

$$2n[Q_n(\hat{\theta}) - Q_n(\hat{\theta}_C)] \xrightarrow{d} \chi_r^2,$$

where r is the number of restrictions (and where here, $Q_n(b) \equiv \frac{1}{n}\sum \log p(x_i, b)$).

Wald statistic (Hayashi 489–91; Greene 159–67; Mahajan 36–8; D&M 278) Consider an estimator $\hat{\theta}_n \equiv \text{argmax}_{b \in \Theta} Q_n(\mathbf{W}, b)$ satisfying:

1. $\sqrt{n}(\hat{\theta} - \theta_0)$ has Taylor expansion

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\Psi^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + o_p;$$

2. $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(\mathbf{0}, \Sigma)$ for some positive definite $\boldsymbol{\Sigma}$;

3. $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution (to *something*);

4. $\Sigma = -\Psi$.

These conditions are satisfied for ML and efficient GMM under certain conditions.

Under null H_0: $\mathbf{a}_{r \times 1}(\theta_0) = 0$, with $\mathbf{A}(\theta) \equiv \frac{\partial \mathbf{a}(\theta)}{\partial \theta'}$ and $\mathbf{A}(\theta_0)$ of full row rank (no redundant restrictions),

$$W \equiv n\mathbf{a}(\hat{\theta})'[\mathbf{A}(\hat{\theta})\widehat{\Sigma}^{-1}\mathbf{A}(\hat{\theta})']^{-1}\mathbf{a}(\hat{\theta}) \xrightarrow{d} \chi_r^2.$$

Caution: Σ is the *inverse* of the asymptotic variance of the estimator.

Note the restricted estimator doesn't enter calculation of the Wald statistic, so the size of resulting tests may be limited. Also, the Wald statistic (unlike the LR or LM statistics) is *not* invariant to the restriction formulation.

Lagrange multiplier statistic (Hayashi 489, 491–93; Greene 159–67; Mahajan 38–9; D&M 275–8) Consider an estimator $\hat{\theta}_n \equiv \text{argmax}_{b \in \Theta} Q_n(\mathbf{W}, b) \in \mathbb{R}^p$ satisfying:

1. $\sqrt{n}(\hat{\theta} - \theta_0)$ has Taylor expansion

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\Psi^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + o_p;$$

2. $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(\mathbf{0}, \Sigma)$ for some positive definite $\boldsymbol{\Sigma}$;

3. $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution (to *something*);

4. $\Sigma = -\Psi$.

These conditions are satisfied for ML and efficient GMM under certain conditions.

Under null H_0: $\mathbf{a}_{r \times 1}(\theta_0) = 0$, with $\mathbf{A}(\theta) \equiv \frac{\partial \mathbf{a}(\theta)}{\partial \theta'}$ and $\mathbf{A}(\theta_0)$ of full row rank (no redundant restrictions), define the constrained estimator as $\hat{\theta}_c \equiv \text{argmax}_{b \in \Theta} Q_n(\mathbf{W}, b)$ s.t. $\mathbf{a}(b) = 0$. Then

$$LM \equiv n\underbrace{\left(\frac{\partial Q_n(\hat{\theta}_c)}{\partial \theta}\right)'}_{1 \times p} \underbrace{\widetilde{\Sigma}^{-1}}_{p \times p} \underbrace{\left(\frac{\partial Q_n(\hat{\theta}_c)}{\partial \theta}\right)}_{1 \times p} \xrightarrow{d} \chi_r^2$$

where $\widetilde{\Sigma}$ is a consistent estimator *under the null* (i.e., constructed using $\hat{\theta}_c$).

Asymptotic significance (Mahajan 39) A test $\delta_n(\cdot)$ has asymptotic significance level α iff $\lim_{n \to \infty} P_P(\delta_n(\mathbf{X}) = 1) \leq \alpha$ for all $P \in \mathcal{P}_0$.

Asymptotic size (Mahajan 39) Asymptotic size (a.k.a. limiting size) of a test $\delta_n(\cdot)$ is $\lim_{n \to \infty} \sup_{P \in \mathcal{P}_0} P_P(\delta_n(\mathbf{X}) = 1)$.

Consistent test (Mahajan 39) A test $\delta_n(\cdot)$ is consistent iff $\lim_{n \to \infty} P_P(\delta_n(\mathbf{X}) = 1) = 1$ for all $P \in \mathcal{P}_1$.

1.13   Time-series concepts

iid white noise $\implies$ stationary mds with finite variance $\implies$ white noise.

Stationarity (Hayashi 98–9; Hansen B1-11; D&M 132) $\{z_i\}$ is (strictly) stationary iff for any finite set of subscripts $i_1, \ldots, i_r$, the joint distribution of $(z_i, z_{i_1}, z_{i_2}, z_{i_r})$ depends only on $i_1 - i, \ldots, i_r - i$ but not on i.

Any (measurable) transformation of a stationary process is stationary.

Covariance stationarity (Hayashi 99–100, 401; Hansen B1-11) $\{z_i\}$ is covariance- (or weakly) stationary iff:

1. $\mu \equiv E[z_i]$ does not depend on i, and

2. $\Gamma_j \equiv \text{Cov}[z_i, z_{i-j}]$ (the j-th order autocovariance) exists, is finite, and depends only on j but not on i.

A strictly stationary process is covariance-stationary as long as the variance and covariances are finite.

LLN for covariance-stationary processes with vanishing autocovariances:

1. If $\lim_{j\to\infty}\gamma_j = 0$, then $\bar{y} \xrightarrow{L_2} \mu$;

2. If $\sum_{j\in\mathbb{Z}}\gamma_j < \infty$, then $\lim_{n\to\infty}\mathrm{Var}(\sqrt{n}\bar{y}) = \sum_{j\in\mathbb{Z}}\gamma_j < \infty$ (called the "long-run variance").

White noise (Hayashi 101; Hansen B1-28) $\{\varepsilon_i\}$ is white noise iff:

1. It is covariance-stationary,

2. $\mathrm{E}[\varepsilon_i] = 0$ for all i (zero mean), and

3. $\Gamma_j \equiv \mathrm{Cov}[\varepsilon_i, \varepsilon_{i-j}] = \mathbf{0}$ for $j \neq 0$ (no serial correlation).

An iid sequence with mean zero and finite variance is an "independent" (i.e., iid) white noise process.

Ergodicity (Hayashi 101–2, 402–5; Hansen B1-12, 26; D&M 132–3) $\{z_i\}$ is ergodic iff it is stationary and for any two bounded functions $f\colon \mathbb{R}^k \to \mathbb{R}$ and $g\colon \mathbb{R}^l \to \mathbb{R}$,

$$\lim_{n\to\infty} \big| \mathrm{E}[f(z_i,\ldots,z_{i+k}) \cdot g(z_{i+n},\ldots,z_{i+n+l})] \big|$$
$$= \big| \mathrm{E}[f(z_i,\ldots,z_{i+k})] \big| \cdot \big| \mathrm{E}[g(z_{i+n},\ldots,z_{i+n+l})] \big|.$$

(Note the RHS needs no limit in n, since by stationarity it doesn't affect $g(\cdot)$.) Intuitively, represents "asymptotic independence": two elements sufficiently far apart are nearly independent.

A sequence of the form $\{Z(z_i,\ldots,z_{i+k})\}$ is also ergodic.

Ergodic Theorem is LLN for ergodic processes: if $\{z_i\}$ ergodic stationary with $\mathrm{E}[z_i] = \mu$, then $\bar{z}_n \equiv \frac{1}{n}\sum_i z_i \xrightarrow{as} \mu$.

Gordin's CLT for zero-mean* ergodic stationary processes: Suppose $\{z_i\}$ is ergodic stationary and

1. $\mathrm{E}[z_t z_t']$ exists and is finite;

2. $\mathrm{E}[z_t | z_{t-j}, z_{t-j-1}, \ldots] \xrightarrow{L_2} 0$ as $j \to \infty$ (i.e., knowing about the distant past does you no good in estimating today);

3. $\sum_{j=0}^{\infty}\sqrt{\mathrm{E}[r_{tj}' r_{tj}]} < \infty$, where $r_{tj} \equiv \mathrm{E}[z_t | z_{t-j}, z_{t-j-1}, \ldots] - \mathrm{E}[z_t | z_{t-j-1}, z_{t-j-2}, \ldots]$ are the "revisions" to conditional expectations as the information set increases.

Then

1. $\mathrm{E}[z_t] = 0$;

2. Autocovariances are absolutely summable (i.e., $\sum_{j\in\mathbb{Z}}|\Gamma_j| < \infty$);

3. $\bar{z}$ is asymptotically normal with long-run variance as in covariance-stationary processes with vanishing covariances:
$$\sqrt{n}\bar{z} \xrightarrow{d} \mathrm{N}\bigg(0, \sum_{j\in\mathbb{Z}}\Gamma_j\bigg).$$

Martingale (Hayashi 102–3; Hansen B1-14; D&M 133) $\{x_i\}$ is a martingale with respect to $\{z_i\}$ (where $x_i \in z_i$) iff $\mathrm{E}[x_i | z_{i-1}, z_{i-2}, \ldots, z_1] = x_{i-1}$ for $i \geq 2$.

Random walk (Hayashi 103; Hansen B1-14) An example of a martingale. Let $\{g_i\}$ be an iid white noise process. The sequence of cumulative sums $\{z_i\}$ where $z_i = g_1 + \cdots + g_i$ is a random walk.

Martingale difference sequence (Hayashi 104; D&M 134; Hansen B1-14) $\{g_i\}$ is an mds iff $\mathrm{E}[g_i | g_{i-1}, g_{i-2}, \ldots, g_1] = 0$ for $i \geq 2$.

The cumulative sum of an mds is a martingale; conversely, the first differences of a martingale are an mds. A mds has no serial correlation (i.e., $\mathrm{Cov}(g_i, g_j) = 0$ for all $i \neq j$).

Ergodic stationary martingale differences CLT (Hayashi 106–7; Hansen B1-15–6) Given a stationary ergodic mds† $\{g_i\}$ with $\mathrm{E}[g_i g_i'] = \Sigma$,
$$\sqrt{n}\bar{g} \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i \xrightarrow{d} \mathrm{N}(0, \Sigma).$$

Lag operator (Hayashi 369–74; Hansen B1-31–3) $L^j x_t \equiv x_{t-j}$. Filter $\alpha(L) \equiv \alpha_0 + \alpha_1 L + \alpha_2 L^2 + \cdots$. Filter applied to a constant is $\alpha(L)c = \alpha(1)c = c\sum_{j=0}^{\infty}\alpha_j$.

1. If coefficients are absolutely summable ($\sum|\alpha_j| < \infty$), and $\{x_t\}$ is covariance-stationary, then $y_t \equiv \alpha(L)x_t$ converges in L_2 and is covariance-stationary; if autocovariances of $\{x_t\}$ are absolutely summable, then so are autocovariances of $\{y_t\}$.

2. As long as $\alpha_0 \neq 0$, the inverse $\alpha(L)^{-1}$ is well defined.

3. If $\alpha(z) = 0 \implies |z| > 1$ (the "stability condition"), then the coefficients of $\alpha(L)^{-1}$ are absolutely summable.

Moving average process (Hayashi 366–8, 402; Hansen B1-28–9, 36–7) $y_t = \mu + \sum_{j=0}^{\infty}\psi_j \varepsilon_{t-j}$, with $\{\varepsilon_t\}$ white noise and $\mathrm{Var}\,\varepsilon_t = \sigma_\varepsilon^2$. Equivalently, $(y_t - \mu) = \psi(L)\varepsilon_t$. Given absolute summability $\sum_{j\in\mathbb{Z}}|\psi_j| < \infty$, a sufficient condition for convergence (in L_2),

1. $\mathrm{E}[y_t] = \mu$;

2. Autocovariances $\gamma_j = \sigma_\varepsilon^2 \sum_{k=0}^{\infty}\psi_{j+k}\psi_k$ (if we have a MA(q) process, $\gamma_j = (\psi_j\psi_0 + \psi_{j+1}\psi_1 + \cdots + \psi_q\psi_{q-j})\sigma_\varepsilon^2$ for $|j| \leq q$);

3. Autocovariances γ_j are absolutely summable (i.e., $\sum_{j\in\mathbb{Z}}|\gamma_j| < \infty$);

4. Long-run variance $\sum_{j\in\mathbb{Z}}\gamma_j = \sigma_\varepsilon^2[\psi(1)]^2$.

5. If $\{\varepsilon_t\}$ is iid white noise, $\{y_t\}$ is ergodic stationary.

CLT for MA(∞): If $\{\varepsilon_t\}$ is *iid* white noise, and absolute summability $\sum_{j\in\mathbb{Z}}|\psi_j| < \infty$ holds, then
$$\sqrt{n}[\bar{y} - \mu] \xrightarrow{d} \mathrm{N}\bigg(0, \sum_{j\in\mathbb{Z}}\gamma_j\bigg) = \mathrm{N}\big(0, \sigma_\varepsilon^2[\psi(1)]^2\big).$$

Autoregressive process of degree one (Hayashi 376–8, 385) $y_t = c + \phi y_{t-1} + \varepsilon_t$, with $\{\varepsilon_t\}$ white noise and $\mathrm{Var}\,\varepsilon_t = \sigma_\varepsilon^2$. Equivalently, $(1 - \phi L)y_t = c + \varepsilon_t$, or $(1 - \phi L)(y_t - \mu) = \varepsilon_t$ where $\mu \equiv c/(1 - \phi)$.

1. If $|\phi| < 1$, then the unique covariance-stationary solution is the MA(∞) $y_t = \mu + \sum_{j=0}^{\infty}\phi^j \varepsilon_{t-j}$; it has absolutely summable coefficients.

   - $\mathrm{E}\,y_t = \mu$,
   - $\gamma_j = \phi^j \sigma_\varepsilon^2/(1 - \phi^2)$,
   - Autocovariances γ_j are absolutely summable,
   - $\rho_j \equiv \gamma_j/\gamma_0 = \phi^j$,
   - Long-run variance $\sum_{j\in\mathbb{Z}}\gamma_j = \sigma_\varepsilon^2/(1 - \phi)^2$.

2. If $|\phi| > 1$, then the unique covariance-stationary solution is a MA(∞) of *future* values of ε.

3. If $|\phi| = 1$ ("unit root"), there is no covariance-stationary solution.

Autoregressive process of degree p (Hayashi 378–80, 385)
$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$
with $\{\varepsilon_t\}$ white noise and $\mathrm{Var}\,\varepsilon_t = \sigma_\varepsilon^2$. Equivalently, $\phi(L)y_t = c + \varepsilon_t$ where $\phi(L) \equiv 1 - \phi_1 L - \cdots - \phi_p L^p$ (note the minus signs). Equivalently, $\phi(L)(y_t - \mu) = \varepsilon_t$ where $\mu \equiv c/\phi(1) = c/(1 - \sum \phi_j)$.

Assuming stability (i.e., $\phi(z) = 0 \implies |z| > 1$),

1. The unique covariance-stationary solution is the MA(∞) $y_t = \mu + \phi(L)^{-1}\varepsilon_t$; it has absolutely summable coefficients;

2. $\mathrm{E}\,y_t = \mu$;

3. Autocovariances γ_j are absolutely summable;

4. Long-run variance $\sum_{j\in\mathbb{Z}}\gamma_j = \sigma_\varepsilon^2/[\phi(1)]^2$.

ARMA process of degree (p, q) (Hayashi 380–3, 385; Hansen B1-30, 43–7) Autoregressive/moving average process with p autoregressive lags and q moving average lags:
$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_0 \varepsilon_t + \cdots + \theta_q \varepsilon_{t-q},$$
with $\{\varepsilon_t\}$ white noise and $\mathrm{Var}\,\varepsilon_t = \sigma_\varepsilon^2$. Equivalently,
$$\phi(L)y_t = c + \theta(L)\varepsilon_t$$

---

*Hansen B1-26 gives a slightly different statement for nonzero-mean ergodic stationary processes.

†This is actually a corollary of Billingsley's stronger CLT (which does *not require* ergodic stationarity) stated in Hansen B1-16.

where $\phi(L) \equiv 1 - \phi_1 L - \cdots - \phi_p L^p$ (note the minus signs) and $\theta(L) \equiv \theta_0 + \theta_1 L + \cdots + \theta_q L^q$. Equivalently,

$$\phi(L)(y_t - \mu) = \theta(L)\varepsilon_t$$

where $\mu \equiv c/\phi(1) = c/(1 - \sum \phi_j)$. Note if $\phi(z)$ and $\theta(z)$ have a common root, we can factor it out and get an ARMA$(p - 1, q - 1)$.

Assuming stability (i.e., $\phi(z) = 0 \implies |z| > 1$),

1. The unique covariance-stationary solution is the MA(∞) $y_t = \mu + \phi(L)^{-1}\theta(L)\varepsilon_t$; it has absolutely summable coefficients;

2. $\mathrm{E}\, y_t = \mu$;

3. Autocovariances γ_j are absolutely summable.

If $\theta(z) = 0 \implies |z| > 1$ (here called the "invertability condition"), we can write as a AR(∞): $\theta(L)^{-1}\phi(L)y_t = (c/\theta(1)) + \varepsilon_t$.

For ARMA$(1, 1)$ $y_t = c + \phi y_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}$, long-run variance is

$$\sum_{j\in\mathbb{Z}} \gamma_j = \sigma_\varepsilon^2 \left(\frac{1-\theta}{1-\phi}\right)^2.$$

For ARMA(p, q) $y_t = c + [\sum_i \phi_i y_{t-i}] + \varepsilon_t - [\sum_j \theta_j \varepsilon_{t-j}]$, long-run variance is

$$\sum_{j\in\mathbb{Z}} \gamma_j = \sigma_\varepsilon^2 \frac{[\theta(1)]^2}{[\phi(1)]^2} = \sigma_\varepsilon^2 \left(\frac{1 - \sum_j \theta_j}{1 - \sum_i \phi_i}\right)^2.$$

Autoregressive conditional heteroscedastic process
(Hayashi 104–5; Hansen B1–38) ARCH(q) is a martingale difference sequence $\{g_i\}$ with conditional variance $\mathrm{Var}[g_i|g_{i-1}, \ldots, g_1] = \zeta + \alpha_1 g_{i-1}^2 + \cdots + \alpha_q g_{i-q}^2$. For example, ARCH$(1)$ is:

$$g_i = \varepsilon_i \sqrt{\zeta + \alpha g_{i-1}^2}$$

and $\{\varepsilon_i\}$ iid with $\mathrm{E}\,\varepsilon_i = 0$ and $\mathrm{Var}\,\varepsilon_i = 1$.

1. $\mathrm{E}[g_i|g_{i-1}, \ldots, g_1] = 0$;

2. $\mathrm{Var}[g_i|g_{i-1}, \ldots, g_1] = \mathrm{E}[g_i^2|g_{i-1}, \ldots, g_1] = \zeta + \alpha g_{i-1}^2$, which depends on the history of the process (own conditional heteroscedasticity);

3. Strictly stationary and ergodic if $|\alpha| < 1$ and g_1 drawn from an appropriate distribution (or process started in infinite past);

4. If the process is stationary, $\mathrm{E}[g_i^2] = \zeta/(1 - \alpha)$.

GARCH process (Hansen B1–38–9) GARCH(p, q) is a sequence $\{r_i\}$ with conditional variance given by

$$\mathrm{Var}[r_i|r_{i-1}, \ldots, r_1] \equiv \sigma_i^2 =$$
$$\omega + \alpha_1 r_{i-1}^2 + \cdots + \alpha_q r_{i-q}^2 + \beta_1 \sigma_{i-1}^2 + \cdots + \sigma_{i-p}^2.$$

For $\{r_i\}$ a GARCH$(1, 1)$ process, define $v_i \equiv r_i^2 - \sigma_i^2$. Then we can write $\{r_i^2\}$ as an ARMA$(1, 1)$ and $\{\sigma_i^2\}$ as an AR(1), with shocks $\{v_i\}$:

$$r_i^2 = \omega + (\alpha_1 + \beta_1) r_{i-1}^2 + v_t - \beta_1 v_{t-1},$$
$$\sigma_i^2 = \omega + (\alpha_1 + \beta_1) \sigma_{i-1}^2 + \alpha_1 v_{t-1}.$$

Local level model (Hansen B1–41–2) X_t is "true" latent variable, Y_t is observed value:

$$Y_t = X_t + \eta_t,$$
$$X_t = X_{t-1} + \varepsilon_t$$

with $\{\eta_t\}$ (measurement error) and $\{\varepsilon_t\}$ independent iid shocks. Y_t is an ARMA$(1, 1)$ with $Y_t = Y_{t-1} + \varepsilon_t + \eta_t - \eta_{t-1}$.[*]

Estimating number of lags (Greene 834–5) Calculate sample autocorrelations and partial autocorrelations using OLS:

- **acf**: For ρ_k, regress y_t on y_{t-k} (and potentially a constant);

- **pacf**: For ρ_k^*, regress y_t on $y_{t-1}, \ldots, y_{t-k}$ (and potentially a constant), and use the coefficient on y_{t-k}.

Typically, an AR(p) will have acf declining monotonically, and pacf irregular out to lag p, after which they will abruptly drop to zero and remain there.

Typically, an MA(q) will have pacf declining monotonically, and acf irregular out to lag q, after which they will abruptly drop to zero and remain there.

An ARMA(p, q) is a mixture of the above.

Estimating AR(p) (Hayashi 392–4, 547–9) Suppose $\{y_t\}$ with $y_t = c + [\sum_{j=1}^p \phi_j y_{t-j}] + \varepsilon_t$, and $\{\varepsilon_t\}$ *iid* white noise (note we need iid for ergodicity and stationarity). Then we can use OLS to regress y_t on $(1, y_{t-1}, \ldots, y_{t-p})'$, since the model satisfies linearity, stationary ergodicity, independence of regressors from errors, conditional homoscedasticity, mds and finite second moments of g, and the rank condition. We have consistent, asymptotically normal estimation of ϕ, as well as consistent variance estimation.

Under Gaussian AR(1), conditional ML for ϕ is numerically equal to OLS estimate, and conditional ML $\hat{\sigma}_{\mathrm{CML}}^2 = \frac{\mathrm{SSR}}{n}$.

Maximum likelihood with serial correlation (Hayashi 543–47) Note that treating a serially correlated process as if it were iid may give consistent ("quasi-") ML estimators, since it is an M-estimator no matter the data's distribution.

1. Exact ML: use the fact that

$$f(y_n, \ldots, y_0) = \frac{f(y_n, \ldots, y_0)}{f(y_{n-1}, \ldots, y_0)} \cdot \frac{f(y_{n-1}, \ldots, y_0)}{f(y_{n-2}, \ldots, y_0)}.$$
$$\cdots \cdot f(y_0)$$
$$= \left[\prod_{t=1}^n f(y_t|y_{t-1}, \ldots, y_0)\right] f(y_0).$$

Log-likelihood is generally nonlinear due to the last term.

2. Conditional ML: use the (linear)

$$\log f(y_n, \ldots, y_0|y_0) = \sum_{t=1}^n \log f(y_t|y_{t-1}, \ldots, y_0).$$

1.14 Ordinary least squares

Ordinary least squares model (Hayashi 4–12, 34, 109–10, 123, 126; Mahajan 6-4–6, 12–3) Let

$$\underbrace{X_i}_{K\times 1} = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{iK} \end{bmatrix}, \underbrace{\beta}_{K\times 1} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix},$$

$$\underbrace{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \underbrace{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \underbrace{X}_{n\times K} = \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix}.$$

The model *may* assume:

1. Linearity (p. 4): $Y = X\beta + \varepsilon$;

2. Strict exogeneity (p. 7–9): $\mathrm{E}[\varepsilon|X] = 0$ (n.b., conditional on regressors for *all* observations); implies:
   - $\mathrm{E}[\varepsilon] = 0$,
   - $\mathrm{E}[x_{jk}\varepsilon_i] = 0$ for all $i = 1, \ldots, n$, $j = 1, \ldots, n$, $k = 1, \ldots, K$,
   - $\mathrm{Cov}(\varepsilon_i, x_{jk}) = 0$ for all $i = 1, \ldots, n$, $j = 1, \ldots, n$, $k = 1, \ldots, K$;

3. No multicolinearity (p. 10): $\mathrm{rank}(X) = K$ with probability 1 (full column rank);

4. Spherical error variance (p. 10–2): $\mathrm{E}[\varepsilon\varepsilon'|X] = \sigma^2 \mathbf{I}_n$, or equivalently,
   - Conditional homoscedasticity: $\mathrm{E}[\varepsilon_i^2|X] = \sigma^2$ for all $i = 1, \ldots, n$,
   - No correlation between observations: $\mathrm{E}[\varepsilon_i\varepsilon_j|X] = 0$ for $i \neq j$;

5. Normal errors (p. 34): $\varepsilon|X \sim \mathrm{N}(0, \sigma^2 \mathbf{I}_n)$, which given above assumptions implies $\varepsilon \perp\!\!\!\perp X$ and $\varepsilon \sim \mathrm{N}(0, \sigma^2 \mathbf{I}_n)$;

6. Ergodic stationarity (p. 109): $\{y_i, X_i\}$ is jointly stationary and ergodic (implies unconditional homoscedasticity, but allows conditional heteroscedasticity);

---

[*]It is not clear to me what the white noise sequence and MA coefficients are that generate this sequence, but since $\varepsilon_t + \eta_t - \eta_{t-1}$ is covariance stationary with two nonzero autocovariances, we should be able to fit one.

7. Predetermined regressors (p. 109–10) All regressors are orthogonal to the contemporaneous error term: $\mathrm{E}[x_{ik}\varepsilon_i] = 0$ for all $i = 1, \ldots, n$, $k = 1, \ldots, K$; also written $\mathrm{E}[g_i] = 0$ where $g_i \equiv X_i\varepsilon_i$;

8. Rank condition (p. 110): The $K \times K$ matrix $\Sigma_{xx} \equiv \mathrm{E}[X_i X_i']$ is nonsingular (and hence finite);

9. g_i is a mds with finite second moments (p. 110): $\{g_i\} \equiv \{X_i\varepsilon_i\}$ is a mds (which implies assumption 7); the $K \times K$ matrix $S \equiv \mathrm{E}[g_i g_i']$ is nonsingular;

10. Finite fourth moments for regressors (p. 123): $\mathrm{E}[(x_{ik}x_{ij})^2]$ exists and is finite for all $k, j = 1, \ldots, K$;

11. Conditional homoscedasticity (p. 126): $\mathrm{E}[\varepsilon_i^2|x_i] = \sigma^2 > 0$ (implies $S = \sigma^2 \Sigma_{xx}$).

Ordinary least squares estimators (Hayashi 16–8, 19, 27–8, 31, 51–2; Mahajan 6-6–11) OLS estimator for β is $b \equiv \mathrm{argmin}_{\tilde{\beta}} \mathrm{SSR}(\tilde{\beta}) \equiv \mathrm{argmin}_{\tilde{\beta}}(Y - X\tilde{\beta})'(Y - X\tilde{\beta})$. FOCs give "normal equations" (a.k.a. "moment equations") $X'Xb = X'Y \iff X'e = 0$ where e is OLS residual $e \equiv Y - Xb$.

By no multicolinearity, $X'X$ is nonsingular (hence positive definite and satisfying SOC), so

$$b = (X'X)^{-1}X'Y = S_{XX}^{-1}s_{XY}$$

minimizes normalized distance function

$$\frac{1}{2n\hat{\sigma}^2}(Y - Z\tilde{\delta})'(Y - Z\tilde{\delta}) = \frac{1}{2n\hat{\sigma}^2}\mathrm{SSR}.$$

1. Unbiased: $\mathrm{E}[b|X] = \beta$ (under assumptions 1–3).

2. Gauss-Markov: b is most efficient among linear unbiased estimators,* i.e., best linear (in Y) unbiased estimator (assumptions 1–4).

3. $\mathrm{Cov}(b, e|X) = 0$, where $e \equiv Y - Xb$ (assumptions 1–4).

4. $\mathrm{Var}[b|X] = \sigma^2(X'X)^{-1}$ (assumptions 1–4); therefore a common estimator is $\widehat{\mathrm{Var}(b|X)} \equiv s^2(X'X)^{-1}$. Under normal error assumption, CRLB

$$\mathcal{I}^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}$$

so b achieves CRLB and is UMVUE—this result is *stronger* than Gauss-Markov, but requires normal errors.

OLS estimator for σ^2 is

$$s^2 \equiv \frac{\mathrm{SSR}}{n - K} = \frac{e'e}{n - K}.$$

Unbiased (assumptions 1–4). Under normal error assumption, $\mathrm{Var}(s^2|X) = 2\sigma^4/(n - K)$ which does not meet CRLB but *is* UMVUE.

Asymptotic properties of OLS estimator (Hayashi 113, 115) OLS estimator b satisfies

1. Consistent: $b \xrightarrow{\mathrm{p}} \beta$ (under assumptions 1, 6–8).

2. Asymptotically normal: $\sqrt{n}(b - \beta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathrm{Avar}(b))$ where $\mathrm{Avar}(b) = \Sigma_{xx}^{-1}S\Sigma_{xx}^{-1}$ (assumptions 1, 6, 8–9).

3. Consistent variance estimation: Assuming existence of a consistent estimator $\widehat{S}$ for $S \equiv \mathrm{E}[g_i g_i']$, estimator $\widehat{\mathrm{Avar}(b)} \equiv S_{xx}^{-1}\widehat{S}S_{xx}^{-1}$ is consistent for $\mathrm{Avar}(b)$ (assumption 6).

OLS error variance estimator s^2 (a.k.a. variance of residuals) is consistent for $\mathrm{E}[\varepsilon_i^2]$, assuming expectation exists and is finite (assumptions 1, 6–8).

Estimating S (Hayashi 123–4, 127–8) If $\hat{\beta}$ (usually b) is consistent, then $\hat{\varepsilon}_i \equiv y_i - x_i'\hat{\beta}$ is consistent. Then under assumptions 1, 6, and 10, a consistent estimator for S is

$$\widehat{S} \equiv \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 x_i x_i'.$$

Under assumptions 1, 6–8, and 11 (note with conditional homoscedasticity, we no longer require finite fourth moments), can use $\widehat{S} \equiv s^2 S_{xx}$ where s is the (consistent) OLS estimate of σ^2.

Biases affecting OLS (Hayashi 188–9, 194–7) OLS estimation inconsistent due to, e.g.:

1. Simultaneity (p. 188–9);

2. Errors-in-variables (e.g., classical measurement error) (p. 194–5);

3. Unobserved variables (p. 196–7).

Maximum likelihood estimator for OLS model (Hayashi 49; Mahajan 6-18–9) Assuming normal errors, the OLS estimate for β is also the ML estimate.

$$\hat{\beta}_{\mathrm{ML}} = b;$$
$$\hat{\sigma}_{\mathrm{ML}}^2 = \frac{1}{n}e'e = \frac{1}{n}\mathrm{SSR} = \frac{n-K}{n}s^2.$$

Best linear predictor (Mahajan 6-2–3; Hayashi 139–40) Population analogue of OLS. Suppose (y, x) a r.v. with $y \in \mathbb{R}$ and $x \in \mathbb{R}^k$. Best linear predictor of y given X is

$$\widehat{E}^*(y|x) \equiv \mathbb{L}(y|x) = x' \mathrm{E}[xx']^{-1}\mathrm{E}[xy].$$

If one regressor in $x \equiv [1 \quad \tilde{x}']'$ is constant, then $\widehat{E}^*(y|x) \equiv \widehat{E}^*(y|1, \tilde{x}) = \mu + \gamma'\tilde{x}$ where $\gamma = \mathrm{Var}[\tilde{x}]^{-1}\mathrm{Cov}[\tilde{x}, y]$ and $\mu = \mathrm{E}[y] - \gamma'\mathrm{E}[\tilde{x}]$.

Fitted value (Hayashi 18) $\widehat{Y} \equiv Xb = PY$. Thus OLS residuals $e = Y - \widehat{Y}$.

Projection matrix (Hayashi 18–9; Mahajan 6-9–10; D&M 292–3) $P \equiv X(X'X)^{-1}X'$. Projects onto space spanned by columns of X ($PX = X$). Fitted values $\widehat{Y} \equiv Xb = PY$. Idempotent and symmetric.

An oblique projection matrix $P^\Omega \equiv X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}$ (and its associated annihilator) are idempotent, but not symmetric.

Annihilator matrix (Hayashi 18–9, 30–1; Mahajan 6-9–10) $M \equiv I - P$. Residuals $e = Y - \widehat{Y} = MY = M\varepsilon$, and thus are orthogonal to X (since $MX = 0$).

Sum of squared residuals $\mathrm{SSR} \equiv e'e = Y'MY = Y'e = e'Y = \varepsilon'M\varepsilon$.

M is idempotent and symmetric. $\mathrm{rank}(M) = \mathrm{tr}(M) = n - K$.

Standard error of the regression (Hayashi 19) $\mathrm{SER} \equiv \sqrt{s^2}$. [a.k.a. standard error of the equation.]

Sampling error (Hayashi 19, 35) $b - \beta = (X'X)^{-1}X'\varepsilon$. Under normal error assumption, $b - \beta|X \sim \mathrm{N}(0, \sigma^2(X'X)^{-1})$.

OLS R^2 (Hayashi 20–1, 67; Mahajan 6-11–2; Greene 250–3) In population, the fraction of variation in Y attributable to variation in X is

$$1 - \frac{\mathrm{Var}\,\varepsilon}{\mathrm{Var}\,y} = \frac{\mathrm{Var}(x'\beta)}{\mathrm{Var}\,y} = (\rho_{y,x'\beta})^2.$$

("Uncentered") sample analogue is: $\widehat{Y}'\widehat{Y}/Y'Y = 1 - e'e/Y'Y$.

Centered: should be used if regressors include a constant;

$$\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{e'e}{\sum(y_i - \bar{y})^2}.$$

Be careful when comparing equations on the basis of the fit (i.e., R^2): the equations must share the same dependent variable to get a meaningful comparison.

Standard error (Hayashi 35–7)

$$\mathrm{SE}(b_k) \equiv \sqrt{s^2[(X'X)^{-1}]_{kk}} = \sqrt{[\widehat{\mathrm{Var}(b|X)}]_{kk}}.$$

Robust standard error (Hayashi 117, 211-2)

$$\mathrm{SE}^*(b_k) \equiv \sqrt{\frac{1}{n}\widehat{\mathrm{Avar}(b_k)}} \equiv \sqrt{\frac{1}{n}[S_{xx}^{-1}\widehat{S}S_{xx}^{-1}]_{kk}}.$$

OLS t test (Hayashi 35–7; Mahajan 6-13–16) Assuming normal errors. Test for $H: \beta_k = \bar{\beta}_k$. Since $b_k - \bar{\beta}_k|X \sim \mathrm{N}(0, \sigma^2[(X'X)^{-1}]_{kk})$ under null, we have

$$z_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2[(X'X)^{-1}]_{kk}}} \sim \mathrm{N}(0, 1).$$

---

*Variance of b is less than or equal to variance of any other linear unbiased estimator of β *in the matrix sense*; i.e., difference is positive semidefinite.

If σ^2 is unknown, use s^2 to get t-ratio:

$$t_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{s^2[(X'X)^{-1}]_{kk}}} \equiv \frac{b_k - \bar{\beta}_k}{\mathrm{SE}(b_k)} \sim t_{n-K}$$

under null. Accept for t_k close to zero—t distribution symmetric about zero.

Level-α confidence interval for $\bar{\beta}_k$ is $b_k \pm \mathrm{SE}(b_k) \cdot t_{\alpha/2}(n-K)$.

OLS robust t test (Hayashi 117–9) Test for $H\colon \beta_k = \bar{\beta}_k$. Under null, we have

$$t_k \equiv \frac{\sqrt{n}(b_k - \bar{\beta}_k)}{\sqrt{\widehat{\mathrm{Avar}(b_k)}}} \equiv \frac{b_k - \bar{\beta}_k}{\mathrm{SE}^*(b_k)} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

OLS F test (Hayashi 39–44, 53; Mahajan 6-16–7, 20–5) Assuming normal errors. Test for $H\colon R_{\#r \times K}\beta_{K \times 1} = r_{\#r \times 1}$, where $\mathrm{rank}(R) = \#r$ (i.e., R has full row rank—there are no redundant restrictions). Accept if $F \sim F_{\#r, n-K}$ (under null) is small, where:

$$
\begin{aligned}
F &\equiv \frac{(Rb - r)'[R(X'X)^{-1}R']^{-1}(Rb-r)/\#r}{s^2} \\
&= (Rb-r)'[R\widehat{\mathrm{Var}(b|X)}R']^{-1}(Rb-r)/\#r \\
&= \frac{(\mathrm{SSR}_R - \mathrm{SSR}_U)/\#r}{\mathrm{SSR}_U/(n-k)} = \frac{n-K}{\#r}(\lambda^{2/n} - 1)
\end{aligned}
$$

(and $\lambda \equiv L_U/L_R = (\mathrm{SSR}_U / \mathrm{SSR}_R)^{-n/2}$ the likelihood ratio). The first and second expressions are based on Wald principle; the third on LR.

F ratio is square of relevant t ratio. F is preferred to multiple t tests—gives an eliptical rather than rectangular confidence region.

OLS robust Wald statistic (Hayashi 118, 122) Test for $H\colon R_{\#r \times K}\beta_{K \times 1} = r_{\#r \times 1}$, where $\mathrm{rank}(R) = \#r$ (i.e., R has full row rank—there are no redundant restrictions). Under null,

$$W \equiv n(Rb - r)'[R\widehat{\mathrm{Avar}(b)}R']^{-1}(Rb - r) \xrightarrow{\mathrm{d}} \chi^2_{\#r}.$$

For $H\colon a_{\#a}(\beta) = 0$, where $A_{\#a \times K}(\beta) \equiv \nabla a(\beta)$ is continuous and of full row rank, under null,

$$W \equiv na(b)'[A(b)\widehat{\mathrm{Avar}(b)}A(b)']^{-1}a(b) \xrightarrow{\mathrm{d}} \chi^2_{\#a}.$$

Generalized least squares (Hayashi 54–9, 415–7; Mahajan 6-25–9; D&M 289–92, 295) Without assumption 4 (spherical error variance), we have $\mathrm{E}[\varepsilon\varepsilon'|X] \equiv \sigma^2 V$ with $V_{n \times n} \neq \mathbf{I}$ assumed to be nonsingular. Gauss-Markov no longer holds; the t and F tests are no longer valid; however, the OLS estimator *is* still unbiased.

By the orthogonal decomposition of V^{-1}, we can write $V^{-1} \equiv C'C$ for some (non-unique C). Writing $\tilde{y} \equiv Cy$, $\tilde{X} \equiv$

CX, $\tilde{\varepsilon} \equiv C\varepsilon$, we get transformed model $\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}$ which satisfies assumptions of classical linear regression model. Thus we have the unbiased, efficient estimator:

$$
\begin{aligned}
\hat{\beta}_{\mathrm{GLS}} &\equiv \underset{\tilde{\beta}}{\mathrm{argmin}}(y - X\tilde{\beta})'V^{-1}(y - X\tilde{\beta}) \\
&\equiv (X'V^{-1}X)^{-1}X'V^{-1}y \\
\mathrm{Var}(\hat{\beta}_{\mathrm{GLS}}|X) &\equiv \sigma^2(X'V^{-1}X)^{-1}.
\end{aligned}
$$

This can also be seen as an IV estimator with instruments $V^{-1}X$.

Note consistency and other attractive properties rely on strict exogeneity of regressors; if they are merely predetermined, the estimator need not be consistent (unlike OLS, which is).

Feasible generalized least squares (Hayashi 59, 133-7, 415–7; Mahajan 6-25–9; D&M 298–301, 295) In practice, the covariance matrix $\sigma^2 V$ is generally not known (even up to a constant). Feasible GLS uses an estimate $\widehat{V}$, and "under reasonable conditions, yields estimates that are not only consistent but also asymptotically equivalent to genuine GLS estimates, and therefore share their efficiency properties. However, even when this is the case, the performance in finite samples of FGLS may be much inferior to that of genuine GLS."

Caution that FGLS is that consistency is *not* guaranteed if the regressors are merely predetermined (rather than strictly exogenous).

Weighted least squares (Hayashi 58–9; Mahajan 6-27–8) GLS where there is no correlation in the error term between observations (i.e., the matrix V is diagonal). GLS divides each element of X and Y by the standard deviation of the relevant error term, and then runs OLS. That is, observations with a higher conditional variance get a lower weight.

Can also think of as

$$\hat{\beta} = \underset{\tilde{\beta}}{\mathrm{argmin}} \sum_i \frac{1}{\sigma_i^2}(y_i - x_i'\tilde{\beta})^2.$$

Thus the "weights" typically passed to OLS software are the reciprocal of the variances (not the reciprocal of the standard deviations).

