ECO5120: Econometric Theory and Application, Fall 2006
Prof. T.L. Chong
HANDOUT 1
PROBABILITY AND DISTRIBUTION THEORY I

Definition 1 *A **random experiment** is an experiment satisfying the following three conditions:*

(i) All possible distinct outcomes are known a priori;
(ii) In any particular trial the outcome is not known a priori;
(iii) It can be repeated under identical conditions.

Definition 2 *The **sample space** Ω is defined to be the set of all possible outcomes of the random experiment.*

Example 3 *When throwing a dice, the sample space is*

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Example 4 *Consider the sum of points when throwing two dices, the sample space will be*

$$\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Definition 5 *An **elementary event** is the element of the sample space Ω.*

Example 6 *When throwing a dice, the element $\{1\}$ is an elementary event.*

Definition 7 *An event E is a subset of the sample space Ω. Every subset is an event. Thus an event may be an empty set, a proper subset of the sample space, or the sample space itself. An elementary event is an event while an event may not be an elementary event.*

Example 8 *Consider the sum of points when throwing two dices, the event that the sum is an even number will be*

$$E = \{2, 4, 6, 8, 10, 12\}.$$

Example 9 *The event that the sum is bigger than* 13 *will be an empty set* ϕ*, we call it a null event.*

Example 10 *The event that the sum is smaller than* 13 *will be* $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$*, or equal the sample space.*

Definition 11 *The collection* $\Im$ *of subsets of* Ω *is called* $\sigma-$***algebra*** *if it satisfies the following properties:*

(i) $\Omega \in \Im$,

(ii) $E \in \Im \Rightarrow E^c \in \Im$, (closure under complementation)

where E^c refers to the complement of E with respect to Ω.

(iii) $E_j \in \Im$, $j = 1, 2, ... \Rightarrow \cup_{j=1}^{\infty} E_j \in \Im$. (closure under countable union)

Example 12 *Consider* $\Omega = \{1, 2, 3\}$*, and let*

$\Im_1 = \{\phi, \Omega, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}\}$;

$\Im_2 = \{\phi, \Omega\}$;

$\Im_3 = \{\phi, \Omega, \{1\}, \{2\}, \{1, 2\}, \{1, 3\}\}$;

It can be verified that $\Im_1$ and $\Im_2$ are $\sigma-$**algebra** but $\Im_3$ is not.

Definition 13 *A* ***probability measure****, denoted by* $P(\cdot)$*, is a real-valued set function that is defined over a* $\sigma-$*algebra* $\Im$ *and satisfies the following properties:*

(i) $P(\Omega) = 1$;

(ii) $E \in \Im \Rightarrow P(E) \geq 0$;

(iii) If $\{E_j\}$ is a countable collection of disjoint sets in $\Im$, then $P\left(\cup_{j=1}^{n} E_j\right) = \sum_{j=1}^{n} P(E_j)$.

Definition 14 *Given a sample space Ω, a $\sigma-$algebra $\Im$ associated with Ω, and a probability measure $P(\cdot)$ defined over $\Im$, we call the triplet $(\Omega, \Im, P)$ a* ***probability space****.*

Definition 15 *The* ***conditional probability*** *of B occurring, given that A has occurred is*

$$\begin{aligned} \Pr(B|A) &= \frac{\Pr(B \cap A)}{\Pr(A)} \qquad if \ \Pr(A) \neq 0; \\ \Pr(B|A) &= 0 \qquad if \ \Pr(A) = 0. \end{aligned}$$

The result implies that

$$\Pr(B \cap A) = \Pr(B|A) \Pr(A).$$

Example 16 *Consider a card game, let H be the event that a "Heart" appears, A be the event that an "Ace" appears.*

$$\Pr(H|A) = \frac{\Pr(H \cap A)}{\Pr(A)} = \frac{1/52}{1/13} = \frac{1}{4}.$$

Definition 17 *Two events A and B are* ***independent*** *if and only if* $\Pr(A \cap B) = \Pr(A) \Pr(B)$. *i.e.* $\Pr(A|B) = \Pr(A)$.

Definition 18 *A* ***random variable*** *on $(\Omega, \Im, P)$ is a real-valued function defined over a sample space Ω, denoted by $X(\omega)$ for $\omega \in \Omega$, such that for any real number x, $\{\omega | X(\omega) < x\} \in \Im$.*

Example 19 *Consider tossing a coin, $\Omega = \{H, T\}$, the $\sigma-$algebra $\Im = \{\phi, \Omega, \{H\}, \{T\}\}$. If we define $X(H) = 1$ and $X(T) = 2$, then X is a random variable. Consider a real number x, if $x = 1.5$, then $\{\omega | X(\omega) < 1.5\} = \{H\} \in \Im$.*

A random variable is always defined relative to some specific $\sigma-$algebra $\Im$. It is *discrete* if its range forms a discrete(countable) set of real number. It is *continuous* if its range forms a continuous(uncountable) set of real numbers and the probability of X equalling any single value in its range is zero.

Definition 20 *Let X be a continuous random variable. The **probability distribution function** of X is defined as $F_x(u) = \Pr(-\infty < X \leq u)$, with $F_x(\infty) = 1$. The **density function** is $f(x) = \dfrac{dF(x)}{dx}$, with $f(x) \geq 0$, and $f(-\infty) = f(\infty) = 0$.*

Definition 21 *The **mean, first moment**, or **expectation** of a random variable X, is defined as:*

$$\begin{aligned} E(X) &= \sum_i x_i P(x_i) \qquad \text{if } X \text{ is discrete} \\ &= \int_{-\infty}^{\infty} x f(x)\, dx \qquad \text{if } X \text{ is continuous} \end{aligned}$$

Definition 22 *The **median** of a random variable X, denoted by m is defined as the value that satisfies $\Pr(X \leq m) = 0.5$.*

Note that median under this definition may not be unique. For example, if X is a continuous random variable uniformly distributed in the region $[0,2] \cup [4,6]$, then m is note unique, as it can be anything from 2 to 4. To ensure uniqueness, we may redefine the median to $\inf_{m \in R} \Pr(X \leq m) = 0.5$.

Definition 23 *The **mode** of a random variable X with density $f(x)$ is defined as $Argmax_{x \in R}\ f(x)$.*

Note that, similar to the median, the mode may not be unique too.

Remark 1 *Note that the three measures of central tendency discussed above, the population mean, population median, and population mode are fixed constants. However, in an empirical sample, the sample mean, sample median*

and sample mode are all random variables whose values vary from sample to sample. Different measures have their own merits and shortcomings. They are the summary statistics of the sample, i.e., by look at their values, one should have a rough picture of what the data should be. Sometimes these measures may not be informative. For example, the sample mean is easily affected by outliers. In the sample $\{1, 2, 3, 4, 1000\}$*, most of the number are small, and the sample mean is 202, which is not informative. In the sample* $\{2, 2, 1000, 1001, 1002, 1003, 1004\}$*, the sample mode equals 2, which is not informative as most of the observations are over 1000.*

Definition 24 *The* ***second moment around the mean*** *or* ***variance*** *of a random variable is*

$$\begin{aligned} Var(X) &= \sum_i (x_i - E(X))^2 P(x_i) \qquad \text{if } X \text{ is discrete} \\ &= \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\, dx \qquad \text{if } X \text{ is continuous} \end{aligned}$$

Definition 25 *Let* X*,* Y *be two continuous random variables. The* ***joint distribution function*** *of* X *and* Y *is defined as* $F(x, y) = \Pr(X \leq x \text{ and } Y \leq y)$*. Their* ***joint density function*** *is* $f(x, y)$*. The relationship between* $F(x, y)$ *and* $f(x, y)$ *is:*

$$\begin{aligned} F(x, y) &= \int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t)\, dsdt, \\ f(x) &= \int_{-\infty}^{\infty} f(x, y)\, dy, \\ f(y) &= \int_{-\infty}^{\infty} f(x, y)\, dx. \end{aligned}$$

Further, $F(-\infty, -\infty) = 0$, $F(\infty, \infty) = 1$, and $f(x, y) \geq 0$. If X and Y are independent, then $F(x, y) = F(x) F(y)$ and $f(x, y) = f(x) f(y)$.

Definition 26 *The **covariance** of two random variables X and Y, is defined to be:*

$$Cov(X,Y) = E(X - E(X))(Y - E(Y)) = E(XY) - E(X)E(Y)$$

where

$$\begin{aligned} E(XY) &= \sum_i x_i y_i P(x_i, y_i) \qquad \text{if } X, Y \text{ are discrete} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dxdy \qquad \text{if } X, Y \text{ are continuous} \end{aligned}$$

$E(XY) = E(X)E(Y)$ if X and Y are independent, i.e., if X and Y are independent, $Cov(X,Y)$ will be equal to zero. However, the reverse is not necessarily true.

Definition 27 *The **correlation coefficient** between X and Y is defined as:*

$$\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}.$$

Theorem 28 ***(Chebyshev's Inequality)***

If X is **any** random variable with finite variance σ^2 and k is a finite positive constant, then

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof. (for continuous random variable)

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \\
&\geq \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)\,dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)\,dx \\
&\geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x)\,dx + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(x)\,dx \\
&= k^2\sigma^2 P(X \leq \mu - k\sigma) + k^2\sigma^2 P(X \geq \mu + k\sigma) \\
&= k^2\sigma^2 P(|X-\mu| \geq k\sigma),
\end{aligned}$$

this implies

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}. \blacksquare$$

Theorem 29 ***(Jensen's Inequality)***

Let $g : R \to R$ be a convex function on an interval $B \subset R$ and let Z be a random variable such that $P(Z \in B) = 1$. Then $g(E(Z)) \leq E(g(Z))$.
Proof. (exercise).

Example 30 *Let $g(z) = |z|$. It follows from Jensen's inequality that $|E(Z)| \leq E|Z|$.*

Example 31 *Let $g(z) = z^2$. It follows from Jensen's inequality that $E^2(Z) \leq E(Z^2)$.*

More demanding materials

Theorem 32 *For random sample of size n from an infinite population which has the value $f(x)$ at x, the probability density of the r^{th}* ***order statistic*** *Y_r is given by*

$$g_r(y_r) = \frac{n!}{(r-1)!\,(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1} f(y_r)\left[\int_{y_r}^{\infty} f(x)\,dx\right]^{n-r}$$

for $y_1 \leq ... \leq y_r \leq ... \leq y_n$.

Proof. Suppose we divide the real line into 3 intervals, $(-\infty, y_r]$, $(y_r, y_r + h]$ and $(y_r + h, \infty)$, then the probability that $r-1$ of the sample values fall into the first interval, one falls into the second interval, and $n-r$ fall into the last interval is

$$\begin{aligned} & \Pr(y_r < Y_r \leq y_r + h) \\ = & \frac{n!}{(r-1)!1!(n-r)!} [\Pr(X \leq y_r)]^{r-1} \Pr(y_r < X \leq y_r + h) [\Pr(X > y_r + h)]^{n-r}. \end{aligned}$$

Let $h \to 0$ and use the facts that $\lim_{h\to 0} \frac{1}{h} \Pr(y_r < X \leq y_r + h) = f(y_r)$ and $\lim_{h\to 0} \frac{1}{h} \Pr(y_r < Y_r \leq y_r + h) = g(y_r)$, we have

$$g_r(y_r) = \frac{n!}{(r-1)!(n-r)!} \left[\int_{-\infty}^{y_r} f(x)\, dx\right]^{r-1} f(y_r) \left[\int_{y_r}^{\infty} f(x)\, dx\right]^{n-r}. \blacksquare$$

Proposition 33 ***(Generalized Chebyshev Inequality or Markov's Inequality)***

Let X be a random variable such that $E|X|^r < \infty$, $p > 0$. Then for every $\varepsilon > 0$,

$$\Pr(|X| \geq \varepsilon) \leq \frac{E|X|^p}{\varepsilon^p}.$$

Proof.

$$\begin{aligned} \varepsilon^p \Pr(|X| \geq \varepsilon) &= \varepsilon^p \int_{|x|\geq\varepsilon} dF(x) \\ &= \int_{|x|\geq\varepsilon} \varepsilon^p dF(x) \\ &\leq \int_{|x|\geq\varepsilon} |x|^p dF(x) \\ &\leq \int_{-\infty}^{\infty} |x|^p dF(x) \\ &= E|X|^p. \blacksquare \end{aligned}$$

Setting $p = 2$ gives the familiar Chebyshev inequality.

Theorem 34 ***(Hölder's Inequality)***

For any $p \geq 1$,

$$E|XY| \leq \|X\|_p \|Y\|_q$$

where
$q = \dfrac{p}{p-1}$ if $p > 1$, and $q = \infty$ if $p = 1$.
$\|X\|_p = (E(|X|^p))^{1/p}$ is the L_p-norm of X.

Proof. (exercise).

Corollary 35 ***(Cauchy-Schwartz Inequality)***

When $p = 2$, the Hölder's Inequality reduced to

$$(E(XY))^2 \leq E\left(X^2\right) E\left(Y^2\right).$$

Theorem 36 ***(Liapunov's Inequality)***

If $r > p > 0$, then

$$\|X\|_r \geq \|X\|_p.$$

Proof. (exercise).

Theorem 37 ***(Minkowski's Inequality)***

For $r \geq 1$,

$$\|X + Y\|_r \leq \|X\|_r + \|Y\|_r.$$

Proof. (exercise).

Theorem 38 ***(Loève's c_r Inequality)***

For $r > 0$,

$$E\left(\left|\sum_{i=1}^{m} X_i\right|^r\right) \leq c_r \sum_{i=1}^{m} E\left(|X_i|^r\right)$$

where $c_r = 1$ when $r \leq 1$ and $c_r = m^{r-1}$ when $r \geq 1$

Proof. (exercise).