Frisch-Waugh Theorem (D&M 19–24; Hayashi 72–4; Greene 245–7) Suppose we partition X into X_1 and X_2 (and therefore β into β_1 and β_2). Then OLS estimators

$$
\begin{aligned}
b_1 &= (X_1'X_1)^{-1}X_1'(Y - X_2 b_2); \\
b_2 &= (X_2'M_1 X_2)^{-1}(X_2'M_1 Y) \\
&= (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'\tilde{Y}
\end{aligned}
$$

where $\tilde{X}_2 \equiv M_1 X_2$ and $\tilde{Y} \equiv M_1 Y$ are the residuals from regressing X_2 and Y respectively on X_1.

In addition, the residuals from regressing Y on X equal the residuals from regressing $\tilde{Y}$ on $\tilde{X}_2$.

Short and long regressions (Mahajan 6-33–7) Suppose we have best linear predictors

$$
\begin{aligned}
\widehat{E}^*(y|x_1, x_2) &= x_1'\beta_1 + x_2'\beta_2, \\
\widehat{E}^*(y|x_1) &= x_1'\delta.
\end{aligned}
$$

Then $\beta_1 = \delta$ iff either $\mathrm{E}[x_1 x_2] = 0$ or $\beta_2 = 0$ since

$$\delta = \beta_1 + \left(\mathrm{E}(x_1 x_1')\right)^{-1}\left(\mathrm{E}(x_1 x_2')\beta_2\right).$$

Note this result is the population analogue of the Frisch-Waugh result for b_1.

1.15 Linear Generalized Method of Moments

This material is also covered (with slightly different notation—notably the exchange of x with z—in Mahajan 7.

Linear GMM model (Hayashi 198–203, 212, 225–6, 406–7) The model may assume:

1. Linearity (p. 198): $y_i = z_i'\delta + \varepsilon_i$ with $z_i, \delta \in \mathbb{R}^L$;

2. Ergodic stationarity (p. 198): $x_i \in \mathbb{R}^K$ instruments; unique, non-constant elements of $\{(y_i, z_i, x_i)\}$ are jointly ergodic stationary (stronger than individually ergodic stationary);

3. Orthogonality (p. 198): All K elements of x_i predetermined (i.e., orthogonal to current error term): $\mathrm{E}[g_i] \equiv \mathrm{E}[x_i \varepsilon_i] = \mathrm{E}[x_i(y_i - z_i'\delta)] = 0$;

4. Rank condition for identification (p. 200): $\Sigma_{xz} \equiv \mathrm{E}[x_i z_i']|_{K \times L}$ has full column rank (i.e., rank L);

   - $K > L$: Overidentified/determined (use GMM estimator),
   - $K = L$: Just/exactly identified/determined (GMM estimator reduces to IV estimator),
   - $K < L$: Not/under identified/determined;

5. Requirements for asymptotic normality (p. 202–3):

   - $\{g_i\} \equiv \{x_i \varepsilon_i\}$ is a martingale difference sequence with $K \times K$ matrix of cross moments $\mathrm{E}[g_i g_i']$ nonsingular;
   - $S \equiv \lim_{n \to \infty} \mathrm{Var}[\sqrt{n}\bar{g}] = \mathrm{Avar}(\bar{g}) = \mathrm{Avar}(\frac{1}{n}\sum g_i) = \mathrm{Avar}(\frac{1}{n}\sum \varepsilon_i^2 x_i x_i')$;
   - By assumption 2 and ergodic stationary mds CLT, $S = \mathrm{E}[g_i g_i'] = \mathrm{E}[\varepsilon_i^2 x_i x_i']$;

6. Finite fourth moments (p. 212): $\mathrm{E}[(x_{ik} z_{il})^2]$ exists and is finite for all k and l;

7. Conditional homoscedasticity (p. 225–6): $\mathrm{E}[\varepsilon_i^2|x_i] = \sigma^2$, which implies $S \equiv \mathrm{E}[g_i g_i'] = \mathrm{E}[\varepsilon_i^2 x_i x_i'] = \sigma^2 \mathrm{E}[x_i x_i'] = \sigma^2 \Sigma_{xx}$. Note:

   - By assumption 5, $\sigma^2 > 0$ and Σ_{xx} nonsingular;

- $\widehat{S} = \hat{\sigma}^2 \frac{1}{n} \sum_i x_i x_i' = \hat{\sigma}^2 S_{xx}$ where $\hat{\sigma}^2$ consistent, so no longer need assumption 6 for consistent estimation of S;

8. $\{g_t\} \equiv \{x_t \varepsilon_t\}$ satisfies Gordin's condition (p. 406–7): it is ergodic stationary and

   (a) $E[g_t g_t']$ exists and is finite;

   (b) $E[g_t | g_{t-j}, g_{t-j-1}, \dots] \xrightarrow{L_2} 0$ as $j \to \infty$;

   (c) $\sum_{j=0}^{\infty} \sqrt{E[r_{tj}' r_{tj}]} < \infty$, where $r_{tj} \equiv E[g_t | g_{t-j}, g_{t-j-1}, \dots] - E[g_t | g_{t-j-1}, g_{t-j-2}, \dots]$;

   and the long-run covariance matrix $S = \text{Avar}(\bar{g}) \equiv \sum_{j \in \mathbb{Z}} \Gamma_j$ is nonsingular.

Instrumental variables estimator (Hayashi 205–6) A method of moments estimator based on K orthogonality (moment) conditions. For just identified GMM model ($K = L$; identification also requires nonsingularity of $\Sigma_{xz} \equiv E[x_i z_i']$),

$$\hat{\delta}_{\text{IV}} \equiv S_{xz}^{-1} s_{xy} \equiv \left(\frac{1}{n} \sum_i x_i z_i' \right)^{-1} \left(\frac{1}{n} \sum_i x_i y_i \right).$$

Numerically equivalent to linear GMM estimator with *any* symmetric positive definite weighting matrix. If instruments are same as regressors ($x_i = z_i$), reduces to OLS estimator. For properties, see GMM estimators.

If one non-constant regressor and one non-constant instrument, $\text{plim}\,\hat{\delta}_{\text{IV}} = \text{Cov}(x, y) / \text{Cov}(x, z)$.

Linear GMM estimator (Hayashi 206–12, 406–7; D&M 219–20) For overidentified case ($K > L$), cannot generally satisfy all K moment conditions, so estimator is

$$\hat{\delta}(\widehat{W}) \equiv \underset{\tilde{\delta}}{\text{argmin}}\, \tfrac{1}{2} g_n(\tilde{\delta})' \widehat{W} g_n(\tilde{\delta})$$
$$\equiv \underset{\tilde{\delta}}{\text{argmin}}\, \tfrac{1}{2n^2} (Y - Z\tilde{\delta})' X \widehat{W} X' (Y - Z\tilde{\delta})$$

with $g_n(\tilde{\delta}) \equiv \frac{1}{n} \sum_i x_i (y_i - z_i' \tilde{\delta})$ and some symmetric positive definite matrix $\widehat{W} \xrightarrow{p} W$ (where $\widehat{W}$ can depend on data, and W also symmetric and positive definite).

$$\hat{\delta}_{\text{GMM}} \equiv (S_{xz}' \widehat{W} S_{xz})^{-1} S_{xz}' \widehat{W} s_{xy}$$
$$\equiv (Z' X \widehat{W} X' Z)^{-1} Z' X \widehat{W} X' Y.$$

If $K = L$ (just identified case), reduces to IV estimator for *any choice* of $\widehat{W}$.

1. Consistency: $\hat{\delta}(\widehat{W}) \xrightarrow{p} \delta$ (under assumptions 1–4).

2. Asymptotic normality: $\sqrt{n}(\hat{\delta}(\widehat{W}) - \delta) \xrightarrow{d} N(0, \text{Avar}(\hat{\delta}(\widehat{W})))$ where

$$\text{Avar}(\hat{\delta}(\widehat{W})) = \underbrace{(\Sigma_{xz}' W \Sigma_{xz})^{-1}}_{-\Psi^{-1}} \underbrace{\Sigma_{xz}' W S W \Sigma_{xz}}_{\Sigma} \underbrace{(\Sigma_{xz}' W \Sigma_{xz})^{-1}}_{-\Psi^{-1}}$$

(assumptions 1–4, 5 or 8). Note sampling error $\hat{\delta}(\widehat{W}) - \delta = (S_{xz}' \widehat{W} S_{xz})^{-1} S_{xz}' \widehat{W} g_n(\delta)$.

3. Consistent variance estimation: If $\widehat{S}$ is a consistent estimator of S, then (assumption 2)

$$\widehat{\text{Avar}(\hat{\delta}(\widehat{W}))} = (S_{xz}' \widehat{W} S_{xz})^{-1} S_{xz}' \widehat{W} \widehat{S} \widehat{W} S_{xz} (S_{xz}' \widehat{W} S_{xz})^{-1}.$$

4. Consistent estimation of S: If $\hat{\delta}$ is consistent, $S \equiv E[g_i g_i']$ exists and is finite, and assumptions 1, 2, (5 implied?] and 6, then

$$\widehat{S} \equiv \tfrac{1}{n} \sum_i \hat{\varepsilon}_i^2 x_i x_i' = \tfrac{1}{n} \sum_i (y_i - z_i' \hat{\delta})^2 x_i x_i' \xrightarrow{p} S.$$

For estimation of S under assumption 8 in place of 5, see Hayashi 407–12 on kernel estimation and VARHAC.

5. Choosing instruments: Adding (correct) instruments generally increases efficiency, but also increases finite-sample bias; "weak" instruments (that are almost uncorrelated with regressors) make asymptotics poorly approximate finite-sample properties.

GMM estimator can be seen as a GLS estimator: GMM minimizes $e' X \widehat{W} X' e = e' \widehat{W}_{\text{GLS}} e$ where $\widehat{W}_{\text{GLS}} \equiv X \widehat{W} X'$.

GMM hypothesis testing (Hayashi 211–2, 222–4; Mahajan 7-7–12; D&M 617–8) Assume 1–5. Under null, Wald principle gives:

1. $H_0 : \delta_l = \bar{\delta}_l$,

$$t_l \equiv \frac{\hat{\delta}_l(\widehat{W}) - \bar{\delta}_l}{\text{SE}_l^*} \xrightarrow{d} N(0, 1).$$

2. $H_0 : R_{\#r \times L} \delta_{L \times 1} = r_{\#r \times 1}$, where $\text{rank}(R) = \#r$ (i.e., R has full row rank—there are no redundant restrictions),

$$W \equiv n(R\hat{\delta}(\widehat{W}) - r)' \left[R[\widehat{\text{Avar}(\hat{\delta}(\widehat{W}))}] R' \right]^{-1} (R\hat{\delta}(\widehat{W}) - r) \xrightarrow{d} \chi_{\#r}^2.$$

3. $H_0 : a_{\#a}(\delta) = 0$, where $A_{\#a \times L}(\delta) \equiv \nabla a(\delta)$ is continuous and of full row rank,

$$W \equiv n a(\hat{\delta}(\widehat{W}))' \left[A(\hat{\delta}(\widehat{W}))[\widehat{\text{Avar}(\hat{\delta}(\widehat{W}))}] A(\hat{\delta}(\widehat{W}))' \right]^{-1} a(\hat{\delta}(\widehat{W})) \xrightarrow{d} \chi_{\#a}^2.$$

(Note W is *not* numerically invariant to the representation of the restriction.)

Distance principle statistic is by

$$J(\hat{\delta}_r(\widehat{S}^{-1}), \widehat{S}^{-1}) - J(\hat{\delta}_u(\widehat{S}^{-1}), \widehat{S}^{-1}) \xrightarrow{d} \chi_{\#a}^2,$$

where J is $2n$ times the minimized objective function:

$$J(\hat{\delta}, \widehat{S}^{-1}) \equiv n g_n(\hat{\delta})' \widehat{S}^{-1} g_n(\hat{\delta})$$
$$\equiv \tfrac{1}{n} (Y - Z\tilde{\delta})' X \widehat{S}^{-1} X' (Y - Z\tilde{\delta})$$

LM statistic is on Majajan 7-7–12 or D&M 617.

Note that unlike the Wald, these statistics—which are numerically equal in finite samples—are only asymptotically chi squared under *efficient* GMM. If the hypothesis' restrictions are linear and the weighting matrix is optimally chosen, they are also numerically equal to the Wald statistic.

Efficient GMM (Hayashi 212–5; D&M 588-9, 597–8, 607–8) Lower bound $\text{Avar}(\hat{\delta}(\widehat{W})) \geq (\Sigma_{xz}' S^{-1} \Sigma_{xz})^{-1}$ is achieved if $\widehat{W}$ chosen so that $W \equiv \text{plim}\,\widehat{W} = S^{-1}$. Two-step efficient GMM:

1. Calculate $\hat{\delta}(\widehat{W})$ for some $\widehat{W}$ (usually S_{xx}^{-1}; use residuals $\hat{\varepsilon}_i \equiv y_i - z_i' \hat{\delta}(\widehat{W})$ to obtain a consistent estimator $\widehat{S}$ of S.

2. Calculate $\hat{\delta}(\widehat{S}^{-1})$.

Note the requirement of estimating $\widehat{S}^{-1}$ harms small-sample properties, since it requires the estimation of fourth moments. Efficient GMM is often outperformed in bias and variance by equally-weighted GMM ($\widehat{W} = \mathbf{I}$) in finite samples.

Testing overidentifying restrictions (Hayashi 217–21; Mahajan 7-13–6; D&M 232–7, 615–6) Assuming overidentification ($K > L$), a consistent estimator $\widehat{S}$ (ensured by assumption 6), and assumptions 1–5, then

$$J \equiv n g_n(\hat{\delta}(\widehat{S}^{-1}))' \widehat{S}^{-1} g_n(\hat{\delta}(\widehat{S}^{-1})) \xrightarrow{d} \chi_{K-L}^2.$$

To test a subset of orthogonality conditions (as long as K_1 remaining—non-suspect—instruments are enough to ensure identification), calculate J_1 statistic using remaining instruments, and $C \equiv J - J_1 \xrightarrow{d} \chi^2(K - K_1)$ (this is a distance principle test, since J is $2n$ times the minimized objective function). In finite samples, use the leading submatrix of $\widehat{S}$ to calculate J_1.

Two-Stage Least Squares (Hayashi 226–31; D&M 215–24) Under assumption 7 (conditional homoscedasticity), efficient GMM becomes 2SLS. Minimizing normalized distance function

$$\frac{1}{2n\hat{\sigma}^2}(Y - Z\tilde{\delta})'P_x(Y - Z\tilde{\delta})$$

yields estimator

$$\hat{\delta}_{2\text{SLS}} \equiv \hat{\delta}(\hat{S}^{-1}) = \hat{\delta}([\sigma^2 S_{xx}]^{-1}) = \hat{\delta}(S_{xx}^{-1})$$
$$= (S_{xz}'S_{xx}^{-1}S_{xz})^{-1}S_{xz}'S_{xx}^{-1}s_{xy}.$$

Note $\hat{\delta}_{2\text{SLS}}$ does not depend on $\hat{\sigma}^2$.

Asymptotic variance $\text{Avar}(\hat{\delta}_{2\text{SLS}}) = \sigma^2(\Sigma_{xz}'\Sigma_{xx}^{-1}\Sigma_{xz})^{-1}$ estimated by

$$\widehat{\text{Avar}(\hat{\delta}_{2\text{SLS}})} \equiv \hat{\sigma}^2(S_{xz}'S_{xx}^{-1}S_{xz})^{-1}$$
$$\equiv \hat{\sigma}^2 n(Z'P_x Z)^{-1}$$

with $\hat{\sigma}^2 \equiv \frac{1}{n}\sum_i (y_i - z_i'\hat{\delta}_{2\text{SLS}})^2$.

If errors are heteroscedastic, 2SLS is inefficient GMM. Robust covariance estimate (since $\widehat{W} = S_{xx}^{-1}$) is

$$\widehat{\text{Avar}(\hat{\delta}_{2\text{SLS}})} =$$
$$(S_{xz}'S_{xx}^{-1}S_{xz})^{-1}S_{xz}'S_{xx}^{-1}\widehat{S}S_{xx}^{-1}S_{xz}(S_{xz}'S_{xx}^{-1}S_{xz})^{-1}.$$

Other characterizations of 2SLS minimize $\|P_x(Y - Z\delta)\|^2 = (Y - Z\delta)'P_x(Y - Z\delta)$, or else note:

$$\hat{\delta}_{2\text{SLS}} = (Z'X(X'X)^{-1}X'Z)^{-1}Z'X(X'X)^{-1}X'Y$$
$$= (Z'P_x Z)^{-1}Z'P_x Y$$

where the projection matrix $P_x \equiv X(X'X)^{-1}X'$. So can view 2SLS as:

1. IV regression of Y (or P_xY) on Z with P_xZ (fitted values from regressing Z on X) as instruments, or;

2. OLS regression of Y (or P_xY) on P_xZ (fitted values from regressing Z on X)—this is why it's called "two-stage least squares."

Caution: Don't view standard errors reported by these regressions are relevant for 2SLS, since they ignore errors from "first stage" (i.e., calculation of fitted values P_xZ).

Durbin-Wu-Hausman test (D&M 237–42; MaCurdy) Suppose we do not know whether certain regressors are endogenous or not, and therefore whether or not they need to be included as instruments. Assuming endogeneity, OLS will be inconsistent and we prefer 2SLS; assuming exogeneity, both OLS and 2SLS estimates will be consistent and we prefer OLS (it is BLUE, and unbiased under appropriate assumptions).

Test by estimating either artificial regression using OLS:

$$Y = Z\beta + P_x Z^*\delta + \hat{\varepsilon}$$
$$Y = Z\beta + M_x Z^*\eta + \tilde{\varepsilon};$$

i.e., including either fitted values or residuals from a first-stage regression of suspect regressors Z^* on known instruments X. If Z^* are exogenous, the coefficient δ or η should be zero.

Interpretation note: although the test is often interpreted as a test for exogeneity, the test is actually for the consistency of the OLS estimate. OLS *can* be consistent even if Z^* are endogenous.

Analogous test will work for nonlinear model *as long as* suspect regressors Z^* enter linearly; if Z^* enter nonlinearly, failure is due to Jensen's inequality.

1.16   Linear multiple equation GMM

See Hayashi 284 for summary of multiple-equation GMM estimators. This material is also covered (with slightly different notation—notably the exchange of x with z—in Mahajan 7.

Linear multiple-equation GMM model (Hayashi 259–265, 269–70, 274–5) Hayashi 270 summarizes comparison to single-equation GMM.

Model has M linear equations, where equation m has L_m regressors and K_m instruments. Assumptions are exactly parallel to single-equation GMM.

1. Each equation must satisfy its (equation-by-equation) rank condition for identification.

2. The vector of *all* $(\mathbf{y}_i, \mathbf{z}_i, \mathbf{x}_i)$ must be *jointly* ergodic stationary—this is stronger than equation-by-equation ergodic stationarity.

3. Asymptotic normality requires that $\{g_i\}$ be a mds with nonsingular second moment, where

$$\underset{\sum K_m \times 1}{\mathbf{g}_i} \equiv \begin{bmatrix} \mathbf{x}_{i1}\varepsilon_{i1} \\ \vdots \\ \mathbf{x}_{iM}\varepsilon_{iM} \end{bmatrix}.$$

By ergodic stationary mds CLT, $S_{\sum K_m \times \sum K_m} = \text{E}[\mathbf{g}_i\mathbf{g}_i']$ where

$$S \equiv \lim_{n\to\infty}\text{Var}[\sqrt{n}\bar{\mathbf{g}}] = \text{Avar}(\bar{\mathbf{g}}) = \text{Avar}(\tfrac{1}{n}\sum\mathbf{g}_i)$$
$$= \begin{bmatrix} \text{E}[\varepsilon_{i1}\varepsilon_{i1}\mathbf{x}_{i1}\mathbf{x}_{i1}'] & \cdots & \text{E}[\varepsilon_{i1}\varepsilon_{iM}\mathbf{x}_{i1}\mathbf{x}_{iM}'] \\ \vdots & \ddots & \vdots \\ \text{E}[\varepsilon_{iM}\varepsilon_{i1}\mathbf{x}_{iM}\mathbf{x}_{i1}'] & \cdots & \text{E}[\varepsilon_{iM}\varepsilon_{iM}\mathbf{x}_{iM}\mathbf{x}_{iM}'] \end{bmatrix}.$$

4. Consistent estimation of S requires finite fourth moments $\text{E}[(x_{imk}z_{ihj})^2]$ for all $k = 1,\ldots,K_m$; $j = 1,\ldots,L_h$; and $m, h = 1,\ldots,M$.

5. Conditional homoscedasticity becomes $\text{E}[\varepsilon_{im}\varepsilon_{ih}|\mathbf{x}_{im},\mathbf{x}_{ih}] = \sigma_{mh}$ for all $m, h = 1,\ldots,M$; in this case

$$S = \begin{bmatrix} \sigma_{11}\text{E}[\mathbf{x}_{i1}\mathbf{x}_{i1}'] & \cdots & \sigma_{1M}\text{E}[\mathbf{x}_{i1}\mathbf{x}_{iM}'] \\ \vdots & \ddots & \vdots \\ \sigma_{M1}\text{E}[\mathbf{x}_{iM}\mathbf{x}_{i1}'] & \cdots & \sigma_{MM}\text{E}[\mathbf{x}_{iM}\mathbf{x}_{iM}'] \end{bmatrix}.$$

Linear multiple-equation GMM estimator (Hayashi 266-7, 270–3) Hayashi 270 summarizes comparison to single-equation GMM. As in single-equation GMM case, $\hat{\delta}_{\text{GMM}}$ minimizes $\frac{1}{2n^2}(Y - Z\tilde{\delta})'X\widehat{W}X'(Y - Z\tilde{\delta})$, hence

$$\hat{\delta}_{\text{GMM}} \equiv (S_{xz}'\widehat{W}S_{xz})^{-1}S_{xz}'\widehat{W}s_{xy}$$
$$\equiv (Z'X\widehat{W}X'Z)^{-1}Z'X\widehat{W}X'Y,$$

(See Hayashi 267 for huge partitioned matrix representation; can also be written with Y, ε, and β stacked vectors, and X and Z block matrices), where

$$\underset{\sum K_m \times \sum L_m}{S_{xz}} \equiv \begin{bmatrix} \frac{1}{n}\sum\mathbf{x}_{i1}\mathbf{z}_{i1}' & & \\ & \ddots & \\ & & \frac{1}{n}\sum\mathbf{x}_{iM}\mathbf{z}_{iM}' \end{bmatrix},$$

$$\underset{\sum K_m \times 1}{s_{xy}} \equiv \begin{bmatrix} \frac{1}{n}\sum\mathbf{x}_{i1}y_{i1} \\ \vdots \\ \frac{1}{n}\sum\mathbf{x}_{iM}y_{iM} \end{bmatrix}.$$

1. If each equation is exactly identified, then choice of weighting matrix doesn't matter, and multiple-equation GMM is numerically equivalent to equation-by-equation IV estimation.

2. Equation-by-equation GMM corresponds to using a block diagonal weighting matrix.

3. Assuming overidentification, efficient multiple-equation GMM will be asymptotically more efficient than equation-by-equation efficient GMM; they will only be asymptotically equivalent if efficient weighting matrix is block diagonal (i.e., if $\text{E}[\varepsilon_{im}\varepsilon_{ih}\mathbf{x}_{im}\mathbf{x}_{ih}'] = 0$ for all $m \neq h$; under conditional homoscedasticity, this is $\text{E}[\varepsilon_{im}\varepsilon_{ih}] = 0$).

Full-Information Instrumental Variables Efficient (Hayashi 275–6; Mahajan 7–16-7) Efficient GMM estimator with:

- Conditional homoscedasticity ($\text{E}(\varepsilon\varepsilon'|X) = \Sigma_{M\times M} \otimes \mathbf{I}_{n\times n}$).

Weighting matrix is $\widehat{S}^{-1}$, where

$$\widehat{S} \equiv \frac{1}{n}\sum_i X_i \widehat{\Sigma} X_i',$$

$$= \begin{bmatrix} \hat\sigma_{11}\frac{1}{n}\sum_i \mathbf{x}_{i1}\mathbf{x}_{i1}' & \cdots & \hat\sigma_{1M}\frac{1}{n}\sum_i \mathbf{x}_{i1}\mathbf{x}_{iM}' \\ \vdots & \ddots & \vdots \\ \hat\sigma_{M1}\frac{1}{n}\sum_i \mathbf{x}_{iM}\mathbf{x}_{i1}' & \cdots & \hat\sigma_{MM}\frac{1}{n}\sum_i \mathbf{x}_{iM}\mathbf{x}_{iM}' \end{bmatrix},$$

and X_i the $\sum K_m \times M$ block-diagonal matrix of instruments for the ith observation; and

$$\widehat{\Sigma} \equiv \frac{1}{n}\sum_i \hat\varepsilon_i \hat\varepsilon_i',$$

or

$$\hat\sigma_{mh} \equiv \frac{1}{n}\sum_i \hat\varepsilon_{im}\hat\varepsilon_{ih}$$

$$= \frac{1}{n}\sum_i (y_{im} - \mathbf{z}_{im}'\hat\delta_m)(y_{ih} - \mathbf{z}_{ih}'\hat\delta_h)$$

for some consistent estimator $\hat\delta_m$ of δ_m (usually 2SLS).

Suppose we have linearity; joint ergodic stationarity; orthogonality; rank condition; mds with finite second moments; conditional homoscedasticity; and $\mathrm{E}[\mathbf{z}_{im}\mathbf{z}_{ih}']$ exists and is finite for all m, h. Then $\hat\delta_{\mathrm{FIVE}}$ is consistent, asymptotically normal, and efficient.

Three-Stage Least Squares (Hayashi 276–9, 308; Mahajan 7–17; MaCurdy) Efficient GMM estimator with:

- Conditional homoscedasticity ($\mathrm{E}(\varepsilon\varepsilon'|X) = \Sigma_{M\times M} \otimes \mathbf{I}_{n\times n}$),
- The set of instruments the same across equations ($X = \mathbf{I}_{M\times M} \otimes \tilde X_{n\times K}$, where $\tilde X$ are the instruments for each equation).

Weighting matrix is $\widehat{S}^{-1} = \widehat{\Sigma}^{-1} \otimes (\frac{1}{n}\sum_i \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i')^{-1} = \widehat{\Sigma}^{-1} \otimes (\frac{1}{n}\tilde X'\tilde X)^{-1}$, where $\widehat{\Sigma} = \frac{1}{n}\sum_i \hat\varepsilon_i\hat\varepsilon_i'$ is the matrix of $\hat\sigma$ calculated as in FIVE using 2SLS residuals. Normalized distance function

$$\frac{1}{2n}(Y - Z\tilde\delta)'(\widehat{\Sigma}^{-1} \otimes P_{\tilde X})(Y - Z\tilde\delta)$$

yields estimator

$$\hat\delta_{\mathrm{3SLS}} = \left[Z'(\widehat{\Sigma}^{-1} \otimes P_{\tilde X})Z\right]^{-1} Z'(\widehat{\Sigma}^{-1} \otimes P_{\tilde X})\mathbf{y}$$
$$= \left[\widehat{Z}'(\widehat{\Sigma}^{-1} \otimes \mathbf{I})\widehat{Z}\right]^{-1} \widehat{Z}'(\widehat{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y},$$

where $\widehat{Z} \equiv (\mathbf{I} \otimes P_{\tilde X})Z$ are the fitted values from the "first-stage" regression of the Z_ms on $\tilde X$.

$$\widehat{\mathrm{Avar}(\hat\delta_{\mathrm{3SLS}})} = n\left[Z'(\widehat{\Sigma}^{-1} \otimes P_{\tilde X})Z\right]^{-1}.$$

Suppose we have linearity; joint ergodic stationarity; orthogonality; rank condition; mds with finite second moments; conditional homoscedasticity; $\mathrm{E}[\mathbf{z}_{im}\mathbf{z}_{ih}']$ exists and is finite for all m, h; and $\mathbf{x}_{im} = \tilde{\mathbf{x}}_i$ (common instruments). Then $\hat\delta_{\mathrm{3SLS}}$ is consistent, asymptotically normal, and efficient.

3SLS is more efficient than multiple-equation 2SLS unless either

1. Σ is diagonal, in which case the two estimators are asymptotically equivalent (although different in finite samples),
2. Each structural equation is exactly identified, in which case the system is exactly identified, and 2SLS and 3SLS are computationally identical in finite samples.

Multiple equation Two-Stage Least Squares (MaCurdy) Efficient GMM estimator with:

- Conditional homoscedasticity *across equations* ($\Sigma = \sigma^2 \mathbf{I}_{M\times M}$),
- Conditional homoscedasticity ($\mathrm{E}(\varepsilon\varepsilon'|X) = \Sigma_{M\times M} \otimes \mathbf{I}_{n\times n} = \sigma^2 \mathbf{I}_{nM\times nM}$),
- The set of instruments the same across equations ($X = \mathbf{I}_{M\times M} \otimes \tilde X_{n\times K}$, where $\tilde X$ are the instruments for each equation).

Weighting matrix doesn't matter up to a factor of proportionality, and is thus $\widehat{S}^{-1} = \mathbf{I} \otimes (\frac{1}{n}\sum_i \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i')^{-1} = \mathbf{I} \otimes (\frac{1}{n}\tilde X'\tilde X)^{-1}$. Normalized distance function

$$\frac{1}{2n}(Y - Z\tilde\delta)'\left(\begin{bmatrix} \hat\sigma_1^2 & & \\ & \ddots & \\ & & \hat\sigma_M^2 \end{bmatrix}^{-1} \otimes P_{\tilde X}\right)(Y - Z\tilde\delta)$$

yields estimator

$$\hat\delta_{\mathrm{2SLS}} = \left[Z'(\mathbf{I} \otimes P_{\tilde X})Z\right]^{-1} Z'(\mathbf{I} \otimes P_{\tilde X})\mathbf{y}$$
$$= [\widehat{Z}'\widehat{Z}]^{-1}\widehat{Z}'\mathbf{y},$$

where $\widehat{Z} \equiv (\mathbf{I}\otimes P_{\tilde X})Z$ are the fitted values from the first-stage regression. The covariance matrix simplifies since $\Sigma = \sigma^2\mathbf{I}$:

$$\widehat{\mathrm{Avar}(\hat\delta_{\mathrm{2SLS}})} = n\hat\sigma^2\left[Z'(\mathbf{I} \otimes P_{\tilde X})Z\right]^{-1}$$
$$= n\hat\sigma^2[\widehat{Z}'\widehat{Z}]^{-1};$$

however, if $\Sigma \neq \sigma^2\mathbf{I}$, we instead get

$$= n[\widehat{Z}'\widehat{Z}]^{-1}\widehat{Z}'(\Sigma \otimes \mathbf{I})\widehat{Z}[\widehat{Z}'\widehat{Z}]^{-1}.$$

This is the same estimator as equation-by-equation 2SLS, and the estimated within-equation covariance matrices are also the same. Joint estimation gives the off-diagonal (i.e., cross-equation) blocks of of the covariance matrix.

Seemingly Unrelated Regressions (Hayashi 279–83, 309; Mahajan 7–17) [a.k.a. Joint Generalized Least Squares] Efficient GMM estimator with:

- Conditional homoscedasticity ($\mathrm{E}(\varepsilon\varepsilon'|X) = \Sigma_{M\times M} \otimes \mathbf{I}_{n\times n}$),
- The set of instruments the same across equations ($X = \mathbf{I}_{M\times M} \otimes \tilde X_{n\times K}$, where $\tilde X$ are the instruments for each equation).
- The set of instruments the union of all the regressors.

Note this *is* the 3SLS estimator; it just has another name when instruments are the union of all regressors. Here, all regressors satisfy "cross" orthogonality: each is predetermined in every other equation. Since regressors are a subset of instruments, residuals used to estimate $\widehat{\Sigma}$ are from OLS.[*] Normalized distance function

$$\frac{1}{2n}(Y - Z\tilde\delta)'(\widehat{\Sigma}^{-1} \otimes \mathbf{I})(Y - Z\tilde\delta)$$

yields estimator

$$\hat\delta_{\mathrm{SUR}} = \left[Z'(\widehat{\Sigma}^{-1} \otimes \mathbf{I})Z\right]^{-1} Z'(\widehat{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y},$$
$$\widehat{\mathrm{Avar}(\hat\delta_{\mathrm{SUR}})} = n\left[Z'(\widehat{\Sigma}^{-1} \otimes \mathbf{I})Z\right]^{-1}.$$

Suppose we have linearity; joint ergodic stationarity; orthogonality; mds with finite second moments; conditional homoscedasticity; and $\mathbf{x}_{im} = \bigcup_m \mathbf{z}_{im}$.[†] Then $\hat\delta_{\mathrm{SUR}}$ is consistent, asymptotically normal, and efficient.

If each equation is just identified, then all the regressors must be the same and SUR is called the "multivariate regression"—it is numerically equivalent to equation-by-equation OLS. If some equation is overidentified, we gain efficiency over equation-by-equation OLS as long as some errors are correlated across equations (i.e., $\sigma_{mh} \neq 0$); we are taking advantage of exclusion restrictions vs. multivariate regression.

Multiple-equation GMM with common coefficients (Hayashi 286–9) When all equations have same coefficient vector $\delta \in \mathbb{R}^L$,

1. Linearity becomes $y_{im} = \mathbf{z}_{im}'\delta + \varepsilon_{im}$ for $m = 1, \ldots, M$; $i = 1, \ldots, n$.

[*]When regressors are a subset of instruments, 2SLS becomes OLS.

[†]We need neither the rank conditions (per Hayashi 285 q. 7) nor assumption that $\mathrm{E}[\mathbf{z}_{im}\mathbf{z}_{ih}']$ exists and is finite for all m, h—it is implied.

2. Rank condition requires full column rank of

$$\underbrace{\Sigma_{xz}}_{\sum K_m \times L} \equiv \begin{bmatrix} \mathrm{E}(\mathbf{x}_{i1}\mathbf{z}_{i1}') \\ \vdots \\ \mathrm{E}(\mathbf{x}_{iM}\mathbf{z}_{iM}') \end{bmatrix};$$

Note here Σ_{xz} is a stacked rather than a block matrix, and the rank condition is weaker than in the standard multiple-equation GMM model—rather than requiring equation-by-equation rank conditions, it suffices for *one* equation to satisfy the rank condition.

$$\hat{\delta}_{\mathrm{GMM}} \equiv (S_{xz}'\widehat{W}S_{xz})^{-1}S_{xz}'\widehat{W}s_{xy}$$

(as in standard multiple-equation GMM estimator), but here

$$S_{xz} \equiv \begin{bmatrix} \frac{1}{n}\sum \mathbf{x}_{i1}\mathbf{z}_{i1}' \\ \vdots \\ \frac{1}{n}\sum \mathbf{x}_{iM}\mathbf{z}_{iM}' \end{bmatrix}, s_{xy} \equiv \begin{bmatrix} \frac{1}{n}\sum \mathbf{x}_{i1}y_{i1} \\ \vdots \\ \frac{1}{n}\sum \mathbf{x}_{iM}y_{iM} \end{bmatrix}.$$

Imposition of conditional homoscedasticity (and other appropriate assumptions as above) gives common-coefficient versions of FIVE, 3SLS, and SUR estimators—the latter is the "random effects" estimator.

Random effects estimator (Hayashi 289–90, 292–3) Efficient multiple-equation GMM (SUR) where all equations have same coefficient vector $\delta \in \mathbb{R}^L$, conditional homoscedasticity applies, and the set of instruments for all equations is the union of all regressors.

We can write the model as $y_i = Z_i\delta + \varepsilon_i$, where

$$\underbrace{y_i}_{M \times 1} = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iM} \end{bmatrix}, \underbrace{Z_i}_{M \times L} = \begin{bmatrix} \mathbf{z}_{i1}' \\ \vdots \\ \mathbf{z}_{iM}' \end{bmatrix}, \underbrace{\varepsilon_i}_{M \times 1} = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iM} \end{bmatrix};$$

or as $\mathbf{y} = Z\delta + \varepsilon$ by stacking the above (note Z is no longer block-diagonal as it was for SUR). Then

$$\hat{\delta}_{\mathrm{RE}} = \left(\frac{1}{n}\sum_i Z_i'\widehat{\Sigma}^{-1}Z_i\right)^{-1}\left(\frac{1}{n}\sum_i Z_i'\widehat{\Sigma}^{-1}y_i\right),$$

$$= [Z'(\widehat{\Sigma}^{-1}\otimes\mathbf{I})Z]^{-1}Z'(\widehat{\Sigma}^{-1}\otimes\mathbf{I})\mathbf{y};$$

$$\mathrm{Avar}(\hat{\delta}_{\mathrm{RE}}) = (\mathrm{E}[Z_i'\Sigma^{-1}Z_i])^{-1};$$

$$\widehat{\mathrm{Avar}(\hat{\delta}_{\mathrm{RE}})} = \left(\frac{1}{n}\sum_i Z_i'\widehat{\Sigma}^{-1}Z_i\right)^{-1},$$

$$= n[Z'(\widehat{\Sigma}^{-1}\otimes\mathbf{I})Z]^{-1}.$$

Pooled OLS (Hayashi 290–3, 309–10) *Inefficient* multiple-equation GMM where all equations have same coefficient vector $\delta \in \mathbb{R}^L$ calculated by pooling all observations and applying OLS. Equivalent to GMM with weighting matrix $\widehat{W} = \mathbf{I}_M \otimes$

$(\frac{1}{n}\sum \mathbf{x}_i\mathbf{x}_i')^{-1}$ rather than $\widehat{\Sigma}^{-1}\otimes(\frac{1}{n}\sum \mathbf{x}_i\mathbf{x}_i')^{-1}$. *Caution:* Running a standard OLS will give incorrect standard errors.

Using matrix notation as in random effects estimator,

$$\hat{\delta}_{\mathrm{pOLS}} = \left(\frac{1}{n}\sum_i Z_i'Z_i\right)^{-1}\left(\frac{1}{n}\sum_i Z_i'y_i\right),$$

$$= (Z'Z)^{-1}Z'\mathbf{y};$$

$$\mathrm{Avar}(\hat{\delta}_{\mathrm{pOLS}}) = (\mathrm{E}[Z_i'Z_i])^{-1}\mathrm{E}[Z_i'\Sigma Z_i](\mathrm{E}[Z_i'Z_i])^{-1};$$

$$\widehat{\mathrm{Avar}(\hat{\delta}_{\mathrm{pOLS}})} =$$

$$\left(\frac{1}{n}\sum_i Z_i'Z_i\right)^{-1}\left(\frac{1}{n}\sum_i Z_i'\widehat{\Sigma}Z_i\right)\left(\frac{1}{n}\sum_i Z_i'Z_i\right)^{-1},$$

$$= n(Z'Z)^{-1}[Z'(\widehat{\Sigma}\otimes\mathbf{I})Z](Z'Z)^{-1}.$$

1.17 ML for multiple-equation linear models

Structural form (D&M 212–3; Hayashi 529–30) Equations each present one endogenous variable as a function of exogenous and endogenous variables (and an error term). It is conventional to write simultaneous equation models so that each endogenous variable appears on the left-hand side of exactly one equation, "but there is nothing sacrosancy about this convention."

Can also be written with the errors a function of all variables ("one error per equation")

Reduced form (D&M 213–4; Hayashi 529–30) Each endogenous variable is on the left side of exactly one equation ("one endogenous variable per equation"), and *only predetermined variables* (exogenous variables and lagged endogenous variables) are on the right side (along with an error term).

Full Information Maximum Likelihood (Hayashi 526–35; MaCurdy) Application of (potentially quasi-) maximum likelihood to 3SLS model with two extensions: iid (dependent variables, regressors, and instruments), and conditionally homoscedastic normally distributed errors. Model requires:

1. Linearity: $y_{tm} = z_{tm}'\delta_m + \varepsilon_{tm}$ for $m = 1, \ldots, M$;

2. Rank conditions: $\mathrm{E}[x_t z_{tm}']$ of full column rank for all m;

3. $\mathrm{E}[x_t x_t']$ nonsingular;

4. The system can be written as a "complete system" of simultaneous equations—let y_t be a vector of all endogenous variables (those in $y_{t1}, \ldots, y_{tM}, z_{t1}, \ldots, z_{tM}$ but *not* in x_t)*:

   • The number of endogenous variables equals the numbers of equations (i.e., $y_t \in \mathbb{R}^M$);

• Structural form

$$\underbrace{\Gamma}_{M\times M}\underbrace{y_t}_{M\times 1} + \underbrace{B}_{M\times K}\underbrace{x_t}_{K\times 1} = \underbrace{\varepsilon_t}_{M\times 1}$$

has nonsingular Γ;

5. $\varepsilon_t | x_t \sim \mathrm{N}(0, \Sigma)$ for positive definite Σ;

6. The vector that collects elements of y_t ($\equiv y_{t1}, \ldots, y_{tM}, z_{t1}, \ldots, z_{tM}$) and x_t is iid (stronger than a jointly ergodic stationary mds);

7. (δ, Γ) is an interior points in a compact parameter space.

We can rewrite complete system as a reduced form

$$y_t = \underbrace{\Pi'}_{-\Gamma^{-1}B}x_t + \underbrace{v_t}_{\Gamma^{-1}\varepsilon_t}.$$

x_t is orthogonal to vector of errors v_t, so this is a multivariate regression model; given our normality assumption, we have

$$y_t | x_t \sim \mathrm{N}(-\Gamma^{-1}B_0 x_t, \Gamma^{-1}\Sigma(\Gamma^{-1})'),$$

and can estimate δ (on which Γ and B depend) and Σ by maximum likelihood. Generally use concentrated ML to first maximize over Σ to get $\widehat{\Sigma}(\delta) = \frac{1}{n}\sum_{t=1}^n (\Gamma y_t + B x_t)(\Gamma y_t + B x_t)'$ and then over δ.

Estimator is consistent and asymptotically normal (and asymptotically equivalent to 3SLS[†]) *even if $\varepsilon_t | x_t$ is not normal*, as long as model is linear in y's and errors (not necessarily in regressors or parameters).

Maximum likelihood for SUR (Hayashi 525–7; MaCurdy P.S. 3 1(v)) A specialization of FIML, since SUR is 3SLS with the set of instruments the set of all regressors. In FIML notation, that means z_{tm} is a subvector of x_t; therefore y_t just collects the (sole remaining endogenous variables) $y_{t1}, \ldots, y_{tM}$ and the structural form parameter $\Gamma = \mathbf{I}$. We get a closed form expression (for some consistent estimator $\widehat{\Sigma}$), unlike for general FIML:

$$\hat{\delta}(\widehat{\Sigma}) = [Z'(\widehat{\Sigma} - 1 \otimes \mathbf{I})Z]^{-1}Z'(\widehat{\Sigma}1 \otimes \mathbf{I})y.$$

LR statistic is valid even if the normality assumption fails to hold.

---

* *Caution:* y_t does not merely collect $y_{t1}, \ldots, y_{tM}$ as usual.

† Per Hayashi 534 and MaCurdy, this asymptotic equivalence does not carry over to nonlinear (in endogenous variables and/or errors) equation systems; nonlinear FIML is more efficient than nonlinear 3SLS.

Limited Information Maximum Likelihood <small>(Hayashi 538–42; MaCurdy)</small> Consider the mth equation in the FIML framework, and partition regressors into endogenous $\tilde{y}_t$ and predetermined $\tilde{x}_t$:

$$y_{tm} = \underbrace{z'_{tm}}_{1\times L_m} \underbrace{\delta_m}_{L_m \times 1} + \varepsilon_t = \underbrace{\tilde{y}'_t}_{1\times M_m} \underbrace{\gamma_m}_{M_m\times 1} + \underbrace{\tilde{x}'_t}_{1\times K_m} \underbrace{\beta_m}_{K_m\times 1} + \varepsilon_t$$

where $L_m = M_m + K_m$.

Combining with the relevant rows from the FIML reduced form $y_t = \Pi' x_t + v_t$ gives a complete $1 + M_m$ equation system:

$$\underbrace{\bar{\Gamma}}_{(1+M_m)\times(1+M_m)} \underbrace{\bar{y}_t}_{(1+M_m)\times 1} + \underbrace{\bar{B}}_{(1+M_m)\times K} \underbrace{x_t}_{K\times 1} = \underbrace{\bar{\varepsilon}_t}_{(1+M_m)\times 1}$$

with

$$\bar{y}_t \equiv \begin{bmatrix} y_{tm} \\ \tilde{y}_t \end{bmatrix}, \quad \bar{\varepsilon}_t \equiv \begin{bmatrix} \varepsilon_{tm} \tilde{v}_t \end{bmatrix},$$

$$\bar{\Gamma} \equiv \begin{bmatrix} 1 & -\gamma_{m'} \\ 0 & \mathbf{I}_{M_m} \end{bmatrix}, \quad \bar{B} \equiv \begin{bmatrix} -\beta'_m & \mathbf{0}' \\ & -\tilde{\Pi}' \end{bmatrix},$$

and $\tilde{x}_t$ assumed to be the first K_m elements of x_t.

We estimate this structural form by ML as in FIML. Estimator is consistent, and asymptotically normal *even if $\varepsilon_t|x_t$ is not normal*, as long as model is linear in y's and errors (not necessarily in regressors or parameters). It has the same asymptotic distribution as 2SLS (again, if specification is linear), but generally outperforms 2SLS in finite samples.

1.18   Unit root processes

Integrated process of order 0 <small>(Hayashi 558)</small> A sequence of r.v.s is an $I(0)$ process if it:

1. is strictly stationary;
2. has long-run variance $\sum_{j\in\mathbb{Z}} \gamma_j \in (0,\infty)$.

It can be written as $\delta + u_t$, where $\{u_t\}$ is a zero-mean stationary process with positive, finite long-run variance.

Integrated process of order d <small>(Hayashi 559)</small> A sequence of r.v.s $\{\xi_t\}$ is an $I(d)$ process if its dth differences $\Delta^d \xi_t \equiv (1-L)^d \xi_t$ are an $I(0)$ process.

Unit root process <small>(Hayashi 559–60, 565-6)</small> An $I(1)$ process, a.k.a. difference-stationary process. If first-differences are mean-zero, the process is called "driftless."

If $\Delta \xi_t \equiv \delta + u_t \equiv \delta + \psi(L)\varepsilon_t$, the process can be written

$$\xi_t = \xi_0 + \delta t + \sum_{s=1}^{t} u_s$$

$$= \underbrace{\delta t}_{\text{time trend}} + \underbrace{\psi(1)\sum_{s=1}^{t}\varepsilon_s}_{\text{stoch. trend}} + \underbrace{\eta_t}_{\text{stationary proc.}} + (\xi_0 - \eta_0)$$

<small>*The last expression assumes $\{\Delta\xi_t\}$ is iid white noise.</small>

where $\eta_t \equiv \alpha(L)\varepsilon_t$ and $\alpha_j = -(\psi_{j+1} + \psi_{j+2} + \cdots)$.

Eventually, the time trend (assuming there is one, i.e., $\delta \neq 0$) dominates the stochastic trend a.k.a. driftless random walk, which dominates which dominates the stationary process.

Brownian motion <small>(Hansen B2-5–7, 18, 21; Hayashi 567–72)</small> The random CADLAG (continuous from right, limit from left) function $W\colon [0,1] \to \mathbb{R}$ (i.e., an element of $\mathbb{D}[0,1]$) such that:

1. $W(0) = 0$ almost surely;
2. For $0 \leq t_1 < t_2 < \cdots < t_k \leq 1$, the random variables $W(t_2) - W(t_1)$, $W(t_3) - W(t_2), \ldots, W(t_k) - W(t_{k-1})$ are independent normals with $W(s) - W(t) \sim \mathrm{N}(0, s-t)$;
3. $W(t)$ is continuous in t with probability 1.

If $\{\xi_t\}$ is a driftless $I(1)$ process, so that $\Delta\xi_t$ is zero-mean $I(0)$ with λ^2 the long-run variance of $\{\Delta\xi_t\}$ and $\gamma_0 \equiv \mathrm{Var}(\Delta\xi_t)$, then

$$\frac{1}{T^2}\sum_{t=1}^{T}\xi_{t-1}^2 \xrightarrow{\mathrm{d}} \lambda^2 \int_0^1 W(r)^2 \, dr$$

$$\frac{1}{T}\sum_{t=1}^{T}\Delta\xi_t \xi_{t-1} \xrightarrow{\mathrm{d}} \tfrac{1}{2}\left(\lambda^2 W(1)^2 - \gamma_0\right)$$

$$\equiv \sigma_{\Delta\xi}^2 \int_0^1 W(u)\, dW(u)$$

jointly (i.e., the $W(\cdot)$ are all the same).*

Function Central Limit Theorem <small>(Hansen B2-9–17)</small> [Donsker's Invariance Principle] A sequence $\{Y_t\}$ (where $Y_0 = 0$) can be mapped into a $\mathbb{D}[0,1]$ function by

$$W_n(u) \equiv Y_{\lfloor un \rfloor}$$

or equivalently by

$$\equiv Y_t \quad \text{for } u \in \left[\tfrac{t}{n}, \tfrac{t+1}{n}\right).$$

If $Y_t \equiv \sum_{s=1}^{t}\varepsilon_s$ for $\{\varepsilon_s\}$ iid white noise, then $n^{-1/2}W_n(u) \equiv n^{-1/2}Y_{\lfloor un \rfloor} \xrightarrow{\mathrm{d}} \sigma_\varepsilon W(u)$, where $W(\cdot)$ is a Brownian motion.

Dickey-Fuller Test <small>(Hayashi 573–7; Hansen B2-2–4)</small> Test whether an AR(1) process has a unit root (the null hypothesis). In case without drift, use OLS to regress $y_t = \rho y_{t-1} + \varepsilon_t$ where $\{\varepsilon_t\}$ iid white noise. Under null,

$$T(\hat{\rho}_T - 1) = \frac{\frac{1}{T}\sum_t Y_{t-1}\varepsilon_t}{\frac{1}{T^2}\sum_t Y_{t-1}^2} \xrightarrow{\mathrm{d}} \frac{\int_0^1 W(u)\, dW(u)}{\int_0^1 W(u)^2\, du} \equiv \mathrm{DF}.$$

Tests with drift or non-zero mean covered in Hayashi.

1.19   Limited dependent variable models

Binary response model <small>(D&M 512–4, 517–21; MaCurdy)</small> Dependent variable $y_t \in \{0,1\}$ and information set at time t is Ω_t. Then

$$\mathrm{E}[y_t|\Omega_t] = \mathrm{Pr}[y_t = 1|\Omega_t] \equiv \underbrace{F(\overbrace{h(x_t, \beta)}^{\text{index}})}_{\text{transformation}}.$$

If $\mathrm{E}[y_t|\Omega_t] = F(x'_t\beta)$, then FOCs for maximizing log likelihood are same as for weighted least squares with weights $[F(1-F)]^{-1/2}$, which is reciprocal of squareroot of error variance.

FOCs may not have a finite solution if data set doesn't identify all parameters.

Probit model <small>(Hayashi 451–2, 460, 466–7,477–8; D&M 514–5; MaCurdy)</small> Binary response model with $\mathrm{E}[y_t|\Omega_t] = \Phi(x'_t\beta)$.

Can also think of as $y_t^* = x'_t\beta + \varepsilon_t$ where $\varepsilon_t \sim \mathrm{N}(0,1)$ and $y_t = \mathbb{I}_{y_t^* > 0}$.

Log likelihood function (QML if non-iid ergodic stationary)

$$\sum_t y_t \Phi(x'_t\beta) + (1 - y_t)(1 - \Phi(x'_t\beta))$$

is concave. MLE is identified iff $\mathrm{E}[x_t x'_t]$ is nonsingular. MLE is consistent and asymptotically normal if identification condition holds, and if $\{y_t, x_t\}$ is ergodic stationary.

Logit model <small>(Hayashi 508–1; D&M 515)</small> Binary response model with $\mathrm{E}[y_t|\Omega_t] = \Lambda(x'_t\beta)$ where $\Lambda(\cdot)$ is the logistic function: $\Lambda(v) \equiv \exp(v)/(1 + \exp(v))$.

Can also think of as $y_t^* = x'_t\beta + \varepsilon_t$ where ε_t have extreme value distribution and $y_t = \mathbb{I}_{y_t^* > 0}$.

Log likelihood function (QML if non-iid ergodic stationary)

$$\sum_t y_t \Lambda(x'_t\beta) + (1 - y_t)(1 - \Lambda(x'_t\beta))$$

is concave. MLE is identified iff $\mathrm{E}[x_t x'_t]$ is nonsingular. MLE is consistent and asymptotically normal if identification condition holds, and if $\{y_t, x_t\}$ is ergodic stationary.

Truncated response model (Hayashi 511–7; D&M 534–7; MaCurdy) Model is $y_t = x_t'\beta_0 + \varepsilon_t$ with $\varepsilon_t|x_t \sim N(0, \sigma_0^2)$, but we only see observations where $y_t > c$ for some known threshold c. Density for each observation is

$$f(y_t|y_t > c) = \frac{f(y_t)}{\Pr(y_t > c)}.$$

Note that if $y_t \sim N(\mu_0, \sigma_0^2)$,

$$E[y_t|y_t > c] = \mu_0 + \sigma_0\lambda(v)$$

$$\text{Var}[y_t|y_t > c] = \sigma_0^2[1 - \lambda'(v)] = \sigma_0^2[1 - \lambda(v)[\lambda(v) - v]]$$

where $v \equiv \frac{c-\mu_0}{\sigma_0}$ and $\lambda(v) \equiv \frac{\phi(v)}{1-\Phi(v)}$ is the inverse Mills ratio. Could estimate

$$E[y_t|y_t > c] = x_t'\beta_0 + \sigma_0\lambda\left(\frac{c - x_t'\beta_0}{\sigma_0}\right)$$

by nonlinear least squares, but ML is preferred because it is asymptotically more efficient (and we have already imposed full distributional assumptions).

Extension to nonlinear index function (in place of $x_t'\beta$) is straightforward.

Tobit model (Hayashi 518–21; D&M 537–42; MaCurdy) [a.k.a. censored response model] All observations are observed, but we do not see y_t if it is less than some known threshold c; "like a truncated regression model combined with a probit model, with the coefficient vectors restricted to be proportional to each other."

Model is

$$y_t = \max\{\underbrace{x_t'\beta_0 + \varepsilon_t}_{\equiv y_t^*}, c\}$$

where $\varepsilon_t|x_t \sim N(0, \sigma_0^2)$ and $\{x_t, y_t\}$ iid. Log-likelihood (a mix of density and mass) is

$$\sum_{y_t > c} \log\left[\frac{1}{\sigma_0}\phi\left(\frac{y_t - x_t'\beta_0}{\sigma_0}\right)\right] + \sum_{y_t = c}\log\Phi\left(\frac{c - x_t'\beta_0}{\sigma_0}\right).$$

Reparametrize to get concavity.

Sample selection model (D&M 542–5; MaCurdy) Rather than truncating according to dependent variable, truncate on another variable correlated with it. Suppose y_t^* (e.g., wage), and z_t^* (e.g., weeks) are latent variables with

$$\begin{bmatrix} y_t^* \\ z_t^* \end{bmatrix} = \begin{bmatrix} x_t'\beta \\ w_t'\gamma \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}, \quad \begin{bmatrix} u_t \\ v_t \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix}\right),$$

and x_t and w_t exogenous or predetermined. We observe z_t (the sign of z_t^*) and $y_t = y_t^*$ if $z_t^* > 0$.

For observations where $z_t^* > 0$ (and hence we observe y_t^*), model becomes

$$y_t = x_{t'}\beta + \rho\sigma\frac{\phi(w_t'\gamma)}{\Phi(w_t'\gamma)} + \varepsilon_t.$$

Maximum Likelihood is best estimation technique (called "generalized Tobit" here), but also sometimes estimated using either

1. Heckman two-step: first use probit estimation on all observations to estimate $\hat{\gamma}$, and then use OLS for observations where $z_t^* > 0$ to estimate $\hat{\beta}$; note the reported covariance matrix in the second step will *not* be correct.

2. Nonlinear least squares: only uses any data on observations where $z_t^* > 0$; generally has identification problems, since estimation is principally based on exploiting nonlinearity of $\phi(\cdot)/\Phi(\cdot)$.

1.20   Inequalities

Bonferroni's Inequality (C&B 1.2.10–11) Bounds below the probability of an intersection in terms of individual events: $P(A \cap B) \geq P(A) + P(B) - 1$; or more generally

$$P\left(\bigcap_{i=1}^n A_i\right) \geq \sum_{i=1}^n P(A_i) - (n-1) = 1 - \sum_{i=1}^n P(A_i^C).$$

Boole's Inequality (C&B 1.2.11) Bounds above the probability of a union in terms of individual events: $P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i)$.

Chebychev's Inequality (C&B 3.6.1–2, 5.8.3; Greene 66) A widely applicable but "necessarily conservative" inequality: for any $r > 0$ and $g(x)$ nonnegative,

$$P(g(X) \geq r) \leq E\,g(X)/r.$$

Letting $g(x) = (x - \mu)^2/\sigma^2$ where $\mu = E\,X$ and $\sigma^2 = \text{Var}\,X$ gives $P(|X - \mu| \geq t\sigma) \leq t^{-2}$.

See C&B 5.8.3 for a (complex) form that applies to sample mean and variance rather than the (presumably unknown) population values.

Markov's Inequality (C&B 3.8.3) Like Chebychev, but imposes conditions on Y to provide more information at equality point. If $P(Y \geq 0) = 1$ (i.e., Y is nonnegative) and $P(Y = 0) < 1$ (i.e., Y not trivially 0), and $r > 0$, then $P(Y \geq r) \leq E\,Y/r$ with equality iff $P(Y = r) = 1 - P(Y = 0)$.

Numerical inequality lemma (C&B 4.7.1) Let a and b be any positive numbers, p and q positive (necessarily > 1) satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then $\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$, with equality iff $a^p = b^q$.

Hölder's Inequality (C&B 4.7.2) Let positive p and q (necessarily > 1) satisfy $\frac{1}{p} + \frac{1}{q} = 1$, then

$$|EXY| \leq E\,|XY| \leq (E\,|X|^p)^{1/p}(E\,|Y|^q)^{1/q}.$$

Cauchy-Schwarz Inequality (C&B 4.7.3; Wikipedia)

$$|EXY| \leq E\,|XY| \leq \sqrt{(E\,X^2)(E\,Y^2)}.$$

This is Hölder's Inequality with $p = q = 2$. Implies that $\text{Cov}(X, Y) \leq \sqrt{\text{Var}(X)\,\text{Var}(Y)}$.

Equality obtains iff X and Y are linearly dependent.

Minkowski's Inequality (C&B 4.7.5)

$$[E\,|X + Y|^p]^{1/p} \leq [E\,|X|^p]^{1/p} + [E\,|Y|^p]^{1/p}$$

for $p \geq 1$. Implies that if X and Y have finite pth moment, then so does $X + Y$. Implies that $E\,|\bar{X}|^p \leq E\,|X|^p$.

Jensen's Inequality (C&B 4.7.7) If $g(\cdot)$ is a convex function, then $E\,g(X) \geq g(E\,X)$.

Equality obtains iff, for every line $a + bx$ tangent to $g(x)$ at $x = E\,X$, $P(g(X) = a + bX) = 1$ (i.e., if $g(\cdot)$ is linear almost everywhere in the support of X).

1.21   Thoughts on MaCurdy questions

Default answer

If we assume _____, we know that _____ is consistent, assuming correct specification and e.g., ergodic stationarity, regressors orthogonal to the contemporary error term, instruments orthogonal to the contemporary error term, appropriate rank condition(s) for identification.

We can test the null hypothesis that _____, which corresponds to _____ = 0, using a Wald test on Model __. The statistic W $\equiv$ _____ is asymptotically distributed $\chi^2(\_)$ under the null, thus we can(not) reject the null hypothesis. Note we have used normal/robust standard errors.

We can also test the null hypothesis using a distance principle test, where model __ is the unrestricted model, and model __ represents the restrictions imposed by the null hypothesis. Using a suitably normalized distance function $Q_T \equiv$ _____, the statistic LR $\equiv 2T(Q_r - Q_u) =$ _____ is asymptotically distributed $\chi^2(\_)$ under the null.

[Note we cannot use the multiple equation 2SLS estimates to conduct a distance principle test, since generating a suitably normalized distance function would require having estimates for the variance of the error for each equation (along with the SSR for each equation).]

[Note that the validity of this distance function requires the restricted and unrestricted models to use the same weighting matrix; since _____ $\neq$ _____, the test statistic is *not* valid. However, given the regression output available, this is the best we can do.]

Thus we can(not) reject the null hypothesis.

If we instead assume _____, our tests are no longer valid; in this case _____ is consistent (again, assuming correct specification), so...

Thoughts on models

1. Two models with and without a restriction is for LR testing.

2. A model that includes only fitted values of an endogenous variable is 2SLS.

3. A model that includes both a (potentially) endogenous variable and its fitted values or residuals is for a DWH test.

4. A model that includes lagged residuals is for testing serial correlation.

Which model and standard errors to use

A less restrictive model will still be consistent and asymptotically more efficient; it is only harmed on finite sample properties; same goes for robust standard errors.

Check for exact identification—models often simplify in these cases.

Look ahead at all the questions; often it's best to make an assumption early, since later questions will ask you to give up that assumption.

1. Heteroscedasticity across observations
   - Parametrize and use GLS.
   - Parametrize and use ML.
   - *Others*?

2. Homoscedasticity (across observations)
   - 3SLS is efficient GMM.
   - FIML is asymptotically equivalent to 3SLS.

3. Homoscedasticity (across observations), uncorrelated errors across equations (i.e., Φ diagonal)
   - 3SLS is efficient GMM; 2SLS is *not* efficient GMM (e.g., *cannot* use minimized distance function for LR testing), but it is asymptotically distributed identically to 3SLS
   - 2SLS should generally have better finite sample properties since it uses a priori knowledge of the form of the S matrix (it assumes that S is $\sigma^2 \mathbf{I}$; even though it is actually only diagonal, the resulting estimator is numerically identical to what we would get if we only assumed diagonal S: equation-by-equation 2SLS).
   - FIML, LIML, 3SLS, and 2SLS are all asymptotically equivalent.

4. Homoscedasticity (across observations), uncorrelated errors across equations, homoscedasticity across equations (i.e., $\Phi = \sigma^2 \mathbf{I}$)
   - 2SLS and 3SLS are both efficient GMM; they are asymptotically equal and identically distributed.

- 2SLS should generally have better finite sample properties since it uses a priori knowledge of the form of the S matrix.
- FIML, LIML, 3SLS, and 2SLS are all asymptotically equivalent.

Hypothesis tests

Wald: basically always possible as long as estimator is consistent and standard errors are calculated appropriately

- Be careful when using formulae from Hayashi; they are generally based on the *asymptotic* variance matrix; regression output doesn't estimate this, it estimates $\frac{1}{n}$ times the asymptotic variance matrix.
- State assumptions that justify consistency of estimation technique, and standard errors used.

Distance principle / LR

- Need to use $2n$ times a *suitably normalized* maximized objective function, i.e., one that satisfies "the property":

$$\text{Var}\left[\sqrt{T} \left. \frac{\partial Q_T}{\partial \theta}\right|_{\theta_0}\right] \stackrel{a}{=} \pm \left. \frac{\partial^2 Q_T}{\partial \theta \, \partial \theta'}\right|_{\theta_0}.$$

- For efficient GMM, $Q_T \equiv \frac{1}{2}(X'e)'\widehat{\Omega}^{-1}(X'e)$ satisfies "the property," where $\widehat{\Omega}$ consistently estimates the variance of $\sqrt{T}X'e$.
- We cannot generally use LR tests in the presence of heteroscedasticity.
- For GMM estimators, must use efficient weighting matrix for LR to be valid.
- Restricted and unrestricted estimates should use same weighting matrix to be valid.
- Including the $2n$ and the normalization, use:

1. MLE: $2 \cdot \text{loglik}$
   Reported value is usually loglik; required normalization is 2.
   - Valid no matter what if model is correctly specified.

2. LS: $\text{SSR}/\hat{\sigma}^2 = e'e/\hat{\sigma}^2$
   Reported value is usually SSR; required normalization is $1/\hat{\sigma}^2$.
   - *Invalid* under heteroscedasticity.
   - Validity under serial correlation???
   - Make sure to use the same estimate $\hat{\sigma}^2$ in both restricted and unrestricted models (typically SSR/n or $\text{SSR}/(n-k)$ from one of the two models).

3. 2SLS (single-equation): $\frac{1}{\hat{\sigma}^2}e'P_z e$
   Reported value is usually $e'P_z e$; required normalization is $1/\hat{\sigma}^2$.
   - *Invalid* under heteroscedasticity (not efficient GMM).
   - Validity under serial correlation???
   - Make sure to use the same estimate for $\hat{\sigma}^2$ in both restricted and unrestricted models (typically SSR/n or $\text{SSR}/(n-k)$ from one of the two models).

4. SUR / JGLS: $e'(\widehat{\Phi}^{-1} \otimes \mathbf{I})e$
   Reported value is usually as desired; no normalization is needed.
   - *Invalid* under heteroscedasticity (across observations–not efficient GMM; however, it's fine with heteroscedasticity across equations).
   - Validity under serial correlation???
   - **Watch out**: this isn't actually valid, since restricted and unrestricted estimation will use a *different* weighting matrix; however, we often comment that this isn't valid, but use it anyway since that's the best output available.

5. 2SLS (multiple-equation):

$$e'\left(\begin{bmatrix} \hat{\sigma}_1^2 & & \\ & \ddots & \\ & & \hat{\sigma}_M^2 \end{bmatrix}^{-1} \otimes P_z\right)e$$

   Reported value is usually $e'(\mathbf{I} \otimes P_z)e$; no normalization can get us a usable distance function.
   - *Invalid* under heteroscedasticity (across observations—not efficient GMM; however, it's fine with heteroscedasticity across equations).
   - *Invalid* if errors are correlated across equations (i.e., Phi is not a diagonal matrix)
   - Validity under serial correlation???
   - I don't think any of the regression programs actually give this output (they typically give $e'(\mathbf{I}\otimes P_z)e$; maximizing this objective function gives the same estimator, but not the same maximized objective function); thus we *can't do LR testing here* here, even though it may be theoretically valid.

6. 3SLS: $e'(\widehat{\Phi}^{-1} \otimes P_z)e$
   Reported value is usually as desired; no normalization is needed.
   - *Invalid* under heteroscedasticity (across observations—not efficient GMM; however, it's fine with heteroscedasticity across equations).
   - Validity under serial correlation???
   - **Watch out**: this isn't actually valid, since restricted and unrestricted estimation will use a *different* weighting matrix; however, we often comment that this isn't valid, but use it anyway since that's the best output available.

Multiple equation systems

1. **Structural equation**: can have multiple endogenous variables; typically written as explicitly solved for one endogenous variable.

2. **Reduced form equation**: only *one* endogenous variable per equation (and one equation per endogenous variable).

Instruments

1. Must be correlated with endogenous variable and uncorrelated with error.

2. Typically things that are predetermined—not meant technically, they were actually determined before the endogenous variable.

3. Predetermined variables, lagged endogenous variables, interactions of the above.

4. In 2SLS, check the quality of instruments' correlation with endogenous variable (note this does *not* check the lack of correlation with error) by looking at the first stage regression:

   - R^2;
   - t-statistics on each instrument;
   - F-statistic for model as a whole, and potentially F-statistic on excluded instruments.

Omitted variables

When we get asked to find out what happens if variables are excluded (i.e., an incorrectly specified model is estimated), a good tool is the Frisch-Waugh Theorem.

Finding probability limits/showing consistency

1. Solve as explicit function of data and use LLN with CMT/Slutsky.

2. Solve as explicit function of data and show bias $\to 0$ and variance $\to 0$.

3. Solve explicitly, find the probability that $|\hat{\theta} - \theta_0| < \varepsilon$, and show the limit of this probability is 1 (per C&B 468).

4. MaCurdy: if the estimator is defined by $L_T(\hat{\theta}) \equiv \frac{1}{T} \sum_t l_t(\hat{\theta}) = 0$, show that $l_t(\theta_0) \sim \text{niid}(0)$, and that it satisfies an LLN so that $L_T(\theta_0) \xrightarrow{\text{p}} 0$.

5. MaCurdy: if the estimator is defined by minimizing $H_T(\hat{\theta})' M_T H_T(\hat{\theta})$ with $H_T \equiv \frac{1}{T} \sum_t h_t$, show that $h_t(\theta_0) \sim \text{niid}(0)$, and that it satisfies an LLN so that $H_T(\theta_0) \xrightarrow{\text{p}} 0$.

6. Aprajit/Hayashi: general consistency theorems with and without compact parameter space.

Finding asymptotic distributions

1. CLTs for

   - iid sample
   - niid sample
   - ergodic stationary process
   - ergodic stationary mds
   - MA(∞)

2. Delta method

3. MLE is asymptotically normal with variance equal to inverse of Fisher info (for a single observation, not joint distribution)

Unreliable standard errors

1. Remember standard errors are *asymptotic estimates* except in Gaussian OLS; therefore finite sample inference may be inaccurate.

2. If you run a multiple-stage regression technique in separate stages (e.g., sticking in fitted values along the way).

3. If you stick something into the regression that doesn't "belong" (e.g., fitted values for a DWH test—although for some reason this one may be OK, inverse Mills for sample selection, . . .).

4. Heteroscedastic errors (when not using robust standard errors).

5. Serially correlated errors (when not using HAC standard errors).

6. Poor instruments (see above).

7. When regression assumptions fail (e.g., using regular standard errors when inappropriate, failure of fourth moment assumptions, . . .).

2 Microeconomics

2.1 Choice Theory

Rational preference relation (Choice 4; MWG 6–7) A binary relation $\succsim$ is a rational preference relation iff it satisfies

1. Completeness: $\forall x, y, x \succsim y \lor y \succsim x$ (NB: implies $x \succsim x$);

2. Transitivity: $\forall x, y, z, (x \succsim y \land y \succsim z) \implies x \succsim z$ (which rules out cycles, except where there's indifference).

If $\succsim$ is rational, then $\succ$ is both irreflexive and transitive; $\sim$ is reflexive, transitive, and symmetric; and $x \succ y \succsim z \implies x \succ z$.

Choice rule (Choice 6) Given a choice set B and preference relation $\succsim$, choice rule $C(B, \succsim) \equiv \{x \in B : \forall y \in B, x \succsim y\}$. This correspondence gives the set of "best" elements of B.

If $\succsim$ is complete and transitive and $|B|$ finite and non-empty, then $C(B, \succsim) \neq \varnothing$.

Revealed preference (Choice 6–8; MWG 11) We observe choices and deduce a preference relation. Consider revealed preferences $C : 2^B \to 2^B$ satisfying $\forall A, C(A) \subseteq A$. Assuming the revealed preference sets are always non-empty, there is a well-defined preference relation $\succsim$ (complete and transitive) satisfying $\forall A, C(A, \succsim) = C(A)$ iff C satisfies HARP (or WARP, and the set of budget sets contains all subsets of up to three elements).

Houthaker's Axiom of Revealed Preferences (Choice 6-88) A set of revealed preferences $C : 2^B \to 2^B$ satisfies HARP iff $\forall x, y \in U \cap V$ such that $x \in C(U)$ and $y \in C(V)$, it is also the case that $x \in C(V)$ and $y \in C(U)$. In other words, suppose two different choice sets both contain x and y; if x is preferred to all elements of one choice set, and y is preferred to all elements of the other, then x is also preferred to all elements of the second, and y is also preferred to all elements of the first.

Weak Axiom of Revealed Preference (MWG 10–11) Choice structure $(\mathcal{B}, C(\cdot))$—where $\mathcal{B}$ is the set of budget sets— satisfies WARP iff $B, B' \in \mathcal{B}$; $x, y \in B$; $x, y \in B'$; $x \in C(B)$; $y \in C(B')$ together imply that $x \in C(B')$. (Basically, HARP, but only for budget sets).

Generalized Axiom of Revealed Preference (Micro P.S. 1.3) A set of revealed preferences $C : \mathcal{A} \to \mathcal{B}$ satisfies GARP if for any sequences $A_1, \ldots, A_n$ and $x_1, \ldots, x_n$ where

1. $\forall i \in \{1, \ldots, n\}, x_i \in A_i$;
2. $\forall i \in \{1, \ldots, n-1\}, x_{i+1} \in C(A_i)$;
3. $x_1 \in C(A_n)$;

then $x_i \in C(A_i)$ for all i. (That is, there are no revealed preference cycles except for revealed indifference.)

Utility function (Choice 9-13; MWG 9) Utility function $u : X \to \mathbb{R}$ represents $\succsim$ on X iff $x \succsim y \iff u(x) \geq u(y)$.

1. This turns choice rule into a maximization problem: $C(B, \succsim) = \operatorname{argmax}_{y \in B} u(y)$.

2. A preference relation can be represented by a utility function only if it is rational (complete and transitive).

3. If $|X|$ is finite, then any rational preference relation $\succsim$ can be represented by a utility function; if $|X|$ is infinite, this is not necessarily the case.

4. If $X \subseteq \mathbb{R}^n$, then $\succsim$ (complete, transitive) can be represented by a continuous utility function iff $\succsim$ is continuous (i.e., $\lim_{n \to \infty} (x_n, y_n) = (x, y)$ and $\forall n, x_n \succsim y_n$ imply $x \succsim y$).

5. The property of representing $\succsim$ on X is ordinal (i.e., invariant to monotone increasing transformations).

Interpersonal comparison (Choice 13) It is difficult to weigh utility tradeoffs between people. Two possible systems are Rawls' "veil of ignorance" (which effectively makes all the choices one person's), and a system of "just noticeable differences" (which suffers transitivity issues).

Continuous preference (Choice 11; MWG 46–7, Micro P.S. 1-5) $\succsim$ on X is continuous if for any sequence $\{(x_n, y_n)\}_{n=1}^\infty$ with $\lim_{n\to\infty}(x_n, y_n) = (x, y)$ and $\forall n$, $x_n \succsim y_n$, we have $x \succsim y$. Equivalently, $\succsim$ is continuous iff for all x, the upper and lower contour sets of x are both closed sets. $\succsim$ is rational and continuous iff it can be represented by a continuous utility function.

Monotone preference (Choice 15–6; MWG 42–3) $\succsim$ is monotone iff $x \geq y \implies x \succsim y$ (i.e., more of something is better).

(MWG uses $x \gg y \implies x \succ y$; this is *not* equivalent.)

Strictly/strongly monotone iff $x > y \implies x \succ y$.

$\succsim$ is (notes) monotone iff $u(\cdot)$ is nondecreasing. $\succsim$ is strictly monotone iff $u(\cdot)$ monotone increasing. Strictly monotone $\implies$ (notes or MWG) monotone. MWG monotone $\implies$ locally non-satiated (on e.g., $\mathbb{R}_+^n$).

Locally non-satiated preference (Choice 15–6; MWG 42–3) $\succsim$ is locally non-satiated on X iff for any $y \in X$ and $\varepsilon > 0$, there exists $x \in X$ such that $\|x - y\| \leq \varepsilon$ and $x \succ y$ (i.e., there are no "thick" indifference curves). $\succsim$ is locally non-satiated iff $u(\cdot)$ has no local maxima in X. Strictly monotone $\implies$ MWG monotone $\implies$ locally non-satiated (on e.g., $\mathbb{R}_+^n$).

Convex preference (Choice 15–6; MWG 44–5) $\succsim$ is convex on X iff (*de facto*, X is a convex set, and) $x \succsim y$ and $x' \succsim y$ together imply that $\forall t \in (0,1)$, $tx + (1-t)x' \succsim y$ (i.e., one never gets worse off by mixing goods). Equivalently, $\succsim$ is convex on X iff the upper contour set of any $y \in X$ (i.e., $\{x \in X : x \succsim y\}$) is a convex set. Can be interpreted as diminishing marginal rates of substitution.

$\succsim$ is strictly convex on X iff X is a convex set, and $x \succsim y$ and $x' \succsim y$ (with $x \neq x'$) together imply that $\forall t \in (0,1)$, $tx + (1-t)x' \succ y$.

$\succsim$ is (strictly) convex iff $u(\cdot)$ is (strictly) quasi-concave.

Homothetic preference (MWG 45, Micro P.S. 1.6) $\succsim$ is homothetic iff for all $\lambda > 0$, $x \succsim y \iff \lambda x \succsim \lambda y$. (MWG uses $\forall \lambda \geq 0$, $x \sim y \implies \lambda x \sim \lambda y$.) A continuous preference relation is homothetic iff it can be represented by a utility function that is homogeneous of degree one (note it can also be represented by utility functions that aren't).

Separable preferences (Choice p.18–9) Suppose $\succsim$ on $X \times Y$ is represented by $u(x, y)$. Then preferences over x do not depend on y iff there exist functions $v : X \to \mathbb{R}$ and $U : \mathbb{R} \times Y \to \mathbb{R}$ such that U is increasing in its first argument and $\forall (x, y)$, $u(x, y) = U(v(x), y)$. Note that this property is asymmetric. Preferences over x given y will be represented by $v(x)$, regardless of y.

Quasi-linear preferences (Choice 20–1; MWG 45) Suppose $\succsim$ on $X = \mathbb{R} \times Y$ is complete and transitive, and that

1. The numeraire good ("good 1") is valuable: $(t, y) \succsim (t', y)$ iff $t \geq t'$;

2. Compensation is possible: For every y, $y' \in Y$, there exists some $t \in \mathbb{R}$ such that $(0, y) \sim (t, y')$;

3. No wealth effects: If $(t, y) \succsim (t', y')$, then for all $d \in \mathbb{R}$, $(t + d, y) \succsim (t' + d, y')$.

Then there exists a utility function representing $\succsim$ of the form $u(t, y) = t + v(y)$ for some $v : Y \to \mathbb{R}$. (Note it can also be represented by utility functions that aren't of this form.) Conversely, any preference relation $\succsim$ on $X = \mathbb{R} \times Y$ represented by a utility function of the form $u(t, y) = t + v(y)$ satisfies the above conditions. (MWG uses slightly different formulation.)

Lexicographic preferences (MWG 46) A preference relation $\succsim$ on $\mathbb{R}^2$ defined by $(x, y) \succsim (x', y')$ iff $x > x'$ or $x = x' \wedge y \geq y'$. Lexicographic preferences are complete, transitive, strongly monotone, and strictly convex; however, they is not continuous and can*not* be represented by any utility function.

2.2 Producer theory

Competitive producer behavior (Producer 1–2) Firms choose a production plan (technologically feasible set of inputs and outputs) to maximize profits. Assumptions include:

1. Firms are price takers (applies to both input and output markets);

2. Technology is exogenously given;

3. Firms maximize profits; should be true as long as

   - The firm is competitive;
   - There is no uncertainty about profits;
   - Managers are perfectly controlled by owners.

Production plan (Producer 4) A vector $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ where an output has $y_k > 0$ and an input has $y_k < 0$.

Production set (Producer 4) Set $Y \subseteq \mathbb{R}^n$ of feasible production plans; generally assumed to be non-empty and closed.

Free disposal (Producer 5) $y \in Y$ and $y' \leq y$ imply $y' \in Y$.

Shutdown (Producer 5) $\mathbf{0} \in Y$.

Nonincreasing returns to scale (Producer 5) $y \in Y$ implies $\alpha y \in Y$ for all $\alpha \in [0, 1]$. Implies shutdown.

Nondecreasing returns to scale (Producer 5, Micro P.S. 2-1) $y \in Y$ implies $\alpha y \in Y$ for all $\alpha \geq 1$. Along with shutdown, implies that $\pi(p) = 0$ or $\pi(p) = +\infty$ for all p.

Constant returns to scale (Producer 5) $y \in Y$ implies $\alpha y \in Y$ for all $\alpha \geq 0$; i.e., nonincreasing *and* nondecreasing returns to scale.

Convexity (Producer 6) y, $y' \in Y$ imply $ty + (1-t)y' \in Y$ for all $t \in [0, 1]$. Vaguely "nonincreasing returns to specialization." If $\mathbf{0} \in Y$, then convexity implies nonincreasing returns to scale. Strictly convex iff for $t \in (0, 1)$, the convex combination is in the interior of Y.

Transformation function (Producer 6, 24) A function $T : \mathbb{R}^n \to \mathbb{R}$ with $T(y) \leq 0 \iff y \in Y$. Can be interpreted as the amount of technical progress required to make y feasible. The set $\{y : T(y) = 0\}$ is the transformation frontier.

Kuhn-Tucker FOC gives necessary condition $\nabla T(y^*) = \lambda p$, which means the price vector is normal to the production possibility frontier at the optimal production plan.

Marginal rate of transformation (Producer 6–7) When the transformation function T is differentiable, MRT between goods k and l is $\mathrm{MRT}_{k,l}(y) \equiv \frac{\partial T(y)}{\partial y_l} / \frac{\partial T(y)}{\partial y_k}$. Measures the extra amount of good k that can be obtained per unit reduction of good l. Equals the slope of the transformation frontier.

Production function (Producer 7) For a firm with only a single output q (and inputs $-z$), defined as $f(z) \equiv \max q$ such that $T(q, -z) \leq 0$. Thus $Y = \{(q, -z) : q \leq f(z)\}$, allowing for free disposal.

Marginal rate of technological substitution (Producer 7) When the production function f is differentiable, MRTS between goods k and l is $\mathrm{MRTS}_{k,l}(z) \equiv \frac{\partial f(z)}{\partial z_l} / \frac{\partial f(z)}{\partial z_k}$. Measures how much of input k must be used in place of one unit of input l to maintain the same level of output. Equals the slope of the isoquant.

Profit maximization (Producer 7–8) The firm's optimal production decisions are given by correspondence $y : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$

$$y(p) \equiv \operatorname*{argmax}_{y \in Y} p \cdot y = \{y \in Y : p \cdot y = \pi(p)\}.$$

Resulting profits are given by

$$\pi(p) \equiv \sup_{y \in Y} p \cdot y.$$

Rationalization: profit maximization functions (Producer 9–11, 13)

1. Profit function $\pi(\cdot)$ is rationalized by production set Y iff $\forall p$, $\pi(p) = \sup_{y \in Y} p \cdot y$.