Exercise 0.1 *Show that $\Im = \{\phi, \Omega\}$ is a σ-algebra.*

Exercise 0.2 *Suppose the business cycle of an economy can be divided into two states, namely, the contraction C, and the expansion E, so that the sample space $\Omega = \{C, E\}$. Find the corresponding $\sigma-$algebra and explain your answer.*

Exercise 0.3 *The Mark Six lottery is a lottery game conducted by HKJC Lotteries Limited using the facilities of The Hong Kong Jockey Club. Since its inception in 1975, the Mark Six has contributed over HK$24 billion to the Hong Kong SAR Government Treasury and the Lotteries Fund, being a fund that supports charitable causes in Hong Kong.*

To win the first prize of the Mark Six, one needs to get 6 numbers correct out of a pool of 49 numbers indexed from 1 to 49. Suppose each number has the same chance of being drawn,

(a) Find the probability of winning the first prize of the Mark Six.
(b) Suppose you have to bet 5 dollars for the first prize of 50,000,000 dollars. If there is only one first prize winner, find the expect gain (or loss) of your game.
(c) Suppose Chinese people have preference over the "lucky" numbers 8, 18, 28, 38, and a large proportion of people like to put these numbers on their Mark-Six tickets. Suppose the amount of money for the first the prize is fixed, and has to be shared among winners. As an rational economic agent, will you avoid these "lucky" numbers when you buy Mark Six? Explain.

Exercise 0.4 *Suppose a continuous random variable X has density function*

$f(x;\theta) = \theta x + .5$ for $-1 < x < 1$.
$f(x;\theta) = 0$ otherwise.
(i) Find values of θ such that $f(x;\theta)$ is a density function.
(ii) Find the mean and median of X.
(iii) Find $\Pr(0.25 \leq X \leq 0.75)$.
(iv) For what value of θ is the variance of X maximized.
(v) Redo (i) to (iv) if
$f(x;\theta) = \theta x^2 (1-x)^3$ for $0 < x < 1$.
$f(x;\theta) = 0$ otherwise.

Exercise 0.5 *Prove that for any two random variables X and Y, $\left|\rho_{xy}\right| \leq 1$.*

Exercise 0.6 *Let X, Y be two independent identical discrete random variable with the probability distribution as follows:*

$X = -1$ with probability $\frac{1}{2}$.
$X = 1$ with probability $\frac{1}{2}$.
$Y = -1$ with probability $\frac{1}{2}$.
$Y = 1$ with probability $\frac{1}{2}$.
Find the distribution of Z if:
a) $Z = X - Y$.
b) $Z = \dfrac{X}{Y}$.
c) $Z = \max\{X, Y\}$.

Exercise 0.7 *If X and Y are two continuous random variables, then $X+Y$ must be continuous too. True or false? Explain.*

Exercise 0.8 *Let X be a random variable with a symmetrical distribution about zero and a finite variance. Give a random variable Y such that X and Y are uncorrelated but not independent.*

Exercise 0.9 *Suppose the joint density of X and Y is given by:*

$f(x,y) = 2$ for $x > 0$, $y > 0$, $x + y < 1$
$f(x,y) = 0$ otherwise
Find
(i) $\Pr\left(X \leq \frac{1}{2} \text{ and } Y \leq \frac{1}{2}\right)$.
(ii) $\Pr\left(X + Y > \frac{2}{3}\right)$.
(iii) $\Pr(X > 2Y)$.

Exercise 0.10 *True/False/Uncertain. Explain. Let X, Y, and Z be three random variables:*

a)If $Cov(X,Z) \neq 0$ and $Cov(Y,Z) \neq 0$, then $Cov(X,Y) \neq 0$.
b)If $Cov(X^2, Y^2) = 0$, then $Cov(X,Y) = 0$.
c)If X and Y are independent and if $E\left(\frac{X}{Y}\right) > 1$, then $\frac{E(X)}{E(Y)} > 1$.

Exercise 0.11 *Let Z_1, Z_2 be independent $N(0,1)$ random variables, let*

$$U = \min\{Z_1, \max\{Z_1, Z_2\}\}.$$

(a) What is the distribution of U?
(b) Find $E(U)$ and $Var(U)$.

Exercise 0.12 *Let X be a continuous random variable which takes values in $(-\infty, \infty)$. Let $Y = \min\{X, X^2\}$, rewrite X in terms of Y.*

Exercise 0.13 *Let X, Y, and Z be three random variables. If $Cov(X,Z) < 0$ and $Cov(Y,Z) < 0$, then $Cov(X,Y)$ must be positive. True or false? Explain.*

Exercise 0.14 *Prove that for a non-negative random variable X, where $F(X)$ is the distribution function of X.*

i) $E(X) = \int_0^\infty (1 - F(x))\, dx$,
ii) $E(X^r) = r\int_0^\infty x^{r-1}(1 - F(x))\, dx$,
where $F(X)$ is the distribution function of X and $r > 0$.

Exercise 0.15 *Let X be a continuous random variable which takes values in $(-\infty, \infty)$. Let $Y = \max\left\{X, \frac{1}{X}\right\}$, rewrite X in terms of Y.*

Exercise 0.16 *Let $F(x)$ and $f(x)$ be the distribution function and the density function of non-negative random variable X.*

a) Suppose

$$\frac{f(x)}{1 - F(x)} = 1 \qquad \text{for all } x$$

Show that $E(X) = 1$.

b) Find the functional form of $f(x)$ such that the ratio

$$\frac{f(x)}{1 - F(x)}$$

does not depends on x.

Exercise 0.17 *Let $X_{\{i\}}$ be the order statistic of i^{th}* ***order statistic*** *with $X_{\{1\}} \leq ... \leq X_{\{i\}} \leq ... \leq X_{\{n\}}$. Find the variance of the simple average $\overline{X}$ and the variance of the trimmed mean $T_1 = \frac{1}{n-1}\sum_{i=2}^{n} X_{\{i\}}$ and $T_2 = \frac{1}{n-2}\sum_{i=2}^{n-1} X_{\{i\}}$ in terms of n.. Intuitively, which variance will be larger? Prove that $Var\left(\overline{X}\right)$ is smaller than the other two variances.*

Exercise 0.18 *Prove the Markov inequality.*

Exercise 0.19 *Prove the Hölder's inequality.*

Exercise 0.20 *Prove the Liapunov's inequality.*

Exercise 0.21 *Prove the Minkowski' inequality.*

Exercise 0.22 *Prove the Loève's c_r inequality.*

ECO5120: Econometric Theory and Application, Fall 2006
Prof. T.L. Chong
HANDOUT 2
PROBABILITY AND DISTRIBUTION THEORY II

Some Commonly Used Probability Distributions

Uniform distribution

$X \sim U(0,1)$ means X is evenly distributed in the interval $[0,1]$, its density function is defined as:

$$\begin{aligned} f(x) &= 1 \qquad \text{for } x \in [0,1] \\ f(x) &= 0 \qquad \text{elsewhere.} \end{aligned}$$

Normal distribution

The normal distribution is the most commonly used distribution, many variables in the real world follow approximately this distribution.

We write a random variable which follows a normal distribution with mean μ and variance σ^2 as $X \sim N(\mu, \sigma^2)$. Its density function is defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \qquad -\infty < x < \infty.$$

Two unique properties associated with normal random variables are that:
(i) If X and Y are normal, then $X + Y$ is also normal.
(ii) If $X$and Y are normal and uncorrelated, then they are independent.

Standardized normal distribution

If $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma}$ follows $N(0,1)$. Its density function is defined as:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right), \qquad -\infty < z < \infty.$$

The distribution function for a standardized normal random variable is defined as

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)\, dx.$$

The Lognormal distribution

If $X = \ln Y \sim N(\mu, \sigma^2)$, then Y follows a lognormal distribution. Its density function is:

$$\begin{aligned} f(y) &= \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln y - \mu}{\sigma}\right)^2\right), \qquad 0 < y < \infty; \\ f(y) &= 0 \qquad \text{elsewhere} \end{aligned}$$

Chi-square distribution

If $Z \sim N(0,1)$, then Z^2 follows Chi-square distribution with degree of freedom equals 1. e.g. If $Z \sim N(0,1)$, then $U = Z^2$ follows χ_1^2. $\Pr(-1 \leq Z \leq 1) = \Pr(0 \leq U \leq 1) \simeq 0.67$, $\Pr(-2 \leq Z \leq 2) = \Pr(0 \leq U \leq 4) \simeq 0.95$, $\Pr(-3 \leq Z \leq 3) = \Pr(0 \leq U \leq 9) \simeq 0.99$. Thus a Chi-square random variable must take non-negative values, and the distribution has a long right tail.

If $Z_1, Z_2, ..., Z_k$ are independent $N(0,1)$, then $U = Z_1^2 + Z_2^2 + ... + Z_k^2$ follows Chi-square distribution with k degrees of freedom, and we write it as χ_k^2. The mean of a Chi-square distribution equals its degrees of freedom. This is because

$$E\left(Z^2\right) = Var(Z) + E^2(Z) = 1 + 0 = 1,$$

and thus

$$E(U) = E\left(Z_1^2 + Z_2^2 + ... + Z_k^2\right) = k.$$

It density function of U is

$$\begin{aligned} f(u) &= \frac{u^{\frac{k-2}{2}} e^{-u/2}}{2^{k/2}\Gamma(k/2)}, \qquad 0 < u < \infty; \\ f(u) &= 0 \qquad \text{elsewhere} \end{aligned}$$

where $\Gamma(n) = (n-1)\Gamma(n-1)$, $\Gamma(1) = 1$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

For large k ($k > 30$), we have the following approximation

$$\Pr(U \leq a) \approx \Phi\left(\sqrt{2a} - \sqrt{2k-1}\right).$$

For example, when $k = 70$ and $a = 85$,

$$\Pr(U \leq 85) \approx \Phi\left(\sqrt{170} - \sqrt{139}\right) = \Phi(1.249) = .8942.$$

The true $\Pr(U \leq 85) = 0.89409$.

Exponential distribution

For $\theta > 0$, a random variable X has an exponential distribution if and only if its density function is given by

$$\begin{aligned} f(x) &= \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right) \qquad \text{for } x \geq 0 \\ f(x) &= 0 \qquad \text{elsewhere.} \end{aligned}$$

The mean of the exponential distribution is θ and the variance is θ^2.

Note that a Chi-square distribution with degrees of freedom equal 2 is identical to an exponential distribution with $\theta = 2$.

Student's t-distribution

If $Z \sim N(0,1)$, U has a χ^2 distribution with k degrees of freedom, and Z and U are independent, then:

$$t = \frac{Z}{\sqrt{U/k}}$$

has a t-distribution with k degrees of freedom. The t distribution was introduced by W. S. Gosset, who published his work under the pen name "Student".

Cauchy distribution

There are many kinds of distributions, most of them have finite mean and variance, and some higher moments also exist, but for some distributions, their mean and variance may not even exist. e.g., Cauchy distribution.

Let Z_1 and Z_2 be independent and follow $N(0,1)$, then the ratio $\frac{Z_1}{Z_2}$ will have a Cauchy distribution. In other words, a *Cauchy distribution is a t-distribution with degrees of freedom equal one.* Its density has the form:

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty$$

and it can be shown that :

$$x = \tan\left(\left(F(x) - \frac{1}{2}\right)\pi\right)$$

The second moment does not exist since

$$\begin{aligned} E\left(X^2\right) &= \int_{-\infty}^{\infty} x^2 \frac{1}{\pi(1+x^2)} dx \\ &= \int_{-\infty}^{\infty} \frac{1+x^2}{\pi(1+x^2)} dx - \int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\pi} dx - \int_{-\infty}^{\infty} f(x)\, dx \\ &= \left[\frac{x}{\pi}\right]_{-\infty}^{\infty} - 1 = \infty \end{aligned}$$

Similarly $E|X| = \infty$.

F-distribution

Let $U \sim \chi^2_m$ and $V \sim \chi^2_n$, and if U and V are independent of each other, then

$$F(m,n) = \frac{U/m}{V/n}$$

has an F-distribution with m and n d.f..

The F-distribution was named after Sir Ronald A. Fisher, a remarkable statistician of this century.

Note 1: As $n \to \infty$, $\dfrac{V}{n} \to E\left(Z^2\right) = 1$, where Z is a standardized normal random variable. Thus,

$$mF\left(m, \infty\right) = U \sim \chi_m^2$$

and

$$F\left(\infty, \infty\right) = 1.$$

Note 2:

$$F\left(1, k\right) = \frac{U}{V/k} = \left(\frac{N\left(0,1\right)}{\sqrt{\chi_k^2/k}}\right)^2 = t_k^2.$$

Beta distribution

Another class of continuous distribution in the zero-one interval is the Beta distribution. Its density function is given by

$$\begin{aligned} f\left(x\right) &= \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} x^{\alpha-1}\left(1-x\right)^{\beta-1} \qquad \text{for } x \in (0,1) \\ f\left(x\right) &= 0 \qquad \text{elsewhere,} \end{aligned}$$

where $\alpha > 0$ and $\beta > 0$. It can be shown that the mean of the Beta distribution is $\dfrac{\alpha}{\alpha+\beta}$ and the variance is $\dfrac{\alpha\beta}{\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)}$.

Poisson distribution

Suppose the random variable X takes discrete values 0,1,2,3,..., if X follows a Poisson distribution with mean λ, then

$$\Pr\left(X = x\right) = \frac{\exp\left(-\lambda\right)\lambda^x}{x!} \qquad \text{for } x = 0, 1, 2, ...$$

where $\lambda > 0$. It can be shown that the mean and the variance of the Poisson distribution are both equal to λ.