2. Supply correspondence $y(\cdot)$ is rationalized by production set Y iff $\forall p$, $y(p) \subseteq \operatorname{argmax}_{y \in Y} p \cdot y$.

3. $\pi(\cdot)$ or $y(\cdot)$ is rationalizable if it is rationalized by some production set.

4. $\pi(\cdot)$ and $y(\cdot)$ are jointly rationalizable if they are both rationalized by the same production set.

We seek a Y that rationalizes both $y(\cdot)$ and $\pi(\cdot)$. Consider an "inner bound": all production plans the firm chooses must be feasible ($Y^I \equiv \bigcup_{p \in P} y(p)$). Consider an "outer bound": Y can only include points that don't give higher profits than $\pi(p)$ ($Y^O \equiv \{y : p \cdot y \leq \pi(p) \text{ for all } p \in P\}$).* A nonempty-valued supply correspondence $y(\cdot)$ and profit function $\pi(\cdot)$ on a price set are jointly rationalized by production set Y iff:

1. $p \cdot y = \pi(p)$ for all $y \in y(p)$ (adding-up);

2. $Y^I \subseteq Y \subseteq Y^O$; i.e., $p \cdot y' \leq \pi(p)$ for all p, p', and all $y' \in y(p')$ (Weak Axiom of Profit Maximization).

If we observe a firm's choices for all positive price vectors on an open convex set P, then necessary conditions for rationalizability include:

1. $\pi(\cdot)$ must be a convex function;

2. $\pi(\cdot)$ must be homogeneous of degree one; i.e., $\pi(\lambda p) = \lambda \pi(p)$ for all $p \in P$ and $\lambda > 0$;

3. $y(\cdot)$ must be homogeneous of degree zero; i.e., $y(\lambda p) = y(p)$ for all $p \in P$ and $\lambda > 0$.

Loss function (Producer 12) $L(p, y) \equiv \pi(p) - p \cdot y$. This is the loss from choosing y rather than the profit-maximizing feasible production plan. The outer bound can be written $Y^0 = \{y : \inf_p L(p, y) \geq 0\}$.

Hotelling's Lemma (Producer 14) $\nabla \pi(p) = y(p)$, assuming differentiable $\pi(\cdot)$. Equivalently, $\nabla_p L(p, y)|_{p=p'} = \nabla \pi(p') - y = 0$ for all $y \in y(p')$. An example of the Envelope Theorem. Implies that if $\pi(\cdot)$ is differentiable at p, then $y(p)$ must be a singleton.

Substitution matrix (Producer 15–6) The Jacobian of the optimal supply function, $Dy(p) = [\partial y_i / \partial p_j]$. By Hotelling's Lemma, $Dy(p) = D^2\pi(p)$ (the Hessian of the profit function), hence the substitution matrix is symmetric. Convexity of $\pi(\cdot)$ implies positive semidefiniteness.

Law of Supply (Producer 16) $(p' - p) \cdot (y(p') - y(p)) \geq 0$; i.e., supply curves are upward-sloping. Law of Supply is the finite-difference equivalent of PSD of substitution matrix. Follows from WAPM ($p \cdot y(p) \geq p \cdot y(p')$).

Rationalization: $y(\cdot)$ **and differentiable** $\pi(\cdot)$ (Producer 15) $y : P \to \mathbb{R}^n$ (the correspondence ensured to be a function by Hotelling's lemma, given differentiable $\pi(\cdot)$) and differentiable $\pi : P \to \mathbb{R}$ on an open convex set $P \subseteq \mathbb{R}^n$ are jointly rationalizable iff

1. $p \cdot y(p) = \pi(p)$ (adding-up);

2. $\nabla \pi(p) = y(p)$ (Hotelling's Lemma);

3. $\pi(\cdot)$ is convex.

*If Y is convex and closed and has free disposal, and $P = \mathbb{R}^n_+ \setminus \{\mathbf{0}\}$, then $Y = Y^O$.

The latter two properties imply WAPM. The second describes the FOC of the maximization problem, the third term describes the second-order condition.

Rationalization: differentiable $y(\cdot)$ (Producer 16) Differentiable $y : P \to \mathbb{R}^n$ on an open convex set $P \subseteq \mathbb{R}^n$ is rationalizable iff

1. $y(\cdot)$ is homogeneous of degree zero;

2. The Jacobian $Dy(p)$ is symmetric and positive semidefinite.

We construct $\pi(\cdot)$ by adding-up, ensure Hotelling's Lemma by symmetry and homogeneity of degree zero, and ensure convexity of $\pi(\cdot)$ by Hotelling's lemma and PSD.

Rationalization: differentiable $\pi(\cdot)$ (Producer 17) Differentiable $\pi : P \to \mathbb{R}$ on a convex set $P \subseteq \mathbb{R}^n$ is rationalizable iff

1. $\pi(\cdot)$ is homogeneous of degree one;

2. $\pi(\cdot)$ is convex.

We construct $y(\cdot)$ by Hotelling's Lemma, and ensure adding-up by homogeneity of degree one; convexity of $\pi(\cdot)$ is given.

Rationalization: general $y(\cdot)$ **and** $\pi(\cdot)$ (Producer 17–9) $y : P \rightrightarrows \mathbb{R}^n$ and $\pi : P \to \mathbb{R}$ on a convex set $P \subseteq \mathbb{R}^n$ are jointly rationalizable iff for any selection $\hat{y}(p) \in y(p)$,

1. $p \cdot \hat{y}(p) = \pi(p)$ (adding-up);

2. (Producer Surplus Formula) For any p, $p' \in P$,

$$\pi(p') = \pi(p) + \int_0^1 (p' - p) \cdot \hat{y}(p + \lambda(p' - p)) \, d\lambda;$$

3. $(p' - p) \cdot (\hat{y}(p') - \hat{y}(p)) \geq 0$ (Law of Supply).

Producer Surplus Formula (Producer 17–20) $\pi(p') = \pi(p) + \int_0^1 (p' - p) \cdot \hat{y}(p + \lambda(p' - p)) \, d\lambda$.

1. "Works in the opposite direction of Hotelling's Lemma: it recovers the firm's profits from its choices, rather than the other way around."

2. If $\pi(\cdot)$ is differentiable, integrating Hotelling's Lemma along the linear path from p to p' gives PSF; however PSF is more general (doesn't require differentiability of $\pi(\cdot)$).

3. As written the integral is along a linear path, but it is actually path-independent.

4. PSF allows calculation of change in profits when price of good i changes by knowing only the supply function for good i; we need not know the prices or supply functions for other goods: $\pi(p_{-i}, b) - \pi(p_{-i}, a) = \int_a^b \hat{y}_i(p_i) \, dp_i$.

Single-output case (Producer 22) For a single-output firm with free disposal, production set described as $\{(q, -z) : z \in \mathbb{R}^m_+, \; q \in [0, f(z)]\}$. With positive output price p, profit-maximization requires $q = f(z)$, so firms maximize $\max_{z \in \mathbb{R}^m_+} pf(z) - w \cdot z$, where $w \in \mathbb{R}^m_+$ input prices.

Cost minimization (Producer 22, Micro P.S. 2-4) For a fixed output level $q \geq 0$, firms minimize costs, choosing inputs according to a conditional factor demand correspondence:

$$c(q, w) \equiv \inf_{z : \, f(z) \geq q} w \cdot z;$$
$$Z^*(q, w) \equiv \underset{z : \, f(z) \geq q}{\operatorname{argmin}} w \cdot z$$
$$= \{z : f(z) \geq q, \text{ and } w \cdot z = c(q, w)\}.$$

Once these problems are solved, firms solve $\max_{q \geq 0} pq - c(q, w)$.

By the envelope theorem, $\frac{\partial c}{\partial w}(w, q) = Z^*(q, w)$.

Rationalization: single-output cost function (Producer 23, Micro P.S. 2-2) Conditional factor demand function $z : \mathbb{R} \times W \rightrightarrows \mathbb{R}^n$ and differentiable cost function $c : \mathbb{R} \times W \to \mathbb{R}$ for a fixed output q on an open convex set $W \subseteq \mathbb{R}^m$ of input prices are jointly rationalizable iff

1. $c(q, w) = w \cdot z(q, w)$ (adding-up);

2. $\nabla_w c(q, w) = z(q, w)$ (Shephard's Lemma);

3. $c(q, \cdot)$ is concave.

Other necessary properties follow from corresponding properties of profit-maximization, e.g.,

1. $c(q, \cdot)$ is homogeneous of degree one in w;

2. $Z^*(q, \cdot)$ is homogeneous of degree zero in w;

3. If $Z^*(q, \cdot)$ is differentiable, then the matrix $D_w Z^*(q, w) = D_w^2 c(q, w)$ is symmetric and negative semidefinite;

4. Under free disposal, $c(\cdot, w)$ is nondecreasing in q;

5. If the production function has nondecreasing (nonincreasing) RTS, the average cost function $c(q, w)/q$ is nonincreasing (nondecreasing) in q;

6. If the production function is concave, the cost function $c(q, w)$ is convex in q.

Monopoly pricing (MWG 384–7) Suppose demand at price p is $x(p)$, continuous and strictly decreasing at all p for which $x(p) > 0$. Suppose the monopolist faces cost function $c(q)$. Monopolist solves $\max_p px(p) - c(x(p))$ for optimal price, or $\max_{q \geq 0} p(q)q - c(q)$ for optimal quantity (where $p(\cdot) = x^{-1}(\cdot)$ is the inverse demand function). Further assumptions:

1. $p(q)$, $c(q)$ continuous and twice differentiable at all $q \geq 0$;

2. $p(0) > c'(0)$ (ensures that supply and demand curves cross);

3. There exists a unique socially optimal output level $q^O \in (0, \infty)$ such that $p(q^o) = c'(q^o)$.

A solution $q^m \in [0, q^o]$ exists, and satisfies FOC $p'(q^m)q^m + p(q^m) = c'(q^m)$. If $p'(q) < 0$, then $p(q^m) > c'(q^m)$; i.e., monopoly price exceeds optimal price.

2.3 Comparative statics

Implicit function theorem (Producer 27–8) Consider $x(t) \equiv \text{argmax}_{x \in X} F(x, t)$. Suppose:

1. F is twice continuously differentiable;

2. X is convex;

3. $F_{xx} < 0$ (strict concavity of F in x; together with convexity of X, this ensures a unique maximizer);

4. $\forall t$, $x(t)$ is in the interior of X.

Then the unique maximizer is given by $F_x(x(t), t) = 0$, and

$$x'(t) = -\frac{F_{xt}(x(t), t)}{F_{xx}(x(t), t)}.$$

Note by strict concavity, the denominator is negative, so $x'(t)$ and $F_{xt}(x(t), t)$ have the same sign.

Envelope theorem (Clayton Producer I 6–8) Given a constrained optimization $v(\theta) = \max_x f(x, \theta)$ such that $g_1(x, \theta) \leq b_1$; $\ldots$; $g_K(x, \theta) \leq b_K$, comparative statics on the value function are given by:

$$\frac{\partial v}{\partial \theta_i} = \frac{\partial f}{\partial \theta_i}\bigg|_{x^*} - \sum_{k=1}^{K} \lambda_k \frac{\partial g_k}{\partial \theta_i}\bigg|_{x^*} = \frac{\partial \mathcal{L}}{\partial \theta_i}\bigg|_{x^*}$$

(for Lagrangian $\mathcal{L}$) for all θ such that the set of binding constraints does not change in an open neighborhood.

Can be thought of as application first of chain rule, and then of FOCs.

Envelope theorem (integral form) (Clayton Producer II 9–10) [a.k.a. Holmstrom's Lemma] Given an optimization $v(q) = \max_x f(x, q)$, the envelope theorem gives us $v'(q) = f'_q(x(q), q)$. Integrating gives

$$v(q_2) = v(q_1) + \int_{q_1}^{q_2} \frac{\partial f}{\partial q}(x(q), q)\, dq.$$

Increasing differences (Producer 30–1, Micro P.S. 2-4') $F: X \times T \to \mathbb{R}$ (with X, $T \subseteq \mathbb{R}$) has ID (a.k.a. weakly increasing differences) iff for all $x' > x$ and $t' > t$, $F(x', t') + F(x, t) \geq F(x', t) + F(x, t')$. Strictly/strongly increasing differences (SID) iff $F(x', t') + F(x, t) > F(x', t) + F(x, t')$.

Assuming $F(\cdot, \cdot)$ is sufficiently smooth, all of the following are equivalent:

1. F has ID;

2. $F_x(x, t)$ is nondecreasing in t for all x;

3. $F_t(x, t)$ is nondecreasing in x for all t;

4. $F_{xt}(x, t) \geq 0$ for all (x, t);

5. $F(x, t)$ is supermodular.

Additional results:

1. If $F(\cdot, \cdot)$ and G both have ID, then for all α, $\beta \geq 0$, the function $\alpha F + \beta G$ also has ID.

2. If F has ID, and $g_1(\cdot)$ and $g_2(\cdot)$ are nondecreasing functions, then $F(g_1(\cdot), g_2(\cdot))$ has ID.

3. Suppose $h(\cdot)$ is twice differentiable. Then $h(x - t)$ has ID in x, t iff $h(\cdot)$ is concave.

Supermodularity (Producer 37) $F: X \to \mathbb{R}^n$ on a sublattice X is supermodular iff for all x, $y \in X$, we have $F(x \wedge y) + F(x \vee y) \geq F(x) + F(y)$.

If X is a product set, $F(\cdot)$ is supermodular iff it has ID in all pairs (x_i, x_j) with $i \neq j$ (holding other variables x_{-ij} fixed).

Submodularity (Producer 41) $F(\cdot)$ is submodular iff $-F(\cdot)$ is supermodular.

Topkis' Theorem (Producer 31–2, 8) If

1. $F: X \times T \to \mathbb{R}$ (with X, $T \subseteq \mathbb{R}$) has ID,

2. $t' > t$,

3. $x \in X^*(t) \equiv \text{argmax}_{\xi \in X} F(\xi, t)$, and $x' \in X^*(t')$, then

$\min\{x, x'\} \in X^*(t)$ and $\max\{x, x'\} \in X^*(t')$. In other words, $X^*(t) \leq X^*(t')$ in strong set order. This implies $\sup X^*(\cdot)$ and $\inf X^*(\cdot)$ are nondecreasing; if $X^*(\cdot)$ is single-valued, then $X^*(\cdot)$ is nondecreasing.

If $F: X \times T \to \mathbb{R}$ (with X a lattice and T fully ordered) is supermodular in x and has ID in (x, t); $t' > t$; and $x \in X^*(t)$ and $x' \in X^*(t')$, then $(x \wedge x') \in X^*(t)$ and $(x \vee x') \in X^*(t')$. In other words, $X^*(\cdot)$ is nondecreasing in t in the stronger set order.

Monotone Selection Theorem (Producer 32) Analogue of Topkis' Theorem for SID. If $F: X \times T \to \mathbb{R}$ with X, $T \in \mathbb{R}$ has SID, $t' > t$, $x \in X^*(t)$, and $x' \in X^*(t')$, then $x' \geq x$.

Milgrom-Shannon Monotonicity Theorem (Producer 34) $X^*(t) \equiv \text{argmax}_{x \in X} F(x, t)$ is nondecreasing in t in SSO for all sets $X \in \mathbb{R}$ iff it has the single-crossing condition (which is non-symmetric): for all $x' > x$ and $t' > t$,

$$F(x', t) \geq F(x, t) \implies F(x', t') \geq F(x, t'), \text{ and}$$
$$F(x', t) > F(x, t) \implies F(x', t') > F(x, t').$$

MCS: robustness to objective function perturbation (Producer 34–5) [Milgrom-Shannon] $X^*(t) \equiv \text{argmax}_{x \in X}[F(x, t) + G(x)]$ is nondecreasing in t in SSO for all functions $G: X \to \mathbb{R}$ iff $F(\cdot)$ has ID. Note Topkis gives *sufficiency* of ID.

Complement inputs (Producer 40–2) Restrict attention to price vectors $(p, w) \in \mathbb{R}^{m+1}$ at which input demand correspondence $z(p, w)$ is single-valued. If production function $f(z)$ is increasing and supermodular, then $z(p, w)$ is nondecreasing in p and nonincreasing in w. That is, supermodularity of the production function implies price-theoretic complementarity of inputs.

If profit function $\pi(p, w)$ is continuously differentiable, then $z_i(p, w)$ is nonincreasing in w_j for all $i \neq j$ iff $\pi(p, w)$ is supermodular in w.

Substitute inputs (Producer 41–2) *Suppose there are only two inputs.* Restrict attention to price vectors $(p, w) \in \mathbb{R}^3_+$ at which input demand correspondence $z(p, w)$ is single-valued. If production function $f(z)$ is increasing and submodular, then $z_1(p, w)$ is nondecreasing in w_2 and $z_2(p, w)$ is nondecreasing in w_1. That is, submodularity of the production function implies price-theoretic substitutability of inputs *in the two input case.*
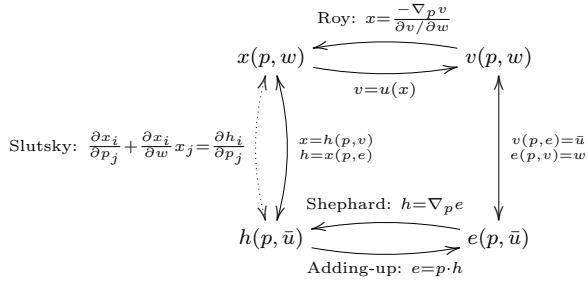
If there are ≥ 3 inputs, feedback between inputs with unchanging prices makes for unpredictable results.

If profit function $\pi(p, w)$ is continuously differentiable, then $z_i(p, w)$ is nondecreasing in w_j for all $i \neq j$ iff $\pi(p, w)$ is submodular in w.

LeChatelier principle (Producer 42–45) Argument (a.k.a Samuelson-LeChatelier principle) that firms react more to input price changes in the long-run than in the short-run, because it has more inputs that it can adjust. Does *not* consistently hold; only holds if each pair of inputs are substitutes everywhere or complements everywhere.

Suppose twice differentiable production function $f(k, l)$ satisfies either $f_{kl} \geq 0$ everywhere, or $f_{kl} \leq 0$ everywhere. Then if wage w_l increases (decreases), the firm's labor demand will decrease (increase), and the decrease (increase) will be larger in the long-run than in the short-run.

2.4 Consumer theory

$$\text{Roy: } x = \frac{-\nabla_p v}{\partial v/\partial w}$$

$$x(p,w) \rightleftharpoons v(p,w)$$
$$v = u(x)$$

Slutsky: $\frac{\partial x_i}{\partial p_j} + \frac{\partial x_i}{\partial w} x_j = \frac{\partial h_i}{\partial p_j}$ $\quad$ $\begin{array}{l} x = h(p,v) \\ h = x(p,e) \end{array}$ $\quad$ $\begin{array}{l} v(p,e) = \bar{u} \\ e(p,v) = w \end{array}$

$$\text{Shephard: } h = \nabla_p e$$

$$h(p,\bar{u}) \rightleftharpoons e(p,\bar{u})$$
$$\text{Adding-up: } e = p \cdot h$$

Budget set (Consumer 2) Given prices p and wealth w, $B(p,w) \equiv \{x \in \mathbb{R}^n_+ : p \cdot x \leq w\}$.

Utility maximization problem (Consumer 1–2, 6–8) $\max_{x \in \mathbb{R}^n_+} u(x)$ such that $p \cdot x \leq w$, or equivalently $\max_{x \in B(p,w)} u(x)$. Assumes:

1. Perfect information,
2. Price taking,
3. Linear prices,
4. Divisible goods.

Construct Lagrangian $\mathcal{L} = u(x) + \lambda(w - p \cdot x) + \sum_i \mu_i x_i$. If u is concave and differentiable, Kuhn-Tucker conditions (FOCs, nonnegativity, complementary slackness) are necessary and sufficient. Thus $\partial u / \partial x_k \leq \lambda p_k$ with equality if $x_k > 0$. For any two goods consumed in positive quantities, $p_j/p_k = \frac{\partial u/\partial x_j}{\partial u/\partial x_k} \equiv \mathrm{MRS}_{jk}$. The Lagrange multiplier on the budget constraint λ is the value in utils of an additional unit of wealth; i.e., the shadow price of wealth or marginal utility of wealth or income.

Indirect utility function (Consumer 3–4) $v(p,w) \equiv \sup_{x \in B(p,w)} u(x)$. Homogeneous of degree zero.

Marshallian demand correspondence (Consumer 3–4) [a.k.a. Walrasian or uncompensated demand] $x \colon \mathbb{R}^n_+ \times \mathbb{R}_+ \rightrightarrows \mathbb{R}^n_+$ with $x(p,w) \equiv \{x \in B(p,w) : u(x) = v(p,x)\} = \operatorname{argmax}_{x \in B(p,w)} u(x)$.

1. Given continuous preferences, $x(p,w) \neq \varnothing$ for $p \gg 0$ and $w \geq 0$.
2. Given convex preferences, $x(p,w)$ is convex-valued.
3. Given strictly convex preferences, $x(p,w)$ is single-valued.
4. $x(p,w)$ is homogeneous of degree zero.

Walras' Law (Consumer 4) Given locally non-satiated preferences:

1. For $x \in x(p,w)$, we have $p \cdot x = w$ (i.e., Marshallian demand is on budget line, and we can replace inequality constraint with equality in consumer problem);

2. For $z \in z(p)$ (where $z(\cdot)$ is excess demand $z(p) \equiv x(p, p \cdot e) - e$), we have $p \cdot z = 0$;

3. $v(p,w)$ is nonincreasing in p and strictly increasing in w.

Expenditure minimization problem (Consumer 9) $\min_{x \in \mathbb{R}^n_+} p \cdot x$ such that $u(x) \geq \bar{u}$; where $\bar{u} > u(0)$ and $p \gg 0$. Finds the cheapest bundle that yields utility at least $\bar{u}$. Equivalent to cost minimization for a single-output firm with production function u.

If $p \gg 0$, $u(\cdot)$ is continuous, and $\exists \hat{x}$ such that $u(\hat{x}) \geq \bar{u}$, then EMP has a solution.

Expenditure function (Consumer 9) $e(p,\bar{u}) \equiv \min_{x \in \mathbb{R}^n_+} p \cdot x$ such that $u(x) \geq \bar{u}$.

Hicksian demand correspondence (Consumer 9) [a.k.a. compensated demand] $h \colon \mathbb{R}^n_+ \times \mathbb{R}_+ \rightrightarrows \mathbb{R}^n_+$ with $h(p,\bar{u}) \equiv \{x \in \mathbb{R}^n_+ : u(x) \geq \bar{u}\} = \operatorname{argmin}_{x \in \mathbb{R}^n_+} p \cdot x$ such that $u(x) \geq \bar{u}$.

Relating Marshallian and Hicksian demand (Consumer 10) Suppose preferences are continuous and locally non-satiated, and $p \gg 0$, $w \geq 0$, $\bar{u} \geq u(0)$. Then:

1. $x(p,w) = h(p, v(p,w))$,
2. $e(p, v(p,w)) = w$,
3. $h(p,\bar{u}) = x(p, e(p,\bar{u}))$,
4. $v(p, e(p,\bar{u})) = \bar{u}$.

Rationalization: h and differentiable e (Consumer 11) Hicksian demand function $h \colon P \times \mathbb{R}_+ \to \mathbb{R}^n_+$ and differentiable expenditure function $e \colon P \times \mathbb{R} \to \mathbb{R}$ on an open convex set $P \subseteq \mathbb{R}^n$ are jointly rationalizable by expenditure-minimization for a given utility level $\bar{u}$ of a monotone utility function iff:

1. $e(p,\bar{u}) = p \cdot h(p,\bar{u})$ (adding-up—together with Shephard's Lemma, ensures $e(\cdot, \bar{u})$ is homogeneous of degree one in prices);
2. $\nabla_p e(p,\bar{u}) = h(p,\bar{u})$ (Shephard's Lemma—equivalent to envelope condition applied to $e(p,\bar{u}) = \min_h p \cdot h$);
3. $e(\cdot, \bar{u})$ is concave in prices.

Rationalization: differentiable h (Consumer 12) A continuously differentiable Hicksian demand function $h \colon P \times \mathbb{R}_+ \to \mathbb{R}^n_+$ on an open convex set $P \subseteq \mathbb{R}^n$ is rationalizable by expenditure-minimization with a monotone utility function iff

1. Hicksian demand is increasing in $\bar{u}$; and
2. The Slutsky matrix

$$D_p h(p,\bar{u}) = \begin{bmatrix} \frac{\partial h_1(p,\bar{u})}{\partial p_1} & \cdots & \frac{\partial h_n(p,\bar{u})}{\partial p_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_1(p,\bar{u})}{\partial p_n} & \cdots & \frac{\partial h_n(p,\bar{u})}{\partial p_n} \end{bmatrix}$$

is symmetric,

3. Slutsky matrix is negative semidefinite (since $e(\cdot, \bar{u})$ is concave in prices),

4. Slutsky matrix satisfies $D_p h(p,\bar{u}) p = 0$ (i.e., $h(\cdot, \bar{u}$ is homogeneous of degree zero in prices).

Rationalization: differentiable x (?) Slutsky matrix can be generated using Slutsky equation. Rationalizability requires Marshallian demand to be homogeneous of degree 0, and the Slutsky matrix to be symmetric and negative semidefinite. [Potentially also positive everywhere and/or increasing in w?]

Rationalization: differentiable e (?) Rationalizability requires e to be:

1. Homogeneous of degree one in prices;
2. Concave in prices;
3. Increasing in $\bar{u}$;
4. Positive everywhere, or equivalently nondecreasing in all prices.

Slutsky equation (Consumer 13–4) Relates Hicksian and Marshallian demand. Suppose preferences are continuous and locally non-satiated, $p \gg 0$, and demand functions $h(p,\bar{u})$ and $x(p,w)$ are single-valued and differentiable. Then for all i, j,

$$\underbrace{\frac{\partial x_i(p,w)}{\partial p_j}}_{\text{Total}} = \underbrace{\frac{\partial h_i(p, u(x(p,w)))}{\partial p_j}}_{\text{Substitution}} - \underbrace{\frac{\partial x_i(p,w)}{\partial w} x_j(p,w)}_{\text{Wealth}},$$

or more concisely, $\frac{\partial x_i}{\partial p_j} = \frac{\partial h_i}{\partial p_j} - \frac{\partial x_i}{\partial w} x_j$.

Derived by differentiating $h_i(p,\bar{u}) = x_i(p, e(p,\bar{u}))$ with respect to p_j and applying Shephard's lemma.

Normal good (Consumer 15) Marshallian demand $x_i(p,w)$ increasing in w. By Slutsky equation, normal goods must be regular.

Inferior good (Consumer 15) Marshallian demand $x_i(p,w)$ decreasing in w.

Regular good (Consumer 15) Marshallian demand $x_i(p,w)$ decreasing in p_i.

Giffen good (Consumer 15) Marshallian demand $x_i(p,w)$ increasing in p_i. By Slutsky equation, Giffen goods must be inferior.

Substitute goods (Consumer 15–6) Goods i and j substitutes iff Hicksian demand $h_i(p,\bar{u})$ is increasing in p_j. Symmetric relationship. In a two-good world, the goods must be substitutes.

Complement goods (Consumer 15–6) Goods i and j complements iff Hicksian demand $h_i(p,\bar{u})$ is decreasing in p_j. Symmetric relationship.

Gross substitute (Consumer 15–7) Good i is a gross substitute for good j iff Marshallian demand $x_i(p,w)$ is increasing in p_j. *Not* necessarily a symmetric relationship.

Gross complement (Consumer 15–7) Good i is a gross complement for good j iff Marshallian demand $x_i(p,w)$ is decreasing in p_j. *Not* necessarily a symmetric relationship.

Engle curve (Consumer 15–6) [a.k.a. income expansion curve] For a given price p, the locus of Marshallian demands at various wealth levels.

Offer curve (Consumer 16–7) [a.k.a. price expansion path] For a given wealth w and prices (for goods other than good i) p_{-i}, the locus of Marshallian demands at various prices p_i.

Roy's identity (Consumer 17–8) Gives Marshallian demand from indirect utility:

$$x_i(p,w) = -\frac{\partial v(p,w)/\partial p_i}{\partial v(p,w)/\partial w},$$

Derived by differentiating $v(p, e(p,\bar{u})) = \bar{u}$ with respect to p and applying Shephard's lemma. Alternately, by applying envelope theorem to utility maximization problem $v(p,w) = \max_{x\,:\,p\cdot x \le w} u(x)$ (giving $\frac{\partial v}{\partial w} = \frac{\partial \mathcal{L}}{\partial w} = \lambda$ and $\frac{\partial v}{\partial p} = \frac{\partial \mathcal{L}}{\partial p} = -\lambda x$.)

Consumer welfare: price changes (Consumer 20-2, 4) Welfare change in utils when prices change from p to p' is $v(p',w) - v(p,w)$, but because utility is ordinal, this is meaningless. More useful to have dollar-denominated measure. So measure amount of additional wealth required to reach some reference utility, generally either previous utility (CV) or new utility (EV).

If preferences are quasi-linear, then CV = EV.

On any range where the good in question is either normal or inferior, $\min\{CV, EV\} \le CS \le \max\{CV, EV\}$.

Compensating variation (Consumer 21, 3) How much less wealth consumer needs to achieve same utility at prices p' as she had at p (*compensating* for price change—consumer faces both new prices *and* new wealth).

$$CV \equiv e(p,\bar{u}) - e(p',\bar{u})$$
$$= w - e(p',\bar{u})$$

which gives—when only price i is changing—the area to the left of the Hicksian demand curve corresponding to the old utility u by the consumer surplus formula and Shephard's Lemma:

$$= \int_{p'_i}^{p_i} \frac{\partial e(p,\bar{u})}{\partial p_i}\, dp_i = \int_{p'_i}^{p_i} h_i(p,\bar{u})\, dp_i.$$

Equivalent variation (Consumer 21, 3) How much additional expenditure is required at old prices p to achieve same utility as consumption at p' (*equivalent* to price change—consumer faces either new prices *or* revised wealth).

$$EV \equiv e(p,\bar{u}') - e(p',\bar{u}')$$
$$= e(p,\bar{u}') - w$$

which gives—when only price i is changing—the area to the left of the Hicksian demand curve corresponding to the new utility u' by the consumer surplus formula and Shephard's Lemma:

$$= \int_{p'_i}^{p_i} \frac{\partial e(p,\bar{u}')}{\partial p_i}\, dp_i = \int_{p'_i}^{p_i} h_i(p,\bar{u}')\, dp_i.$$

Marshallian consumer surplus (Consumer 23–4) Area to the left of Marshallian demand curve: $CS \equiv \int_{p'_i}^{p_i} x_i(p,w)\, dp_i$.

Price index (Consumer 25–6) Given a basket of goods consumed at quantity x given price p, and quantity x' given price p',

1. Laspeyres index: $\frac{p'\cdot x}{p\cdot x} = \frac{p'\cdot x}{e(p,\bar{u})}$ (basket is old purchases). Overestimates welfare effects of inflation due to substitution bias.

2. Paasche index: $\frac{p'\cdot x'}{p\cdot x'} = \frac{e(p',\bar{u}')}{p\cdot x'}$ (basket is new purchases). Underestimates welfare effects of inflation.

3. Ideal index: $\frac{e(p',\bar{u})}{e(p,\bar{u})}$ for some fixed utility level $\bar{u}$, generally either $u(x)$ or $u(x')$; the percentage compensation in the wealth of a consumer with utility $\bar{u}$ needed to make him as well off at the new prices as he was at the old ones.

Paasche $\le$ Ideal $\le$ Laspeyres. Substitution biases result from using the same basket of goods at new and old prices. Include

1. New good bias,

2. Outlet bias.

Aggregating consumer demand (Consumer 29–32)

1. Can we predict aggregate demand knowing only aggregate wealth (not distribution)? True iff indirect utility functions take Gorman form: $v_i(p,w_i) = a_i(p) + b(p)w_i$ with the same function $b(\cdot)$ for all consumers.

2. Can aggregate demand be explained as though there were a single "positive representative consumer"?

3. (If 2 holds), can the welfare of the representative consumer be used as a proxy for some welfare aggregate of individual consumers? (i.e., Do we have a "normative representative consumer"?)

2.5 Choice under uncertainty

Lottery (Uncertainty 2–4) A vector of probabilities adding to 1 assigned to each possible outcome (prize). The set of lotteries for a given prize space is convex.

Preference axioms under uncertainty (Uncertainty 5–6) In addition to (usual) completeness and transitivity, assume preferences are:

1. Continuous: For any p, p', $p'' \in \mathcal{P}$ with $p \succsim p' \succsim p''$, there exists $\alpha \in [0,1]$ such that $\alpha p + (1-\alpha)p'' \sim p'$.

2. Independent: [a.k.a. substitution axiom] For any p, p', $p'' \in \mathcal{P}$ and $\alpha \in [0,1]$, we have $p \succsim p' \iff \alpha p + (1-\alpha)p'' \succsim \alpha p' + (1-\alpha)p''$.

Expected utility function (Uncertainty 6–10) Utility function $u\colon \mathcal{P} \to \mathbb{R}$ has expected utility form iff there are numbers $(u_1, \ldots, u_n)$ for each of the n (certain) outcomes such that for every $p \in \mathcal{P}$, $u(p) = \sum_i p_i \cdot u_i$.

Equivalently, for any p, $p' \in \mathcal{P}$, $\alpha \in [0,1]$, we have $u(\alpha p + (1-\alpha)p') = \alpha U(p) + (1-\alpha)U(p')$.

Unlike a more general utility function, an expected utility functions is not merely ordinal—it is *not* invariant to *any* increasing transformation, only to affine transformations. If $u(\cdot)$ is an expected utility representation of $\succsim$, then $v(\cdot)$ is also an expected utility representation of $\succsim$ iff $\exists a \in \mathbb{R}$, $\exists b > 0$ such that $v(p) = a + bu(p)$ for all $p \in \mathcal{P}$.

Preferences can be represented by an expected utility function iff they are complete, transitive, *and* satisfy continuity and independence (assuming $|\mathcal{P}| < \infty$; otherwise we also need the "sure thing principle"). Obtains since both require indifference curves to be parallel straight lines.

Bernoulli utility function (Uncertainty 12) Assuming prize space $\mathcal{X}$ is an interval on the real line, Bernoulli utility function $u\colon \mathcal{X} \to \mathbb{R}$ assumed increasing and continuous.

von Neumann-Morgenstern utility function (Uncertainty 12) An expected utility representation of preferences over lotteries characterized by a cdf over prizes $\mathcal{X}$ (an interval on the real line). If $F(x)$ is the probability of receiving less than or equal to x, and $u(\cdot)$ is the Bernoulli utility function, then vN-M utility function $U(F) \equiv \int_{\mathbb{R}} u(x)\, dF(x)$.

Risk aversion (Uncertainty 12–4) A decision maker is (strictly) risk-averse iff for any non-degenerate lottery $F(\cdot)$ with expected value $E_F = \int_{\mathbb{R}} x\, dF(x)$, the lottery δ_{E_F} which pays E_F for certain is (strictly) preferred to F.

Stated mathematically, $\int u(x)\, dF(x) \le u(\int_{\mathbb{R}} x\, dF(x))$ for all $F(\cdot)$, which by Jensen's inequality obtains iff $u(\cdot)$ is concave.

The following notions of $u(\cdot)$ being "more risk-averse" then $v(\cdot)$ are equivalent:

1. $F \succsim_u \delta_x \implies F \succsim_v \delta_x$ for all F and x.*
2. Certain equivalent $c(F,u) \le c(F,v)$ for all F.

---

*Note this does *not* mean $F \succsim_u G \implies F \succsim_v G$ where G is also a risky prospect—this would be a stronger version of "more risk averse."

3. $u(\cdot)$ is "more concave" than $v(\cdot)$; i.e., there exists an increasing concave function $g(\cdot)$ such that $u = g \circ v$.

4. Arrow-Pratt coefficient $A(x, u) \geq A(x, v)$ for all x.

Certain equivalent (Uncertainty 13–4) $c(F, u)$ is the certain payout such that $\delta_{c(F,u)} \sim_u F$, or equivalently $u(c(F, u)) = \int_{\mathbb{R}} u(x) \, dF(x)$. Given risk aversion (i.e., concave u), $c(F, u) \leq \mathrm{E}_F$.

Absolute risk aversion (Uncertainty 14–6) For a twice differentiable Bernoulli utility function $u(\cdot)$, the Arrow-Pratt coefficient of absolute risk aversion is $A(x) \equiv -u''(x)/u'(x)$.

$u(\cdot)$ has decreasing (constant, increasing) absolute risk aversion iff $A(x)$ is decreasing $(\ldots)$ in x. Under DARA, if I will gamble \$10 when poor, I will gamble \$10 when rich.

Since $R(x) = xA(x)$, we have IARA $\implies$ IRRA.

Certain equivalent rate of return (Uncertainty 16) A proportionate gamble pays tx where t is a non-negative random variable with cdf F. The certain equivalent rate of return is $\mathrm{cr}(F, x, u) \equiv \hat{t}$ where $u(\hat{t}x) = \int u(tx) \, dF(t)$.

Relative risk aversion (Uncertainty 16) For a twice differentiable Bernoulli utility function $u(\cdot)$, the coefficient of relative risk aversion is $R(x) \equiv -xu''(x)/u'(x) = xA(x)$.

$u(\cdot)$ has decreasing (constant, increasing) relative risk aversion iff $R(x)$ is decreasing $(\ldots)$ in x. An agent exhibits DRRA iff certain equivalent rate of return $\mathrm{cr}(F, x)$ is increasing in x. Under DRRA, if I will invest 10% of my wealth in a risky asset when poor, I will invest 10% when rich.

Since $R(x) = xA(x)$, we have DRRA $\implies$ DARA.

First-order stochastic dominance (Uncertainty 17–8) cdf G first-order stochastically dominates cdf F iff $G(x) \leq F(x)$ for all x.

Equivalently, for every nondecreasing function $u \colon \mathbb{R} \to \mathbb{R}$, $\int u(x) \, dG(x) \geq \int u(x) \, dF(x)$.

Equivalently, we can construct G as a compound lottery starting with F and followed by (weakly) upward shifts.

Second-order stochastic dominance (Uncertainty 18–21) cdf G second-order stochastically dominates cdf F (where F and G have the same mean*) iff for every x, $\int_{-\infty}^{x} G(y) \, dy \leq \int_{-\infty}^{x} F(y) \, dy$.

Equivalently, for every (nondecreasing?) concave function $u \colon \mathbb{R} \to \mathbb{R}$, $\int u(x) \, dG(x) \geq \int u(x) \, dF(x)$.

Equivalently, we can construct F as a compound lottery starting with G and followed by mean-preserving spreads.

Demand for insurance (Uncertainty 21–3) A risk-averse agent with wealth w faces a probability of p of incurring a loss L. She can insure against this loss by buying a policy that will pay out a in the event the loss occurs, at cost qa.

---

*If $\mathrm{E}\,F > \mathrm{E}\,G$, there is always a concave utility function that will prefer F to G.

If insurance is actuarially fair ($q = p$), the agent fully insures ($a^* = L$) for all wealth levels. If $p < q$, the agent's insurance coverage a^* will decrease (increase) with wealth if the agent has decreasing (increasing) absolute risk aversion.

Portfolio problem (Uncertainty 23–5) A risk-averse agent with wealth w must choose to allocate investment between a "safe" asset that returns r and a risky asset that pays return z with cdf F.

If risk-neutral, the agent will invest all in the asset with higher expected return (r or $\mathrm{E}\,z$). If (strictly) risk-averse, she will invest at least some in the risky asset as long as its real expected return is positive. (To see why, consider marginal utility to investing in the risky asset at investment $a = 0$.)

If u is more risk-averse than v, then u will invest less in the risky asset than v for any initial level of wealth. An agent with decreasing (constant, increasing) absolute risk aversion will invest more (same, less) in the risky asset at higher levels of wealth.

Subjective probabilities (Uncertainty 26–8) We relax the assumption that there are objectively correct probabilities for various states of the world to be realized. If preferences over acts (bets) satisfy a set of properties "similar in spirit" to the vN-M axioms (completeness, transitivity, something like continuity, the sure thing principle, and two axioms that have the flavor of substitution), then decision makers' choices are consistent with some utility function and some prior probability distribution (Savage 1954).

Savage requires an exhaustive list of possible states of the world. No reason to assume different decision makers are using the same implied probability distribution over states, although we often make a "common prior" assumption, which implies that "differences in opinion are due to differences in information."

2.6 General equilibrium

Walrasian model (G.E. 3–4, 5) An economy $\mathcal{E} \equiv ((u^i, e^i)_{i \in \mathcal{I}})$ comprises:

1. L commodities (indexed $l \in \mathcal{L} \equiv \{1, \ldots, L\}$);
2. I agents (indexed $i \in \mathcal{I} \equiv \{1, \ldots, I\}$), each with
   - Endowment $e^i \in \mathbb{R}_+^L$, and
   - Utility function $u^i \colon \mathbb{R}_+^L \to \mathbb{R}$.

Given market prices $p \in \mathbb{R}_+^L$, each agent chooses consumption to maximize utility given a budget constraint: $\max_{x \in \mathbb{R}_+^L} u^i(x)$ such that $p \cdot x \leq p \cdot e^i$, or equivalently $p \in \mathcal{B}^i(p) \equiv \{x \colon p \cdot x \leq p \cdot e^i\}$.

We often assume (some or all of):

1. $\forall i$, $u^i(\cdot)$ is continuous;
2. $\forall i$, $u^i(\cdot)$ is increasing; i.e., $u^i(x') > u^i(x)$ whenever $x' \gg x$;
3. $\forall i$, $u^i(\cdot)$ is concave;
4. $\forall i$, $e^i \gg 0$;

Walrasian equilibrium (G.E. 4, 17, 24–9) A WE for economy $\mathcal{E}$ is a vector of prices and allocations $(p, (x^i)_{i \in \mathcal{I}})$ such that:

1. Agents maximize their utilities: $\max_{x \in \mathcal{B}^i(p)} u^i(x)$ for all $i \in \mathcal{I}$;
2. Markets clear: $\sum_{i \in \mathcal{I}} x_l^i = \sum_{i \in \mathcal{I}} e_l^i$ for all $l \in \mathcal{L}$, or equivalently $\sum_{i \in \mathcal{I}} x^i = \sum_{i \in \mathcal{I}} e^i$.

Under assumptions 1–4 above, a WE exists (proof using fixed point theorem). WE are not generally unique, but are locally unique (and there are an odd number of them). Price adjustment process ("tatonnement") may not converge to an equilibrium.

Feasible allocation (G.E. 4) An allocation $(x^i)_{i \in \mathcal{I}} \in \mathbb{R}_+^{IL}$ is feasible iff $\sum_{i \in \mathcal{I}} x^i \leq \sum_{i \in \mathcal{I}} e^i$.

Pareto optimality (G.E. 5) A feasible allocation $x \equiv (x^i)_{i \in \mathcal{I}}$ for economy $\mathcal{E}$ is Pareto optimal iff there is no other feasible allocation $\hat{x}$ such that $u^i(\hat{x}^i) \geq u^i(x^i)$ for all $i \in \mathcal{I}$ with strict inequality for some $i \in \mathcal{I}$.

Edgeworth box (G.E. 6–10) Graphical representation of the two-good, two-person exchange economy. Bottom left corner is origin for one consumer; upper right corner is origin for other consumer (with direction of axes reversed). Budget line has slope $-p_1/p_2$, and passes through endowment e.

1. Locus of Marshallian demands for each consumer as relative prices shift is her *offer curve*. WE are intersections of the two consumers' offer curves.
2. Set of PO allocations is locus of points of tangency between the two consumers' indifference curves, generally a path from the upper right to lower left of the Edgeworth box.
3. Portion of Pareto set that lies between the indifference curves that pass through e is the *contract curve*: PO outcomes preferred by both consumers to their endowments.

First Welfare Theorem (G.E. 11) If $\forall i$, $u^i(\cdot)$ is increasing (i.e., $u^i(x') > u^i(x)$ whenever $x' \gg x$) and $(p, (x^i)_{i \in \mathcal{I}})$ is a WE, then the allocation $(x^i)_{i \in \mathcal{I}}$ is PO. Note implicit assumptions such as

1. All agents face same prices;
2. All agents are price takers;
3. Markes exist for all goods, and individuals can freely participate;

4. Prices are somehow arrived at.

Proof by adding results of Walras' Law across consumer at potentially Pareto-improving allocation. Result shows that allocation cannot be feasible.

Second Welfare Theorem (G.E. 11-3) If allocation $(e^i)_{i\in\mathcal{I}}$ is PO and

1. $\forall i$, $u^i(\cdot)$ is continuous;

2. $\forall i$, $u^i(\cdot)$ is increasing; i.e., $u^i(x') > u^i(x)$ whenever $x' \gg x$;

3. $\forall i$, $u^i(\cdot)$ is concave;

4. $\forall i$, $e^i \gg 0$;

then there exists a price vector $p \in \mathbb{R}^L_+$ such that $(p, (e^i)_{i\in\mathcal{I}})$ is a WE.

Note this does *not* say that starting from a given endowment, every PO allocation is a WE. Thus decentralizing a PO allocation is not simply a matter of identifying the correct prices—large-scale redistribution may be required as well.

Proof by separating hyperplane theorem. Consider the set of changes to total endowment that strictly improve every consumer's utility; by concavity of $u^i(\cdot)$, this set is convex. Separate this set from 0, and show that the resulting prices are nonnegative, and that at these prices, e^i maximizes each consumer's utility.

Excess demand (G.E. 18) $z^i(p) \equiv x^i(p, p\cdot e^i) - e^i$, where x^i is the agent's Marshallian demand. Walras' Law gives $p \cdot z^i(p) = 0$.

Aggregate excess demand is $z(p) \equiv \sum_{i\in\mathcal{I}} z^i(p)$. If $z(p) = 0$, then $(p, (x^i(p, p\cdot e^i))_{i\in\mathcal{I}})$ is a WE.

Sonnenschein-Mantel-Debreu Theorem (G.E. 30) Let $B \subseteq \mathbb{R}^L_{++}$ be open and bounded, and $f: B \to \mathbb{R}^L$ be continuous, homogeneous of degree zero, and satisfy $p \cdot z(p) = 0$ for all p. Then there exist an economy $\mathcal{E}$ with aggregate excess demand function $z(p)$ satisfying $z(p) = f(p)$ on B.

Interpretation is that without special assumptions, pretty much any any comparative statics result could be obtained in a GE model. However, Brown and Matzkin show that if we can observe endowments as well as prices, GE *may* be testable.

Gross substitutes property (G.E. 32–5) A Marshallian demand function $x(p)$ satisfies the gross substitutes property if for all k, whenever $p'_k > p_k$ and $p'_{-k} = p_{-k}$, then $x_{-k}(p') > x_{-k}(p)$; i.e., all pairs of goods are (strict) gross substitutes. Implies that excess demand function satisfies gross substitutes. If every individual satisfies gross substitutes, then so does aggregate excess demand.

If aggregate excess demand satisfies gross substitutes,

1. The economy has at most one (price-normalized) WE.

2. If $z(p^*) = 0$ (i.e., p^* are WE prices), then for any p not colinear with p^*, we have $p^* \cdot z(p) > 0$.

3. The tatonnement process $\frac{dp}{dy} = \alpha z(p(t))$ with $\alpha > 0$ converges to WE prices for any initial condition $p(0)$.

4. Any change that raises the excess demand for good k will increase the equilibrium price of good k.

Incomplete markets (Jackson) Under incomplete markets, Walrasian equilibria may:

1. Fail to be efficient, and even fail to be constrained efficient (i.e., there may be more efficient outcomes that are feasible under restricted trading regime);

2. Fail to exist ("although only rarely");

3. Have prices/trades that depend on the resolution of uncertainty.

Rational expectations equilibrium (Jackson) Suppose l (primitive) goods, state space S (with $|S| < \infty$), and n agents each of whom has

- endowment $e_i: S \to \mathbb{R}^l_+$,

- preferences $u_i: \mathbb{R}^l_+ \times S \to \mathbb{R}$,

- information I_i (including information contained in e_i and u_i).

The allocations $x_i: S \to \mathbb{R}^l_+$ (or equivalently $x_i \in \mathbb{R}^{|S|\cdot l}_+$) and prices $p: S \to \mathbb{R}^l_+$ are an REE iff:

1. Information revelation: x_i is measurable with respect to $I_i \vee I_{\text{prices}}$ for all i;

2. Market clearing: $\sum_i x_i(s) \leq \sum_i e_i(s)$ for all $s \in S$;

3. Optimizing behavior: $x_i(s) \in \text{argmax}_x u_i[x_i(s)]$ such that x_i is measurable with respect to $I_i \vee I_{\text{prices}}$ and $p(s) \cdot x_i(s) \leq p(s) \cdot e_i(s)$ for all i.

2.7 Games

Game tree (Bernheim 2–5) Description of a game comprising:

1. Nodes,

2. A mapping from nodes to the set of players,

3. Branches,

4. A mapping from branches to the set of action labels,

5. Precedence (a partial ordering),

6. A probability distribution over branches for all notes that map to nature.

We assume:

1. There is a finite number of nodes;

2. There is a unique root—a node that has no predecessors and is a predecessor for everything else;

3. There is a unique path (following precedence) from the root to each terminal node (those nodes without successors).

We also add:

1. Information: a partition over the set of nodes such that
   - The same player makes the decisions at all nodes within any element of the partition;
   - The same actions are available at all nodes within any element of the partition;
   - No element of the partition contains both a node and its predecessor.

2. Payoffs: a mapping from terminal nodes to a vector of utilities.

Perfect recall (Bernheim 6) Informally, a player never forgets either a decision he made in the past or information that he possessed when making a previous decision.

Perfect information (Bernheim 7) Every information set is a singleton.

Complete information (Bernheim 13, 73–4) Each player knows the payoffs received by every player at every terminal node.

Per Harsanyi, a game of incomplete information can be written as a game of imperfect information by adding nature as a player whose choices determine hidden characteristics. The probability governing Nature's decisions is taken to be common knowledge. NE of this "Bayesian game" is a Bayesian NE.

Strategy (Bernheim 7–8) A mapping that assigns a feasible action to all information sets for which a player is the decision-maker (i.e., a complete contingent plan). Notation is:

1. S_i the set of player i's feasible strategies;

2. $S \equiv \prod_j S_j$ the set of feasible strategy profiles;

3. $S_{-i} \equiv \prod_{j\neq i} S_j$ the set of feasible strategy profiles for every player but i.

Payoff function (Bernheim 8) $g_i: S \to \mathbb{R}$ gives player i's expected utility if everyone plays according to $s \in S$.

Normal form (Bernheim 8) A description of a game as a collection of players $\{1, \dots, I\}$, a strategy profile set S, and a payoff function $g: S \to \mathbb{R}^I$ where $g(s) = (g_1(s), \dots, g_I(s))$.

Revelation principle (Bernheim 79) In a game with externalities, when the mechanism designer doesn't have information about preferences, agents will have an incentive to understate or exaggerate their preferences. The revelation principle states that in searching for an optimal mechanism within a much broader class, the designer can restrict attention to direct revelation mechanisms (those that assume agents have revealed their true preferences) for which truth-telling is an optimal strategy for each agent.

Proper subgame (Bernheim 91) Consider a node t in an extensive form game, with information set $h(t)$ and successors $S(t)$. Then $\{t\} \cup S(t)$ (along with associated mappings from information sets to players, from branches to action labels, and from terminal notes to payoffs) is a proper subgame iff

1. $h(t) = \{t\}$ (the information set for t is a singleton); and
2. $\forall t' \in S(t)$, we have $h(t') \subseteq S(t)$ (every player knows we are at t).

System of beliefs (Bernheim 94) Given decision nodes X, information sets H, including the information set $h(t)$ containing $t \in X$, and $\phi(h)$ the player who makes the decision at information set $h \in H$, a system of beliefs is a mapping $\mu \colon X \to [0,1]$ such that $\forall h \in H$ we have $\sum_{t \in h} \mu(t) = 1$. That is, a set of probability distributions over nodes in each information set.

Strictly mixed strategy (Bernheim 98) Behavior strategy profile δ is strictly mixed if every action at every information set is selected with strictly positive probability.

Note since every information set is reached with strictly positive probability, one can completely infer a system of beliefs using Bayes' rule.

2.8 Dominance

Dominant strategy (Bernheim 13) s_i is a (strictly) dominant strategy iff for all $\hat{s} \in S$ with $\hat{s}_i \neq s_i$, we have $g_i(s_i, \hat{s}_{-i}) > g_i(\hat{s}_i, \hat{s}_{-i})$.

Dominated strategy (Bernheim 15) s_i is a (strictly) dominated strategy iff there exists some probability distribution ρ over $S_i \equiv \{s_i^1, \ldots, s_i^M\}$ such that for all $s_{-i} \in S_{-i}$, we have

$$\sum_{m=1}^{M} \rho^m g_i(s_i^m, s_{-i}) > g_i(\hat{s}_i, \hat{s}_{-i}).$$

Iterated deletion of dominated strategies (Bernheim 15–6) We iteratively delete (*strictly*) dominated strategies from the game. Relies on common knowledge of rationality (i.e., everyone is, everyone knows, everyone knows everyone knows, ...). If this yields a unique outcome, the game is "dominance solvable." The order of deletion is irrelevant.

For two player games, strategies that survive iterative deletion of dominated strategies are precisely are precisely the set of rationalizable strategies. This equivalence holds for games with more than two players only if we do not insist on independence in defining rationalizability; if we *do* insist on independence, the set of rationalizable strategies is smaller.

Weakly dominated strategy (Bernheim 20) $\hat{s}_i$ is a weakly dominated strategy iff there exists some probability distribution ρ over $S_i \equiv \{s_i^1, \ldots, s_i^M\}$ such that for all $s_{-i} \in S_{-i}$, we have

$$\sum_{m=1}^{M} \rho^m g_i(s_i^m, s_{-i}) \geq g_i(\hat{s}_i, s_{-i}),$$

with strict inequality for some $s_{-i} \in S_{-i}$.

We *cannot* iteratively delete weakly dominated strategies; unlike for strict domination, the order of deletion matters.

2.9 Equilibrium concepts

PSNE $\subseteq$ rationalizable $\subseteq$ ISD. Rationalizable strategies are a best response based on *some* prior. PSNE are best responses based on a *common* prior.

Normal form equilibrium concepts (static games): THPE $\subseteq$ MSNE; PSNE $\subseteq$ MSNE; MSNE are made up of rationalizable strategies. BNE are MSNE of "extended game" that includes nature choosing types.

Extensive form equilibrium concepts (dynamic games)[*]:
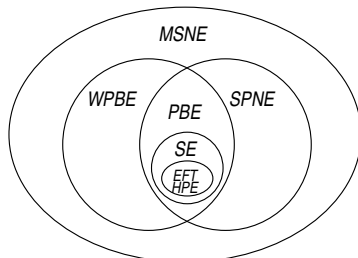
Rationalizable strategy (Bernheim 27)

1. A 1-*rationalizable strategy* is a best response to some (independent)[†] probability distribution over other players' strategies.
2. A k-*rationalizable strategy* is a best response to some (independent) probability distribution over other players' $(k-1)$-rationalizable strategies.
3. A *rationalizable strategy* is k-rationalizable for all k.

For two player games, rationalizable strategies are precisely those that survive iterative deletion of dominated strategies. This equivalence holds for games with more than two players only if we do not insist on independence in defining rationalizability; if we *do* insist on independence, the set of rationalizable strategies is smaller.

Pure strategy Nash equilibrium (Bernheim 29–33) $s^* \in S$ is a PSNE iff for all $s \in S$, we have $g_i(s_i^*, s_{-i}^*) \geq g_i(s_i, s_{-i}^*)$. The strategies played in a PSNE are all rationalizable.

A finite game of perfect information has a PSNE [Zermelo]. Proved using backwards induction.

If $S_1, \ldots, S_I$ are compact, convex Euclidean sets and g_i is continuous in s and quasiconcave in s_i, then there exists a PSNE. By Berge's Theorem, the best response correspondence is upper hemi-continuous. By quasiconcavity of g_i, the best response correspondence is convex valued. Thus by Kakutani's Fixed Point Theorem, it has a fixed point.

Mixed strategy Nash equilibrium (Bernheim 56–9) We define a new game where the strategy space is the set of probability distributions over (normal form) strategies in the original game (i.e., the set of *mixed strategies*). A MSNE of the original game is any PSNE of this new game. Players must be indifferent between playing any strategies included (with strictly positive probability) in a MSNE.

Another approach is to consider randomization over actions at each information set (*behavior strategies*). However, Kuhn's Theorem assures us that for any game of perfect recall there are mixed strategies that yields the same distribution over outcomes as any combination of behavioral strategies; also it gives that there are behavioral strategies that yield the same distribution over outcomes as any combination of mixed strategies.

Every finite game has a MSNE.

Trembling-hand perfection (Bernheim 65–7) There is always some risk that another player will make a "mistake." Bernheim notes include two equivalent rigorous definitions; original definitions are for finite games, but there is an extension available for infinite games. Key notes:

1. In a THPE, no player selects a weakly dominated strategy with positive probability.
2. For two player games, an MSNE is a THPE iff no player selects a weakly dominated strategy with positive probability.
3. For more than two player games, the set of THPE may be smaller than the set of MSNE where no player selects a weakly dominated strategy with positive probability; if we allow correlations between the trembles of different players, the sets are the same.

Correlated equilibrium (Bernheim 70) In a finite game $(\{S_i\}_{i=1}^I, g)$, a probability distribution δ^* over S is a CE iff for all i and s_i chosen with strictly positive probability, s_i solves

$$\max_{s_i' \in S_i} \mathrm{E}_{s_{-i}}[g_i(s_i', s_{-i}) | s_i, \delta^*];$$

i.e., player i has no incentive to defect from any strategy s_i, assuming that other players respond per δ^*.

---

[*]Unclear whether PBE is the intersection of WPBE and SPNE, or a *subset* of the intersection.

[†]Question is whether to allow other players' randomized strategies to be correlated with each other.

Bayesian Nash equilibrium (Bernheim 73–4) Per Harsanyi, we write our game of incomplete information as a game of imperfect information where nature selects hidden characteristics. A BNE is a MSNE of this "Bayesian game." A pure strategy BNE is a PSNE of the Bayesian game.

Characterized by a set of decision rules that determine each player's strategy contingent on his type.

Subgame perfect Nash equilibrium (Bernheim 92) A MSNE in behavior strategies δ^* is a SPNE iff for every proper subgame, the restriction of δ^* to the subgame forms a MSNE in behavior strategies.

For finite games, we can find SPNE using backwards induction on subgames of extensive form.

A SPNE need not be a WPBE, and a WPBE need not be a SPNE.

Sequentially rational strategy (Bernheim 94) Behavior strategy profile δ is sequentially rational given a system of beliefs μ iff for all information sets $h \in H$, the actions for player $\phi(h)$ at $h \cup \left[\bigcup_{t \in h} S(t) \right]$ are optimal starting from h given an initial probability over h governed by μ, and given that other players will adhere to $\delta_{-\phi(h)}$.

Weak perfect Bayesian equilibrium (Bernheim 93–5) Implausible equilibria can still be SPNE, because we lack beliefs at each information set. Behavior strategy profile δ^* *and system of beliefs* μ^* are a WPBE iff:

1. δ^* is sequentially rational given beliefs μ^*, and

2. Where possible, μ^* is computed from δ^* using Bayes' rule; i.e., for any information set h with $\Pr(h|\delta^*) > 0$, for all $t \in h$,
$$\mu^*(t) = \frac{\Pr(t|\delta^*)}{\Pr(h|\delta^*)}.$$

Note only restriction on out-of-equilibrium beliefs is that they exist. A SPNE need not be a WPBE, and a WPBE need not be a SPNE.

Perfect Bayesian equilibrium (Bernheim 93–5) (δ^*, μ^*) is a PBE if it a WPBE in all proper subgames. This ensures it is also a SPNE.

Consistent strategy (Bernheim 98) Behavior strategy profile δ is consistent given a system of beliefs μ iff there exists a sequence of strictly mixed behavior strategy profiles $\delta_n \to \delta$ such that $\mu_n \to \mu$, where μ_n is generated from δ_n by Bayes' rule.

Sequential equilibrium (Bernheim 98) (δ^*, μ^*) is a SE if it is sequentially rational and consistent.

SE places additional restriction on beliefs vs. WPBE, hence SE $\subseteq$ WPBE; also, can show that SE are SPNE, so an SE is also a PBE.

Extensive form trembling hand perfection (Bernheim 102–5) "Agent normal form" is the normal form that would obtain if each player selected a different agent to make her decisions at every information set, and all of a player's agents acted independently with the object of maximizing the player's payoffs. An EFTHPE is a THPE in the agent normal form of the game.

EFTHPE $\subseteq$ SE; for generic finite games, they are the same.

2.10 Basic game-theoretic models

Other models and examples are throughout Bernheim lecture notes.

Cournot competition (Bernheim 11, 46–50) Firms simultaneously choose quantity. Inverse demand is $P(Q)$ (monotonically decreasing); cost to firm i of producing quantity q_i is $c_i(q_i)$ where $c_i(0) = 0$. Normal form:

1. Strategies: $S = \mathbb{R}_+^I$;

2. Payouts: $g_i(s) = P(\sum_j s_j)s_i - c_i(s_i)$.

To ensure PSNE existence, we need quasiconcavity of g_i in s_i (generally don't worry about unboundedness of strategy set). Sufficient conditions are $c_i(\cdot)$ convex and $P(\cdot)$ concave. The former rules out increasing returns to scale. The latter "is not a conventional property of demand functions, but *is* satisfied for linear functions."

Note:

1. Characterized by strategic substitutes; i.e., best response curves are downward sloping.

2. Production is spread among firms.

Bertrand competition (Bernheim 11, 37) Firms simultaneously choose price; consumers purchase from low-price firm. Demand is $Q(p)$ (monotonically decreasing); cost to firm i of producing quantity q_i is $c_i(q_i)$ where $c_i(0) = 0$. Normal form (two-firm case):

1. Strategies: $S = \mathbb{R}_+^I$;

2. Payouts:

$$g_i(s) = \begin{cases} 0, & s_i > s_{-i}; \\ s_i Q(s_i) - c_i(Q(s_i)), & s_i < s_{-i}; \\ \frac{1}{2}\left[s_i Q(s_i) - c_i(Q(s_i)) \right], & s_i = s_{-i}. \end{cases}$$

Note:

1. One firm case is monopoly (in which case demand matters); more than one yields perfectly competitive outcome since only the marginal cost of the second-most efficient firm matters—the properties of the demand curve are then irrelevant.

2. Characterized by strategic complements; i.e., best response curves are upward sloping.

3. All production is done by most efficient firm.

Bertrand competition—non-spatial differentiation (Bernheim 38–40) Firm i chooses price for good i. Demand for good i is given by $Q(p_i, p_{-1})$. Strategic complements; i.e., best response curves are upward sloping.

Bertrand competition—horizontal differentiation (Bernheim 40–2) a.k.a. Hotelling spatial location model. Consumers are indexed by $\theta \in [0, 1]$, representing location. Each consumer purchases zero or one unit, with payoff 0 if no purchase, and $v - p_i - t(x_i - \theta)^2$ from purchasing a type x_i good at price p_i. v is value of good, t is unit transport cost.

If firms cannot choose product type x_i, prices are strategic complements; i.e., best response curves are upward sloping.

Bertrand competition—vertical differentiation (Bernheim 42–5) Consumers are indexed by $\theta \in [0, 1]$, representing value attached to quantity. Each consumer purchases zero or one unit, with payoff 0 if no purchase, and $\theta v_i - p_i$ from purchasing a quality x_i good at price p_i.

Sequential competition (Bernheim 107–12) First one firm selects price/quantity, then the other firm follows. The leader always does (weakly) better than in the simultaneous choice model; whether the follower does better or worse than in simultaneous choice depends whether there are strategic complements or substitutes. This also determines which firm does better in the sequential choice model.

Herfindahl-Hirshman index (Bernheim 37) $\mathcal{H} \equiv 10000 \times \sum_i \alpha_i^2$, where α_i is the market share of firm i. When all N firms evenly split market, $\mathcal{H} = 10000/N$.

Lerner index (Bernheim 37) $\mathcal{L} \equiv \sum_i \alpha_i \mathcal{L}_i$, where $\mathcal{L}_i \equiv \frac{p_i - c_i'(q_i)}{p_i}$ is firm i's margin and α_i is the market share of firm i.

Monopolistic competition (Bernheim 135–8)

1. Products are distinct, and each firm faces a downward-sloping demand curve;

2. The decisions of any given firm have negligible effects on any other firm (note this does *not* hold for the Hotelling model, where firms have a measurable effect on their neighbors);

3. There is free entry, with zero profits.

Can formalize as a vector of N differentiated commodities (for N large) and a numeraire good y, where representative consumer has utility

$$u(x, y) = y + g\left(\sum_{i=1}^{N} f(x_i) \right)$$

and the curvature of $f(\cdot)$ gives the extent of substitutability of the goods.

Relationship between equilibrium variety and optimal variety is dictated by:

1. When a product is added, revenues generated fall short of incremental consumer surplus because firms can't perfectly price discriminate; this biases towards too little entry.

2. Firms don't take into account the effect of introducing a product on the profits of others; if goods are substitutes, this biases towards too much entry.

If goods are complements, these biases reinforce and we have too little variety relative to social optimum.

Entry deterrence (Bernheim) An incumbent can take some action with long-term commitment prior to the arrival of an entrant that makes entry less attractive. For example:

1. Selecting to produce a middle-of-the-road product (Hotelling model); note the deterrent action is the same as what the firm would do if it faced no entry with certainty (entry is "blockaded.")

2. Producing a proliferating range of products (e.g., RTE cereal).

3. Preemptive investment in production capacity; shifts a portion of marginal cost to a sunk cost. Note entrant may also limit capacity to reduce threat to incumbent.

Vickrey auction (Bernheim 22–3, 82–3) a.k.a. second-price auction. I bidders simultaneously submit sealed bids for an item that each values at v_i. Winning bidder pays second highest bid. Bidding $p_i = v_i$ weakly dominates any other pure strategy, and is not weakly dominated by any other strategy. Analysis does not depend on complete vs. incomplete information—everyone bidding their valuation is a BNE.

First-price sealed bid auction (Bernheim 83–7) I bidders simultaneously submit sealed bids for an item that each values at v_i. Winning bidder pays his own bid. Symmetric BNE gives a decision rule where each bids below his valuation.

Realized revenues typically differ from Vickrey (second-price sealed bid) auction, but expected revenue is the same *given independent private valuations* (per the revenue equivalence theorem).

English auction (Bernheim 87) a.k.a. ascending price auction Posted price of good is slowly increased until only one bidder remains. Staying in until posted price exceeds valuation is a weakly dominated strategy; outcomes are equivalent to Vickrey (second-price sealed bid) auction.

Dutch auction (Bernheim 87) a.k.a. descending price auction Posted price of good is slowly decreased until a bidder buys. Outcomes are equivalent to first-price sealed bid auction.

Public good (Bernheim 51–4) A non-rival and non-excludable good.

Spence signaling model (Bernheim 218–35) Workers choose education level; education does nothing for worker's productivity, but is less costly on margin for more productive workers. All workers have same outside opportunities. Three types of WPBE:

1. Separating equilibria: different types choose different education levels.

2. Pooling equilibria: different types choose same education levels. (Note pooling equilibria with strictly positive levels of education are highly inefficient—education adds nothing to productivity nor does it differentiate workers.)

3. Hyprids: some members of differently-typed groups pool, others separate.

Equilibrium satisfies equilibrium dominance condition a.k.a. the "intuitive criterion" iff whenever an individual gets some level of education that no one should in equilbrium, no low-productivity worker should ever choose, and some high-productivity worker could conceivably choose, the firm assumes he is a high-productivity worker with certainty.

2.11 Repeated games with complete information

Infinitely repeated game (Bernheim 164–5) A *supergame* formed by (potentially) infinite repetitions of some *stage game* (i.e., the game played in each period). Note there need not actually be infinite repetitions, but there must be a nonzero possibility in each stage that the game will continue.

Given absence of terminal nodes, need mapping from strategies to expected payoffs, e.g.,

1. Discounted payoffs: $u_i(v_i) = \sum_{t=0}^{\infty} \delta^t v_i(t)$, where the discount factor may reflect both time preference and the probability of continuation.

2. Average payoffs: $u_i(v_i) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} v_i(t)$.

3. Overtaking criterion: a strategy is preferred iff there is some T beyond which all partial sums of stage game payoffs exceed the corresponding partial sum for other strategies; n.b., only a partial ordering.

Feasible payoff (Bernheim 169–70) The convex hull of payoff vectors from pure strategy profiles in the stage game. Note this is potentially larger than the set of payoffs achievable through mixed strategies, since we allow for correlations.

Individually rational payoff (Bernheim 170–1) (Stage game) payoff vectors where each player receives at least his minmax payoff

$$\pi_i^m \equiv \min_{\delta_{-i}} \max_{\delta_i} \pi_i(\delta).$$

Folk Theorem (Bernheim 171–2, 5, 7–8) Consider a supergame formed by repeating a finite stage game an infinite number of times; suppose players use the average payoff criterion. Then the set of *feasible* and *individually rational* payoffs is precisely the set of average payoffs for Nash equilibria (which need not be SPNE).

1. "Anything can happen;" this makes comparative statics are problematic.

2. Inability to write binding contracts is not very damaging; anything attainable through a contract can also be obtained through some self-enforcing agreement (if there is no discounting).

For discounted payoffs, all feasible payoffs that strictly exceed minmax payoffs for every player are the average payoff for some NE and all discount rates sufficiently close to 1.

Subject to some technical conditions, versions of the folk theorem (with and without discounting) hold for SPNE.

Nash reversion (Bernheim 176–7) a.k.a. "grim trigger" strategy. Players attempt to support cooperation, reverting to a static (stage-game) equilibrium as punishment if someone deviates. If a Nash reversion strategy is a NE, then it is a SPNE.

Finitely repeated game (Bernheim 179–80) If there is a unique (up to payoffs) NE for the stage game, there is a unique SPNE for the repeated game, consisting of the repeated stage game equilibrium. However, cooperation may be possible when the stage game has multiple NE.

Stick-and-carrot equilibrium (Bernheim 185–6) If a firm strays from the (cooperative) equilibrium path, all firms *including itself* punish it for one period. If any firm does not participate in the punishment (including the punished firm itself), it gets punished again in the next period. The punishment is the "stick;" the fact that punishment will end as soon as the firm punishes itself is the "carrot."

2.12 Collusion and cooperation

Core (Jackson; MWG 653–4) The core is the set of allocations not *blocked* by any coalition. A coalition will block an allocation if there is some allocation feasible for the coalition such that each member is strictly better off ("strong blocking"), or that every member is weakly better off and some member is strictly better off ("weak blocking").

Core allocations must be Pareto optimal.

The set of Walrasian equilibria is a subset of the core (basically, the FWT) since there are no externalities. In other settings, the core could be empty due to externalities.

Core convergence (Jackson; MWG 655–7) As we increase the number of "replicas" of a pure exchange economy, the core of the replicated economy converges to (equal treatment replicas of) the W.E. of the original economy.

Nontransferable utility game (Jackson) Feasible allocations for a coalition T must be listed explicitly.

Transferable utility game (Jackson; MWG 676) Feasible allocations for a coalition S are $V(S) \equiv \mathbb{R}_+^{|S|}$. Generally assume:

1. The normalization $V(\varnothing) = 0$;

2. $V(I) \geq V(S) + V(I \setminus S)$ for all $I \subseteq S$, where I is the set of all players;

3. Superadditivity (a stronger version of the prior assumption): $V(S \cup T) \geq V(S) + V(T)$ when $S \cap T = \varnothing$.

In a TU game, an allocation can be strictly blocked iff it can be weakly blocked; the weak and strong cores are therefore identically

$$\{x \in R^{|I|} : \sum_i x_i = V(I) \text{ and } \forall S \subseteq I, \sum_i x_i \geq V(S)\}.$$

Shapley value (Jackson; MWG 673, 679–81) Given a characteristic function $V : 2^I \to \mathbb{R}_+$, the Shapley value is value function:

$$\phi_i^{\mathrm{SV}}(V) \equiv \sum_{S \subseteq I \text{ s.t. } i \notin S} \frac{|S|!\,(|I| - |S| - 1)!}{|I|!} \big[V(S \cup \{i\}) - V(S)\big].$$

That is, the average marginal value that i contributes over all possible orderings in which members could be added to a coalition. It is a normative allocation, and can be seen as a positive description under certain kinds of bargaining regimes.

The Shapley value is the unique value function satisfying:

1. Symmetry: If we relabel agents, the Shapley values are relabeled accordingly.

2. Carrier: $T \subseteq I$ is a carrier iff $V(S \cap T) = V(S)$ for all $S \subseteq I$; if T is a carrier, then $\sum_{i \in T} \phi_i(V) = V(I) = V(T)$.

3. Dummy: i is a dummy iff $V(S \cup \{i\}) = V(S)$ for all $S \subseteq I$; if i is a dummy, $\phi_i(V) = 0$. Note Carrier $\implies$ Dummy.

4. Additivity: $\phi(V + W) = \phi(V) + \phi(W)$; implies that $\phi(\lambda V) = \lambda \phi(V)$. Convenient, but not mathematically clear why this should hold.

Simple game (Jackson) A TU game where

1. $V(S) \in \{0, 1\}$,

2. $S \subseteq T \implies V(S) \leq V(T)$,

3. $V(S) = 1 \implies V(I \setminus S) = 0$,

4. $V(I) = 1$.

In a simple game, i is a "veto player" iff $V(S) = 1 \implies i \in S$. The core is nonempty iff there is at least one veto player. Shapley values are then:

$$\phi_i^{\mathrm{SV}} = \begin{cases} \frac{1}{\# \text{ of veto players}}, & i \text{ is a veto player}; \\ 0, & \text{otherwise.} \end{cases}$$

If the set of veto players is a carrier, $\phi^{\mathrm{SV}} \in$ core.

Convex game (Jackson; MWG 683) A TU game $V(\cdot)$ is convex iff $S \subseteq T$ and $i \notin T$ implies

$$V(S \cup \{i\}) - V(S) \leq V(T \cup \{i\}) - V(T);$$

i.e., there are *increasing* marginal returns as coalitions grow.

For a convex game, the core is nonempty; the Shapley value is in the core.

One-sided matching (Jackson) Players are each allocated one object; each player has ordinal preferences over objects. Given strict preferences, we can find the unique (weak) core allocation using the Top Trading Cycles algorithm:

1. Look for cycles among individuals' top choices;

2. Assign each member of cycles her top choice;

3. Return to step 1, with already assigned individuals/objects removed from consideration.

We can support this core as a competitive equilibrium by assigning a common price to the objects in each cycle as they are assigned. The price for cycles assigned in the same round can vary across cycles, but must be strictly lower than the prices for objects assigned in previous rounds.

The algorithm is strategy-proof: truth telling dominates.

Two-sided matching (Jackson) Two sets M and W, which may have different cardinalities. Every individual is either matched to someone in the other set, or remains unmatched.

Can find two core allocations (here strict = weak) using Gale-Shapley [a.k.a. Deferred acceptance] algorithm: Suppose M "propose," then

1. Each M proposes to his favorite W to whom he has not yet proposed (unless he would prefer to stay single);

2. Each W who has been proposed to accepts her favorite proposal (unless she would prefer to stay single)—this could involve breaking an engagement she made in a previous round;

3. Return to step 1, where the only unengaged M propose (i.e., those who either had a proposal rejected last round, or had an engagement broken).

The core is a lattice; Gale-Shapley algorithm yields "best" for proposing group (and "worst" for other group). Thus if we get same result from M-proposing and W-proposing algorithms, core is a singleton. Note this result relies on:

- Two-sided matching (cf, roommate matching);

- One-to-one matching (cf, firm and workers);

- Strict preferences.

The algorithm is *not* strategy-proof: there may be profitable manipulations (lying by rejecting proposals), but they are typically difficult to implement.

2.13 Principal-agent problems

Moral hazard (Jackson) "Hidden action" problems. Agent takes (non-contractable) action e, outcome π has some distribution that varies based on e. Generally assume risk-neutral principal and risk-averse (and/or liability-limited) agent, so that optimal solution is not merely "selling the firm to the agent."

Principal structures optimal contract—payment $w(\pi)$ as a function of realized outcome—that, for a desired effort level e^*,

1. Maximizes principal's expected payoff: $\mathrm{E}[\pi - w(\pi)|e^*]$;

2. Satisfies the agent's participation constraint (i.e., individual rationality constraint): $\mathrm{E}[u(w(\pi))|e^*] - g(e^*) \geq \bar{u}$, where $g(\cdot)$ is the agent's cost of effort, and $\bar{u}$ is his reservation utility;

3. Satisfies the agent's incentive compatibility constraint: $e^* \in \mathrm{argmax}_e \mathrm{E}[u(w(\pi))|e] - g(e)$.

Adverse selection (Jackson) "Hidden information" problems. Can be mitigated using, e.g., warantees, repeated interaction, reputation mechanisms, signaling/screening.

Signaling (Jackson) Agents choose a costly signal (which per Spence is often assumed to be meaningless other than for its informational value), and then principals Bertrand bid for contracts with agents.

A multitude of pure strategy sequential equilibria typically exist—both pooling (various types choose same signal) and separating (various types choose different signals).

Screening (Jackson) Bertrand principals announce a schedule of contracts they are willing to engage in (pairs of signals and wages), and then agents choose a contract (and hence a signal).

No pooling equilibria can exist, and there is only one separating equilibrium that can exist (but may not). Thus in contrast with signaling, where our problem was a multitude of PSNE, screening games may have no PSNE.

2.14 Social choice theory

Voting rule (Jackson) n individuals must choose among a set A of alternatives. Each individual has a complete, transitive preference relation $\succ_i$ over A that does not have indifference.

A voting rule $R(\succ) \equiv R(\succ_1, \ldots, \succ_n)$ gives a social welfare ordering over A (which in general allows indifference). The corresponding strict social welfare ranking is $P(\succ)$.

Neutrality (Jackson) For $A = \{a, b\}$ (i.e., $|A| = 2$), consider $\succ$ and $\succ'$ with $\succ_i \neq \succ_i'$. Then $R(\cdot)$ is neutral (over alternatives) iff $aR(\succ)b \iff bR(\succ')a$.

Anonymity (Jackson) Let $\pi(\cdot)$ be a permutation over $\{1, \ldots, n\}$. Define $\succ_i' \equiv \succ_{\pi(i)}$. Then $R(\cdot)$ is anonymous (over individuals) iff $aR(\succ)b \iff aR(\succ')b$.

Monotonicity (Jackson) Consider $\succ$ and $\succ'$ with $\succ_j = \succ_j'$ for all $j \neq i$, and $\succ_i = a$ while $\succ_i' = b$. Then $R(\cdot)$ satisfies monotonicity iff $bR(\succ)a \implies bR(\succ')a$.

$R(\cdot)$ satisfies strict monotonicity iff $bR(\succ)a \implies bP(\succ')a$.

May's Theorem (Jackson) Let $A = \{a, b\}$ (i.e., $|A| = 2$). Then $R(\cdot)$ is complete, transitive (which has no bite for $|A| = 2$), neutral, anonymous, and satisfies *strict* monotonicity iff it is majority rule.

If we replace strict monotonicity with monotonicity, we get that $R(\cdot)$ must be a quota rule.

Unanimity (Jackson) $\mathbb{R}(\cdot)$ is unanimous iff $a \succ_i b \forall i$ implies that $aP(\succ)b$.

Arrow's independence of irrelevant alternatives (Jackson) Consider $\succ$ and $\succ'$ with $a \succ_i b \iff a \succ_i' b$ for all i.

$\mathbb{R}(\cdot)$ satisfies AIIA iff $aR(\succ)b \iff aR(\succ')b$; i.e., $R(\succ)$ over $\{a, b\}$ only depends on $\succ$ over $\{a, b\}$, not over any other ("irrelevant") alternatives.

Arrow's Theorem (Jackson; Micro P.S. 4) Let $|A| \geq 3$. Then $R(\cdot)$ is complete, transitive, unanimous, and satisfies AIIA iff it is dictatorial (i.e., if there is some i such that $\forall \succ$, $aP(\succ)b \iff a \succ_i b$).

The theorem is tight; i.e., giving up any of completeness, transitivity, unanimity, or AIIA allows a non-dictatorial $R(\cdot)$.

Condorcet winner (Jackson) $a \in A$ is a Condorcet winner iff it is majority preferred to every other alternative.

Condorcet-consistent rule (Jackson) A voting rule that picks the Condorcet winner if there is one.

Gibbard-Satterthwaite Theorem (Jackson) Let $3 \leq |A| < \infty$, and $F(\cdot)$ be a social choice function (note we do not require a full ordering). Then $F(\cdot)$ has range A (implied by unanimity) and is strategy-proof (i.e., DSIC) iff it is dictatorial.

Vickrey-Clarke-Groves mechanism (Jackson) Let $\theta_i \in \Theta_i$ be the type of individual i, and $\hat{\theta}_i$ be his announced type. Suppose $d(\theta)$ makes ex post efficient decisions. Then a mechanism with transfers

$$t_i(\hat{\theta}) = \sum_{j \neq i} \left[u_j(d(\hat{\theta}), \hat{\theta}_j) \right] + x_i(\hat{\theta}_{-i})$$

for any $x_i(\cdot)$ is dominant-strategy incentive compatible (DSIC).

Conversely, if (d, t) is DSIC, $d(\cdot)$ is ex post efficient, and Θ_i is sufficiently rich such that $\forall v_i \in \{v_i \colon D \to \mathbb{R}\}$ there exists θ_i such that $v_i(\cdot) = u_i(\cdot, \theta_i)$, then $t(\cdot)$ must satisfy the above condition.