Mixtures

Mixtures of densities occur in some models. For example, one might have a random variable X distributed as $N(0, Y)$ where Y comes from a Poisson process, i.e., the variance of X is a draw from another density. Such models occur with asset price changes (X). The variable Y is the number of news items coming in on any given day so that the variance (volatility) of X depends upon the amount of news becoming available.

Theorem 39 *Let X and Y be two random variables, the unconditional expectation of Y is the expectation of the conditional expectation of Y given X. i.e.,*

$$E(Y) = E_x(E(Y|X)).$$

This theorem is called the **Law of Iterated Expectation**.

For example, suppose given X, the random variable Y is normally distributed with mean X and variance 1, i.e. $Y|X \sim N(X, 1)$. Now suppose X is uniformly distributed in the zero-one interval. Then without knowing the value of X, the unconditional expectation of Y will be

$$E(Y) = E_x(E(Y|X)) = E_x(X) = 0.5.$$

Definition 40 *Other measures often used to describe a probability distribution are*

$$\begin{aligned}
\textbf{Skewness} &= E\left[(X-\mu)^3\right], \\
\textbf{Skewness coefficient} &= \frac{E\left[(X-\mu)^3\right]}{\sigma^3}, \\
\textbf{Kurtosis} &= E\left[(X-\mu)^4\right], \\
\textbf{Degree of excess} &= \frac{E\left[(X-\mu)^4\right]}{\sigma^4} - 3.
\end{aligned}$$

Skewness is a measure of the asymmetry of a distribution. For symmetric distributions, skewness=0. Kurtosis is a measure of the thickness of the

tails of the distribution. For a Normal distribution, the skewness =0 and the degree of excess=0. Thus, when checking whether a random variable is normally distributed, it will be helpful to see if both the skewness and the degree of excess are zero.

Although the moments of most distributions can be determined directly by evaluating the necessary integrals or sums, there is an alternative procedure which sometimes provides considerable simplifications.

Definition 41 *The **probability-generating function** of a **discrete** random variable X, where it exists, is given by*

$$P(t) = E\left(t^X\right) = \sum_{j=0}^{\infty} t^j \Pr(X = j).$$

The p.g.f. can be used to compute the mean of a discrete random variable X. The first derivative of $P(t)$ evaluated at $t = 1$ is the mean of X. i.e.,

$$E(X) = \left.\frac{dP(t)}{dt}\right|_{t=1} = P'(1).$$

We also have

$$Var(X) = P''(1) + P'(1) - [P'(1)]^2.$$

Note that p.g.f. is for discrete random variables only. For continuous random variables, one has to use the moment generating function.

Definition 42 *The **moment-generating function** of a random variable X, where it exists, is given by*

$$M_X(t) = E\left(e^{tX}\right) = E\left(\sum_{j=0}^{\infty} \frac{(tX)^j}{j!}\right) = \sum_{j=0}^{\infty} \frac{t^j}{j!} E\left(X^j\right).$$

The result utilizes the Taylor's expansion that

$$e^z = \left(\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n\right)^z = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{nz} = \sum_{j=0}^{\infty}\frac{Z^j}{j!}.$$

We say that a moment-generating function exists if there exists a positive constant b such that $M_X(t) < \infty$ for $t \leq b$. If it exists, it is unique and completely characterizes the distribution of the random variable X.

$$\begin{aligned} M_X(t) &= \sum_i e^{tx_i} P(x_i) \qquad \text{if } X \text{ is discrete;} \\ &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \qquad \text{if } X \text{ is continuous.} \end{aligned}$$

$$\frac{d^r M_X(t)}{dt^r}|_{t=0} = E(X^r).$$

Theorem 43 *If a and b are constant, then*

1. $M_{X+a}(t) = e^{at} M_X(t)$;
2. $M_{bX}(t) = M_X(bt)$;
3. $M_{a+bX}(t) = e^{at} M_X(bt)$.

Moment-generating function plays an important role in determining the probability distribution or density of a function of random variables when the function is a linear combination of n *independent* random variables. The method is based on the theorem that the moment-generating function of the sum of n independent random variables equals the product of their moment-generating function.

Theorem 44 *If $X_1, X_2, ...,$ and X_n are **independent** random variables and $Y = X_1 + X_2 + ... + X_n$, then*

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

where $M_{X_i}(t)$ is the value of the moment-generating function of X_i at t.

This suggests a simple approach to analyzing the distribution of independent sums. The difficulty is that the method is not universal, since the m.g.f. is not defined for every distribution. Considering the series expansion of e^{tx}, all the moments of X must evidently exist. The solution to this problem is to replace the variable t by it, where i is the imaginary number, $\sqrt{-1}$.

Definition 45 *The **characteristic function** of a random variable X is defined as*

$$\Phi_X(t) = E\left(e^{itX}\right) = \int_{-\infty}^{\infty} e^{itx} dF(x).$$

It is defined for any distribution because by the modulus inequality for complex random variables, we have

$$\begin{aligned} \left|E\left(e^{itX}\right)\right| &\leq E\left|e^{itX}\right| \\ &= E\left[\sqrt{(\cos(tX) + i\sin(tX))(\cos(tX) - i\sin(tX))}\right] \\ &= E\left(\sqrt{\cos^2(tX) + \sin^2(tX)}\right) = E\left(\sqrt{1}\right) = 1 < \infty. \end{aligned}$$

Note that the second step is utilizing the facts that $e^{i\theta} = \cos(\theta) + i\sin(\theta)$ and that the modulus of a complex number $a + bi$ is given by $|a + bi| = \sqrt{(a+bi)(a-bi)} = \sqrt{a^2 + b^2}$.

Theorem 46 *If $E\left(|X|^k\right) < \infty$, then*

$$\frac{d^k \Phi_X(t)}{dt^k}|_{t=0} = i^k E\left(X^k\right).$$

Theorem 47 *If a and b are constant, X and Y are independent random variables, then*

1. $\Phi_{a+bX}(t) = e^{iat}\Phi_X(bt)$;
2. $\Phi_{X+Y}(t) = \Phi_X(t)\Phi_Y(t)$.

Transformation of Random Variable:

Sometimes we would like to find the density function of a function of a particular variable X. Suppose the distribution function and density function of X are $F_X(x)$ and $f_X(x)$ respectively, what is the density function of $Y = g(X)$?

Theorem 48 *Let $Y = g(X)$ where $g(\cdot)$ is a strictly monotonic, differentiable function (ensuring the inverse function $X = g^{-1}(Y)$ exists), X is a continuous random variable with density function $f_X(x)$ and $g^{-1}(Y)$ is the inverse function for g. Then*

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|$$

where the last element is the Jacobian of the transformation.

Proof. Denote the capital X as a r.v., and small letter x be a particular value. We know that

$$\begin{aligned}\Pr(X \leq x) &= \Pr(g(X) \leq g(x)) \\ &= \Pr(Y \leq y).\end{aligned}$$

In other words

$$F_X(x) = F_Y(y).$$

Differentiate with respect to x and use the fact that F_X and F_Y must take non-negative values, we have:

$$f_X(x) = f_Y(y)\left|\frac{dy}{dx}\right|$$

or

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|.$$

Plug in $x = g^{-1}(y)$,

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right|. \blacksquare$$

Example 49 *If $X = \ln Y \sim N(\mu, \sigma^2)$, then $Y = \exp X$ will follow a log-normal distribution. To find its density function, note that $\left|\frac{dx}{dy}\right| = \left|\frac{1}{y}\right|$, $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$, $g^{-1}(y) = \ln y$. We have*

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dx}{dy}\right| = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln y - \mu}{\sigma}\right)^2\right)\frac{1}{y}.$$

Exercise 0.23 *Let Z_1, Z_2 be independent $N(0,1)$ random variables, let*

$$U = \min\{Z_1, \max\{Z_1, Z_2\}\}.$$

(a) What is the distribution of U?
(b) Find $E(U)$ and $Var(U)$.

Exercise 0.24 *Show that for a normally distributed random variable, the skewness and the degree of excess are zero.*

Exercise 0.25 *Suppose you know that, conditional upon Z, X is distributed as $N(0, Z)$. Find $E(X^2)$, $E(X^3)$, $E(X^4)$ and determine the degree of excess and the kurtosis in X if*

a) Z is a Poisson distributed random variable;
b) Z is a $U(0,1)$ random variable.

Exercise 0.26 *Let $Z = XY$ where $Y \sim N(0, \sigma_Y^2)$, the conditional expectation of X given Y is zero while the variance of X conditional upon Y is $2Y^2$. Find the probability that the variance Z, conditional upon Y, exceeds the unconditional variance.*

Exercise 0.27 *Let Z_1,..., Z_k, Z_{k+1} be independent $N(0,1)$ random variables, let*

$$U = Z_1^2 + Z_2^2 + Z_3^2 + ... + Z_{k-1}^2 + Z_k^2$$

a) What is the distribution of U? Find $E(U)$.

b) What are the distributions of $\frac{Z_{k+1}}{\sqrt{U/k}}$ and $\frac{Z_{k+1}^2}{U/k}$?

c) If we define another random variable $V = U - Z_{k+1}^2$, then V must have a Chi-square distribution with degrees of freedom $k-1$, true or false? Explain.

Exercise 0.28 *What is the density functions of $Y = X^2$ and $Y = (2X-1)^{1/3}$ if:*

a. $X \sim U(0,1)$,
b. $X \sim N(0,1)$.

Exercise 0.29 *Let X and Y be two independent standardized normal random variables. Show that*

i) $\text{Cov}(X, \max\{X,Y\}) = 0.5$.
ii) $\text{Cov}(X, \min\{X,Y\}) = 0.5$
iii) $\text{Cov}(\min\{X,Y\}, \max\{X,Y\}) = \frac{1}{\pi}$.
iv) $\text{Var}(\max\{X,Y\}) = 1 - \frac{1}{\pi}$.
v) $\text{Var}(\min\{X,Y\}) = 1 - \frac{1}{\pi}$.

Exercise 0.30 *True/False/Uncertain. Explain.*

(a) There exists a random variable X such that $E(X) = 2$ and $E(X^2) = 1$.

(b) The Chi-square distribution with 2 degrees of freedom is an exponential distribution with mean 2.

(c) If X follows a $N(0, \sigma^2)$ distribution, then $E(X^4) = 3\sigma^4$.

(d) For a random variable X, $Cov(X, X) = Var(X)$.

(e) The difference of two independent Chi-square random variables is still a Chi-square random variable.

(f) For an exponential random variable X with density $f(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)$, the median is larger than the mean.

Exercise 0.31 *True/False/Uncertain. Explain.*

(a) There exists a random variable X such that $E(X) > E(X^2)$.

(b) The $F(1, n)$ distribution will approach a Chi-square distribution with 1 degree of freedom when n tends to infinity.

(c) If X follows a $U(0, 1)$ distribution, then $E(X) > Var(X)$.

(d) For a random variable X, $Cov(-X, -X) = Var(X)$.

Exercise 0.32 *Let X be a continuous random variable which takes values in $(-\infty, \infty)$. Let $Y = \max\left\{X, \frac{1}{X}\right\}$, rewrite X in terms of Y.*

Exercise 0.33 *Let $\{X_i\}_{i=1}^n$ be independent $U(-0.5, 0.5)$ random variables. Let $X_{\{i\}}$ be the order statistic of i^{th} **order statistic** with $X_{\{1\}} \leq ... \leq X_{\{i\}} \leq ... \leq X_{\{n\}}$. Find the variance of the sample average $\overline{X}$ and the variance of the trimmed mean $T_1 = \frac{1}{n-1}\sum_{i=2}^{n} X_{\{i\}}$ and $T_2 = \frac{1}{n-2}\sum_{i=2}^{n-1} X_{\{i\}}$ in terms of n. Intuitively, which variance will be larger? Prove that $Var\left(\overline{X}\right)$ is smaller than the other two variances.*

Exercise 0.34 *Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be independent $N(0, 1)$ random variables. Let X and Y be independent of each other. Let $W_i = \max\{X_i, Y_i\}$, $Z_i = \max\left\{\frac{X_i - \overline{X}}{se(X)}, \frac{Y_i - \overline{Y}}{se(Y)}\right\}$, where $\overline{X}$ and $\overline{Y}$ are the sample average of X and Y respectively, $se(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$, $se(Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$. Let $\overline{W}$ and $\overline{Z}$ be the sample average of W and Z respectively. Find*

(a) $\mathrm{Var}\left(\overline{W}\right)$;

(b) $\mathrm{Var}\left(\overline{Z}\right)$.

Exercise 0.35 *Let X_1, X_2 be independent $N(0,1)$ random variables conditional upon U, which is uniform (0,1). If $Z = UX_1 + (1-U)X_2$, find*

a) The conditional distribution of Z given U.

b) $E(Z)$, $Var(Z)$.

Exercise 0.36 *Suppose you know that, conditional upon Z, X is distributed as $U(0,Z)$. If Z is a $U(0,1)$ random variable, find $E(X)$ and $E(X^2)$.*

Exercise 0.37 *Show that the probability generating function for of a Poisson random variable X with mean λ is*

$$P(t) = \exp(\lambda(t-1))$$

Find the mean and the variance of X via $P(t)$.

Exercise 0.38 *Explain why there can be no random variable X for which the moment generating function $M_X(t) = \dfrac{t}{1-t}$.*

Exercise 0.39 *Can there be a random variable X for which the moment generating function $M_X(t) = t$? Explain.*

Exercise 0.40 *Let Z_1, Z_2 be independent $N(0,1)$ random variables, let*

$$U = Z_1^2 + Z_2^2$$

a) What is the distribution of U?

b) Find $E(U)$ and $Var(U)$.

c) If we define another random variable $V = 2Z_1Z_2$, find $E(V)$ and $Var(V)$.

d) What is the distribution of $\dfrac{U+V}{2}$?