Note that Gibbard-Satterthwaite says that a general DSIC mechanism must be dictatorial. However, here we have restricted ourselves to quasilinear preferences, and get a DSIC mechanism without being dictatorial. However, although it reaches an efficient decision it is *not* always balanced (payments do not total to zero), and hence is not typically overall efficient.

Pivotal mechanism (Jackson) An example of a VCG mechanism where

$$t_i(\hat{\theta}) = \sum_{j \neq i} \left[u_j(d(\hat{\theta}), \hat{\theta}_j) \right] - \max_d \sum_{j \neq i} u_j(d, \hat{\theta}_j).$$

That is, the transfers are the externality imposed on others by choosing $d(\hat{\theta})$ instead of $d(\hat{\theta}_{-i})$.

Ensures feasibility, since transfers are always negative. However, budget generally doesn't balance—there is a need to "burn money."

3 Macroeconomics

3.1 Models

Two period intertemporal choice model (Nir 1-11) Maximize $U(C_0, C_1)$ subject to $C_0 + S_0 \leq Y_0$ and $C_1 \leq RS_0$. The constraints can be rewritten $C_0 + \frac{1}{R}C_1 \leq Y_0$.

Neoclassical growth model (discrete time) (Nir 3-3–5, 11-3, 8, 10, 16–28, 12-12–26) (a.k.a. Ramsey model) Maximize $\sum_t \beta^t U(c_t)$ subject to:

1. $c_t + k_{t+1} \leq F(k_t, n_t) + (1-\delta)k_t$ for all t (RHS can be denoted $f(k_t, n)$, or $f(k_t)$, since the lack of disutility of working ensures that $n_t = 1$);

2. $k_{t+1} \geq 0$ for all t;

3. $c_t \geq 0$ for all t;

4. k_0 is given.

We further assume $U(\cdot)$ satisfies Inada conditions; that the production function is constant returns to scale and satisfies $F_k' > 0$, $F_{kk}'' < 0$, $F_n' > 0$, and $F_{nn}'' < 0$; and TVC $\lim_{t \to \infty} \beta^t U'(c_t) f'(k_t) k_t = 0$.

The problem can be rewritten in "SL canonical form" as maximizing $\sum_t \beta^t U(f(k_t) - k_{t+1})$ subject to $k_{t+1} \in [0, f(k_t)]$, the given k_0, and conditions on $f(\cdot)$ and $U(\cdot)$ as above. In functional equation form, we write $V(k) \equiv \max_{k' \in [0, f(k)]} U(f(k) - k') + \beta V(k')$ (again, with additional conditions as above).

Steady state satisfies $\beta f'(k^*) = 1$. Note the utility function does *not* affect the steady state (although it will affect dynamics).

Linearization of the intereuler

$$u'(f(k_t) - k_{t+1}) - \beta f'(k_{t+1}) u'(f(k_{t+1}) - k_{t+2}) = 0$$

about steady state gives:

$$\begin{bmatrix} k_{t+2} - k_* \\ k_{t+1} - k_* \end{bmatrix} \approx \begin{bmatrix} 1 + \frac{1}{\beta} + \beta \frac{U'(c_*)}{U''(c_*)} f''(k_*) & -\frac{1}{\beta} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} k_{t+1} - k_* \\ k_t - k_* \end{bmatrix}$$

The optimal policy function $g(k)$ satisfies:

1. $g(\cdot)$ is single-valued, since the value function is strictly concave.

2. $g(0) = 0$ since that is the only feasible choice.

3. $g(k)$ is in the interior of $\gamma(k)$, since otherwise exactly one of $U'(c)$ and $U'(c')$ would be infinite, violating the Intereuler condition.

4. $g'(k) > 0$, since as k increases, marginal cost of saving goes down, while marginal benefit stays same.

5. There is unique $k^* > 0$ such that $g(k^*) = k^*$.

6. $k > k^* \implies k > g(k) > k^*$, and $k < k^* \implies k < g(k) < k^*$; i.e., capital moves closer to (without crossing over) the steady-state level.

7. The sequence $k_0, g(k_0), g(g(k_0)), \ldots$ displays monotonic and global convergence.

Endowment economy, Arrow-Debreu (Nir 13-4–6) At every period household gets y_t units of good; there is no storage, and trading in all commodities (goods delivered in each period) takes place in period 0. Maximize $\sum \beta^t U(c_t)$ subject to $\sum p_t c_t \leq \sum p_t y_t$ (budget constraint), and ensure $y_t = c_t$ (market clearing).

Endowment economy, sequential markets (Nir 13-7–9) At every period household gets y_t units of good; there is no storage, and trading in "assets" (loans across periods) takes place each period. Maximize $\sum \beta^t U(c_t)$ subject to $c_t + a_{t+1} \leq y_t + R_t a_t$ (budget constraint), and ensure $y_t = c_t$, $a_{t+1} = 0$ (market clearing).

Production economy, Arrow-Debreu (Nir 13-10–6) Household owns capital rented for production by competitive firms, which also rent labor from households. Capital depreciates at rate δ. There is no storage, and trading in all commodities (goods delivered, capital rental, and labor in each period) takes place in period 0.

1. Households maximize $\sum \beta^t U(c_t)$ subject to $\sum p_t[c_t + k_{t+1}] \leq \sum p_t[r_t k_t + (1-\delta)k_t + n_t w_t]$.
2. Firms maximize (in each period) $p_t F(k_t, n_t) - p_t r_t k_t - p_t w_t n_t$.
3. Markets clear: $c_t + k_{t+1} = F(k_t, n_t) + (1-\delta)k_t$.

Note the rental rate of capital (r_t) and wage rate (w_t) are measured in units of consumption good.

Production economy, sequential markets (Nir 13-25) As above, household owns capital rented for production by competitive firms, which also rent labor from households. Capital depreciates at rate δ. There is no storage, and each period, goods delivered, capital rental, and labor are traded. Note there is no inter-period trading, so we do not need a price for goods.

1. Households maximize $\sum \beta^t U(c_t)$ subject to $c_t + k_{t+1} \leq r_t k_t + (1-\delta)k_t + n_t w_t$.
2. Firms maximize (in each period) $F(k_t, n_t) - r_t k_t - w_t n_t$.
3. Markets clear: $c_t + k_{t+1} = F(k_t, n_t) + (1-\delta)k_t$.

Formulated recursively (assuming $n=1$),

1. Households have $V(k, K) = \max_{c,k'}[U(c) + \beta V(k', K')]$ subject to $c + k' = R(K)k + W(K)$; where K is aggregate capital, and $R(K)$ and $W(K)$ are the rental and wage rates as functions of aggregate capital. We further require the "forecast" of future capital to be rational: $K' = G(K)$, where $G(\cdot)$ is the optimal aggregate policy function.
2. Firms maximize $F(K) - RK - W$, which gives FOC $R(K) = F_k(K) + 1 - \delta$ and $W(K) = F_n(K)$.
3. Markets clear: $C + K' = F(K) + (1-\delta)K$.
4. Consistency: $G(K) = g(K, K)$.

Overlapping generations (Nir 17-15–6, 17-32) In an endowment OLG economy, a competitive equilibrium is Pareto optimal iff $\sum_{t=0}^{\infty} 1/p_t = \infty$ where the interest rate is $R_t = p_t/p_{t+1}$.

In an an OLG economy with production, a competitive equilibrium is dynamic efficient iff $F'_K(K_*) + 1 - \delta \geq 1$.

Real Business Cycles model (Nir 18-7–14) Benchmark RBC model is NCGM with endogenous hours, ownership of firms and capital by households, and stochastic productivity A_t multiplying labor (so production function is $F(k_t, A_t n_t)$). We specified:

1. CRRA utility: $U(c, n) = \frac{c^{1-\sigma}-1}{1-\sigma} - \frac{n^{1+\chi}}{1+\chi}$;
2. Cobb-Douglas production: $y_t = A_t k_t^\alpha n_t^{1-\alpha}$;
3. Productivity given by $\log A_t = \rho \log A_{t-1} + \varepsilon_t$ with ε_t distributed iid normal.

Optimal taxation—primal approach (Nir 10) Solve HH problem and combine FOC and budget constraint with government budget constraint to get one expression that ties together allocations but has *no prices or taxes*. Then government maximizes HH utility over *allocations* subject to this constraint.

Note that in the case with capital,* the problem is not stationary; i.e., we cannot use our typical dynamic programming tools, and need to worry about the government's ability to commit to a future tax policy.

3.2 Imperfect competition

Imperfect competition model (Nir 21) Consider two stages of production: final good manufacturer operates competitively, purchasing inputs from intermediate good producers, who are small enough to take general price levels and aggregate demand as given, but whose products are imperfect substitutes to the production process.

Imperfect competition: final goods (Nir 21) Final good producer has Dixit-Stiglitz production technology:

$$Y = \left[\int_0^1 Q_j^\rho \, dj\right]^{1/\rho}$$

which is constant elasticity of substitution ($\frac{1}{1-\rho}$ between all pairs of inputs) and constant returns to scale. $\rho \in (0,1)$, with $\rho \to 1$ corresponding to a competitive market where all inputs are perfect substitutes.

Conditional factor demands are therefore

$$Q_j(Y) = \left[\frac{p_j}{\lambda}\right]^{\frac{1}{\rho-1}} Y = \left[\frac{p_j}{P}\right]^{\frac{1}{\rho-1}} Y,$$

where the Lagrange multiplier on $[\int_0^1 Q_j^\rho \, dj]^{1/\rho} \geq Y$ also satisfies $\lambda = E^*/Y \equiv P$, the (aggregate) average cost. Solving for this gives

$$\lambda = P = \left[\int_0^1 p_j^{\frac{\rho}{\rho-1}}\right]^{\frac{\rho-1}{\rho}}.$$

The own-price elasticity of demand is $\eta \equiv \frac{1}{\rho-1}$, ignoring the effects through P.

Imperfect competition: intermediate goods (Nir 21) Intermediate goods producers have identical production technologies (affected by same shocks): $Q_j = z k_j^\alpha n_j^{1-\alpha} - \phi$, where $\phi > 0$ is overhead cost (since there aren't really economic profits). Has constant marginal cost, and therefore increasing returns to scale (because of ϕ).

Monopolistic problem is to maximize over prices $pQ(p) - c(Q(p))$; first-order condition gives markup $\equiv p/\text{MC} = \frac{\eta}{1+\eta} = \frac{1}{\rho}$.

3.3 General concepts

Cobb-Douglas production function (Macro P.S. 1) $F(K, N) = zK^\alpha N^{1-\alpha}$. Constant returns to scale. Fraction α of output goes to capital, and $1 - \alpha$ to labor.

Constant relative risk aversion utility function (Nir 14-15, 18-72–4) $U(c) = (c^{1-\sigma} - 1)/(1-\sigma)$, with $\sigma > 0$ (reduced to log utility with $\sigma = 1$;† empirically we expect $\sigma \in [1,5]$). Relative risk aversion is σ. Higher risk aversion also corresponds to a higher desire to smooth consumption over time.

Additively separability with stationary discounting (Nir 1-24) General assumption that $U(C_0, C_1) = u(C_0) + \beta u(C_1)$ with $\beta \in (0,1)$.

Intertemporal Euler equation (Nir 1-28, 11-5) $u'(c_t) = \beta R u'(c_{t+1})$, where R is the return on not-consuming (e.g., $f'(k_{t+1}) \equiv F'_k(k_{t+1}, n_{t+1}) + (1-\delta)$). This first-order difference equation is given by FOCs of the maximization problem for two consecutive time periods.

Intertemporal elasticity of substitution (Nir 1-31, Max notes) $-d\log\frac{C_0}{C_1}/d\log R = -d\log\frac{C_0}{C_1}/d\log\frac{U'(C_0)}{U'(C_1)}$.

For CRRA utility function $U(c) = c^{1-\sigma}/(1-\sigma)$, intereuler is $C_0^{-\sigma} = \beta C_1^{-\sigma} R$ so IEoS is $\frac{1}{\sigma} = \text{RRA}^{-1}$.

General equilibrium (Nir 2-11) A competitive equilibrium is a set of *allocations* and *prices* such that

1. Actors (i.e., households, firms, …) maximize their objectives;
2. Markets clear; and
3. Resource constraints are satisfied.

Inada conditions (Nir 3-5) We generally assume utility satisfies Inada conditions, which imply that the nonnegativity constraint on consumption will not bind:

1. $U(0) = 0$;
2. U is continuously differentiable;

---

*We generally need to impose restrictions on first-period taxation, since otherwise the government will tax capital very highly then (since it's supplied perfectly inelastically).

†Note the Taylor approximation of $c^{1-\sigma} - 1$ about $\sigma = 1$ is $c^{1-\sigma} - 1 \approx (c^0 - 1) - c^0(\log c)(\sigma - 1) = (1-\sigma)\log c$. Thus as $\sigma \to 1$, we have $U(c) = (c^{1-\sigma} - 1)/(1-\sigma) \approx \log c$.

3. $U'(x) > 0$ (U is strictly increasing);

4. $U''(x) \leq 0$ (U is strictly concave);

5. $\lim_{x \to 0+} U'(x) = \infty$;

6. $\lim_{x \to \infty} U'(x) = 0$.

Transversality condition (Macro P.S. 2) In NCGM, $\lim_{t \to \infty} \beta^t U'(c_t) f'(k_t) k_t = 0$.

No Ponzi condition (Max notes) Credit constraint on agents, which should never bind but suffice to prevent the optimality of a "doubling" strategy. For example, $\forall t$, $a_t \geq -\bar{k}$ for some $\bar{k}$.

Dynamic systems (Nir 12-4–11, Macro P.S. 5-4) Let the sequence $\{x_t\}$ evolve according to $x_{t+1} = W x_t$. Using eigen decomposition $W = P \Lambda P^{-1}$, giving the "decoupled system" $P^{-1} x_{t+1} = \Lambda P^{-1} x_t$. Then if $\tilde{x}_t \equiv P^{-1} x_t$, we have $\tilde{x}_{ti} = \lambda_i^t \tilde{x}_{0i}$ and hence $x_t = P \Lambda^t \tilde{X}_0$.

For 2×2 case, if

1. $|\lambda_1| > 1$ and $|\lambda_2| > 1$, then the system "explodes" unless $\tilde{x}_{0,1} = \tilde{x}_{0,2} = 0$ (source).

2. $|\lambda_1| < 1$ and $|\lambda_2| < 1$, then the system converges for any $\tilde{x}_{0,1}$ and $\tilde{x}_{0,2}$ (sink).

3. $|\lambda_1| < 1$ and $|\lambda_2| > 1$, then the system converges for any $\tilde{x}_{0,1}$ as long as $\tilde{x}_{0,2} = 0$ (saddle path).

The speed of convergence for a state variable equals one minus the slope of the optimal policy function $(1 - g'(k))$.

First Welfare Theorem (Nir 13-27) The competitive equilibrium is Pareto optimal. Then if we know our solution to the social planner problem is the unique Pareto optimal solution, we know it must equal the competitive equilibrium.

Ramsey equilibrium (Nir 20-3, final solutions) A Ramsey equilibrium is allocations, prices, and taxes such that:

1. Households maximize utility subject to their budget constraints, taking prices and taxes as given, and

2. Government maximizes households' utility while financing government expenditures (i.e., meeting its budget constraint).

3.4 Dynamic programming mathematics

Metric space (Nir 6-3–4) A set S and a distance function $\rho \colon S \times S \to \mathbb{R}$ such that:

1. $\forall x, y \in S$, $\rho(x, y) \geq 0$;

2. $\forall x, y \in S$, $\rho(x, y) = 0 \iff x = y$;

3. $\forall x, y \in S$, $\rho(x, y) = \rho(y, x)$;

4. Triangle inequality: $\forall x, y, z \in S$, $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.

Normed vector space (Nir 6-5–6) A vector space S and a norm $\|\cdot\| \colon S \to \mathbb{R}$ such that:

1. $\forall x \in S$, $\|x\| \geq 0$;

2. $\forall x \in S$, $\|x\| = 0 \iff x = 0$;

3. Triangle inequality: $\forall x, y \in S$, $\|x + y\| \leq \|x\| + \|y\|$.

4. Scalar multiplication: $\forall x \in S$, $\forall \alpha \in \mathbb{R}$, $\|\alpha x\| = |\alpha| \, \|x\|$.

Note any normed vector space induces a metric space with $\rho(x, y) \equiv \|x - y\|$.

Supremum norm ($\mathbb{R}^n$) (Nir 6-7–8) $\|\cdot\|_s \colon \mathbb{R}^n \to \mathbb{R}$ with $\|x\|_s \equiv \sup_{i=1,\ldots,n} |x^i|$.

Euclidean norm (Nir 6-9) $\|\cdot\|_E \colon \mathbb{R}^n \to \mathbb{R}$ with

$$\|x\|_E \equiv \sqrt[n]{\sum_{i=1}^{n} |x_i|^n}.$$

Continuous function (Nir 6-10; Micro math 3) $f \colon S \to \mathbb{R}$ is continuous at x iff $\forall y \in S$ and $\forall \varepsilon > 0$, there exists $\delta > 0$ such that $\|y - x\| < \delta \implies |f(y) - f(x)| < \varepsilon$.

Equivalently, iff for all sequences x_n converging to x, the sequence $f(x_n)$ converges to $f(x)$.

Supremum norm (real-valued functions) (Nir 6-7–8) Let $C(X)$ denote the set of bounded, continuous functions from X to $\mathbb{R}$. Then $\|\cdot\|_s \colon C(X) \to \mathbb{R}$ with $\|f\|_s \equiv \sup_{x \in X} |f(x)|$.

Convergent sequence (Nir 7-4) The sequence $\{x_i\}_{i=0}^{\infty}$ in S converges to $x \in S$ (or equivalently, the sequence has limit x) iff $\forall \varepsilon > 0$, there exists n such that $i > n \implies \|x_i - x\| < \varepsilon$. In other words, there is a point in the sequence beyond which all elements are arbitrarily close to the limit.

Cauchy sequence (Nir 7-9) The sequence $\{x_i\}_{i=0}^{\infty}$ in S is Cauchy iff $\forall \varepsilon > 0$, there exists n such that $i > n$ and $j > n$ together imply that $\|x_i - x_j\| < \varepsilon$. In other words, there is a point in the sequence beyond which all elements are arbitrarily close to each other. Every convergent sequence is Cauchy (shown via triangle inequality), but not every Cauchy sequence is convergent.

Complete metric space (Nir 7-10–16) A metric space (S, ρ) is complete iff every Cauchy sequence in S converges *to some point in S*. Importantly, if $C(X)$ is the set of bounded, continuous functions from X to $\mathbb{R}$, and $\|\cdot\|_s \colon C(X) \to \mathbb{R}$ is the sup norm, then $(C(X), \|\cdot\|_s)$ is a complete metric space.

Contraction (Nir 8-3–4) If (S, ρ) is a metric space, the operator $T \colon S \to S$ is a contraction of modulus $\beta \in (0, 1)$ iff $\forall x, y \in S$, we have $\rho(Tx, Ty) \leq \beta \rho(x, y)$. That is, T brings any two elements closer together. Every contraction is a continuous mapping.

Contraction Mapping Theorem (Nir 8-5) [a.k.a. Banach fixed point theorem.] If T is a contraction on a *complete* metric space, then

1. T has exactly one fixed point V^* such that $TV^* = V^*$; and

2. The sequence $\{V_i\}$ where $V_{i+1} = TV_i$ converges to V^* from any starting V_0.

Contraction on subsets (Nir 8-11) Suppose T is a contraction on a complete metric space (S, ρ), with fixed point $V^* = TV^*$. Further suppose $Y \subseteq S$ is a closed set, that $Z \subseteq Y$, and that $\forall y \in Y$, $Ty \in Z$. Then $V^* \in Z$.

Blackwell's sufficient conditions for a contraction (Nir 8-12–3) Blackwell gives sufficient (not necessary) conditions for an operator to be a contraction on the metric space $B(\mathbb{R}^n, \mathbb{R})$ (the set of bounded functions $\mathbb{R}^n \to \mathbb{R}$) with the sup norm. An operator T is a contraction if it satisfies

1. Monotonicity: if $\forall x$, $f(x) \leq g(x)$, then $\forall x$, $Tf(x) \leq Tg(x)$; and

2. Discounting: there exists $\beta \in (0, 1)$ such that for all $\alpha \geq 0$, $f(\cdot)$, and x, we have $T(f(x) + \alpha) \leq T(f(x)) + \beta \alpha$ [slight abuse of notation].

Principle of Optimality (Nir 9) Under certain conditions, the solutions to the following two problems are the same:

1. Sequence problem: $W(x_0) = \max_{\{x_{t+1}\}} \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1})$ such that x_0 given and $\forall t$, $x_{t+1} \in \Gamma(x_t)$.

2. Functional equation: $V(x) = \max_{x' \in \Gamma(x)}[F(x, x') + \beta V(x')]$.

Assume

1. $\Gamma(x)$ is nonempty for all x;

2. For all initial conditions x_0 and feasible plans $\{x_t\}$, the limit $u(\{x_t\}) \equiv \lim_{n \to \infty} \beta^t F(x_t, x_{t+1})$ exists (although it may be $\pm \infty$).

Then:

1. If $W(x_0)$ is the supremum over feasible $\{x_t\}$ of $u(\{x_t\})$, then W satisfies the FE.

2. Any solution V to the FE that satisfies boundedness condition $\lim_{n \to \infty} \beta^n V(x_n) = 0$ is a solution to the SP.

3. Any feasible plan $\{x_t^*\}$ that attains the supremum in the SP satisfies $W(x_t^*) = F(x_t^*, x_{t+1}^*) + \beta w(x_{t+1}^*)$ for $\forall t$.

4. Any feasible plan $\{x_t^*\}$ that satisfies $W(x_t^*) = F(x_t^*, x_{t+1}^*) + \beta w(x_{t+1}^*)$ for $\forall t$ and $\lim_{n \to \infty} \sup \beta^t W(x_t^*) \leq 0$ attains the supremum in the SP. [?]

So approach is: solve FE, pick the unique solution that satisfies boundeness condition, construct a plan from the policy corresponding to this solution, check the limit condition to make sure this plan is indeed optimal for the SP.

Dynamic programming: bounded returns (Nir 10) Given the SP/FE as above, assume:

1. x takes on values in a convex subset of $\mathbb{R}^l$, and $\Gamma(x)$ is nonempty and compact-valued for all x. This implies assumption 1 above.

2. The function F is bounded and continuous; $\beta \in (0,1)$. Together with assumption 1, this implies assumption 2 above.

3. For each x', the function $F(\cdot, x')$ is increasing in each of its first arguments.

4. $\Gamma(\cdot)$ is monotone; i.e., $x_1 \leq x_2 \implies \Gamma(x_1) \subseteq \Gamma(x_2)$.

5. F is strictly concave.

6. Γ is convex; i.e., if $x_1' \in \Gamma(x_1)$ and $x_2' \in \Gamma(x_2)$, then $\theta x_1' + (1-\theta)x_2' \in \Gamma(\theta x_1 + (1-\theta)x_2)$.

7. F is continuously differentiable on the interior of the set on which it is defined.

Then

1. Under assumptions 1 and 2, the operator T where $TV(x) \equiv \max_{x' \in \Gamma(x)}[F(x, x') + \beta V(x')]$ maps bounded continuous functions into bounded continuous functions, is a contraction (and therefore has a unique fixed point), and is the value function for the corresponding FE.

2. Under assumptions 1–4, the value function V (unique fixed point of T as defined above) is strictly increasing.

3. Under assumptions 1–2 and 5–6, the value function V is strictly concave, and the corresponding optimal policy correspondence g is a continuous function.

4. Under assumptions 1–2 and 5–7, the value function V is differentiable (by Benveniste-Scheinkman) and satisfies envelope condition $V'(x_0) = \frac{\partial F}{\partial x}|_{(x_0, g(x_0))}$.

Benveniste-Scheinkman Theorem (Nir 10-25) Suppose

1. $X \subseteq \mathbb{R}^l$ is a convex set;

2. $V: X \to \mathbb{R}$ is a concave function;

3. $x_0 \in \text{interior}(X)$; and D is a neighborhood of x_0, with $D \subseteq X$;

4. $Q: D \to R$ is a concave differentiable function with $Q(x_0) = V(x_0)$ and $\forall x \in D, \, Q(x) \leq V(x)$.

Then $V(\cdot)$ is differentiable at x_0 with $V'(x_0) = Q'(x_0)$.

The envelope condition of a value function is sometimes called the Benveniste-Scheinkman condition.

3.5 Continuous time

Dynamic systems in continuous time (Nir 14-2–3, 7) In $\mathbb{R}$, the system

$$\dot{x}_t = a x_t \equiv \frac{dx_t}{dt} = ax \iff \frac{dx_t}{x} = a \, dt$$
$$\iff \log x_t = at + c$$
$$\iff x_t = x_0 e^{at}$$

converges iff $\text{real}(a) < 0$.

In $\mathbb{R}^n$, the system $\dot{x}_t = A x_t \equiv \dot{\tilde{x}}_t = \Lambda \tilde{x}_t$ where $\tilde{x}_t = P^{-1} x_t$ and the eigen decomposition is $A = P\Lambda P^{-1}$. Therefore

$$\tilde{x}_t = e^{\Lambda t} \tilde{x}_0 = \begin{bmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{bmatrix} \tilde{x}_0.$$

For 2×2 case (recall $\det A = \prod \lambda_i$ and $\text{tr} A = \sum \lambda_i$),

$\det A < 0$:	Saddle path
$\det A > 0, \text{tr} A > 0$:	Unstable
$\text{tr} A < 0$:	Sink

Linearization in continuous time (Nir 14-6, Max notes) $\dot{x}_t = f(x_t) \implies \dot{x}_t \approx D_f(x_*)(x_t - x_*)$. [cf discrete time $x_{t+1} = g(x_t) \implies x_{t+1} - x_* \approx D_g(x_*)(x_t - x_*)$.]

NCGM in continuous time (Nir 14-8–15) Maximize $\int_{t=0}^{\infty} e^{-\rho t} U(c_t) \, dt$ such that $\dot{k}_t = W_t + R_t k_t - c_t - \delta k_t$. The Hamiltonian is given by:

$$H \equiv e^{-\rho t}\left[U(c_t) + \lambda_t [\underbrace{W_t + R_t k_t - c_t - \delta k_t}_{\text{Budget constraint w/o } \dot{k}_t}] \right]$$

$$\equiv e^{-\rho t} \; U(c_t) + \mu_t [W_t + R_t k_t - c_t - \delta k_t].$$

First order conditions $\frac{\partial H}{\partial c_t} = 0$ and $\frac{\partial H}{\partial k_t} = -\frac{d}{dt}(e^{-\rho t}\lambda_t) = e^{-\rho t}(\rho \lambda_t - \dot{\lambda}_t)$ (or equivalently, $\frac{\partial H}{\partial k_t} = -\dot{\mu}_t$), along with TVC $\lim_{t \to \infty} e^{-\rho t} \lambda_t k_t = 0$ (or equivalently, $\lim_{t \to \infty} \mu_t k_t = 0$), characterize the solution. Solving gives:

$$\dot{c}_t = \frac{U'(c_t)}{U''(c_t)}(\delta - R_t + \rho)$$
$$\dot{k}_t = \underbrace{W_t + R_t k_t}_{= y_t \text{ by CRS}} - c_t - \delta k_t.$$

Imposing functional forms for $U(\cdot)$ and $F(\cdot)$ and log-linearizing allows us to get a continuous time dynamic system for $\hat{k}_t$ and $\hat{c}_t$.

Log-linearization (Nir 14-24, Max notes) Write every variable x as $e^{\log x}$ and then linearize in $\log x$ about steady state $\log x_*$. Use notation $\hat{x} \equiv \log x - \log x_* \approx \frac{x - x_*}{x_*}$. For example,

1. $x \approx x_*(1 + \hat{x})$;

2. $xy \approx x_* y_*(1 + \hat{x} + \hat{y})$;

3. $x^\alpha y^\beta \approx x_*^\alpha y_*^\beta(1 + \alpha \hat{x} + \beta \hat{y})$;

4. $f(x) \approx f(x_*) + f'(x_*)x_* \hat{x} = f(x_*)(1 + \eta \hat{x})$.

Given $Y = F(K, L)$, log-linearization gives $\hat{Y} = \eta_{YK} \hat{K} + \eta_{YL} \hat{L}$ where η_{YX} is the elasticity of Y with respect to X:

$$\eta_{YX} \equiv \frac{F_X'(K_*, L_*) X_*}{F(K_*, L_*)}.$$

Note that $\dot{\hat{x}} \equiv \frac{d}{dt}(\log x - \log x_*) = \frac{d}{dt}(\log x) = \dot{x}/x$.

3.6 Uncertainty

Uncertainty: general setup (Nir 16-3–6) $s \equiv \{y_1, \ldots, y_n\}$ set of possible states of economy. s_t realization at time t. $s^t \equiv (s_0, \ldots, s_t)$ "history" at time t. $\pi(\cdot)$ probability function.

Often assume a Markov process: $\pi(s_{t+1}|s^t) = \pi(s_{t+1}|s_t)$; i.e., only the previous period matters. Process described by a transition matrix P where $P_{ij} = \pi(s_{t+1} = y_j | s_t = y_i)$ with $\sum_j P_{ij} = 1$ for all i. "Invariant distribution" is eigenvector $\pi_{1 \times n} = \pi P$.

Lucas tree (Nir 16-22–4) Assume von Neumann-Morgenstern utility function, thus objective function

$$\sum_{t=0}^{\infty} \sum_{z^t} \beta^t \pi(z^t) U(c_t(z^t)) = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U(c_t(z^t)).$$

Price (at time 0) of a claim that delivers one unit at time t given history z^t is $p_t(z^t)$. Budget constraint (Arrow-Debreu) is

$$\sum_{t=0}^{\infty} \sum_{z^t} p_t(z^t) c_t(z^t) \leq \sum_{t=0}^{\infty} \sum_{z^t} p_t(z^t) y_t(z^t)$$

where y_t is a (stochastic) endowment. Normalizing $p_0 = 1$ gives FOC

$$p_t(z^t) = \frac{\beta^t \pi(z^t) U'(y_t(z^t))}{\lambda} = \beta^t \pi(z^t) \frac{U'(y_t(z^t))}{U'(y_0)}.$$

Risk-free bond pricing (Nir 16-25–6) An asset purchased at time t that pays 1 consumption unit at time $t+1$. Price is:

$$q_t^{\text{RF}}(z^t) = \frac{\sum_{z_{t+1}} p_{t+1}(z_{t+1}, z^t)}{p_t(z^t)}$$
$$= \beta \sum_{z_{t+1}} \frac{\pi(z_{t+1}, z^t)}{\pi_t(z^t)} \frac{U'(y_{t+1}(z_{t+1}, z^t))}{U'(y_t(z^t))}$$
$$= \beta \frac{\mathbb{E}_t[U'(y_{t+1}(z_{t+1}, z^t))]}{U'(y_t(z^t))}.$$

Stock pricing (Nir 16-26–7) An asset purchased at time t that pays dividend $d_t(z^t)$. Price is:

$$q_t^{\text{tree}}(z^t) = \frac{\sum_{s=t+1}^{\infty} \sum_{z^s} p_s(z^s) d_s(z^s)}{p_t(z^t)}$$

$$= \mathrm{E}_t \sum_{s=t+1}^{\infty} \beta^{s-t} \frac{U'(y_s(z^s))}{U'(y_t(z^t))} d_s(z^s).$$

3.7 Manuel Amador

Lucas cost of busines cycles (Manuel) We find that the welfare impact of eliminating the business cycle is very small ($< 0.04\%$), especially compared to the impact of raising the growth rate of consumption. Complaints include the facts that:

1. A representative agent is bad for measuring utility and estimating the variance in consumption.

2. The model treats shocks as transitory; treating consumption as persistent results in significantly higher "cost" estimates.

3. Uncertainty could also affect the consumption *growth rate*, which requires an endogenous growth model (not the NCGM).

Incomplete markets: finite horizon (Manuel) Suppose a two period model with uncertain income in the second period, and incomplete markets in which only a risk-free bond is available at return R. The consumer chooses savings a to maximize $u(y_0 - a) + \beta \mathrm{E} u(y_1 + Ra)$, yielding FOC (intereuler)

$$u'(y_0 - a^*) = \beta R \mathrm{E} u'(y_1 + Ra^*).$$

If instead there were no uncertainty in the second period, the analogous FOC would be

$$u'(y_0 - \hat{a}) = \beta R u'(\mathrm{E} y_1 + R\hat{a}).$$

By Jensen's inequality, optimal savings is higher in the uncertain world ($a^* > \hat{a}$) as long as $u''' > 0$.

Thus in a two period model, the difference between savings in complete and incomplete markets depends on the sign of u'''.

Incomplete markets: infinite horizon (Manuel) Again, we suppose uncertain income, and the availability of only a risk-free bond. Taking FOC and envelope condition of

$$V(x) = \max_{a \in [\phi, x]} u(x - a) + \sum_{s \in S} \beta \pi(s) V[Ra + y(s)]$$

gives that $V'(x) \geq \beta R \mathrm{E}[V'(x')]$. Thus if $\beta R \geq 1$, we have $V'(x)$ is a nonnegative supermartingale and converges to a finite value by Dobb's Convergence Theorem. Thus x converges, but cannot converge to a finite value since

$x^{\infty} = Ra_*(x^{\infty}) + y$, the RHS of which is stochastic. So $x \to \infty$.

Thus in an infinite-horizon model, complete and incomplete markets with $\beta R = 1$ necessarily look different.

Hall's martingale hypothesis (Manuel) Suppose $u(c) = Ac - Bc^2$ (which may be a second-order Taylor expansion of the true utility function), and that only a risk-free bond is available, with $R\beta = 1$. Intereuler gives that $\mathrm{E} c_{t+1} = c_t$; that is, $\{c_t\}$ follows an AR(1) process.

However, many models have c today as a predictor of c tomorrow, even without the same intereuler (e.g., the Aiyagari model). Tests are always rejected, but are generally run with aggregate data, with resulting problems of

1. Goods aggregation (i.e., how do you define consumption given a basket of goods?),

2. Agent aggregation (i.e., even if intereuler holds for each of hetrogeneous agents, it may not hold in aggregate),

3. Time aggregation (i.e., how do you deal with data that is in discrete time period "chunks"?).

The presence of durable goods introduces an MA component to $\{c_t\}$, making it an ARMA process.

One-sided lack of commitment (Manuel) One reason markets might be incomplete is a lack of commitment. Suppose risk-averse consumuer can borrow from a risk-neutral lender, but can't commit to repayment. Suppose $R\beta = 1$; lender takes stochastic endowment and gives consumer consumption $c(s)$ today, promising consumer future lifetime value $w(s)$. Thus value to lender is:

$$P(v) = \max_{\{c(s), w(s)\}_s} \sum_{s \in S} \pi(s) \big[y(s) - c(s) + \beta P(w(s)) \big]$$

such that

$$[\text{PK } (\mu)]: \sum_{s \in S} \pi(s) \big[u(c(s)) + \beta w(s) \big] \geq v,$$

$$[\text{IC } (\lambda(s))]: u(c(s)) + \beta w(s) \geq u(y(s)) + \beta \underbrace{v_{\text{autarky}}}_{\mathrm{E} \frac{u(y(s'))}{1-\beta}}.$$

FOCs and envelope condition give that

$$\frac{1}{u'(c(s'))} = \frac{1}{u'(c(s))} + \lambda(s'),$$

so consumption is weakly increasing over time. This is not consistent with G.E., where we therefore must have $R\beta < 1$.

Two-sided lack of commitment (Manuel) [Kocherlakota] Suppose two risk-averse consumuers who each get a stochastic

endowment that has no aggregate shock. Neither can commit to remain in an insurance contract. Value to insurer in excess of autarky is:

$$Q(\Delta_0, s_0) =$$
$$\max_{c, \{\Delta(s)\}_s} u(1 - c) - u(1 - y(s_0)) + \beta \sum_{s \in S} \pi(s) Q(\Delta(s), s)$$

such that

$$[\text{PK } (\mu)]: u(c) - u(y(s_0)) + \beta \sum_{s \in S} \pi(s) \Delta(s) \geq \Delta_0,$$

$$[\text{IC-1 } (\beta\pi(s)\lambda(s))]: \Delta(s) \geq 0,$$
$$[\text{IC-2 } (\beta\pi(s)\theta(s))]: Q(\Delta(s), s) \geq 0.$$

FOCs and envelope condition give that

$$Q'(\Delta_0, s_0) = \lambda(s) + (1 + \theta(s)) Q'(\Delta(s), s),$$

thus either consumption stays the same, increases the the lowest level that satisfies IC-1, or falls to the highest level that satisfies IC-2.

Bulow-Rogoff defaultable debt model (Manuel) An agent (who need only have monotone preferences) cannot commit to repay a risk-neutral insurer; if he defaults, he will only be able to save in a complete-markets "Swiss bank account" that pays expected return R.

If wealth

$$W(s^t) \equiv \sum_{\tau \geq t} \sum_{s^\tau} \pi(s^\tau | s^t) \frac{y(s^\tau)}{R^{\tau - t}}$$

is finite for all s^t, and we impose no Ponzi condition (natural borrowing constraint) that debt

$$D(s^t) \equiv \sum_{\tau \geq t} \sum_{s^\tau} \pi(s^\tau | s^t) \frac{p(s^\tau)}{R^{\tau - t}} \leq W(s^t)$$

for all s^t, then debt will always be nonpositive—if the agent were ever a borrower, he would do better to defect.

Thus to get (e.g., international) debt, we need either

1. Infinite wealth in some state s^t,
2. A limited ability to save in nearly-complete markets, or
3. Lenders' ability to punish.

Hyperbolic discounting (Manuel) Agent has a higher discount rate between today and tomorrow than between tomorrow and following periods (time-inconsistent). Often parameterized as "β-δ discounting":

$$u(c_t) + \beta \sum_{\tau=1}^{\infty} \delta^\tau u(c_{t+\tau}).$$

Caution: Suppose an agent only likes candy bars when he's healthy. If he knows he's healthy today, he may prefer 1 today to 2 tomorrow, but 2 the day after tomorrow to 1 tomorrow. This may wind up *looking* like hyperbolic discounting, but isn't.

3.8   John Taylor

Taylor Rule (Taylor) A policy rule (or family of policy rules given different coefficients) for the nominal interest rate as a function of (four-quarter average) inflation and the GDP output gap (expressed as a percentage deviation from trend):

$$i = \pi + 0.5y + 0.5(\pi - 2) + 2$$
$$= \underbrace{1.5}_{>1}\,\pi + 0.5y + 1.$$

The "greater than one" principal suggests that the coefficient on inflation should exceed one.

Granger causality (Taylor) p_t Granger causes y_t iff the error from predicting y_t based on lagged y and p ($\sigma^2_{\text{prediction}}(y_t|p_{t-1}, p_{t-2}, \ldots, y_{t-1}, y_{t-2}, \ldots)$) is less than the prediction error from predicting y_t based on lagged y only ($\sigma^2_{\text{prediction}}(y_t|y_{t-1}, y_{t-2}, \ldots)$). Note this is not a "philosophical" point, and does not address causation vs. correlation issues.

That is, the hypothesis that p does *not* Granger cause y is the hypothesis that coefficients on lagged ps are all zero.

Okun's Law (Taylor)

$$\frac{Y - Y^*}{Y} = -2.5(u - u^*)$$

where Y^* is potential GDP and u^* is the natural rate of unemployment.

Cagan money demand (Taylor) $m_t - p_t = ay_t - bi_t$, where money supply m, price level p, and output gap y are all logs. LHS is real money balance; when interest rate is higher, demand for money is lower (semi-log specification).

Note Lucas prefers log-log specification $m_t - p_t = ay_t - b\log(i_t)$.

Rational expectations models (Taylor) Variables in system are a function of exogenous shocks, lags, *and* expected future values of variables. General solution method:

1. Set up in vector notation (may require time shifting and substitution).

2. Use method of undetermined coefficients to get a (deterministic) difference equation in the coefficients (γ_i) on the MA(∞) representation of variables.

3. Guess a form for the particular solution, and solve for coefficients.

4. Cholesky decompose the coefficient matrix ($A = H\Lambda H^{-1}$) in the homogenous part of the deterministic difference equation to get a decoupled equation (in $\mu_i \equiv H^{-1}\gamma_i^{\text{H}}$).

5. Apply stability condition to pin down some element(s) of μ_0.

6. Use $\gamma_0 = H\mu_0 + \mu_0^{\text{P}}$ as another equation to solve for solution.

Cagan model (Taylor) Rational expectations model given by:

$$m_t - p_t = -\beta(\hat{p}_{t+1} - p_t).$$

Derived from Cagan money demand equation with

1. Either zero coefficient on y_t, or $y_t = 0$ for all t, and

2. r_t ignored in $i_t = r_t + \hat{\pi}_{t+1} = r_t + \hat{p}_{t+1} - p_t$.

Price reactions to money supply shocks are not quite one-for-one, since there's an anticipated future deflation which causes people to hold some of the extra money.

Dornbush model (Taylor) Rational expectations model given by:

$$m_t - p_t = -\alpha(\hat{e}_{t+1} - e_t)$$
$$p_t - p_{t-1} = \beta(e_t - p_t).$$

(Note higher e is currency *de*preciation.)