Exercise 0.41 *Describe how to generate an exponential distribution from a $U[0,1]$ distribution.*

Exercise 0.42 *Consider a random variable X whose density function is*

$$\begin{aligned} f(x) &= a_0 + a_1 x + a_2 x^2 \qquad \text{for x} \in [0,1] \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

a) What are the restrictions on a_0, a_1 and a_2 for $f(x)$ to be a density function.

b) Consider a random variable U, with

$$U = a_0 X + \frac{1}{2} a_1 X^2 + \frac{1}{3} a_2 X^3.$$

If U has a Uniform zero-one distribution, what is the distribution of X?

c) Describe how to use GAUSS to generate a random variable with density

$$\begin{aligned} f(x) &= \frac{1}{6} + x + x^2 \qquad \text{for } x \in [0,1]; \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

d) Describe how to use GAUSS to generate a random variable with density

$$\begin{aligned} f(x) &= 2 - x - 3x^2 + 2x^3 \qquad \text{for } x \in [0,1]; \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Exercise 0.43 *Let $X \sim U(0,1)$, suppose $g(X) \sim N(0,1)$, what is the functional form of $g(X)$?*

Exercise 0.44 *Find the lower limit of the following probability using the Chebychev inequality:*

i) $\Pr(-4 < x < 4)$ where x is a $N(0, 3^2)$;

ii) $\Pr(0 < x < 16)$ where x is a Chi-square random variable with 8 degrees of freedom.

Exercise 0.45 *Find the moment generating function of the uniform (0,1) random variable.*

Exercise 0.46 *For an exponential random variable X with density*

$$f(x) = \frac{1}{2}\exp\left(-\frac{x}{2}\right),$$

the median is larger than the mode. True or False? Explain.

Exercise 0.47 *Greene, Chapter 3, Exercises 14, 18, 25-30.*

ECO5120: Econometric Theory and Application, Fall 2006
Prof. T.L. Chong
HANDOUT 3
LARGE-SAMPLE THEORY

One of the objectives of econometrics is to estimate the unknown population parameters via various estimators. For example, we can estimate the population mean μ via the sample mean $\overline{X}_n$, where n is the sample size. In this case, $\overline{X}_n$ will be a random (stochastic) sequence in n, in the sense that its values are different from sample to sample. We want to examine if this random sequence converges to the true mean when n is very large. Before studying the random sequences, we first introduce the concept of deterministic sequences. A deterministic sequence b_n is basically a function of n, where n is a positive integer.

Definition 50 *A **real sequence** is a mapping from N to R, where $N = \{n : n = 1, 2, ...\}$ is the set of natural numbers, and R is the real line.*

Limits of Sequences

Let $\{c_n\}_{n=1}^{\infty}$ be a sequence of real numbers. The sequence is said to converge to a constant c if for any $\varepsilon > 0$, there exists an N such that $|c_n - c| < \varepsilon$ whenever $n \geq N$; This is indicated as

$$\lim_{n\to\infty} c_n = c.$$

or equivalently,

$$c_n \to c \text{ as } n \to \infty.$$

Example 51 *If* $c_n = \dfrac{1}{n}$, $\lim\limits_{n\to\infty} c_n = 0$.

Example 52 *If* $c_n = \sum_{i=1}^{n} \dfrac{1}{i}$, $\lim\limits_{n\to\infty} c_n = \infty$.

Proof. For $n > 2^k$, we have

$$c_n = \sum_{i=1}^{n} \frac{1}{i} = 1+\frac{1}{2}+\frac{1}{3}+...+\frac{1}{n} > 1+\frac{1}{2}+\left(\frac{1}{3}+\frac{1}{4}\right)+\left(\frac{1}{5}+...+\frac{1}{8}\right)+...+\left(\frac{1}{2^{k-1}+1}+...+\frac{1}{2^k}\right).$$

Now since

$$\frac{1}{2^{k-1}+1}+...+\frac{1}{2^k} = \sum_{i=1}^{2^{k-1}} \frac{1}{2^{k-1}+i} > \sum_{i=1}^{2^{k-1}} \frac{1}{2^{k-1}+2^{k-1}} = \frac{2^{k-1}}{2^k} = \frac{1}{2}.$$

Thus every term in the parenthesis is bigger than $\frac{1}{2}$ and we have

$$c_n > 1 + \frac{k}{2}.$$

Therefore, as $k \to \infty$, we have $n \to \infty$ and $c_n \to \infty$.

Example 53 *If* $c_n = \left(1+\frac{a}{n}\right)^n$, $\lim_{n\to\infty} c_n = \exp(a)$.

Definition 54 *A sequence of deterministic matrices* $\boldsymbol{C}_n$ *converges to* $\boldsymbol{C}$ *if each element of* $\boldsymbol{C}_n$ *converges to the corresponding element of* $\boldsymbol{C}$.

Example 55 *If* $\mathbf{C}_n = \begin{pmatrix} \frac{1}{n} & \frac{n}{e^n} \\ \frac{\ln n}{n} & \left(1-\frac{1}{n}\right)^n \end{pmatrix}$, *then* $\mathbf{C} = \begin{pmatrix} 0 & 0 \\ 0 & e^{-1} \end{pmatrix}$.

Definition 56 *The* ***supremum*** *of the sequence, denoted by*

$$\sup_{n\geq 1} c_n$$

is the **least upper bound (l.u.b.)** of the sequence, i.e., the smallest number, say, α, such that $c_n \leq \alpha$, for all n.

Definition 57 *The* ***infimum*** *of the sequence, denoted by*

$$\inf_{n\geq 1} c_n$$

is the **greatest lower bound (g.l.b.)** of the sequence, i.e., the largest number, say, α, such that $c_n \geq \alpha$, for all n.

Example 58 *If* $c_n = \dfrac{1}{n}$, *then* $\sup_{n\geq 1} c_n = 1$ *and* $\inf_{n\geq 1} c_n = 0$.

Definition 59 *The sequence* $\{c_n\}_{n=1}^{\infty}$ *is said to be a* ***monotone non-increasing*** *sequence if*

$$c_{n+1} \leq c_n, \text{ for all } n.$$

and it is said to be a **monotone non-decreasing** sequence if

$$c_{n+1} \geq c_n, \text{ for all } n.$$

Definition 60 *Let* $\{c_n\}_{n=1}^{\infty}$ *be a sequence of real numbers and let*

$$\begin{aligned} a_n &= \sup_{k\geq n} c_k, \\ b_n &= \inf_{k\geq n} c_k. \end{aligned}$$

Then, the sequences $\{a_n\}$, $\{b_n\}$ are, respectively, monotone non-increasing and non-decreasing, and their limits are said to be the limit superior and limit inferior of the original sequence and are denoted, respectively, by

$$\limsup, \quad \liminf, \quad \text{or } \overline{\lim}, \ \underline{\lim}.$$

Thus we write

$$\begin{aligned} \lim_{n\to\infty} a_n &= \lim_{n\to\infty} \sup_{k\geq n} c_k, \\ \lim_{n\to\infty} b_n &= \lim_{n\to\infty} \inf_{k\geq n} c_k. \end{aligned}$$

The limsup is the eventual upper bound of a sequence, and the liminf is the eventual lower bound of a sequence.

Proposition 61 *Let* $\{c_n\}_{n=1}^{\infty}$ *be a sequence of real numbers, then*

$$\limsup c_n \geq \liminf c_n.$$

Definition 62 *Let* $\{c_n\}_{n=1}^{\infty}$ *be a sequence of real numbers, then its limit exists if and only if*

$$\limsup c_n = \liminf c_n.$$

Example 63 *If* $c_n = (-1)^n$, *then* $\limsup c_n = 1$, $\liminf c_n = -1$. *The limit does not exist.*

Example 64 *If* $c_n = \left(-\frac{1}{n}\right)^n$, *then* $\limsup c_n = 0$, $\liminf c_n = 0$. *The limit exists and is equal to zero.*

Rates of Convergence

Definition 65 *The sequence* $\{b_n\}$ *is* ***at most of order*** $\boldsymbol{n^{\lambda}}$, *denoted* $O\left(n^{\lambda}\right)$, *if and only if for* ***some*** *real number* Δ, $0 < \Delta < \infty$, *there exists a finite integer* N *such that for all* $n \geq N$, $\left|\frac{b_n}{n^{\lambda}}\right| < \Delta$. *In other words,* $\{b_n\}$ *is* $O\left(n^{\lambda}\right)$ *if* $\frac{b_n}{n^{\lambda}}$ *is eventually bounded as* n *becomes large.*

Definition 66 *The sequence* $\{b_n\}$ *is* ***of order smaller than*** $\boldsymbol{n^{\lambda}}$, *denoted* $o\left(n^{\lambda}\right)$, *if and only if for* ***every*** *real number* δ, $0 < \delta < \infty$, *there exists a finite integer* $N(\delta)$ *such that for all* $n \geq N(\delta)$, $\left|\frac{b_n}{n^{\lambda}}\right| < \delta$. *i.e.* $b_n = o\left(n^{\lambda}\right)$ *if* $\frac{b_n}{n^{\lambda}} \to 0$ *as* n *becomes large.*

Obviously, if $\{b_n\}$ is $o\left(n^{\lambda}\right)$, then $\{b_n\}$ is $O\left(n^{\lambda}\right)$.

In particular, $\{b_n\}$ is $O(1)$ if b_n is eventually bounded, where as $\{b_n\}$ is $o(1)$ if $b_n \to 0$.

Example 67 *Let* $b_n = 4 + 2n + 6n^2$. *Then* $\{b_n\}$ *is* $O(n^2)$ *and* $o\left(n^{2+\delta}\right)$ *for every* $\delta > 0$.

Example 68 *Let* $b_n = (-1)^n$. *Then* $\{b_n\}$ *is* $O(1)$ *and* $o\left(n^{\delta}\right)$ *for every* $\delta > 0$.

Example 69 *Let* $b_n = \exp(n)$.*In this case,* $b_n = O(\exp(n))$.

Proposition 70 *Let* $\{a_n\}_{n=1}^{\infty}$ *and* $\{b_n\}_{n=1}^{\infty}$ *be sequences of real number.*

(*i*) If $\{a_n\}$ is $O\left(n^{\lambda}\right)$ and $\{b_n\}$ is $O(n^{\mu})$, then $a_n b_n$ is $O\left(n^{\lambda+\mu}\right)$ and $(a_n + b_n)$ is $O(n^{\kappa})$, where $\kappa = \max[\lambda, \mu]$.

(*ii*) If $\{a_n\}$ is $o\left(n^{\lambda}\right)$ and $\{b_n\}$ is $o(n^{\mu})$, then $a_n b_n$ is $o\left(n^{\lambda+\mu}\right)$ and $(a_n + b_n)$ is $o(n^{\kappa})$.

(*iii*) If $\{a_n\}$ is $O\left(n^{\lambda}\right)$ and $\{b_n\}$ is $o(n^{\mu})$, then $a_n b_n$ is $o\left(n^{\lambda+\mu}\right)$ and $(a_n + b_n)$ is $O(n^{\kappa})$.

Definition 71 *The stochastic sequence* $\{b_n(\omega)\}$ *is* ***at most of order*** n^{λ} ***in probability****, denoted* $O_p\left(n^{\lambda}\right)$*, if there exists an* $O(1)$ *nonstochastic sequence* a_n *such that* $\dfrac{b_n(\omega)}{n^{\lambda}} - a_n \xrightarrow{p} 0$.

When a sequence $\{b_n(\omega)\}$ is $O_p\left(n^{\lambda}\right)$, we say it is **bounded in probability**.

Definition 72 *The sequence* $\{b_n(\omega)\}$ *is* ***of order smaller than*** n^{λ} ***in probability****, denoted* $o_p\left(n^{\lambda}\right)$*, if* $\dfrac{b_n(\omega)}{n^{\lambda}} \xrightarrow{p} 0$.

Proposition 73 *Let* $\{a_n\}_{n=1}^{\infty}$ *and* $\{b_n\}_{n=1}^{\infty}$ *be sequences of random number.*

(*i*) If $\{a_n\}$ is $O_p\left(n^{\lambda}\right)$ and $\{b_n\}$ is $O_p\left(n^{\mu}\right)$, then $a_n b_n$ is $O_p\left(n^{\lambda+\mu}\right)$ and $(a_n + b_n)$ is $O_p\left(n^{\kappa}\right)$, where $\kappa = \max[\lambda, \mu]$;

(*ii*) If $\{a_n\}$ is $o_p\left(n^{\lambda}\right)$ and $\{b_n\}$ is $o_p\left(n^{\mu}\right)$, then $a_n b_n$ is $o_p\left(n^{\lambda+\mu}\right)$ and $(a_n + b_n)$ is $o_p\left(n^{\kappa}\right)$;

(*iii*) If $\{a_n\}$ is $O_p\left(n^{\lambda}\right)$ and $\{b_n\}$ is $o_p\left(n^{\mu}\right)$, then $a_n b_n$ is $o_p\left(n^{\lambda+\mu}\right)$ and $(a_n + b_n)$ is $O_p\left(n^{\kappa}\right)$.

Various Modes of Convergence

Let X_n be a sequence of random variables and c be a constant.

Definition 74 *X_n is said to **converge to** c **in probability** if*

$$\lim_{n\to\infty} \Pr\left(|X_n - c| > \epsilon\right) = 0$$

for any $\epsilon > 0$. We write $X_n \xrightarrow{p} c$ or $\text{plim} X_n = c$.

Example 75 *Let X_n be a sequence of random variables such that X_n equals either 0 or n, with probability $\left(1 - \frac{1}{n}\right)$ and $\frac{1}{n}$ respectively. Then $X_n \xrightarrow{p} 0$.*
Proof.

$$\begin{aligned}
\Pr\left(|X_n - 0| > \epsilon\right) &= \Pr\left(|X_n| > \epsilon\right) \\
&= \Pr\left(0 > \epsilon\right)\Pr\left(X_n = 0\right) + \Pr\left(n > \epsilon\right)\Pr\left(X_n = n\right) \\
&= 0 \times \left(1 - \frac{1}{n}\right) + \Pr\left(n > \epsilon\right)\frac{1}{n} \\
&= \frac{\Pr\left(n > \epsilon\right)}{n}. \\
\lim_{n\to\infty} \Pr\left(|X_n - 0| > \epsilon\right) &= \lim_{n\to\infty} \frac{\Pr\left(n > \epsilon\right)}{n} = 0. \blacksquare
\end{aligned}$$

In the rest of this handout, we let a, b, and c be constants and $g(\cdot)$ a real-valued continuous function taking finite values.

Theorem 76 *If* $X_n \xrightarrow{p} c$, *then* $g(X_n) \xrightarrow{p} g(c)$.

Proof. (exercise).

Theorem 77 *If* $X_n \xrightarrow{p} a$, $Y_n \xrightarrow{p} b$, *then*

(i) $X_n + Y_n \xrightarrow{p} a + b$;
(ii) $X_n Y_n \xrightarrow{p} ab$;
(iii) $\dfrac{X_n}{Y_n} \xrightarrow{p} \dfrac{a}{b}$ if $b \neq 0$.

Proof. (exercise).

Definition 78 *If* X_n *has a c.d.f.* $F_n(x)$, *it* ***converges in distribution*** *to a random variable* X *with cumulative distribution function* $F(x)$ *if*

$$\lim_{n\to\infty} |F_n(x) - F(x)| = 0$$

at *all continuity points* of $F(x)$. We say that $X_n \xrightarrow{d} X$.

Generally speaking, if the distance between $F_n(x)$ and $F(x)$ converges to zero in the domain of X, we say that X_n converge in distribution to X. The reason for saying "at all continuity points" is to allow the case where the distance between $F_n(x)$ and $F(x)$ does not converge to zero at some points. Consider the following example.

Example 79 *Let*

$$\begin{aligned} F(x) &= 0 \qquad x < 0 \\ &= 1 \qquad x \geq 0 \end{aligned}$$

and let

$$\begin{aligned} F_n(x) &= 0 \quad x < -\frac{1}{n} \\ &= \frac{1}{2} + \frac{n}{2}x \quad -\frac{1}{n} \leq x \leq \frac{1}{n} \\ &= 1 \quad x > \frac{1}{n}. \end{aligned}$$

In this case, note that $F(x)$ is discontinuous at $x = 0$. The distance between $F_n(x)$ and $F(x)$ does not converge to zero at $x = 0$ since $F_\infty(0) = \frac{1}{2}$ and $F(0) = 1$. By adding the phase "at all continuity points", we exclude the point $x = 0$ and the definition above allows X_n to converge in distribution to X.

Thus, the limiting distribution of a sequence of random variable, if exists, cannot in general be determined by the limit of the c.d.f..

Theorem 80 ***(Mann and Wald)*** *Let $g(\cdot)$ be a continuous function, if $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.*

Proof. (exercise).

Example 81 *Let X_n be a random variable which follows a t-distribution with degrees of freedom n and let $g(x) = x^2$. Since*

$$g(X_n) = t_n^2 = F(1, n),$$

$$X_n \xrightarrow{d} X = N(0, 1)$$

and

$$g(X) = [N(0, 1)]^2 = \chi^2(1),$$

the above theorem says that

$$F(1, n) \xrightarrow{d} \chi^2(1).$$

Theorem 82 *If $X_n \xrightarrow{d} X$, and $Y_n \xrightarrow{p} 0$, then $X_n Y_n \xrightarrow{p} 0$.*

Proof. (exercise).

Theorem 83 ***(Slutsky)*** *If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} c$, then*

(i) $X_n + Y_n \xrightarrow{d} X + c$;
(ii) $X_n Y_n \xrightarrow{d} cX$;
(iii) $\dfrac{X_n}{Y_n} \xrightarrow{d} \dfrac{X}{c}$ if $c \neq 0$.

Proof.

(i) Since $X_n \xrightarrow{d} X$, we have $X_n + c \xrightarrow{d} X + c$. Further, $X_n + Y_n - (X_n + c) = Y_n - c \xrightarrow{p} 0$, thus $X_n + Y_n$ and $X_n + c$ have the same limiting distribution which is $X + c$.

(ii) Since $X_n \xrightarrow{d} X$, we have $cX_n \xrightarrow{d} cX$. Further, $X_n Y_n - cX_n = X_n (Y_n - c) \xrightarrow{p} 0$. Thus $X_n Y_n$ and cX_n have the same limiting distribution which is cX.

(iii) Since $X_n \xrightarrow{d} X$, we have $\dfrac{X_n}{c} \xrightarrow{d} \dfrac{X}{c}$. Further, $\dfrac{X_n}{Y_n} - \dfrac{X_n}{c} = X_n \left(\dfrac{1}{Y_n} - \dfrac{1}{c} \right) \xrightarrow{p} 0$. Thus $\dfrac{X_n}{Y_n}$ and $\dfrac{X_n}{c}$ have the same limiting distribution which is $\dfrac{X}{c}$.

Definition 84 *A sequence of random variables X_n are an* ***Independent and Identical Distributed (i.i.d.)*** *if all the X_n have the same distribution and X_i does not depend on X_j for any $i \neq j$.*

Theorem 85 ***Weak Law of Large Numbers (Linchine)***

If $\{X_i\}_{i=1}^n$ are i.i.d. with finite mean μ and finite variance σ^2, the sample average $\overline{X} = \dfrac{1}{n} \sum_{i=1}^n X_i$ converges to the true mean μ as the sample size n goes to infinity.

Proof. By Chebyshev's inequality,

$$\Pr\left(\left|\overline{X}-\mu\right| \geq k\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2}.$$

Putting $\varepsilon = \dfrac{k\sigma}{\sqrt{n}}$,

$$\begin{aligned}\Pr\left(\left|\overline{X}-\mu\right| \geq \varepsilon\right) &\leq \frac{\sigma^2}{\varepsilon^2 n} \\ &\to 0 \quad \text{as } n \to \infty \text{ for any } \varepsilon > 0.\end{aligned}$$

Thus, $\overline{X} \xrightarrow{p} \mu$. ■

We call it a weak law as we make a very strong assumption that X_n are i.i.d.. Remember, the power of a law or a theorem depends on the assumptions that you make, the weaker the assumptions (the less you assume), the higher the power of your theorem.

Theorem 86 ***Central Limit Theorem (Lindeberg-Lévy)***

If $\{X_i\}_{i=1}^n$ is an i.i.d. sequence with mean μ and variance σ^2, then as $n \to \infty$,

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i \xrightarrow{d} N(0,1).$$

where $Y_i = \dfrac{X_i - \mu}{\sigma}$.

Proof. (exercise).

This is the simplest version of the Central Limit Theorem, which states that if $\{X_i\}_{i=1}^n$ are i.i.d. with finite mean μ and finite variance σ^2, the sample average $\overline{X}$ converges in distribution to a normal distribution with mean μ and variance $\dfrac{\sigma^2}{n}$, as the sample size n goes to infinity. It is a **powerful** theorem because X_i can be **any** distributions. Most of the statistical inference and hypothesis testing are based on this theorem.

Note that we assume total independence and only require the existence of the first and second moments in the above simplest versions. There are many different versions of Law of Large Numbers and Central Limit Theorem generated from the trade-off between degrees of dependence and the moment requirements. In other words, we may allow X_i and X_j to be slightly dependent, but we may require the existence of higher moments of X, i.e., if we permit X to be dependent, we may need to assume $E(X^r) < \infty$ for some $r > 2$.

Example 87 *Let X_1 and X_2 be two independent random variables distributed as*

$$\Pr(X_i = -1) = \Pr(X_i = 1) = \frac{1}{2}$$

where $i = 1, 2$.

Then the distribution of

$$\overline{X} = \frac{X_1 + X_2}{2}$$

will be

$$\begin{aligned} \Pr\left(\overline{X} = -1\right) &= \Pr(X_1 = -1 \text{ and } X_2 = -1) \\ &= \Pr(X_1 = -1)\Pr(X_2 = -1) \\ &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}. \end{aligned}$$

$$\begin{aligned} \Pr\left(\overline{X} = 0\right) &= \Pr(\{X_1 = -1 \text{ and } X_2 = 1\} \text{ or } \{X_1 = 1 \text{ and } X_2 = -1\}\ ) \\ &= \Pr(X_1 = -1)\Pr(X_2 = 1) + \Pr(X_1 = 1)\Pr(X_2 = -1) \\ &= \frac{1}{2}. \end{aligned}$$

$$\begin{aligned}\Pr\left(\overline{X}=1\right) &= \Pr\left(X_1=1 \text{ and } X_2=1\right)\\ &= \Pr\left(X_1=1\right)\Pr\left(X_2=1\right)\\ &= \frac{1}{2}\times\frac{1}{2}=\frac{1}{4}.\end{aligned}$$

Note that although X_1 and X_2 are evenly distributed, $\overline{X}$ is not evenly distributed but has a bell-shape distribution. As the number of observations tends to infinity, $\overline{X}$ will have a normal distribution.

More demanding materials

We discuss the case where X_n converges in probability to a constant c in the previous section. Actually, the concept of convergence in probability is more than that. X_n can also be converge in probability to a random variable X. The following theorem states that if X_n converges in probability to a random variable X, than it must converge in distribution to X.

Theorem 88 $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$.

Proof. For $\epsilon > 0$, we have

$\Pr\left(X_n \leq x\right)$
$= \Pr\left(\{X_n \leq x\} \cap \{|X_n - X| \leq \epsilon\}\right) + \Pr\left(\{X_n \leq x\} \cap \{|X_n - X| > \epsilon\}\right)$
$\leq \Pr\left(X \leq x+\epsilon\right) + \Pr\left(|X_n - X| > \epsilon\right),$

where the events whose probabilities appear on the right-hand side of the inequality contain the corresponding events on the left.

$\Pr\left(|X_n - X| > \epsilon\right) \to 0$ by hypothesis, and hence

$$\limsup \Pr\left(X_n \leq x\right) \leq \Pr\left(X \leq x+\epsilon\right).$$

Similarly

$\Pr\left(X \leq x-\epsilon\right)$
$= \Pr\left(\{X \leq x-\epsilon\} \cap \{|X_n - X| \leq \epsilon\}\right) + \Pr\left(\{X \leq x-\epsilon\} \cap \{|X_n - X| > \epsilon\}\right)$

$\leq \Pr(X_n \leq x) + \Pr(|X_n - X| > \epsilon)$,
and so

$$\Pr(X \leq x - \epsilon) \leq \liminf \Pr(X_n \leq x).$$

Since $\liminf \Pr(X_n \leq x) \leq \limsup \Pr(X_n \leq x)$, we have

$$\Pr(X \leq x - \epsilon) \leq \liminf \Pr(X_n \leq x) \leq \limsup \Pr(X_n \leq x) \leq \Pr(X \leq x + \epsilon).$$

Further, not also that

$$\Pr(X \leq x - \epsilon) \leq \Pr(X \leq x) \leq \Pr(X \leq x + \epsilon).$$

Since ϵ is arbitrary, we let it go to zero such that

$$\Pr(X \leq x) = \liminf \Pr(X_n \leq x) = \limsup \Pr(X_n \leq x).$$

Since

$$\lim \Pr(X_n \leq x) = \Pr(X \leq x)$$

Thus the limiting distribution of X_n exists and is the same as that of X.

■

Note that convergence in probability implies convergence in distribution, but the converse does not necessarily hold. The following is the relationship among the four concepts of convergence.

Relation between four modes of convergence

$$\begin{array}{ccccc} & & a.s. & & \\ & & \downarrow & & \\ r.m. & \rightarrow & p & \rightarrow & d \end{array}$$

Theorem 89 *If $X_n \xrightarrow{d} X$, and $|Y_n - X_n| \xrightarrow{p} 0$, then the limiting distribution of Y_n exists and is the same as that of X_n.*

Proof. Let $\epsilon > 0$,

$$\begin{aligned}
&\Pr(Y_n \le x) = \Pr(Y_n + X_n \le x + X_n) = \Pr(X_n \le x + X_n - Y_n) \\
&= \Pr(\{X_n \le x + X_n - Y_n, |X_n - Y_n| < \epsilon\} \cup \{X_n \le x + X_n - Y_n, |X_n - Y_n| \ge \epsilon\}) \\
&\le \Pr(X_n \le x + X_n - Y_n, |X_n - Y_n| < \epsilon) + \Pr(X_n \le x + X_n - Y_n, |X_n - Y_n| \ge \epsilon) \\
&\le \Pr(X_n \le x + \epsilon) + \Pr(|X_n - Y_n| \ge \epsilon).
\end{aligned}$$

$$\begin{aligned}
&\limsup \Pr(Y_n \le x) \\
&\le \limsup \Pr(X_n \le x + \epsilon) + \limsup \Pr(|X_n - Y_n| \ge \epsilon) \\
&= \lim \Pr(X_n \le x + \epsilon) + 0 \\
&= \Pr(X \le x + \epsilon).
\end{aligned}$$

Similarly

$$\begin{aligned}
&\Pr(X_n \le x - \epsilon) = \Pr(X_n + Y_n \le x - \epsilon + Y_n) = \Pr(Y_n \le x - \epsilon + Y_n - X_n) \\
&= \Pr(\{Y_n \le x - \epsilon + Y_n - X_n, |Y_n - X_n| < \epsilon\} \cup \{Y_n \le x - \epsilon + Y_n - X_n, |Y_n - X_n| \ge \epsilon\}) \\
&\le \Pr(Y_n \le x - \epsilon + Y_n - X_n, |Y_n - X_n| < \epsilon) + \Pr(Y_n \le x - \epsilon + Y_n - X_n, |Y_n - X_n| \ge \epsilon) \\
&\le \Pr(Y_n \le x) + \Pr(|Y_n - X_n| \ge \epsilon).
\end{aligned}$$

$$\begin{aligned}
\liminf \Pr(X_n \le x - \epsilon) &\le \liminf \Pr(Y_n \le x) + \liminf \Pr(|Y_n - X_n| \ge \epsilon) \\
\lim \Pr(X_n \le x - \epsilon) &\le \liminf \Pr(Y_n \le x) + 0 \\
\Pr(X \le x - \epsilon) &\le \liminf \Pr(Y_n \le x).
\end{aligned}$$

Thus

$$\Pr(X \le x - \epsilon) \le \liminf \Pr(Y_n \le x) \le \limsup \Pr(Y_n \le x) \le \Pr(X \le x + \epsilon)$$

Let $\epsilon \to 0$,

$$\lim \Pr(Y_n \le x) = \Pr(X \le x)$$

Thus the limiting distribution of Y_n exists and is the same as that of X_n.