Prices adjust gradually, but the exchange rate overshoots. By first equation, if money supply is shocked there must be expected future currency appreciation ($e\downarrow$); but there must be long run depreciation ($e\uparrow$) by second equation.

Philips Curve (Taylor) Connection between inflation and either GDP gap or unemployment (by Okun's Law).

1. Traditional (short-run):

$$\pi = \varphi_{\text{sr}}y = -2.5\varphi_{\text{sr}}(u - u^*).$$

2. Expectations-augmented (long-run):

$$\pi = \hat{\pi} + \varphi_{\text{lr}}y = \hat{\pi} - 2.5\varphi_{\text{lr}}(u - u^*).$$

Time inconsistency (Taylor) Three types of policy plans π_2 that aim to maximize social welfare $S(x_1, x_2, \pi_1, \pi_2)$ where π are policies and x are responses (including expectations of policy):

1. Consistent (discretion): Take x_1 as given when choosing π_2.

2. Optimal (rule-based): Recognize that x_1 responds to π_2.

3. Inconsistent (cheating): Promise optimal, but then name consistent.

Lucas supply function (Taylor) $y_t = \phi(\pi_t - \hat{\pi}_t) + \varepsilon_t$. Can be seen directly from Expectations-augmented Philips curve.

4   Mathematics

4.1   General mathematical concepts

Elasticity (Wikipedia) Elasticity of $f(x)$ is

$$\left|\frac{\partial \log f(x)}{\partial \log x}\right| = \left|\frac{\partial f}{\partial x} \cdot \frac{x}{f(x)}\right| = \left|\frac{\partial f/f}{\partial x/x}\right| = \left|\frac{xf'(x)}{f(x)}\right|.$$

Euler's law (Producer 14) If f is differentiable, it is homogeneous of degree k iff $p \cdot \nabla f(p) = kf(p)$.

One direction proved by differentiating $f(\lambda p) = \lambda^k f(p)$ with respect to λ, and then setting $\lambda = 1$. Implies that if f is homogeneous of degree one, then ∇f is homogeneous of degree zero.

Strong set order (Producer 32) $A \leq B$ in the strong set order iff for all $a \in A$ and $b \in B$ with $a \geq b$, then $a \in B$ and $b \in A$. Equivalently, every element in $A \setminus B$ is $\leq$ every element in $A \cap B$, which is $\leq$ every element in $B \setminus A$.

Meet (Producer 36) For x, $y \in \mathbb{R}^n$, the meet is $x \wedge y \equiv (\min\{x_1, y_1\}, \ldots, \min\{x_n, y_n\})$. More generally, on a partially ordered set, $x \wedge y$ is the greatest lower bound of x and y.

Join (Producer 36) For x, $y \in \mathbb{R}^n$, the meet is $x \vee y \equiv (\max\{x_1, y_1\}, \ldots, \max\{x_n, y_n\})$. More generally, on a partially ordered set, $x \vee y$ is the least upper bound of x and y.

Sublattice (Producer 37) A set X is a sublattice iff $\forall x, y \in X$, we have $x \wedge y \in X$ and $x \vee y \in X$. Any sublattice in $\mathbb{R}^n$ can be described as an intersection of sets of the forms

1. A product set $X_1 \times \cdots \times X_n$; or

2. A set $\{(x_1, \ldots, x_n) : x_i \leq g(x_j)\}$, where $g(\cdot)$ is an increasing function.

Orthants of Euclidean space (?)

1. $\mathbb{R}^n_+ \equiv \{\mathbf{x} : \mathbf{x} \geq \mathbf{0}\} \equiv \{\mathbf{x} : x_i \geq 0 \,\forall i\}$, which includes the axes and $\mathbf{0}$.

2. $\{\mathbf{x} : \mathbf{x} > \mathbf{0}\} \equiv \{\mathbf{x} : x_i \geq 0 \,\forall i\} \setminus \mathbf{0}$, which includes the axes, but not $\mathbf{0}$.

3. $\mathbb{R}^n_{++} \equiv \{\mathbf{x} : \mathbf{x} \gg \mathbf{0}\} \equiv \{\mathbf{x} : x_i > 0 \,\forall i\}$, which includes neither the axes nor $\mathbf{0}$.

Sum of powers (?)

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4}$$

$$\sum_{i=1}^{n} i^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

Geometric series (?)

$$\sum_{i=0}^{n} aR^i = \frac{a(1-R^{n+1})}{1-R}$$

$$\sum_{i=0}^{\infty} R^i = \frac{a}{1-R} \quad (\text{if } |R| < 1)$$

4.2 Set theory

Supremum limit (Math camp) $\limsup_n A_n \equiv \overline{\lim}_n A_n \equiv \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$. Set of all elements appearing in an infinite number of A_i.

Infimum limit (Math camp) $\liminf_n A_n \equiv \underline{\lim}_n A_n \equiv \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$. Set of all elements appearing in all but a finite number of A_i.

Partition (Math camp) $\{A_\alpha\}$ such that $A_\alpha \subseteq A$ and

1. Non-empty: $A_\alpha \neq \varnothing$;
2. Exhaustive: $\bigcup A_\alpha = A$;
3. Non-overlapping: $A_i \neq A_j \implies A_i \cap A_j = \varnothing$.

Convex set (Math camp) $\forall x$ and $x' \in S$, $\forall t \in [0,1]$, $tx + (1-t)x' \in S$ (i.e., the linear combination of any two points in S is also in S.)

4.3 Binary relations

Complete (Math camp) $aRb \vee bRa$.

Transitive (Math camp) $aRb \wedge bRc \implies aRc$.

Symmetric (Math camp) $aRb \iff bRa$.

Negative transitive (Math camp) $aRb \implies cRb \vee aRc$.

Reflexive (Math camp) aRa.

Irreflexive (Math camp) $a \neq b \implies aRb$.

Asymmetric (Math camp) $aRb \wedge b \neq c \implies cRa$.

Equivalence relation (Math camp) A binary relation that is reflexive, symmetric, and transitive.

4.4 Functions

Function (Math camp) A binary relation that is

1. *Well* defined on its range: $\forall x$, $\exists y$ such that xRy;
2. *Uniquely* defined on its range: $xRy \wedge xRz \implies y = z$.

Surjective (Math camp) Range is Y. (a.k.a. "onto.")

Injective (Math camp) $x \neq x' \implies f(x) \neq f(x')$.

Bijective (Math camp) Both surjective and injective. (a.k.a. "one-to-one.")

Monotone (C&B p. 50) $g(x)$ monotone on its domain iff either $u > v \implies g(u) > g(v)$ (increasing), or $u < v \implies g(u) < g(v)$ (decreasing). A monotone function is one-to-one and onto from its domain to its image (note, not necessarily to its range).

Convex function (Hansen 2-22) $g \colon \mathbb{R}^n \to \mathbb{R}$ is convex iff $\forall x, y \in \mathbb{R}^n$, $\forall \lambda \in [0,1]$, $\lambda g(x) + (1-\lambda)g(y) \geq g(\lambda x + (1-\lambda)y)$.

Binomial theorem (C&B 3.2.2) For any $x, y \in \mathbb{R}$ and $n \in \mathbb{Z}$, $n \geq 0$, then $(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$. Special cases: $1 = (p + (1-p))^n = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i}$ and $2^n = \sum_{i=0}^{n} \binom{n}{i}$.

Gamma function (C&B p. 99) $\Gamma(\alpha) \equiv \int_0^{\infty} t^{\alpha-1} e^{-t} \, dt = \int_0^1 [\log(1/t)]^{\alpha-1} \, dt$ on $\alpha > 0$. (Γ is also defined everywhere else except for $0, -1, -2, \dots$.)

1. $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$, when $\alpha > 0$.
2. $\Gamma(n) = (n-1)!$ for integer $n > 0$.
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
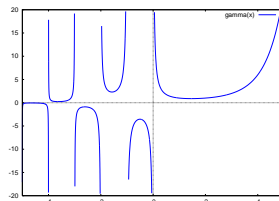
Beta function (C&B eq. 3.3.17) $B(\alpha, \beta) \equiv \Gamma(\alpha) \cdot \Gamma(\beta) / \Gamma(\alpha + \beta)$.

Logistic function (D&M 515; Metrics P.S. 5-4a)

$$\Lambda(x) \equiv (1 + e^{-x})^{-1} \equiv \frac{e^x}{1 + e^x}.$$

Inverse is $\Lambda^{-1}(\mu) \equiv \ell(\mu) \equiv \log[\mu/(1-\mu)]$. First derivative

$$\lambda(x) \equiv \Lambda'(x) \equiv \frac{e^x}{(1+e^x)^2} = \Lambda(x)(1 - \Lambda(x)) = \Lambda(x)\Lambda-x.$$

Taylor series (C&B 5.5.20–1) If $g(x)$ has derivatives of order r (i.e., $g^{(r)}(x)$ exists), then for any constant a the Taylor polynomial of order r about a is:

$$g(x) \approx T_r(x) \equiv \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!} (x - a)^i.$$

The remainder from the approximation Taylor approximation, $g(x) - T_r(x)$ equals $\int_a^x g^{(r+1)}(t) \cdot (x-t)^r / r! \, dt$. This error always tends to 0 faster than the highest-order explicit term—i.e., $\lim_{x \to a} [g(x) - T_r(x)]/(x-a)^r = 0$.

Derivative of power series (Hansen 2-29) If $g(t) \equiv a_0 + a_1 t + a_2 t^2 + \cdots = \sum_i a_i t^i$, then

$$\left. \frac{\partial^k}{\partial t^k} g(t) \right|_{t=0} = a_k k!.$$

This is useful for calculating moments if we can express the mgf as a power series.

Taylor series examples (?) $f(x) \approx f(x') + \nabla f(x') \cdot (x - x')$. Note that for concave (convex) $f(\cdot)$, the LHS is weakly less (greater) than the RHS.

For small ε, we have:

$$e^{n\varepsilon} \approx 1 + n\varepsilon;$$
$$(1 + \alpha\varepsilon)^n \approx 1 + \alpha n \varepsilon;$$
$$\log(1 + \varepsilon) \approx \varepsilon;$$
$$\frac{a^\varepsilon - 1}{\varepsilon} \approx \log a.$$

Second-order is

$$f(x) \approx f(x') + \nabla f(x') \cdot (x - x') + \tfrac{1}{2}(x-x') \cdot \nabla^2 f(x')(x - x').$$

Mean value expansion (Mahajan 3-13; Hayashi 470–1) If $h \colon \mathbb{R}^p \to \mathbb{R}^q$ is a continuously differentiable function, then $h(\cdot)$ admits a mean value expansion about θ

$$h(b) = h(\theta) + \frac{\partial h(\bar{b})}{\partial b}(b - \theta)$$

where $\bar{b}$ is a value lying between b and θ (note the MVT applies to individual elements of $h(\cdot)$, so $\bar{b}$ actually differs from element to element in the vector equation).

4.5 Correspondences

Correspondence (Micro math 5) $\phi\colon X \rightrightarrows Y$ is a mapping from elements of X to subsets of Y (i.e., for $x \in X$, we have $\phi(x) \subseteq Y$).

Lower semi-continuity (Micro math 7; Clayton G.E. I) $\phi\colon X \rightrightarrows Y$ is lower semi-/hemi-continuous at $x \in X$ iff for every open set $G \subseteq Y$ containing $\phi(x)$, there is an open set $U(x) \subseteq X$ containing x such that if $x' \in U(x)$, then $\phi(x') \cap G \neq \varnothing$.

Intuitively, any element in $\phi(x)$ can be "approached" from all directions.

For a single-valued correspondence (i.e., a function), lsc $\iff$ usc $\iff$ continuous.

Upper semi-continuity (Micro math; Clayton G.E. I; Bernheim 32) $\phi\colon X \rightrightarrows Y$ is upper semi-/hemi-continuous at $x \in X$ iff for every open set $G \subseteq Y$ containing $\phi(x)$, there is an open set $U(x) \subseteq X$ containing x such that if $x' \in U(x)$, then $\phi(x') \subset G$.

Identically, if for any sequences $x_t \to x$ and $y_t \to y$ with $y_t \in \phi(x_t)$ for all t, we have $y \in \phi(x)$.

Intuitively, $\phi(x)$ does not "suddenly contain new points;" i.e., the graph of the function is a closed set.

For a single-valued correspondence (i.e., a function), usc $\iff$ lsc $\iff$ continuous.

Brouwer's Fixed Point Theorem (Clayton G.E. I) If a function $f\colon A \to A$ is continuous, and A is compact and convex, then $f(x) = x$ for some fixed point $x \in A$.

Kakutani's Fixed Point Theorem (Clayton G.E. I; Bernheim 32) Suppose S is compact and convex. $\gamma\colon S \rightrightarrows S$ is convex-valued and upper semi-continuous. Then there exists a fixed point $s^* \in S$ such that $s^* \in \gamma(s^*)$.

4.6 Linear algebra

Matrix (Amemiya Ch. 11) Matrix $\mathbf{A}_{n \times m} \equiv \{a_{ij}\}$ has n rows and m columns.

1. Square iff $n = m$.
2. Transpose $\mathbf{A}' \equiv \{a_{ji}\}$ is $m \times n$.
3. Symmetric iff square and $\forall i,j$, $a_{ij} = a_{ji}$, or equivalently iff $\mathbf{A} = \mathbf{A}'$.
4. Diagonal iff $\forall i \neq j$, $a_{ij} = 0$ (i.e., nondiagonal elements are 0).

Matrix multiplication (Amemiya Ch. 11) If c is a scalar, $c\mathbf{A} = \mathbf{A}c = \{ca_{ij}\}$. If $\mathbf{A}$ is $n \times m$ and $\mathbf{B}$ is $m \times r$, then $\mathbf{C} = \mathbf{AB}$ is an $n \times r$ matrix with $c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$. If $\mathbf{AB}$ is defined, then $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

Determinant (Amemiya Ch. 11; Greene 23–6) $|\mathbf{A}| \equiv \sum_i (-1)^{i+j} a_{ij} |\mathbf{A}_{ij}|$, where j is an arbitrarily chosen integer and $\mathbf{A}_{ij}$ is the matrix that deletes the ith row and jth column from $\mathbf{A}$.

This is the volume of the parallelotope with edges along the columns of $\mathbf{A}$.

1. Determinant of a 2×2 matrix $\mathbf{A}$ is $a_{11}a_{22} - a_{12}a_{21}$.
2. Determinant of a 3×3 matrix $\mathbf{A}$ is $a_{11}a_{22}a_{33} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13}$.
3. $|\mathbf{A}| = |\mathbf{A}'|$.
4. For diagonal $\mathbf{A}$, $|\mathbf{A}| = \prod_i a_{ii}$; thus $|\mathbf{I}| = 1$.
5. $|c\mathbf{A}_{n \times n}| = c^n |\mathbf{A}|$.
6. If $\mathbf{A}$, $\mathbf{B}$ both $n \times n$, then $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$; together with $|\mathbf{I}| = 1$, this gives $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$.
7. If any column/row of $\mathbf{A}$ is all zeroes, $|\mathbf{A}| = 0$; switching two adjacent columns/rows changes the sign of the determinant; if two c/rs are identical, the determinant is zero.
8. For any eigenvalue λ, $|\mathbf{A} - \lambda\mathbf{I}| = 0$.
9. $|\mathbf{A}_{n \times n}| = \prod_i \lambda_i$ (i.e., determinant is the product of eigenvalues).

Matrix inverse (Amemiya Ch. 11; Greene 30–1) $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ for any matrix $\mathbf{A}$ such that $|\mathbf{A}| \neq 0$.

1. If $\mathbf{A}$, $\mathbf{B}$ both $n \times n$, and $|\mathbf{A}| \neq 0$, $|\mathbf{B}| \neq 0$, then $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
2. If $\mathbf{A}^{-1}$ exists, then $(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}$.
3. The inverse of a 2×2 matrix is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{|\mathbf{A}|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

4. For diagonal $\mathbf{A}$, the inverse is diagonal with elements a_{ii}^{-1}.

Orthogonal matrix (Amemiya Ch. 11) A square matrix $\mathbf{A}_{n \times n}$ is orthogonal iff $\mathbf{A}' = \mathbf{A}^{-1}$. This corresponds to having columns orthogonal and normalized.

Trace (Amemiya Ch. 11; Metrics section; Greene 41–2) $\text{tr}(\mathbf{A}_{n \times n}) \equiv \sum_i a_{ii}$ (i.e., the sum of the diagonal elements).

1. $\text{tr}(c\mathbf{A}) = c\,\text{tr}(\mathbf{A})$.
2. $\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A})$.
3. $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$.
4. $\text{tr}(\mathbf{A}_{n \times m} \mathbf{B}_{m \times n}) = \text{tr}(\mathbf{BA})$.
5. Trace is the sum of eigenvalues: $\text{tr}(\mathbf{A}_{n \times n}) = \sum_i \lambda_i$.

6. $\text{tr}(\mathbf{A}_{n \times m} \mathbf{A}') = \text{tr}(\mathbf{A}'\mathbf{A}) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2 = \sum_{j=1}^{m} a_j' a_j = \sum_{j=1}^{m} \text{tr}(a_j a_j')$, where a_i are the columns of $\mathbf{A}$.

Caution: in general, $\text{tr}(\mathbf{AB}) \neq \text{tr}(\mathbf{A}) \cdot \text{tr}(\mathbf{B})$.

Linear independence, rank (Amemiya Ch. 11; Greene 20–3, 39) Vectors $\{\mathbf{x}_i\}$ linearly independent iff $\sum_i c_i \mathbf{x}_i = 0 \implies \forall i$, $c_i = 0$.

1. The following are equivalent and called nonsingularity of square $\mathbf{A}$:
   - $|\mathbf{A}| \neq 0$, $\mathbf{A}^{-1}$ exists;
   - $\mathbf{A}$ is column independent or row independent;
   - $\forall \mathbf{y}$, $\exists \mathbf{x}$, $\mathbf{A}\mathbf{x} = \mathbf{y}$;
   - $\mathbf{A}\mathbf{x} = 0 \implies \mathbf{x} = 0$.
2. Column rank (maximum number of linearly independent columns) equals row rank, which implies $\text{rank}\,\mathbf{A} = \text{rank}\,\mathbf{A}'$.
3. $\mathbf{A}_{n \times m}$ is full rank iff rank equals $\min(n, m)$; a square matrix is full rank iff it is nonsingular.
4. $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}'\mathbf{A}) = \text{rank}(\mathbf{AA}')$; in particular, $\mathbf{A}_{n \times m}$ with $m \leq n$ (i.e., "tall") is full rank iff $\mathbf{A}'\mathbf{A}$ is nonsingular.
5. $\mathbf{A}$ is nonsingular $\implies \text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{B})$.
6. $\text{rank}(\mathbf{A}_{n \times n})$ equals the number of nonzero eigenvalues (not necessarily distinct).

Eigen decomposition (Mathworld) For any square $\mathbf{A}$ with $\mathbf{\Lambda}$ a diagonal matrix containing eigenvalues of $\mathbf{A}$, and $\mathbf{P}$ a matrix whose columns are the eigenvectors of $\mathbf{A}$ (ordered as in $\mathbf{\Lambda}$), then $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$.

Orthogonal decomposition (Amemiya Ch. 11; Greene 37–8) a.k.a. spectral decomposition. For any symmetric $\mathbf{A}$, $\exists \mathbf{H}_{n \times n}$, $\mathbf{\Lambda}_{n \times n}$ such that $\mathbf{A} = \mathbf{H}\mathbf{\Lambda}\mathbf{H}'$, where $\mathbf{H}'\mathbf{H} = \mathbf{I}$ (i.e., $\mathbf{H}$ is orthogonal) and $\mathbf{\Lambda}$ is diagonal. Note that $\mathbf{H}'\mathbf{A}\mathbf{H} = \mathbf{\Lambda}$.

This is just the eigen decomposition for a symmetric matrix; here the matrix of eigenvectors is orthogonal.

Matrix squareroot (Greene 42) Given a positive definite symmetric matrix $\mathbf{A} = \mathbf{H}\mathbf{\Lambda}\mathbf{H}'$, we can get a (symmetric!) matrix squareroot $\mathbf{A}^{1/2} \equiv \mathbf{H}\mathbf{\Lambda}^{1/2}\mathbf{H}'$, where $\mathbf{\Lambda}^{1/2}$ is the diagonal matrix that takes the squareroot of diagonal elements of $\mathbf{\Lambda}$.

Eigenvector, eigenvalue (Amemiya Ch. 11; Woodford 670–2) Note some results may only hold for symmetric matrices (not made explicit in Amemiya). Let the orthogonal decomposition of symmetric $\mathbf{A}$ be $\mathbf{A} = \mathbf{H}\mathbf{\Lambda}\mathbf{H}'$, where $\mathbf{H}$ is orthogonal and $\mathbf{\Lambda}$ is diagonal.

1. Diagonal elements of $\mathbf{\Lambda} \equiv \mathbf{D}(\lambda_i)$ are eigenvalues (a.k.a. characteristic roots) of $\mathbf{A}$; columns of $\mathbf{H}$ are the corresponding eigenvectors (a.k.a. characteristic vectors).

2. If $\mathbf{h}$ is the eigenvector corresponding to λ, then $\mathbf{Ah} = \lambda\mathbf{h}$ and $|\mathbf{A} - \lambda\mathbf{I}| = 0$. (This is an alternate definition for eigenvectors/eigenvalues.)

3. Matrix operations $f(\mathbf{A})$ can be reduced to the corresponding scalar operation: $f(\mathbf{A}) = \mathbf{HD}[f(\lambda_i)]\mathbf{H}'$.

4. $\text{rank}(\mathbf{A}_{n\times n})$ equals the number of nonzero eigenvalues (not necessarily distinct).

5. $\mathbf{A}_{n\times m}\mathbf{B}_{m\times n}$ and $\mathbf{BA}$ have the same nonzero eigenvalues.

6. Symmetric $\mathbf{A}_{n\times n}$ and $\mathbf{B}_{n\times n}$ can be diagonalized by the same orthogonal matrix $\mathbf{H}$ iff $\mathbf{AB} = \mathbf{BA}$.

7. $|\mathbf{A}_{n\times n}| = \prod_i \lambda_i$ (i.e., determinant is the product of eigenvalues).

8. $\text{tr}(\mathbf{A}_{n\times n}) = \sum_i \lambda_i$ (i.e., trace is the sum of eigenvalues).

9. $\mathbf{A}$ and $\mathbf{A}^{-1}$ have same eigenvectors and inverse eigenvalues.

A 2×2 matrix $\mathbf{A}$ has two explosive eigenvalues (outside the unit circle) iff either

- Case I:

$$|\mathbf{A}| + \text{tr}\,\mathbf{A} \geq -1,$$
$$|\mathbf{A}| - \text{tr}\,\mathbf{A} \geq -1,$$
$$|\mathbf{A}| \qquad \leq \quad 1.$$

- Case II:

$$|\mathbf{A}| + \text{tr}\,\mathbf{A} \leq -1,$$
$$|\mathbf{A}| - \text{tr}\,\mathbf{A} \leq -1,$$
$$|\mathbf{A}| \qquad \leq \quad 1 \quad \text{(trivially)}.$$

Positive definiteness, &c. (Amemiya Ch. 11; Greene 46–9) Symmetric $\mathbf{A}$ is positive definite iff $\forall \mathbf{x} \neq 0$, $\mathbf{x}'\mathbf{Ax} > 0$. Also written $\mathbf{A} > \mathbf{0}$; if $\mathbf{A} - \mathbf{B}$ is positive definite, write $\mathbf{A} > \mathbf{B}$. If equality isn't strict, $\mathbf{A}$ is positive semidefinite (a.k.a. nonnegative definite). Negative definite and negative semidefinite (a.k.a. nonpositive definite) similarly defined.

1. A symmetric matrix is positive definite iff its eigenvalues are all positive; similar for other (semi-)definiteness.

2. If $\mathbf{A} > \mathbf{0}$, all diagonal elements are > 0 (consider a quadratic form with a unit vector).

3. $\mathbf{B}'\mathbf{B} \geq \mathbf{0}$; if $\mathbf{B}$ has full column rank, then $\mathbf{B}'\mathbf{B} > \mathbf{0}$.

4. $\mathbf{A} > \mathbf{0} \implies \mathbf{A}^{-1} > \mathbf{0}$ (i.e., the inverse of a positive definite matrix is positive definite).

5. $\forall \mathbf{B}_{n\times k}$ "tall" (i.e., $n \geq k$), $\mathbf{A}_{n\times n} \geq \mathbf{0} \implies \mathbf{B}'\mathbf{AB} \geq \mathbf{0}$. If $\mathbf{B}$ is full rank (i.e., $\text{rank}(\mathbf{B}) = k$), then $\mathbf{A}_{n\times n} > \mathbf{0} \implies \mathbf{B}'\mathbf{AB} > \mathbf{0}$.

6. For $\mathbf{A}_{n\times n}$ and $\mathbf{B}_{n\times n}$ both positive definite, $\mathbf{A} \geq \mathbf{B} \iff \mathbf{B}^{-1} \geq \mathbf{A}^{-1}$, and $\mathbf{A} > \mathbf{B} \iff \mathbf{B}^{-1} > \mathbf{A}^{-1}$.

7. $\mathbf{A} \geq \mathbf{0} \implies |\mathbf{A}| \geq 0$ and $\mathbf{A} > \mathbf{0} \implies |\mathbf{A}| > 0$—note the natural analogues for negative (semi)definite matrices do *not* hold.

Orthogonal completion (Hansen 5-24) For a "wide" matrix $\mathbf{A}_{n\times k}$ (i.e., $n < k$) with full row rank (i.e., all rows are linearly independent), we can construct a (non-unique) $(k-n)\times k$ matrix $\mathbf{A}_\perp$ such that:

1. The rows of $\mathbf{A}$ and $\mathbf{A}_\perp$ are linearly independent (i.e., $|(\mathbf{A}', \mathbf{A}'_\perp)| \neq 0$;

2. The rows of $\mathbf{A}_\perp$ are orthogonal to the rows of $\mathbf{A}$ (i.e., $\mathbf{A}_\perp\mathbf{A}' = \mathbf{0}_{k-n\times n}$, or identically, $\mathbf{A}\mathbf{A}'_\perp = \mathbf{0}_{n\times k-n}$).

For a "tall" matrix $\mathbf{B}_{n\times k}$ (i.e., $n > k$) with full column rank (i.e., all columns are linearly independent), we can construct a (non-unique) $n \times (n-k)$ matrix $\mathbf{B}_\perp$ such that:

1. The columns of $\mathbf{B}$ and $\mathbf{B}_\perp$ are linearly independent (i.e., $|(\mathbf{B}, \mathbf{B}_\perp)| \neq 0$;

2. The columns of $\mathbf{B}_\perp$ are orthogonal to the columns of $\mathbf{B}$ (i.e., $\mathbf{B}'_\perp\mathbf{B} = \mathbf{0}_{n-k\times k}$, or identically, $\mathbf{B}'\mathbf{B}_\perp = \mathbf{0}_{k\times n-k}$).

Idempotent matrix (Hansen 5-33–4; Amemiya 37; Hayashi 30–1, 36–7, 244-5) $\mathbf{P}$ is idempotent iff $\mathbf{PP} = \mathbf{P}$.

1. If $\mathbf{P}$ is idempotent, then so is $\mathbf{I} - \mathbf{P}$.

2. Every eigenvalue of a (symmetric?) idempotent matrix is either 0 or 1.

3. A symmetric and idempotent matrix is positive semidefinite.

4. For any $\mathbf{X}_{n\times k}$, we can find an idempotent projection matrix (that projects onto the subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}$): $\mathbf{P_X} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

5. For any $\mathbf{X}_{n\times k}$ tall (i.e., $n > k$), we can find an idempotent projection matrix (that projects onto the subspace of $\mathbb{R}^n$ *not* spanned by columns of $\mathbf{X}$): $\mathbf{I} - \mathbf{P_X} = \mathbf{P_{X_\perp}}$.

6. $\text{rank}(\mathbf{P}) = \text{tr}(\mathbf{P})$ if $\mathbf{P}$ is idempotent.

Matrix derivatives (Ying handout; Greene 51–53; MaCurdy) [Note derivative convention may be transpose of MWG.]

1. $\frac{\partial}{\partial\mathbf{x}}\mathbf{Ax} = \mathbf{A}'$.

2. $\frac{\partial}{\partial\mathbf{x}}\mathbf{x}'\mathbf{Ax} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$.

3. $\frac{\partial}{\partial\mathbf{A}}\mathbf{x}'\mathbf{Ax} = \mathbf{xx}'$.

4. $\frac{\partial}{\partial\mathbf{A}}\log|\mathbf{A}| = \mathbf{A}'^{-1}$.

5. $\frac{\partial}{\partial\mathbf{A}^{-1}}\log|\mathbf{A}| = -\mathbf{A}'$.

6. $\frac{\partial}{\partial w}\log|\mathbf{A}| = |\mathbf{A}|^{-1} \cdot \frac{\partial}{\partial w}|\mathbf{A}| = \text{tr}[\mathbf{A}^{-1} \cdot \frac{\partial}{\partial w}\mathbf{A}]$ (may require symmetric $\mathbf{A}$?).

7. $\frac{\partial}{\partial w}\mathbf{A}^{-1} = -\mathbf{A}^{-1}[\frac{\partial}{\partial w}\mathbf{A}]\mathbf{A}^{-1}$ (may require symmetric $\mathbf{A}$?).

Partitioned matrices (Hayashi 670–673; Greene 32–3)

$$\left|\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}\right| = |\mathbf{A}_{22}| \cdot |\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}|$$
$$= |\mathbf{A}_{11}| \cdot |\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|;$$

$$\begin{bmatrix} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{M1} & \cdots & \mathbf{A}_{MN} \end{bmatrix}\begin{bmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_N \end{bmatrix} = \begin{bmatrix} \sum_i \mathbf{A}_{1i}\mathbf{c}_i \\ \vdots \\ \sum_i \mathbf{A}_{Mi}\mathbf{c}_i \end{bmatrix};$$

$$\begin{bmatrix} \mathbf{A}_{11} & & \\ & \ddots & \\ & & \mathbf{A}_{MM} \end{bmatrix}\begin{bmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_N \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{c}_1 \\ \vdots \\ \mathbf{A}_{MM}\mathbf{c}_M \end{bmatrix};$$

$$\mathbf{A}_{\text{diag}}\mathbf{B}_{\text{diag}} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} & & \\ & \ddots & \\ & & \mathbf{A}_{MM}\mathbf{B}_{MM} \end{bmatrix};$$

$$\mathbf{A}'_{\text{diag}}\mathbf{B}\mathbf{A}_{\text{diag}}$$
$$= \begin{bmatrix} \mathbf{A}'_{11}\mathbf{B}_{11}\mathbf{A}_{11} & \cdots & \mathbf{A}'_{11}\mathbf{B}_{1M}\mathbf{A}_{MM} \\ \vdots & \ddots & \vdots \\ \mathbf{A}'_{MM}\mathbf{B}_{M1}\mathbf{A}_{11} & \cdots & \mathbf{A}'_{MM}\mathbf{B}_{MM}\mathbf{A}_{MM} \end{bmatrix};$$

$$\mathbf{A}'_{\text{diag}}\mathbf{Bc}$$
$$= \begin{bmatrix} \mathbf{A}'_{11}\mathbf{B}_{11}\mathbf{c}_1 + \ldots + \mathbf{A}'_{11}\mathbf{B}_{1M}\mathbf{c}_M \\ \vdots \\ \mathbf{A}'_{MM}\mathbf{B}_{M1}\mathbf{c}_1 + \ldots + \mathbf{A}'_{MM}\mathbf{B}_{MM}\mathbf{c}_M \end{bmatrix};$$

$$\mathbf{A}_{\text{diag}}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & & \\ & \ddots & \\ & & \mathbf{A}_{MM}^{-1} \end{bmatrix}.$$

The inverse of a 2×2 partitioned matrix is:

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1}$$
$$= \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_2\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_2 \\ -\mathbf{F}_2\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{F}_2 \end{bmatrix},$$

where $\mathbf{F}_2 \equiv (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}$. The upper left block can also be written as $\mathbf{F}_1 \equiv (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}$

Kronecker product (Hayashi 673; Greene 34–5) For $\mathbf{A}_{M\times N}$ and $\mathbf{B}_{K\times L}$,

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1N}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{M1}\mathbf{B} & \cdots & a_{MN}\mathbf{B} \end{bmatrix}$$

is $MK \times NL$. Operation is *not* commutative.

1. $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$ (assuming matrix multiplications are conformable);

2. $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$;

3. $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$;

4. $\text{tr}(\mathbf{A}_{M \times M} \otimes \mathbf{B}_{K \times K}) = \text{tr}(\mathbf{A}) \cdot \text{tr}(\mathbf{B})$;

5. $|\mathbf{A}_{M \times M} \otimes \mathbf{B}_{K \times K}| = |\mathbf{A}|^M \cdot |\mathbf{B}|^K$.

Deviation from means (Hansen 5-34–5; Greene 14–5) $\mathbf{M}_\iota \mathbf{x}$ is the vector of $x_i - \bar{\mathbf{x}}$, and $\mathbf{x}'\mathbf{M}_\iota \mathbf{x} = \sum_i (x_i - \bar{\mathbf{x}})^2$, where

$$\mathbf{M}_\iota \equiv \mathbf{I} - \iota(\iota'\iota)^{-1}\iota' = \mathbf{I} - \frac{1}{n}\iota\iota'$$

is the symmetric idempotent matrix with $1 - \frac{1}{n}$ on diagonal, and $-\frac{1}{n}$ off diagonal.*

---

5   References

Amemiya   Amemiya: Introduction to Econometrics and Statistics
Bernheim   Bernheim: Economics 203 notes
C&B   Casella, Berger: Statistical Inference, 2 ed.
Choice   Levin, Milgrom, Segal: Introduction to Choice Theory notes
Clayton   Featherstone: Economics 202 section notes
Consumer   Levin, Milgrom, Segal: Consumer Theory notes
D&M   Davidson, MacKinnon: Estimation and Inference in Econometrics
G.E.   Levin: General Equilibrium notes
Greene   Greene: Econometric Analysis, 3 ed.
Hansen   Hansen: Economics 270/271 notes
Hayashi   Hayashi: Econometrics
Jackson   Jackson: Economics 204 lectures

MaCurdy   MaCurdy: Economics 272 notes
Mahajan   Mahajan: Economics 270/271 notes
Manuel   Amador: Economics 211 lectures
Math camp   Fan: Math camp lectures
MathWorld   `mathworld.wolfram.com`
Max notes   Floetotto: Economics 210 section
Micro Math   Levin, Rangel: Useful Math for Microeconomists notes
MWG   Mas-Collell, Whinston, Green: Microeconomic Theory
Nir   Jaimovich: Economics 210 notes
Producer   Levin, Milgrom, Segal: Producer Theory notes
Taylor   Taylor: Economics 212 notes
Uncertainty   Levin: Choice Under Uncertainty notes
Wikipedia   `wikipedia.com`
Woodford   Woodford: Interest and Prices

---

*In Greene, $\mathbf{M}_\iota$ is called $\mathbf{M}^0$.

Index

Standard normal critical values

(two-tailed: e.g., 5% of observations from N(0,1) will fall outside ±1.96)
(one-tailed: e.g., 5% of observations from N(0,1) will fall above 1.64)

	50%	25%	10%	5%	2.5%	Significance 1%	0.5%	0.25%	0.1%	0.05%	0.025%	0.01%
two-tailed	0.67	1.15	1.64	1.96	2.24	2.58	2.81	3.02	3.29	3.48	3.66	3.89
one-tailed	(0.00)	0.67	1.28	1.64	1.96	2.33	2.58	2.81	3.09	3.29	3.48	3.72

Chi-squared critical values

(one-tailed: e.g., 5% of observations from $\chi^2(1)$ will exceed 3.84)

df	50%	25%	10%	5%	2.5%	Significance 1%	0.5%	0.25%	0.1%	0.05%	0.025%	0.01%
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88	9.14	10.83	12.12	13.41	15.14
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60	11.98	13.82	15.20	16.59	18.42
3	2.37	4.11	6.25	7.81	9.35	11.34	12.84	14.32	16.27	17.73	19.19	21.11
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86	16.42	18.47	20.00	21.52	23.51
5	4.35	6.63	9.24	11.07	12.83	15.09	16.75	18.39	20.52	22.11	23.68	25.74
6	5.35	7.84	10.64	12.59	14.45	16.81	18.55	20.25	22.46	24.10	25.73	27.86
7	6.35	9.04	12.02	14.07	16.01	18.48	20.28	22.04	24.32	26.02	27.69	29.88
8	7.34	10.22	13.36	15.51	17.53	20.09	21.95	23.77	26.12	27.87	29.59	31.83
9	8.34	11.39	14.68	16.92	19.02	21.67	23.59	25.46	27.88	29.67	31.43	33.72
10	9.34	12.55	15.99	18.31	20.48	23.21	25.19	27.11	29.59	31.42	33.22	35.56
11	10.34	13.70	17.28	19.68	21.92	24.72	26.76	28.73	31.26	33.14	34.98	37.37
12	11.34	14.85	18.55	21.03	23.34	26.22	28.30	30.32	32.91	34.82	36.70	39.13
13	12.34	15.98	19.81	22.36	24.74	27.69	29.82	31.88	34.53	36.48	38.39	40.87
14	13.34	17.12	21.06	23.68	26.12	29.14	31.32	33.43	36.12	38.11	40.06	42.58
15	14.34	18.25	22.31	25.00	27.49	30.58	32.80	34.95	37.70	39.72	41.70	44.26
16	15.34	19.37	23.54	26.30	28.85	32.00	34.27	36.46	39.25	41.31	43.32	45.92
17	16.34	20.49	24.77	27.59	30.19	33.41	35.72	37.95	40.79	42.88	44.92	47.57
18	17.34	21.60	25.99	28.87	31.53	34.81	37.16	39.42	42.31	44.43	46.51	49.19
19	18.34	22.72	27.20	30.14	32.85	36.19	38.58	40.88	43.82	45.97	48.08	50.80
20	19.34	23.83	28.41	31.41	34.17	37.57	40.00	42.34	45.31	47.50	49.63	52.39
21	20.34	24.93	29.62	32.67	35.48	38.93	41.40	43.78	46.80	49.01	51.17	53.96
22	21.34	26.04	30.81	33.92	36.78	40.29	42.80	45.20	48.27	50.51	52.70	55.52
23	22.34	27.14	32.01	35.17	38.08	41.64	44.18	46.62	49.73	52.00	54.22	57.07
24	23.34	28.24	33.20	36.42	39.36	42.98	45.56	48.03	51.18	53.48	55.72	58.61
25	24.34	29.34	34.38	37.65	40.65	44.31	46.93	49.44	52.62	54.95	57.22	60.14
26	25.34	30.43	35.56	38.89	41.92	45.64	48.29	50.83	54.05	56.41	58.70	61.66
27	26.34	31.53	36.74	40.11	43.19	46.96	49.64	52.22	55.48	57.86	60.18	63.16
28	27.34	32.62	37.92	41.34	44.46	48.28	50.99	53.59	56.89	59.30	61.64	64.66
29	28.34	33.71	39.09	42.56	45.72	49.59	52.34	54.97	58.30	60.73	63.10	66.15
30	29.34	34.80	40.26	43.77	46.98	50.89	53.67	56.33	59.70	62.16	64.56	67.63
31	30.34	35.89	41.42	44.99	48.23	52.19	55.00	57.69	61.10	63.58	66.00	69.11
32	31.34	36.97	42.58	46.19	49.48	53.49	56.33	59.05	62.49	65.00	67.44	70.57
33	32.34	38.06	43.75	47.40	50.73	54.78	57.65	60.39	63.87	66.40	68.87	72.03
34	33.34	39.14	44.90	48.60	51.97	56.06	58.96	61.74	65.25	67.80	70.29	73.48
35	34.34	40.22	46.06	49.80	53.20	57.34	60.27	63.08	66.62	69.20	71.71	74.93
36	35.34	41.30	47.21	51.00	54.44	58.62	61.58	64.41	67.99	70.59	73.12	76.36
37	36.34	42.38	48.36	52.19	55.67	59.89	62.88	65.74	69.35	71.97	74.52	77.80
38	37.34	43.46	49.51	53.38	56.90	61.16	64.18	67.06	70.70	73.35	75.92	79.22
39	38.34	44.54	50.66	54.57	58.12	62.43	65.48	68.38	72.05	74.73	77.32	80.65
40	39.34	45.62	51.81	55.76	59.34	63.69	66.77	69.70	73.40	76.09	78.71	82.06
41	40.34	46.69	52.95	56.94	60.56	64.95	68.05	71.01	74.74	77.46	80.09	83.47
42	41.34	47.77	54.09	58.12	61.78	66.21	69.34	72.32	76.08	78.82	81.47	84.88
43	42.34	48.84	55.23	59.30	62.99	67.46	70.62	73.62	77.42	80.18	82.85	86.28
44	43.34	49.91	56.37	60.48	64.20	68.71	71.89	74.93	78.75	81.53	84.22	87.68
45	44.34	50.98	57.51	61.66	65.41	69.96	73.17	76.22	80.08	82.88	85.59	89.07
46	45.34	52.06	58.64	62.83	66.62	71.20	74.44	77.52	81.40	84.22	86.95	90.46
47	46.34	53.13	59.77	64.00	67.82	72.44	75.70	78.81	82.72	85.56	88.31	91.84
48	47.34	54.20	60.91	65.17	69.02	73.68	76.97	80.10	84.04	86.90	89.67	93.22
49	48.33	55.27	62.04	66.34	70.22	74.92	78.23	81.38	85.35	88.23	91.02	94.60
50	49.33	56.33	63.17	67.50	71.42	76.15	79.49	82.66	86.66	89.56	92.37	95.97

Standard normal critical values

% of obs. from N(0,1)…			% of obs. from N(0,1)…			% of obs. from N(0,1)…			% of obs. from N(0,1)…		
x	above x	outside ±x	x	above x	outside ±x	x	above x	outside ±x	x	above x	outside ±x
0.00	50.0%	100.0%	1.00	15.9%	31.7%	2.00	2.3%	4.6%	3.00	0.13%	0.27%
0.02	49.2%	98.4%	1.02	15.4%	30.8%	2.02	2.2%	4.3%	3.02	0.13%	0.25%
0.04	48.4%	96.8%	1.04	14.9%	29.8%	2.04	2.1%	4.1%	3.04	0.12%	0.24%
0.06	47.6%	95.2%	1.06	14.5%	28.9%	2.06	2.0%	3.9%	3.06	0.11%	0.22%
0.08	46.8%	93.6%	1.08	14.0%	28.0%	2.08	1.9%	3.8%	3.08	0.10%	0.21%
0.10	46.0%	92.0%	1.10	13.6%	27.1%	2.10	1.8%	3.6%	3.10	0.097%	0.19%
0.12	45.2%	90.4%	1.12	13.1%	26.3%	2.12	1.7%	3.4%	3.12	0.090%	0.18%
0.14	44.4%	88.9%	1.14	12.7%	25.4%	2.14	1.6%	3.2%	3.14	0.084%	0.17%
0.16	43.6%	87.3%	1.16	12.3%	24.6%	2.16	1.5%	3.1%	3.16	0.079%	0.16%
0.18	42.9%	85.7%	1.18	11.9%	23.8%	2.18	1.5%	2.9%	3.18	0.074%	0.15%
0.20	42.1%	84.1%	1.20	11.5%	23.0%	2.20	1.4%	2.8%	3.20	0.069%	0.14%
0.22	41.3%	82.6%	1.22	11.1%	22.2%	2.22	1.3%	2.6%	3.22	0.064%	0.13%
0.24	40.5%	81.0%	1.24	10.7%	21.5%	2.24	1.3%	2.5%	3.24	0.060%	0.12%
0.26	39.7%	79.5%	1.26	10.4%	20.8%	2.26	1.2%	2.4%	3.26	0.056%	0.11%
0.28	39.0%	77.9%	1.28	10.0%	20.1%	2.28	1.1%	2.3%	3.28	0.052%	0.10%
0.30	38.2%	76.4%	1.30	9.7%	19.4%	2.30	1.1%	2.1%	3.30	0.048%	0.097%
0.32	37.4%	74.9%	1.32	9.3%	18.7%	2.32	1.0%	2.0%	3.32	0.045%	0.090%
0.34	36.7%	73.4%	1.34	9.0%	18.0%	2.34	0.96%	1.9%	3.34	0.042%	0.084%
0.36	35.9%	71.9%	1.36	8.7%	17.4%	2.36	0.91%	1.8%	3.36	0.039%	0.078%
0.38	35.2%	70.4%	1.38	8.4%	16.8%	2.38	0.87%	1.7%	3.38	0.036%	0.072%
0.40	34.5%	68.9%	1.40	8.1%	16.2%	2.40	0.82%	1.6%	3.40	0.034%	0.067%
0.42	33.7%	67.4%	1.42	7.8%	15.6%	2.42	0.78%	1.6%	3.42	0.031%	0.063%
0.44	33.0%	66.0%	1.44	7.5%	15.0%	2.44	0.73%	1.5%	3.44	0.029%	0.058%
0.46	32.3%	64.6%	1.46	7.2%	14.4%	2.46	0.69%	1.4%	3.46	0.027%	0.054%
0.48	31.6%	63.1%	1.48	6.9%	13.9%	2.48	0.66%	1.3%	3.48	0.025%	0.050%
0.50	30.9%	61.7%	1.50	6.7%	13.4%	2.50	0.62%	1.2%	3.50	0.023%	0.047%
0.52	30.2%	60.3%	1.52	6.4%	12.9%	2.52	0.59%	1.2%	3.52	0.022%	0.043%
0.54	29.5%	58.9%	1.54	6.2%	12.4%	2.54	0.55%	1.1%	3.54	0.020%	0.040%
0.56	28.8%	57.5%	1.56	5.9%	11.9%	2.56	0.52%	1.0%	3.56	0.019%	0.037%
0.58	28.1%	56.2%	1.58	5.7%	11.4%	2.58	0.49%	0.99%	3.58	0.017%	0.034%
0.60	27.4%	54.9%	1.60	5.5%	11.0%	2.60	0.47%	0.93%	3.60	0.016%	0.032%
0.62	26.8%	53.5%	1.62	5.3%	10.5%	2.62	0.44%	0.88%	3.62	0.015%	0.029%
0.64	26.1%	52.2%	1.64	5.1%	10.1%	2.64	0.41%	0.83%	3.64	0.014%	0.027%
0.66	25.5%	50.9%	1.66	4.8%	9.7%	2.66	0.39%	0.78%	3.66	0.013%	0.025%
0.68	24.8%	49.7%	1.68	4.6%	9.3%	2.68	0.37%	0.74%	3.68	0.012%	0.023%
0.70	24.2%	48.4%	1.70	4.5%	8.9%	2.70	0.35%	0.69%	3.70	0.011%	0.022%
0.72	23.6%	47.2%	1.72	4.3%	8.5%	2.72	0.33%	0.65%	3.72	0.0100%	0.020%
0.74	23.0%	45.9%	1.74	4.1%	8.2%	2.74	0.31%	0.61%	3.74	0.0092%	0.018%
0.76	22.4%	44.7%	1.76	3.9%	7.8%	2.76	0.29%	0.58%	3.76	0.0085%	0.017%
0.78	21.8%	43.5%	1.78	3.8%	7.5%	2.78	0.27%	0.54%	3.78	0.0078%	0.016%
0.80	21.2%	42.4%	1.80	3.6%	7.2%	2.80	0.26%	0.51%	3.80	0.0072%	0.014%
0.82	20.6%	41.2%	1.82	3.4%	6.9%	2.82	0.24%	0.48%	3.82	0.0067%	0.013%
0.84	20.0%	40.1%	1.84	3.3%	6.6%	2.84	0.23%	0.45%	3.84	0.0062%	0.012%
0.86	19.5%	39.0%	1.86	3.1%	6.3%	2.86	0.21%	0.42%	3.86	0.0057%	0.011%
0.88	18.9%	37.9%	1.88	3.0%	6.0%	2.88	0.20%	0.40%	3.88	0.0052%	0.010%
0.90	18.4%	36.8%	1.90	2.9%	5.7%	2.90	0.19%	0.37%	3.90	0.0048%	0.0096%
0.92	17.9%	35.8%	1.92	2.7%	5.5%	2.92	0.18%	0.35%	3.92	0.0044%	0.0089%
0.94	17.4%	34.7%	1.94	2.6%	5.2%	2.94	0.16%	0.33%	3.94	0.0041%	0.0081%
0.96	16.9%	33.7%	1.96	2.5%	5.0%	2.96	0.15%	0.31%	3.96	0.0037%	0.0075%
0.98	16.4%	32.7%	1.98	2.4%	4.8%	2.98	0.14%	0.29%	3.98	0.0034%	0.0069%

Standard normal critical values
% of observations from χ2(1)…

x	above x	x	above x	x	above x	x	above x	x	above x	x	above x
0.000	100.0%	1.250	26.4%	2.500	11.4%	3.750	5.3%	5.000	2.5%	6.250	1.2%
0.025	87.4%	1.275	25.9%	2.525	11.2%	3.775	5.2%	5.025	2.5%	6.275	1.2%
0.050	82.3%	1.300	25.4%	2.550	11.0%	3.800	5.1%	5.050	2.5%	6.300	1.2%
0.075	78.4%	1.325	25.0%	2.575	10.9%	3.825	5.0%	5.075	2.4%	6.325	1.2%
0.100	75.2%	1.350	24.5%	2.600	10.7%	3.850	5.0%	5.100	2.4%	6.350	1.2%
0.125	72.4%	1.375	24.1%	2.625	10.5%	3.875	4.9%	5.125	2.4%	6.375	1.2%
0.150	69.9%	1.400	23.7%	2.650	10.4%	3.900	4.8%	5.150	2.3%	6.400	1.1%
0.175	67.6%	1.425	23.3%	2.675	10.2%	3.925	4.8%	5.175	2.3%	6.425	1.1%
0.200	65.5%	1.450	22.9%	2.700	10.0%	3.950	4.7%	5.200	2.3%	6.450	1.1%
0.225	63.5%	1.475	22.5%	2.725	9.9%	3.975	4.6%	5.225	2.2%	6.475	1.1%
0.250	61.7%	1.500	22.1%	2.750	9.7%	4.000	4.6%	5.250	2.2%	6.500	1.1%
0.275	60.0%	1.525	21.7%	2.775	9.6%	4.025	4.5%	5.275	2.2%	6.525	1.1%
0.300	58.4%	1.550	21.3%	2.800	9.4%	4.050	4.4%	5.300	2.1%	6.550	1.0%
0.325	56.9%	1.575	20.9%	2.825	9.3%	4.075	4.4%	5.325	2.1%	6.575	1.0%
0.350	55.4%	1.600	20.6%	2.850	9.1%	4.100	4.3%	5.350	2.1%	6.600	1.0%
0.375	54.0%	1.625	20.2%	2.875	9.0%	4.125	4.2%	5.375	2.0%	6.625	1.0%
0.400	52.7%	1.650	19.9%	2.900	8.9%	4.150	4.2%	5.400	2.0%	6.650	0.99%
0.425	51.4%	1.675	19.6%	2.925	8.7%	4.175	4.1%	5.425	2.0%	6.675	0.98%
0.450	50.2%	1.700	19.2%	2.950	8.6%	4.200	4.0%	5.450	2.0%	6.700	0.96%
0.475	49.1%	1.725	18.9%	2.975	8.5%	4.225	4.0%	5.475	1.9%	6.725	0.95%
0.500	48.0%	1.750	18.6%	3.000	8.3%	4.250	3.9%	5.500	1.9%	6.750	0.94%
0.525	46.9%	1.775	18.3%	3.025	8.2%	4.275	3.9%	5.525	1.9%	6.775	0.92%
0.550	45.8%	1.800	18.0%	3.050	8.1%	4.300	3.8%	5.550	1.8%	6.800	0.91%
0.575	44.8%	1.825	17.7%	3.075	8.0%	4.325	3.8%	5.575	1.8%	6.825	0.90%
0.600	43.9%	1.850	17.4%	3.100	7.8%	4.350	3.7%	5.600	1.8%	6.850	0.89%
0.625	42.9%	1.875	17.1%	3.125	7.7%	4.375	3.6%	5.625	1.8%	6.875	0.87%
0.650	42.0%	1.900	16.8%	3.150	7.6%	4.400	3.6%	5.650	1.7%	6.900	0.86%
0.675	41.1%	1.925	16.5%	3.175	7.5%	4.425	3.5%	5.675	1.7%	6.925	0.85%
0.700	40.3%	1.950	16.3%	3.200	7.4%	4.450	3.5%	5.700	1.7%	6.950	0.84%
0.725	39.5%	1.975	16.0%	3.225	7.3%	4.475	3.4%	5.725	1.7%	6.975	0.83%
0.750	38.6%	2.000	15.7%	3.250	7.1%	4.500	3.4%	5.750	1.6%	7.000	0.82%
0.775	37.9%	2.025	15.5%	3.275	7.0%	4.525	3.3%	5.775	1.6%	7.025	0.80%
0.800	37.1%	2.050	15.2%	3.300	6.9%	4.550	3.3%	5.800	1.6%	7.050	0.79%
0.825	36.4%	2.075	15.0%	3.325	6.8%	4.575	3.2%	5.825	1.6%	7.075	0.78%
0.850	35.7%	2.100	14.7%	3.350	6.7%	4.600	3.2%	5.850	1.6%	7.100	0.77%
0.875	35.0%	2.125	14.5%	3.375	6.6%	4.625	3.2%	5.875	1.5%	7.125	0.76%
0.900	34.3%	2.150	14.3%	3.400	6.5%	4.650	3.1%	5.900	1.5%	7.150	0.75%
0.925	33.6%	2.175	14.0%	3.425	6.4%	4.675	3.1%	5.925	1.5%	7.175	0.74%
0.950	33.0%	2.200	13.8%	3.450	6.3%	4.700	3.0%	5.950	1.5%	7.200	0.73%
0.975	32.3%	2.225	13.6%	3.475	6.2%	4.725	3.0%	5.975	1.5%	7.225	0.72%
1.000	31.7%	2.250	13.4%	3.500	6.1%	4.750	2.9%	6.000	1.4%	7.250	0.71%
1.025	31.1%	2.275	13.1%	3.525	6.0%	4.775	2.9%	6.025	1.4%	7.275	0.70%
1.050	30.6%	2.300	12.9%	3.550	6.0%	4.800	2.8%	6.050	1.4%	7.300	0.69%
1.075	30.0%	2.325	12.7%	3.575	5.9%	4.825	2.8%	6.075	1.4%	7.325	0.68%
1.100	29.4%	2.350	12.5%	3.600	5.8%	4.850	2.8%	6.100	1.4%	7.350	0.67%
1.125	28.9%	2.375	12.3%	3.625	5.7%	4.875	2.7%	6.125	1.3%	7.375	0.66%
1.150	28.4%	2.400	12.1%	3.650	5.6%	4.900	2.7%	6.150	1.3%	7.400	0.65%
1.175	27.8%	2.425	11.9%	3.675	5.5%	4.925	2.6%	6.175	1.3%	7.425	0.64%
1.200	27.3%	2.450	11.8%	3.700	5.4%	4.950	2.6%	6.200	1.3%	7.450	0.63%
1.225	26.8%	2.475	11.6%	3.725	5.4%	4.975	2.6%	6.225	1.3%	7.475	0.63%

Standard normal critical values
% of observations from χ2(2)…

x	above x	x	above x	x	above x	x	above x	x	above x	x	above x
0.00	100.0%	2.50	28.7%	5.00	8.2%	7.50	2.4%	10.00	0.67%	12.50	0.19%
0.05	97.5%	2.55	27.9%	5.05	8.0%	7.55	2.3%	10.05	0.66%	12.55	0.19%
0.10	95.1%	2.60	27.3%	5.10	7.8%	7.60	2.2%	10.10	0.64%	12.60	0.18%
0.15	92.8%	2.65	26.6%	5.15	7.6%	7.65	2.2%	10.15	0.63%	12.65	0.18%
0.20	90.5%	2.70	25.9%	5.20	7.4%	7.70	2.1%	10.20	0.61%	12.70	0.17%
0.25	88.2%	2.75	25.3%	5.25	7.2%	7.75	2.1%	10.25	0.59%	12.75	0.17%
0.30	86.1%	2.80	24.7%	5.30	7.1%	7.80	2.0%	10.30	0.58%	12.80	0.17%
0.35	83.9%	2.85	24.1%	5.35	6.9%	7.85	2.0%	10.35	0.57%	12.85	0.16%
0.40	81.9%	2.90	23.5%	5.40	6.7%	7.90	1.9%	10.40	0.55%	12.90	0.16%
0.45	79.9%	2.95	22.9%	5.45	6.6%	7.95	1.9%	10.45	0.54%	12.95	0.15%
0.50	77.9%	3.00	22.3%	5.50	6.4%	8.00	1.8%	10.50	0.52%	13.00	0.15%
0.55	76.0%	3.05	21.8%	5.55	6.2%	8.05	1.8%	10.55	0.51%	13.05	0.15%
0.60	74.1%	3.10	21.2%	5.60	6.1%	8.10	1.7%	10.60	0.50%	13.10	0.14%
0.65	72.3%	3.15	20.7%	5.65	5.9%	8.15	1.7%	10.65	0.49%	13.15	0.14%
0.70	70.5%	3.20	20.2%	5.70	5.8%	8.20	1.7%	10.70	0.47%	13.20	0.14%
0.75	68.7%	3.25	19.7%	5.75	5.6%	8.25	1.6%	10.75	0.46%	13.25	0.13%
0.80	67.0%	3.30	19.2%	5.80	5.5%	8.30	1.6%	10.80	0.45%	13.30	0.13%
0.85	65.4%	3.35	18.7%	5.85	5.4%	8.35	1.5%	10.85	0.44%	13.35	0.13%
0.90	63.8%	3.40	18.3%	5.90	5.2%	8.40	1.5%	10.90	0.43%	13.40	0.12%
0.95	62.2%	3.45	17.8%	5.95	5.1%	8.45	1.5%	10.95	0.42%	13.45	0.12%
1.00	60.7%	3.50	17.4%	6.00	5.0%	8.50	1.4%	11.00	0.41%	13.50	0.12%
1.05	59.2%	3.55	16.9%	6.05	4.9%	8.55	1.4%	11.05	0.40%	13.55	0.11%
1.10	57.7%	3.60	16.5%	6.10	4.7%	8.60	1.4%	11.10	0.39%	13.60	0.11%
1.15	56.3%	3.65	16.1%	6.15	4.6%	8.65	1.3%	11.15	0.38%	13.65	0.11%
1.20	54.9%	3.70	15.7%	6.20	4.5%	8.70	1.3%	11.20	0.37%	13.70	0.11%
1.25	53.5%	3.75	15.3%	6.25	4.4%	8.75	1.3%	11.25	0.36%	13.75	0.10%
1.30	52.2%	3.80	15.0%	6.30	4.3%	8.80	1.2%	11.30	0.35%	13.80	0.10%
1.35	50.9%	3.85	14.6%	6.35	4.2%	8.85	1.2%	11.35	0.34%	13.85	0.098%
1.40	49.7%	3.90	14.2%	6.40	4.1%	8.90	1.2%	11.40	0.33%	13.90	0.096%
1.45	48.4%	3.95	13.9%	6.45	4.0%	8.95	1.1%	11.45	0.33%	13.95	0.093%
1.50	47.2%	4.00	13.5%	6.50	3.9%	9.00	1.1%	11.50	0.32%	14.00	0.091%
1.55	46.1%	4.05	13.2%	6.55	3.8%	9.05	1.1%	11.55	0.31%	14.05	0.089%
1.60	44.9%	4.10	12.9%	6.60	3.7%	9.10	1.1%	11.60	0.30%	14.10	0.087%
1.65	43.8%	4.15	12.6%	6.65	3.6%	9.15	1.0%	11.65	0.30%	14.15	0.085%
1.70	42.7%	4.20	12.2%	6.70	3.5%	9.20	1.01%	11.70	0.29%	14.20	0.083%
1.75	41.7%	4.25	11.9%	6.75	3.4%	9.25	0.98%	11.75	0.28%	14.25	0.080%
1.80	40.7%	4.30	11.6%	6.80	3.3%	9.30	0.96%	11.80	0.27%	14.30	0.078%
1.85	39.7%	4.35	11.4%	6.85	3.3%	9.35	0.93%	11.85	0.27%	14.35	0.077%
1.90	38.7%	4.40	11.1%	6.90	3.2%	9.40	0.91%	11.90	0.26%	14.40	0.075%
1.95	37.7%	4.45	10.8%	6.95	3.1%	9.45	0.89%	11.95	0.25%	14.45	0.073%
2.00	36.8%	4.50	10.5%	7.00	3.0%	9.50	0.87%	12.00	0.25%	14.50	0.071%
2.05	35.9%	4.55	10.3%	7.05	2.9%	9.55	0.84%	12.05	0.24%	14.55	0.069%
2.10	35.0%	4.60	10.0%	7.10	2.9%	9.60	0.82%	12.10	0.24%	14.60	0.068%
2.15	34.1%	4.65	9.8%	7.15	2.8%	9.65	0.80%	12.15	0.23%	14.65	0.066%
2.20	33.3%	4.70	9.5%	7.20	2.7%	9.70	0.78%	12.20	0.22%	14.70	0.064%
2.25	32.5%	4.75	9.3%	7.25	2.7%	9.75	0.76%	12.25	0.22%	14.75	0.063%
2.30	31.7%	4.80	9.1%	7.30	2.6%	9.80	0.74%	12.30	0.21%	14.80	0.061%
2.35	30.9%	4.85	8.8%	7.35	2.5%	9.85	0.73%	12.35	0.21%	14.85	0.060%
2.40	30.1%	4.90	8.6%	7.40	2.5%	9.90	0.71%	12.40	0.20%	14.90	0.058%
2.45	29.4%	4.95	8.4%	7.45	2.4%	9.95	0.69%	12.45	0.20%	14.95	0.057%

Standard normal critical values
% of observations from χ2(3)…

x	above x	x	above x	x	above x	x	above x	x	above x	x	above x
0.00	100.0%	2.50	47.5%	5.00	17.2%	7.50	5.8%	10.00	1.9%	12.50	0.59%
0.05	99.7%	2.55	46.6%	5.05	16.8%	7.55	5.6%	10.05	1.8%	12.55	0.57%
0.10	99.2%	2.60	45.7%	5.10	16.5%	7.60	5.5%	10.10	1.8%	12.60	0.56%
0.15	98.5%	2.65	44.9%	5.15	16.1%	7.65	5.4%	10.15	1.7%	12.65	0.55%
0.20	97.8%	2.70	44.0%	5.20	15.8%	7.70	5.3%	10.20	1.7%	12.70	0.53%
0.25	96.9%	2.75	43.2%	5.25	15.4%	7.75	5.1%	10.25	1.7%	12.75	0.52%
0.30	96.0%	2.80	42.3%	5.30	15.1%	7.80	5.0%	10.30	1.6%	12.80	0.51%
0.35	95.0%	2.85	41.5%	5.35	14.8%	7.85	4.9%	10.35	1.6%	12.85	0.50%
0.40	94.0%	2.90	40.7%	5.40	14.5%	7.90	4.8%	10.40	1.5%	12.90	0.49%
0.45	93.0%	2.95	39.9%	5.45	14.2%	7.95	4.7%	10.45	1.5%	12.95	0.47%
0.50	91.9%	3.00	39.2%	5.50	13.9%	8.00	4.6%	10.50	1.5%	13.00	0.46%
0.55	90.8%	3.05	38.4%	5.55	13.6%	8.05	4.5%	10.55	1.4%	13.05	0.45%
0.60	89.6%	3.10	37.6%	5.60	13.3%	8.10	4.4%	10.60	1.4%	13.10	0.44%
0.65	88.5%	3.15	36.9%	5.65	13.0%	8.15	4.3%	10.65	1.4%	13.15	0.43%
0.70	87.3%	3.20	36.2%	5.70	12.7%	8.20	4.2%	10.70	1.3%	13.20	0.42%
0.75	86.1%	3.25	35.5%	5.75	12.4%	8.25	4.1%	10.75	1.3%	13.25	0.41%
0.80	84.9%	3.30	34.8%	5.80	12.2%	8.30	4.0%	10.80	1.3%	13.30	0.40%
0.85	83.7%	3.35	34.1%	5.85	11.9%	8.35	3.9%	10.85	1.3%	13.35	0.39%
0.90	82.5%	3.40	33.4%	5.90	11.7%	8.40	3.8%	10.90	1.2%	13.40	0.38%
0.95	81.3%	3.45	32.7%	5.95	11.4%	8.45	3.8%	10.95	1.2%	13.45	0.38%
1.00	80.1%	3.50	32.1%	6.00	11.2%	8.50	3.7%	11.00	1.2%	13.50	0.37%
1.05	78.9%	3.55	31.4%	6.05	10.9%	8.55	3.6%	11.05	1.1%	13.55	0.36%
1.10	77.7%	3.60	30.8%	6.10	10.7%	8.60	3.5%	11.10	1.1%	13.60	0.35%
1.15	76.5%	3.65	30.2%	6.15	10.5%	8.65	3.4%	11.15	1.1%	13.65	0.34%
1.20	75.3%	3.70	29.6%	6.20	10.2%	8.70	3.4%	11.20	1.1%	13.70	0.33%
1.25	74.1%	3.75	29.0%	6.25	10.0%	8.75	3.3%	11.25	1.0%	13.75	0.33%
1.30	72.9%	3.80	28.4%	6.30	9.8%	8.80	3.2%	11.30	1.0%	13.80	0.32%
1.35	71.7%	3.85	27.8%	6.35	9.6%	8.85	3.1%	11.35	1.00%	13.85	0.31%
1.40	70.6%	3.90	27.2%	6.40	9.4%	8.90	3.1%	11.40	0.97%	13.90	0.30%
1.45	69.4%	3.95	26.7%	6.45	9.2%	8.95	3.0%	11.45	0.95%	13.95	0.30%
1.50	68.2%	4.00	26.1%	6.50	9.0%	9.00	2.9%	11.50	0.93%	14.00	0.29%
1.55	67.1%	4.05	25.6%	6.55	8.8%	9.05	2.9%	11.55	0.91%	14.05	0.28%
1.60	65.9%	4.10	25.1%	6.60	8.6%	9.10	2.8%	11.60	0.89%	14.10	0.28%
1.65	64.8%	4.15	24.6%	6.65	8.4%	9.15	2.7%	11.65	0.87%	14.15	0.27%
1.70	63.7%	4.20	24.1%	6.70	8.2%	9.20	2.7%	11.70	0.85%	14.20	0.26%
1.75	62.6%	4.25	23.6%	6.75	8.0%	9.25	2.6%	11.75	0.83%	14.25	0.26%
1.80	61.5%	4.30	23.1%	6.80	7.9%	9.30	2.6%	11.80	0.81%	14.30	0.25%
1.85	60.4%	4.35	22.6%	6.85	7.7%	9.35	2.5%	11.85	0.79%	14.35	0.25%
1.90	59.3%	4.40	22.1%	6.90	7.5%	9.40	2.4%	11.90	0.77%	14.40	0.24%
1.95	58.3%	4.45	21.7%	6.95	7.4%	9.45	2.4%	11.95	0.76%	14.45	0.24%
2.00	57.2%	4.50	21.2%	7.00	7.2%	9.50	2.3%	12.00	0.74%	14.50	0.23%
2.05	56.2%	4.55	20.8%	7.05	7.0%	9.55	2.3%	12.05	0.72%	14.55	0.22%
2.10	55.2%	4.60	20.4%	7.10	6.9%	9.60	2.2%	12.10	0.70%	14.60	0.22%
2.15	54.2%	4.65	19.9%	7.15	6.7%	9.65	2.2%	12.15	0.69%	14.65	0.21%
2.20	53.2%	4.70	19.5%	7.20	6.6%	9.70	2.1%	12.20	0.67%	14.70	0.21%
2.25	52.2%	4.75	19.1%	7.25	6.4%	9.75	2.1%	12.25	0.66%	14.75	0.20%
2.30	51.3%	4.80	18.7%	7.30	6.3%	9.80	2.0%	12.30	0.64%	14.80	0.20%
2.35	50.3%	4.85	18.3%	7.35	6.2%	9.85	2.0%	12.35	0.63%	14.85	0.19%
2.40	49.4%	4.90	17.9%	7.40	6.0%	9.90	1.9%	12.40	0.61%	14.90	0.19%
2.45	48.4%	4.95	17.5%	7.45	5.9%	9.95	1.9%	12.45	0.60%	14.95	0.19%

Standard normal critical values
% of observations from χ2(4)…

x	above x	x	above x	x	above x	x	above x	x	above x	x	above x
0.00	100.0%	2.50	64.5%	5.00	28.7%	7.50	11.2%	10.00	4.0%	12.50	1.4%
0.05	100.0%	2.55	63.6%	5.05	28.2%	7.55	11.0%	10.05	4.0%	12.55	1.4%
0.10	99.9%	2.60	62.7%	5.10	27.7%	7.60	10.7%	10.10	3.9%	12.60	1.3%
0.15	99.7%	2.65	61.8%	5.15	27.2%	7.65	10.5%	10.15	3.8%	12.65	1.3%
0.20	99.5%	2.70	60.9%	5.20	26.7%	7.70	10.3%	10.20	3.7%	12.70	1.3%
0.25	99.3%	2.75	60.0%	5.25	26.3%	7.75	10.1%	10.25	3.6%	12.75	1.3%
0.30	99.0%	2.80	59.2%	5.30	25.8%	7.80	9.9%	10.30	3.6%	12.80	1.2%
0.35	98.6%	2.85	58.3%	5.35	25.3%	7.85	9.7%	10.35	3.5%	12.85	1.2%
0.40	98.2%	2.90	57.5%	5.40	24.9%	7.90	9.5%	10.40	3.4%	12.90	1.2%
0.45	97.8%	2.95	56.6%	5.45	24.4%	7.95	9.3%	10.45	3.3%	12.95	1.2%
0.50	97.4%	3.00	55.8%	5.50	24.0%	8.00	9.2%	10.50	3.3%	13.00	1.1%
0.55	96.8%	3.05	54.9%	5.55	23.5%	8.05	9.0%	10.55	3.2%	13.05	1.1%
0.60	96.3%	3.10	54.1%	5.60	23.1%	8.10	8.8%	10.60	3.1%	13.10	1.1%
0.65	95.7%	3.15	53.3%	5.65	22.7%	8.15	8.6%	10.65	3.1%	13.15	1.1%
0.70	95.1%	3.20	52.5%	5.70	22.3%	8.20	8.5%	10.70	3.0%	13.20	1.0%
0.75	94.5%	3.25	51.7%	5.75	21.9%	8.25	8.3%	10.75	3.0%	13.25	1.0%
0.80	93.8%	3.30	50.9%	5.80	21.5%	8.30	8.1%	10.80	2.9%	13.30	0.99%
0.85	93.2%	3.35	50.1%	5.85	21.1%	8.35	8.0%	10.85	2.8%	13.35	0.97%
0.90	92.5%	3.40	49.3%	5.90	20.7%	8.40	7.8%	10.90	2.8%	13.40	0.95%
0.95	91.7%	3.45	48.6%	5.95	20.3%	8.45	7.6%	10.95	2.7%	13.45	0.93%
1.00	91.0%	3.50	47.8%	6.00	19.9%	8.50	7.5%	11.00	2.7%	13.50	0.91%
1.05	90.2%	3.55	47.0%	6.05	19.5%	8.55	7.3%	11.05	2.6%	13.55	0.89%
1.10	89.4%	3.60	46.3%	6.10	19.2%	8.60	7.2%	11.10	2.5%	13.60	0.87%
1.15	88.6%	3.65	45.5%	6.15	18.8%	8.65	7.0%	11.15	2.5%	13.65	0.85%
1.20	87.8%	3.70	44.8%	6.20	18.5%	8.70	6.9%	11.20	2.4%	13.70	0.83%
1.25	87.0%	3.75	44.1%	6.25	18.1%	8.75	6.8%	11.25	2.4%	13.75	0.81%
1.30	86.1%	3.80	43.4%	6.30	17.8%	8.80	6.6%	11.30	2.3%	13.80	0.80%
1.35	85.3%	3.85	42.7%	6.35	17.4%	8.85	6.5%	11.35	2.3%	13.85	0.78%
1.40	84.4%	3.90	42.0%	6.40	17.1%	8.90	6.4%	11.40	2.2%	13.90	0.76%
1.45	83.5%	3.95	41.3%	6.45	16.8%	8.95	6.2%	11.45	2.2%	13.95	0.75%
1.50	82.7%	4.00	40.6%	6.50	16.5%	9.00	6.1%	11.50	2.1%	14.00	0.73%
1.55	81.8%	4.05	39.9%	6.55	16.2%	9.05	6.0%	11.55	2.1%	14.05	0.71%
1.60	80.9%	4.10	39.3%	6.60	15.9%	9.10	5.9%	11.60	2.1%	14.10	0.70%
1.65	80.0%	4.15	38.6%	6.65	15.6%	9.15	5.7%	11.65	2.0%	14.15	0.68%
1.70	79.1%	4.20	38.0%	6.70	15.3%	9.20	5.6%	11.70	2.0%	14.20	0.67%
1.75	78.2%	4.25	37.3%	6.75	15.0%	9.25	5.5%	11.75	1.9%	14.25	0.65%
1.80	77.2%	4.30	36.7%	6.80	14.7%	9.30	5.4%	11.80	1.9%	14.30	0.64%
1.85	76.3%	4.35	36.1%	6.85	14.4%	9.35	5.3%	11.85	1.9%	14.35	0.63%
1.90	75.4%	4.40	35.5%	6.90	14.1%	9.40	5.2%	11.90	1.8%	14.40	0.61%
1.95	74.5%	4.45	34.9%	6.95	13.9%	9.45	5.1%	11.95	1.8%	14.45	0.60%
2.00	73.6%	4.50	34.3%	7.00	13.6%	9.50	5.0%	12.00	1.7%	14.50	0.59%
2.05	72.7%	4.55	33.7%	7.05	13.3%	9.55	4.9%	12.05	1.7%	14.55	0.57%
2.10	71.7%	4.60	33.1%	7.10	13.1%	9.60	4.8%	12.10	1.7%	14.60	0.56%
2.15	70.8%	4.65	32.5%	7.15	12.8%	9.65	4.7%	12.15	1.6%	14.65	0.55%
2.20	69.9%	4.70	31.9%	7.20	12.6%	9.70	4.6%	12.20	1.6%	14.70	0.54%
2.25	69.0%	4.75	31.4%	7.25	12.3%	9.75	4.5%	12.25	1.6%	14.75	0.52%
2.30	68.1%	4.80	30.8%	7.30	12.1%	9.80	4.4%	12.30	1.5%	14.80	0.51%
2.35	67.2%	4.85	30.3%	7.35	11.9%	9.85	4.3%	12.35	1.5%	14.85	0.50%
2.40	66.3%	4.90	29.8%	7.40	11.6%	9.90	4.2%	12.40	1.5%	14.90	0.49%
2.45	65.4%	4.95	29.2%	7.45	11.4%	9.95	4.1%	12.45	1.4%	14.95	0.48%

Standard normal critical values
% of observations from χ2(5)…

x	above x	x	above x	x	above x	x	above x	x	above x	x	above x
0.00	100.0%	2.50	77.6%	5.00	41.6%	7.50	18.6%	10.00	7.5%	12.50	2.9%
0.05	100.0%	2.55	76.9%	5.05	41.0%	7.55	18.3%	10.05	7.4%	12.55	2.8%
0.10	100.0%	2.60	76.1%	5.10	40.4%	7.60	18.0%	10.10	7.2%	12.60	2.7%
0.15	100.0%	2.65	75.4%	5.15	39.8%	7.65	17.7%	10.15	7.1%	12.65	2.7%
0.20	99.9%	2.70	74.6%	5.20	39.2%	7.70	17.4%	10.20	7.0%	12.70	2.6%
0.25	99.8%	2.75	73.8%	5.25	38.6%	7.75	17.1%	10.25	6.8%	12.75	2.6%
0.30	99.8%	2.80	73.1%	5.30	38.0%	7.80	16.8%	10.30	6.7%	12.80	2.5%
0.35	99.7%	2.85	72.3%	5.35	37.5%	7.85	16.5%	10.35	6.6%	12.85	2.5%
0.40	99.5%	2.90	71.5%	5.40	36.9%	7.90	16.2%	10.40	6.5%	12.90	2.4%
0.45	99.4%	2.95	70.8%	5.45	36.3%	7.95	15.9%	10.45	6.3%	12.95	2.4%
0.50	99.2%	3.00	70.0%	5.50	35.8%	8.00	15.6%	10.50	6.2%	13.00	2.3%
0.55	99.0%	3.05	69.2%	5.55	35.2%	8.05	15.4%	10.55	6.1%	13.05	2.3%
0.60	98.8%	3.10	68.5%	5.60	34.7%	8.10	15.1%	10.60	6.0%	13.10	2.2%
0.65	98.6%	3.15	67.7%	5.65	34.2%	8.15	14.8%	10.65	5.9%	13.15	2.2%
0.70	98.3%	3.20	66.9%	5.70	33.7%	8.20	14.6%	10.70	5.8%	13.20	2.2%
0.75	98.0%	3.25	66.2%	5.75	33.1%	8.25	14.3%	10.75	5.7%	13.25	2.1%
0.80	97.7%	3.30	65.4%	5.80	32.6%	8.30	14.0%	10.80	5.5%	13.30	2.1%
0.85	97.4%	3.35	64.6%	5.85	32.1%	8.35	13.8%	10.85	5.4%	13.35	2.0%
0.90	97.0%	3.40	63.9%	5.90	31.6%	8.40	13.6%	10.90	5.3%	13.40	2.0%
0.95	96.6%	3.45	63.1%	5.95	31.1%	8.45	13.3%	10.95	5.2%	13.45	2.0%
1.00	96.3%	3.50	62.3%	6.00	30.6%	8.50	13.1%	11.00	5.1%	13.50	1.9%
1.05	95.8%	3.55	61.6%	6.05	30.1%	8.55	12.8%	11.05	5.0%	13.55	1.9%
1.10	95.4%	3.60	60.8%	6.10	29.7%	8.60	12.6%	11.10	4.9%	13.60	1.8%
1.15	95.0%	3.65	60.1%	6.15	29.2%	8.65	12.4%	11.15	4.8%	13.65	1.8%
1.20	94.5%	3.70	59.3%	6.20	28.7%	8.70	12.2%	11.20	4.8%	13.70	1.8%
1.25	94.0%	3.75	58.6%	6.25	28.3%	8.75	11.9%	11.25	4.7%	13.75	1.7%
1.30	93.5%	3.80	57.9%	6.30	27.8%	8.80	11.7%	11.30	4.6%	13.80	1.7%
1.35	93.0%	3.85	57.1%	6.35	27.4%	8.85	11.5%	11.35	4.5%	13.85	1.7%
1.40	92.4%	3.90	56.4%	6.40	26.9%	8.90	11.3%	11.40	4.4%	13.90	1.6%
1.45	91.9%	3.95	55.7%	6.45	26.5%	8.95	11.1%	11.45	4.3%	13.95	1.6%
1.50	91.3%	4.00	54.9%	6.50	26.1%	9.00	10.9%	11.50	4.2%	14.00	1.6%
1.55	90.7%	4.05	54.2%	6.55	25.6%	9.05	10.7%	11.55	4.2%	14.05	1.5%
1.60	90.1%	4.10	53.5%	6.60	25.2%	9.10	10.5%	11.60	4.1%	14.10	1.5%
1.65	89.5%	4.15	52.8%	6.65	24.8%	9.15	10.3%	11.65	4.0%	14.15	1.5%
1.70	88.9%	4.20	52.1%	6.70	24.4%	9.20	10.1%	11.70	3.9%	14.20	1.4%
1.75	88.3%	4.25	51.4%	6.75	24.0%	9.25	9.9%	11.75	3.8%	14.25	1.4%
1.80	87.6%	4.30	50.7%	6.80	23.6%	9.30	9.8%	11.80	3.8%	14.30	1.4%
1.85	86.9%	4.35	50.0%	6.85	23.2%	9.35	9.6%	11.85	3.7%	14.35	1.4%
1.90	86.3%	4.40	49.3%	6.90	22.8%	9.40	9.4%	11.90	3.6%	14.40	1.3%
1.95	85.6%	4.45	48.7%	6.95	22.4%	9.45	9.2%	11.95	3.5%	14.45	1.3%
2.00	84.9%	4.50	48.0%	7.00	22.1%	9.50	9.1%	12.00	3.5%	14.50	1.3%
2.05	84.2%	4.55	47.3%	7.05	21.7%	9.55	8.9%	12.05	3.4%	14.55	1.2%
2.10	83.5%	4.60	46.7%	7.10	21.3%	9.60	8.7%	12.10	3.3%	14.60	1.2%
2.15	82.8%	4.65	46.0%	7.15	21.0%	9.65	8.6%	12.15	3.3%	14.65	1.2%
2.20	82.1%	4.70	45.4%	7.20	20.6%	9.70	8.4%	12.20	3.2%	14.70	1.2%
2.25	81.4%	4.75	44.7%	7.25	20.3%	9.75	8.3%	12.25	3.2%	14.75	1.1%
2.30	80.6%	4.80	44.1%	7.30	19.9%	9.80	8.1%	12.30	3.1%	14.80	1.1%
2.35	79.9%	4.85	43.4%	7.35	19.6%	9.85	8.0%	12.35	3.0%	14.85	1.1%
2.40	79.1%	4.90	42.8%	7.40	19.3%	9.90	7.8%	12.40	3.0%	14.90	1.1%
2.45	78.4%	4.95	42.2%	7.45	18.9%	9.95	7.7%	12.45	2.9%	14.95	1.1%