■

Definition 90 *A sequence of random variables* $\{X_n\}$ *is said to converge to* c ***almost surely*** *if*

$$\Pr\left(\lim_{n\to\infty} X_n = c\right) = 1,$$

or equivalently, for every $\epsilon > 0$,

$$\lim_{N\to\infty} \Pr\left(\sup_{n\geq N} |X_n - c| > \epsilon\right) = 0.$$

We write $X_n \xrightarrow{a.s.} c$.

In other words, let $\Phi \subseteq \Omega$ be the set of outcomes such that, for every $\omega \in \Phi$, $X_n(\omega) \to c$ as $n \to \infty$. If the probability measure $\Pr(\Phi) = 1$, the sequence is said to converge almost surely, or converge with probability one.

Theorem 91 $X_n \xrightarrow{a.s.} c \Rightarrow X_n \xrightarrow{p} c$.

Proof. Since

$$|X_N - c| > \epsilon \Rightarrow \sup_{n\geq N} |X_n - c| > \epsilon,$$

we have

$$\begin{aligned} \Pr(|X_N - c| > \epsilon) &\leq \Pr\left(\sup_{n\geq N} |X_n - c| > \epsilon\right) \\ \lim_{N\to\infty} \Pr(|X_N - c| > \epsilon) &\leq \lim_{N\to\infty} \Pr\left(\sup_{n\geq N} |X_n - c| > \epsilon\right) \end{aligned}$$

Since both are non-negative, we have

$$\lim_{N\to\infty} \Pr\left(\sup_{n\geq N} |X_n - c| > \epsilon\right) = 0$$

implies

$$\lim_{N\to\infty} \Pr(|X_N - c| > \epsilon) = 0.$$

However, the converse is not necessarily true. Thus, almost sure convergence implies convergence in probability. A sequence that converges in probability always contains a subsequence that converges almost surely. Convergence in probability allows more erratic behavior in the converging sequence than almost sure convergence, and by simply disregarding the erratic elements of the sequence we can obtain an almost surely convergent subsequence.

Example 92 *Convergence in probability does not imply convergence almost surely, to give a counter-example, let the sample space be* $\Omega = [0,1]$.

Let

$$\Omega_n = \left[0, A_{n-1} + \frac{1}{n} - 1\right] \cup \left[A_{n-1}, \min\left\{1, A_{n-1} + \frac{1}{n}\right\}\right]$$

where
$A_0 = 0$,
$A_n = \left(\sum_{i=1}^{n} \frac{1}{i}\right) - \text{integer}\left(\sum_{i=1}^{n} \frac{1}{i}\right)$ for $n \geq 1$.
e.g.
$A_1 = \left(\sum_{i=1}^{1} \frac{1}{i}\right) - \text{integer}\left(\sum_{i=1}^{1} \frac{1}{i}\right) = 1 - 1 = 0.$
$A_2 = \left(\sum_{i=1}^{2} \frac{1}{i}\right) - \text{integer}\left(\sum_{i=1}^{2} \frac{1}{i}\right) = 1\frac{1}{2} - 1 = \frac{1}{2}.$

$$\Omega_1 = \left[0, A_0 + \frac{1}{1} - 1\right] \cup \left[A_0, \min\left\{1, A_0 + \frac{1}{1}\right\}\right]$$
$$= [0,0] \cup [0, \min\{1,1\}] = [0,1].$$

$$\Omega_2 = \left[0, A_1 + \frac{1}{2} - 1\right] \cup \left[A_1, \min\left\{1, A_1 + \frac{1}{2}\right\}\right]$$
$$= \left[0, -\frac{1}{2}\right] \cup \left[0, \min\left\{1, \frac{1}{2}\right\}\right] = \phi \cup \left[0, \frac{1}{2}\right] = \left[0, \frac{1}{2}\right].$$

$$\Omega_3 = \left[0, A_2 + \frac{1}{3} - 1\right] \cup \left[A_2, \min\left\{1, A_2 + \frac{1}{3}\right\}\right]$$
$$= \left[0, \frac{1}{2} + \frac{1}{3} - 1\right] \cup \left[\frac{1}{2}, \min\left\{1, \frac{1}{2} + \frac{1}{3}\right\}\right] = \phi \cup \left[\frac{1}{2}, \min\left\{1, \frac{5}{6}\right\}\right] = \left[\frac{1}{2}, \frac{5}{6}\right],$$

and so on.

$$\begin{aligned} X_n(\omega) &= 1 \qquad \text{for } \omega \in \Omega_n, \\ X_n(\omega) &= 0 \qquad \text{otherwise.} \end{aligned}$$

e.g.,

$$X_1(0) = 1, X_2(0) = 1, X_3(0) = 0, X_4(0) = 1, X_5(0) = 0, ...$$

$$X_1\left(\frac{1}{12}\right) = 1, X_2\left(\frac{1}{12}\right) = 1, X_3\left(\frac{1}{12}\right) = 0, X_4\left(\frac{1}{12}\right) = 1, X_5\left(\frac{1}{12}\right) = 1, ...$$

$$X_1(1) = 1, X_2(1) = 0, X_3(1) = 0, X_4(1) = 1, X_5(1) = 0, ...$$

We have for $0 < \varepsilon < 1$,

$$\Pr(|X_n| > \varepsilon) = \frac{1}{n}$$

and

$$\lim_{n\to\infty} \Pr(|X_n - 0| > \varepsilon) = 0.$$

Thus X_n converges in probability to zero.

However, X_n does not converge almost surely to zero. To see this, note that Ω_n keeps moving to the right until, passes through the point $\omega = 1$ and starts again from $\omega = 0$. Thus, for any given ω and n, there exists a finite constant $N > n$ such that $X_N(\omega) = 1$.

Thus

$$\Pr\left\{\lim_{n\to\infty} X_n(\omega) = 0\right\} = 0$$

and X_n does not converge to 0 almost surely since

$$\Pr\left(\lim_{n\to\infty} X_n = 0\right) \neq 1.$$

Definition 93 *X_n is said to* ***converge in the r^{th} mean*** *to c if*

$$\lim_{n\to\infty} E\left(|X_n - c|^r\right) = 0$$

for some $r > 0$. We write $X_n \overset{r.m.}{\to} c$.

When $r = 2$, the convergence is said to occur in **quadratic mean**, denoted $X_n \overset{q.m.}{\to} c$.

A useful property of convergence in the r^{th} mean is that it implies convergence in the s^{th} mean for $s < r$.

Theorem 94 *If $X_n \overset{r.m.}{\to} c$ and $r > s$, then $X_n \overset{s.m.}{\to} c$.*

Proof. Let $g(z) = z^q$, $q < 1$, $z \geq 0$. Then g is concave. set $z = |X_n - c|$ and $q = \dfrac{s}{r}$. From Jensen's inequality,

$$E\left(|X_n - c|^s\right) = E\left(\left\{|X_n - c|^r\right\}^{s/r}\right) \leq \left\{E\left(|X_n - c|^r\right)\right\}^{s/r}.$$

Since $E\left(|X_n - c|^r\right) \to 0$, it follows that $E\left(|X_n - c|^s\right) \to 0$, $X_n \overset{s.m.}{\to} c$. ■

Convergence in the r^{th} mean is a stronger convergence concept than convergence in probability, and in fact implies convergence in probability. To show this, we use the generalized Chebyshev inequality.

Theorem 95 *If $X_n \overset{r.m.}{\to} c$ for some $r > 0$, then $X_n \overset{p}{\to} c$.*

Proof. Let $Z = X_n - c$, and apply the Generalized Chebyshev inequality, for every $\epsilon > 0$,

$$\Pr\left(|X_n - c| \geq \epsilon\right) \leq \frac{E\left|X_n - c\right|^r}{\epsilon^r}$$

Since $X_n \xrightarrow{r.m.} c$, we have $E|X_n - c|^r \to 0$, and as a result $\Pr(|X_n - c| \geq \epsilon) \to 0$. Thus we have $X_n \xrightarrow{p} c$. ■

Exercise 0.48 *True/False.*

(a) There exists a moment generating function $M(t) = t + 1$.

(b) If X follows a $N(0, \sigma^2)$ distribution, then $E(X^3) = E(X^5)$.

(c) For any random variables X and Y, $Cov(X, Cov(X, Y)) = 0$.

(d) $\lim_{n\to\infty} 2 = \infty$.

(e) $\lim_{n\to\infty} \sum_{j=0}^{n} \frac{j}{2^{j+1}} = 1$.

(f) The sequence $c_n = n^2 - 10n + 25$ is a monotone non-decreasing sequence.

(g) Let U_n be a Chi-square random variable with n degrees of freedom, then $\frac{U_n}{n} = O_p(1)$.

(h) $O_p(1) - O_p(1) = 0$.

(i) If X_n follows a uniform distribution $U\left(0, \frac{1}{n}\right)$, then X_n converges in probability to zero.

(j) Convergence in distribution implies convergence in probability .

(k) The Weak Law of Large Number holds for all random variables.

(l) Convergence in probability implies almost sure convergence.

(m) An estimator is consistent if it is asymptotically unbiased.

Exercise 0.49 *The probability generating function of a discrete random variable X is given by*

$$P(t) = E\left(t^X\right) = \sum_{j=0}^{\infty} t^j \Pr(X = j).$$

If a random variable X follows a distribution

$$\Pr(X = j) = \frac{1}{2^{j+1}} \qquad \text{for } j = 0, 1, 2, ...$$

(a) Find the probability generating function.

(b) Find the mean and the variance of X via the probability generating function.

Exercise 0.50 *For a Uniform random variable $U(0, a)$, the variance is smaller than a. True or False? Explain.*

Exercise 0.51 *Let X be a continuous random variable which takes values in $(0, \infty)$. Let $Y = \max\left\{X + \frac{1}{X}, 1\right\}$, rewrite X in terms of Y.*

Exercise 0.52 *To show the Law of Large Number, consider the random experiment of throwing a dice T times. Let X_t be the outcome at the t trial, $t = 1, 2, .., T$. Let $\overline{X}$ be the sample average of these X_t.*

(a) What is the population mean of the outcome for throwing a dice infinite number of times?

(b) What possible values will $\overline{X}$ take if $T = 1$? $T = 2$? $T = 3$?

(c) Try the experiment yourself, record the value of $\overline{X}$ and plot a diagram to indicates its behavior as T getting large from 1 to 30. Does $\overline{X}$ converge to 3.5?

Exercise 0.53 *To show the Central limit Theorem, let use consider the random experiment in the previous exercise of throwing a dice T times.*

(a) Try the experiment yourself, by using $T = 30$. Record the value of $\overline{X}$.

(b) Throw the dice for another 30 times, record the value of $\overline{X}$, does the value of $\overline{X}$ different from the previous one?

(c) Keep repeating part (d) until you collects 20 values of $\overline{X}$, i.e. you have 18 more rounds to go.

(d) Plot the histogram (the frequency diagram) of $\overline{X}$ for the range 0 to 6, with each increment equal 0.1.

(e) Repeat part (d) by finding another 4 classmates and pool the result of 100 values of $\overline{X}$. If you do not want to approach other classmates, do the experiment yourself 100 times then.

Exercise 0.54 *Use GAUSS to generate* 36 *random numbers from the uniform distribution* $U(0,1)$*; calculate the sample mean, and repeat this procedure* 100 *times. Thus you will have 100 sample means, say,* $\overline{X}_1$, $\overline{X}_2$,...,$\overline{X}_{100}$. *Define a variable* $Y_t = \sqrt{36}\left(\overline{X}_t - 0.5\right)$, $t = 1, 2, ..., 100$. *Now make two frequency tables of* Y_t *with the length of each interval* 0.01 *and* 0.1 *respectively. Plot the two histograms.*

Exercise 0.55 *Suppose* X_t $(t = 1, 2, ..., 100)$ *is a discrete random variable which takes* 0 *with probability* $\frac{1}{2}$*, and* 1 *with probability* $\frac{1}{2}$*. Write a simple GAUSS procedure to generate* X_t*. (Hint: think of how to generate such a discrete random variable from a uniform (0,1) distribution.)*

Exercise 0.56 *True/False/Uncertain. Explain.*

(a) $O_p(1) \times o_p(1) = o_p(1)$;

(b) $\lim_{n\to\infty}\left(\dfrac{2n + \ln 2}{2n - \ln 2}\right)^{-n} = 2$;

(c) The Central Limit Theorem states that the sample average of an. i.i.d. random sequence with finite mean and finite variance has a normal distribution when the sample size goes to infinity;

(d) Convergence in probability implies convergence in distribution;

(e) The sequence $c_n = \dfrac{e^n}{n}$ is a monotone non-increasing sequence.

(f) The sequence $c_n = n^2 - 10n$ is a monotone non-decreasing sequence.

(g) Convergence in probability implies almost sure convergence.

(h) The Weak Law of Large Number states that the sample average of an. i.i.d. random sequence has a normal distribution when the sample size goes to infinity.

Exercise 0.57 *Find the limits of the following sequences*

(a) $\lim_{n\to\infty} \left(1 + \dfrac{\ln 2}{n}\right)^n$;

(b) $\lim_{n\to\infty} \left(\dfrac{2n + \ln 2}{2n - \ln 2}\right)^n$;

(c) $\lim_{n\to\infty} \dfrac{n^2 + n - 1}{n^2 - n - 1}$.

Exercise 0.58 *If* $c_n = x^n$, *find* $\sup_{n\geq 1} c_n$, $\inf_{n\geq 1} c_n$, $\limsup c_n$ *and* $\liminf c_n$ *for*

(a) $x < -1$;

(b) $x = -1$;

(c) $-1 < x < 0$;

(d) $0 < x < 1$;

(e) $x = 1$;

(f) $x > 1$.

Exercise 0.59 *Show that the limsup and liminf can also be defined as*

$$\limsup_{n\to\infty} \sup_{k\geq n} c_k = \inf_n \sup_{k\geq n} c_k$$

and

$$\liminf_{n\to\infty} \inf_{k\geq n} c_k = \sup_n \inf_{k\geq n} c_k.$$

Exercise 0.60 *True or false?*

(a) $\limsup (a_n + b_n) = \limsup a_n + \limsup b_n$;

(b) $\limsup (a_n b_n) = (\limsup a_n)(\limsup b_n)$.

If true, prove it; If false, give a counter example.

Exercise 0.61 *Find the limits of*

a)

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i(i+1)}.$$

b)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=-n}^{n} \frac{1}{1+\exp(-i)}.$$

Exercise 0.62 *Find the supremum and infimum of the following sets:*

a) $(-10, 10)$
b) $(-\infty, \infty)$
c) $\cap_{i=1}^{\infty} \left\{\frac{1}{i}\right\}$
d) $\cup_{i=1}^{\infty} \left\{\frac{1}{i}\right\}$
e) $\cap_{i=1}^{\infty} \left[\frac{1}{i}, 1\right]$
f) $\cup_{i=1}^{\infty} \left[\frac{1}{i}, 1\right]$
g) $\left\{x : x = \frac{1}{i}, i = 1, 2, ...\right\}$
h) $\left\{x : x = -\frac{1}{i}, i = 1, 2, ...\right\}$
i) $\{x : x = i, i = 1, 2, ...\}$
j) $\{x : x = -i, i = 1, 2, ...\}$

Exercise 0.63 (*i*) *Find the limsup and liminf of the following sequences and determine if the limits of these sequences exist.*

a) $\{a_n : a_n = (-1)^n,\ n \geq 1\}$
b) $\left\{b_n : b_n = (-1)^n \frac{2}{n},\ n \geq 1\right\}$
c) $\{c_n : c_n = a_n - b_n,\ n \geq 1\}$

(ii) Which of the above sequences is(are) $O(1)$, and which is(are) $o(1)$?

Exercise 0.64 *Suppose we have a sample* $\{X_i\}_{i=1}^n$ *which drawn from a U(0,1) distribution. Let*

$Z_n = \max\{X_1, X_2,...,X_n\}$,
$Y_n = \min\{X_1, X_2,...,X_n\}$,

$F_{Z_n}(z) = \Pr(Z_n \leq z)$ be the distribution function of Z_n,
$f_{Z_n}(z)$ be the density function of Z_n,
$F_{Y_n}(y) = \Pr(Y_n \leq y)$ be the distribution function of Y_n,
$f_{Y_n}(y)$ be the density function of Y_n.

a) Find $F_{Z_n}(z)$, $f_{Z_n}(z)$, $F_{Y_n}(y)$, $f_{Y_n}(y)$.

b) Find the expectation of the range $E(Z_n - Y_n)$ and calculate its limit when $n \to \infty$.

c) Plot $F_{Z_n}(z)$ for $n = 1, 2, 3, 4$.

d) Plot $F_{Y_n}(y)$ for $n = 1, 2, 3, 4$.

e) Suppose Z_n converges in distribution to Z, which has a distribution function $F_Z(z)$, find $F_Z(z)$.

f) Suppose Y_n converges in distribution to Y, which has a distribution function $F_Y(y)$, find $F_Y(y)$.

Exercise 0.65 *Let* $\{X_i\}_{i=1}^n$ *be independent discrete random variables which take values zero or one with probability half and half.*

Let
$Z_n = X_1 \times X_2 \times X_3 \times ... \times X_n$,
$F_{Z_n}(z) = \Pr(Z_n \leq z)$ be the distribution function of Z_n,
(a) Plot $F_{Z_n}(z)$ against z for $n = 1, 2$.

(b) Find the expectation $E(Z_n)$ and calculate its limit when $n \to \infty$.

(c) Does Z_n converge in probability to zero? Prove or disprove.
(d) Does Z_n converge in distribution to zero? Prove or disprove.
(e) Does Z_n converge almost surely to zero? Prove or disprove.

Exercise 0.66 *Let X_1, X_2,..., X_n be i.i.d. random variables with finite mean μ and finite variance $0 < \sigma^2 < \infty$, then the sample **mode** has a normal distribution when the sample size tends to infinity. True or False? Explain.*

Exercise 0.67 *Find the limits of the following sequences:*

(a) $c_1 = \frac{2}{1} \times \frac{2}{3}, c_2 = \frac{2}{1} \times \frac{2}{3} \times \frac{4}{3} \times \frac{4}{5}, c_3 = \frac{2}{1} \times \frac{2}{3} \times \frac{4}{3} \times \frac{4}{5} \times \frac{6}{5} \times \frac{6}{7}, c_4 = ...;$
(b) $c_1 = 2, c_2 = 2 \times \frac{2}{\sqrt{2}}, c_3 = 2 \times \frac{2}{\sqrt{2}} \times \frac{2}{\sqrt{2+\sqrt{2}}}, c_4 = ...;$
(c) $c_n = \sum_{i=1}^{n} (-1)^{i-1} \frac{1}{2i-1};$
(d) $c_n = \sum_{i=1}^{n} \frac{1}{i^4};$

Exercise 0.68 *The sequence of Fibonacci numbers is given by*

$$f_0 = 0, f_1 = 1, f_n = f_{n-1} + f_{n-2} \text{ for } n \geq 2.$$

(a) Write down $\frac{f_n}{f_{n-1}}$ for $n = 2, 3, 4, 5$.
(b) Find

$$\lim_{n \to \infty} \frac{f_n}{f_{n-1}}.$$

(c) Solve the close form solution of f_n in terms of n.

Exercise 0.69 *Find the limits of the following sequences:*

(a) $c_1 = \sqrt{1}, c_2 = \sqrt{1+\sqrt{1}}, c_3 = \sqrt{1+\sqrt{1+\sqrt{1}}}, c_4 = ...;$
(b) $c_1 = 1, c_2 = 1 + \frac{1}{1+1}, c_3 = 1 + \frac{1}{1+\frac{1}{1+1}}, c_4 = ...;$
(c) $c_n = \sum_{i=1}^{n} \frac{1}{i(i+1)};$

(d) $c_n = \sum_{i=1}^{n} \frac{1}{i^2}$;
(e) $c_n = \sum_{i=1}^{n} (-1)^{i-1} \frac{1}{i}$;
(f) $c_n = \frac{n^2}{2^n}$;
(g) $c_n = \sum_{k=1}^{n} \frac{1}{k} - \ln n$

Exercise 0.70 *Let*

$$y_0 = \sqrt{2} - 1,$$

$$y_n = \frac{1 - \sqrt[4]{1 - y_{n-1}^4}}{1 + \sqrt[4]{1 - y_{n-1}^4}},$$

$$\alpha_0 = 6 - 4\sqrt{2},$$

$$\alpha_n = (1 + y_n)^4 \alpha_{n-1} - 2^{2n+1} y_n \left(1 + y_n + y_n^2\right).$$

Calculate the value of $\frac{1}{\alpha_{15}}$.

Exercise 0.71 [***Difficult***]*For a random sample of size* $2m+1$ *from the population with mean* μ*, median* $\widetilde{\mu}$ *and variance* σ^2*: Let* $y_1, y_2, ..., y_{2m+1}$ *be the order statistic in ascending order.*

a) Prove that a central limit theorem exists for the median y_{m+1}, i.e. for large m, the median y_{m+1} is approximately normal with mean equal to the population median $\widetilde{\mu}$ and variance $\frac{1}{8\left[f\left(\widetilde{\mu}\right)\right]^2 m}$.

b) Is there any central limit theorem(s) applied to y_1 and/or y_{2m+1}? If yes, prove it. If not, give counter-examples.

c) Find the density functions and the expectations of the order statistic y_1, y_{m+1}, y_{2m+1} from the following populations:
(i) $U(0,1)$
(ii) $N(0,1)$
(iii) Exponential distribution with mean 1.

Exercise 0.72 [***Difficult***]*Let* $C(2n)$ *be the number of ways of decomposing the number 2n (n=1,2,3....) into the sum of two prime numbers. For example,* $24 = 5 + 19 = 7 + 17 = 11 + 13$*, so* $C(24) = 3$*.*

(a) Find $C(2n)$ for $n = 11, 12, ..., 30$.

(b) Find $\lim_{n\to\infty} \frac{C(2n)}{2n}$

(c) Prove or disprove the conjecture that $C(2n) > 0$ for all $n = 2, 3, 4, ...$

(d) Prove or disprove the conjecture that $C(6n) \geq C(6n \pm 2)$ for all $n = 4, 5, 6, ...$

ECO5120: Econometric Theory and Application, Fall 2006
Prof. T.L. Chong
HANDOUT 4
POINT ESTIMATION

Population and sample are very different concepts. We would like to uncover the population mean (μ) and the population variance (σ^2), but there are not sufficient resources to do a detailed study on the population. Even in the case of throwing a dice, we do not know whether the dice is leaded or not. What we normally do is to draw a sample from the population. A sample is a subset of a population. Hopefully, we can retrieve information about the population from a sample when the sample size is large enough. Having a sample, we can construct estimators to estimate the mean and variance of a population.

Definition 96 *An **estimator** is a rule or formula that tells us how to estimate a population quantity, such as the population mean and population variance.*

An estimator is often constructed by exploiting the sample information. Thus, it is usually a random variable since it takes different values under different samples. An estimator has a mean, a variance and a distribution.

Definition 97 *An **estimate** is the numerical value taken by an estimator, it usually depends on the sample drawn.*

Example 98 *Suppose we have a sample of size T, the sample mean*

$$\overline{X} = \frac{X_1 + X_2 + ... + X_T}{T}$$

is an estimator of the population mean.

If $\overline{X}$ turns out to be 3.4, the 3.4 is an estimate of the population mean. Thus, the estimate differs from sample to sample.

Example 99 *The statistic*

$$\widetilde{X} = \frac{X_1 + X_2 + ... + X_{T-1}}{T}$$

is also an estimator of the population mean. Conventionally, $\overline{X}$ denotes the sample mean, we may use $\widetilde{X}$, $\widehat{X}$, X^*, etc. to denote other estimators.

Example 100 *An weighted average*

$$\widetilde{X} = w_1 X_1 + w_2 X_2 + ... + w_T X_T, \qquad where \sum_{i=1}^{T} w_i = 1$$

is also estimator of the population mean.

Example 101 *A single observation X_1 is also estimator of the population mean.*

$$X^* = \frac{X_1^2 + X_2^2 + ... + X_T^2}{T}$$

is also an estimator of the population mean.

Example 102 *A constant, for example, 3.551 is also an estimator of the population mean. In this case, 3.551 is both an estimator and an estimate. Note that when we use a constant as an estimator, the sample has no role. No matter what sample we draw, the estimator and the estimate are always equal to 3.551.*

Thus, there are a lot of estimators for the population mean. What criteria should be used to evaluate a good estimator? In choosing the best estimator, we usually use criteria such as linearity, unbiasedness and efficiency.

The first criterion is linearity, an linear estimator is by construction simpler than a nonlinear estimator. The mean and variance of a linear estimator are easy to evaluate.

Definition 103 *An estimator $\widehat{X}$ is **linear** if it is a linear combination of the sample observations. i.e.*

$$\widehat{X} = a_1 X_1 + a_2 X_2 + ... + a_T X_T,$$

where a_t $(t = 1, 2, ..., T)$ are constants. They can be negative, larger than 1, and some of them can be zero.

However, if all a_t are zero, then $\widehat{X}$ is no longer an estimator.

Thus, estimators in the first four examples are linear, while estimators in the last two examples are not linear.

We reduce the sets of all possible estimators by just focusing on linear estimators. Still, there are plenty of linear estimators, so how should they be compared? Here, we introduce the concept of unbiasedness.

Definition 104 *An estimator $\widehat{X}$ is **unbiased** if $E\left(\widehat{X}\right) = \mu$, where μ is the true mean of the random variable X.*

It is important to realize that any single observation from a sample is unbiased. i.e.,

$$E(X_t) = \mu, \qquad t = 1, 2, ..., T.$$

This is because if an observation is drawn from a population, the best and most reasonable guess of its value is the true mean (μ) of the population.

For an estimator constructed by using two or more observations, whether it is unbiased depends on the way it is constructed.

Example 105 *If X_t $(t = 1, 2, ..., T)$ are random variables with $E(X_t) = \mu$ and $Var(X_t) = \sigma^2$. Show that:*

(a) $\overline{X} = \dfrac{\sum_{t=1}^{T} X_t}{T}$ is an unbiased estimator for μ.

(b) $\widehat{\sigma}^2 = \dfrac{\sum_{t=1}^{T} \left(X_t - \overline{X}\right)^2}{T-1}$ is an unbiased estimator for σ^2.

Solution:(a)

$$E\left(\overline{X}\right) = E\left(\frac{1}{T}\sum_{t=1}^{T} X_t\right) = \frac{1}{T}\sum_{t=1}^{T} E\left(X_t\right) = \frac{1}{T}\sum_{t=1}^{T} \mu = \mu. \blacksquare$$

(b)

$$\begin{aligned} E\left(\widehat{\sigma}^2\right) &= E\left(\frac{\sum_{t=1}^{T}(X_t - \overline{X})^2}{T-1}\right) \\ &= E\left(\frac{\sum_{t=1}^{T} X_t^2 - T\overline{X}^2}{T-1}\right) \\ &= \frac{\sum_{t=1}^{T} E\left(X_t^2\right) - TE\left(\overline{X}^2\right)}{T-1} \\ &= \frac{T\left(\sigma^2 + \mu^2\right) - T\left(\sigma^2 \ / \ T + \mu^2\right)}{T-1} \\ &= \sigma^2. \blacksquare \end{aligned}$$

Example 106 *Consider a regression model*

$$Y_t = \beta_0 + \beta_1 X_t + u, \qquad u_t \sim i.i.d.\left(0, \sigma^2\right)$$

$$\widehat{u}_t = Y_t - \widehat{\beta}_0 \ - \widehat{\beta}_1 X_t$$

is the estimated residual. Show that

$$\widehat{\sigma}^2 = \frac{\sum_{t=1}^{T} \widehat{u}_t^2}{T-2}$$

is an unbiased estimator for σ^2.

Solution: We only have to show that $E\left(\sum_{t=1}^{T} \widehat{u}_t^2\right) = (T-2)\,\sigma^2$. Note that

$$E\left(\sum_{t=1}^{T} \widehat{u}_t^2\right) = E\left(\sum_{t=1}^{T}\left(Y_t - \widehat{\beta}_0 \ - \widehat{\beta}_1 X_t\right)^2\right)$$

$$= E\left(\sum_{t=1}^{T}\left(Y_t - \overline{Y} - \widehat{\beta}_1\left(X_t - \overline{X}\right)\right)^2\right)$$

$$= E\left(\sum_{t=1}^{T}\left(\beta_0+\beta_1 X_t+u_t-\left(\beta_0+\beta_1\overline{X}+\overline{u}\right)-\widehat{\beta}_1\left(X_t-\overline{X}\right)\right)^2\right)$$
$$= E\left(\sum_{t=1}^{T}\left(u_t-\overline{u}-\left(\widehat{\beta}_1-\beta_1\right)\left(X_t-\overline{X}\right)\right)^2\right)$$
$$= E\left[\sum_{t=1}^{T}\left(u_t-\overline{u}\right)^2+\left(\widehat{\beta}_1-\beta_1\right)^2\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2-2\left(\widehat{\beta}_1-\beta_1\right)\sum_{t=1}^{T}\left(X_t-\overline{X}\right)\left(u_t-\overline{u}\right)\right]$$
$$= E\left[\sum_{t=1}^{T}\left(u_t-\overline{u}\right)^2+\left(\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)u_t}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}\right)^2\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2-2\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)u_t}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)X_t}\sum_{t=1}^{T}\left(X_t-\overline{X}\right)u_t\right]$$
$$= \sum_{t=1}^{T}E\left(u_t-\overline{u}\right)^2-\frac{E\left[\sum_{t=1}^{T}\left(X_t-\overline{X}\right)u_t\right]^2}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}$$
$$= \sum_{t=1}^{T}E\left(u_t^2\right)-TE\left(\overline{u}^2\right)-\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2E\left(u_t^2\right)+2\sum_{i=1}^{T-1}\sum_{j>i}^{T}\left(X_i-\overline{X}\right)\left(X_j-\overline{X}\right)E(u_iu_j)}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}$$
$$= \sum_{t=1}^{T}\sigma^2-T\left(\frac{\sigma^2}{T}\right)-\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2\sigma^2+2\sum_{i=1}^{T-1}\sum_{j>i}^{T}\left(X_i-\overline{X}\right)\left(X_j-\overline{X}\right)(0)}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}$$
$$= (T-1)\sigma^2-\sigma^2=(T-2)\sigma^2. \ \blacksquare$$

We further reduce the sets of all possible estimators by just focusing on linear and unbiased estimators. However, if there are plenty of linear and unbiased estimators, how do we select the best estimator linear unbiased estimator? We introduce the concept of efficiency.

Definition 107 *An estimator $\widehat{X}$ is more **efficient** than another estimator X^* if $Var\left(\widehat{X}\right)<Var\left(X^*\right)$.*

Example 108 *If we just look at the efficiency criteria, estimator in the last example is the most efficient estimator since the variance of a constant is zero. However, it is neither linear nor unbiased. A constant as an estimator actually gives us no information about the population mean. Thus, despite the fact that it is efficient, it is not a good estimator.*

Definition 109 *An estimator $\widehat{X}$ is **consistent** estimator of the population mean μ if it converges to the μ as the sample size goes to infinity.*

A necessary condition for an estimator to be consistent is that $Var\left(\widehat{X}\right) \rightarrow 0$ as the sample size goes to infinity. If the estimator truly reveals the value of the population mean μ, the variation of this estimator should be small when the sample is very large. In the extreme case, when the sample size is infinity, the estimator should have no variation at all.

An unbiased estimator with this condition satisfied can be considered as a consistent estimator. If the estimator is biased, it can be consistent too, provided that the bias and the variance of this estimator both go to zero as the sample size goes to infinity.

Consistency is a rather difficult concept. It is very important for an estimator to be consistent, as what we finally want to know is the information of the population parameters. If an estimator is inconsistent, it tells us nothing about the population no matter how large the sample is.

One of the consistent estimator for the population mean is the sample mean

$$\overline{X} = \frac{X_1 + X_2 + ... + X_T}{T}.$$

Note that it is unbiased as

$$\begin{aligned} E\left(\overline{X}\right) &= E\left(\frac{X_1 + X_2 + ... + X_T}{T}\right) \\ &= \frac{E\left(X_1\right) + E\left(X_2\right) + ... + E\left(X_T\right)}{T} \\ &= \frac{\mu + \mu + ... + \mu}{T} = \frac{T\mu}{T} = \mu. \end{aligned}$$

Second, suppose the variance of X_t, $Var\left(X_t\right) = \sigma^2 < \infty$ for $t = 1, 2, ...T$, then

$$\begin{aligned} Var\left(\overline{X}\right) &= Var\left(\frac{X_1 + X_2 + ... + X_T}{T}\right) \\ &= \frac{1}{T^2} Var\left(X_1 + X_2 + ... + X_T\right) \\ &= \frac{1}{T^2}\left[Var\left(X_1\right) + Var\left(X_2\right) + ... + Var\left(X_T\right)\right] \\ &= \frac{1}{T^2}\left[\sigma^2 + \sigma^2 + ... + \sigma^2\right] \\ &= \frac{1}{T^2}\left[T\sigma^2\right] = \frac{\sigma^2}{T} \to 0 \qquad \text{as } T \to \infty. \end{aligned}$$

Be careful that consistency and unbiasedness do not imply each other.

An estimator can be biased but consistent. Consider the estimator in example 2,

$$\widetilde{X} = \frac{X_1 + X_2 + ... + X_{T-1}}{T}.$$

For any given value of sample size T,

$$E\left(\widetilde{X}\right) = \frac{T-1}{T}\mu \neq \mu,$$

The bias is

$$\frac{1}{T}\mu$$

which goes to zero as $T \to \infty$. Thus, we say $\widetilde{X}$ is biased in finite sample but is **asymptotically unbiased**.

Note also that as $T \to \infty$,

$$Var\left(\widetilde{X}\right) = Var\left(\frac{X_1 + X_2 + ... + X_{T-1}}{T}\right) = \frac{T-1}{T^2}\sigma^2 = \left(\frac{1}{T} - \frac{1}{T^2}\right)\sigma^2 \to 0.$$

Since both the bias and the variance of $\widetilde{X}$ go to zero, $\widetilde{X}$ is a consistent estimator.

An estimator can also be unbiased but inconsistent. A single observation as an estimator for the population mean. It is unbiased. However, it is

inconsistent as we just use one observation from a sample of size T, no matter how large T is, thus increasing the number of other observations cannot improve the precision of this estimator.

Maximum Likelihood Estimation

The principle of maximum likelihood provides a mean of choosing an asymptotically efficient estimator for a set of parameters.

Let $\{y_t\}_{i=1}^{T}$ be i.i.d. random variable with joint density $f(y_1, y_{2...}, y_T; \theta)$, where $\theta = (\theta_1, \theta_2, ...\theta_K)'$. Since the sample values have been observed and therefore fixed number, we regard $f(y_t; \theta)$ as a function of θ.

Definition 110 *Let $y = (y_1, y_{2...}, y_T)'$, we defined the **likelihood function** as*

$$L(y; \theta) = f(y_1, y_2, ..., y_T; \theta) = \prod_{t=1}^{T} f(y_t; \theta).$$

and the **log-likelihood function** is defined as $\ln L(y; \theta)$.

The maximum likelihood estimator $\widehat{\theta}_{ML}$ is the estimator that maximizes the likelihood function. Since logarithmic function is a strictly monotonic function, $\widehat{\theta}_{ML}$ also maximizes the log-likelihood function.

$$\widehat{\theta}_{ML} = \arg\max L(y; \theta) = \arg\max (\ln L(y; \theta)).$$

Example 111 *Consider a random sample of 10 observations from a Poisson distribution $y_1, y_2, ..., y_{10}$.*

The density of each observation is

$$f(y_t; \theta) = \frac{\theta^{y_t} \exp(-\theta)}{y_t!},$$

with

$$E(y_t) = \theta,$$

$$Var\left(y_t\right)=\theta.$$

$$\begin{aligned} L\left(y;\theta\right) &= f\left(y_1,y_2,...,y_{10};\theta\right) \\ &= \prod_{t=1}^{10} f\left(y_t;\theta\right). \\ &= \prod_{t=1}^{10} \frac{\theta^{y_t}\exp\left(-\theta\right)}{y_t!} \\ &= \frac{\theta^{y_1+y_2...+y_{10}}\exp\left(-10\theta\right)}{\prod_{t=1}^{10} y_t!}. \end{aligned}$$

$$\ln L\left(y;\theta\right)=\ln\frac{\theta^{y_1+y_2...+y_{10}}\exp\left(-10\theta\right)}{\prod_{t=1}^{10} y_t!}=\left(\sum_{t=1}^{10} y_t\right)\ln\theta-10\theta-\ln\left(\prod_{t=1}^{10} y_t!\right)$$

$$\widehat{\theta}_{ML}=\arg\max\left(\ln L\left(y;\theta\right)\right)$$

First order condition,

$$\frac{\partial}{\partial\theta}\ln L\left(y;\theta\right)=\frac{\sum_{t=1}^{10} y_t}{\theta}-10=0.$$

$$\widehat{\theta}_{ML}=\frac{\sum_{t=1}^{10} y_t}{10}.$$

Definition 112 *The scores S is a K by 1 vector defined as*

$$S=\frac{\partial}{\partial\theta}\ln L\left(y;\theta\right)=\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\ln f\left(y_t;\theta\right).$$

Example 113 *For the Poisson random variable in previous example,*

$$\begin{aligned} S &= \frac{\partial}{\partial\theta}\ln L(y;\theta) \\ &= \frac{\partial}{\partial\theta}\left(\left(\sum_{t=1}^{10} y_t\right)\ln\theta - 10\theta - \ln\left(\prod_{t=1}^{10} y_t!\right)\right) \\ &= \frac{\sum_{t=1}^{10} y_t}{\theta} - 10. \end{aligned}$$

Theorem 114 *The scores have zero expectation when the density for y_t is correctly specified.*

Proof. $E(S) = E\left[\frac{\partial}{\partial\theta}\ln L(y;\theta)\right] = E\left[\frac{1}{L(y;\theta)}\frac{\partial}{\partial\theta}L(y;\theta)\right]$

$$\begin{aligned} &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[\frac{1}{L(y;\theta)}\frac{\partial}{\partial\theta}L(y;\theta)\right]L(y;\theta)\,dy_1\cdots dy_T \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}L(y;\theta)\,dy_1\cdots dy_T \\ &= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}L(y;\theta)\,dy_1\cdots dy_T \\ &= \frac{\partial}{\partial\theta}[1] = 0. \ \blacksquare \end{aligned}$$

Note that if the density is misspecified, say suppose the true joint density is $g(y_1, y_{2\ldots}, y_T;\theta)$, then the expectation of score will not be zero in general since

$$E(S) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[\frac{1}{L(y;\theta)}\frac{\partial}{\partial\theta}L(y;\theta)\right]g(y_1, y_{2\ldots}, y_T;\theta)\,dy_1\cdots dy_T \neq 0.$$

Example 115 *For the Poisson random variable in previous example,*

$$E(S) = E\left(\frac{\sum_{t=1}^{10} y_t}{\theta} - 10\right) = \frac{\sum_{t=1}^{10} E(y_t)}{\theta} - 10 = \frac{\sum_{t=1}^{10}\theta}{\theta} - 10 = 0.$$

Definition 116 *Fisher's Information Matrix $I_{\theta\theta}$ is the variance-covariance matrix of the scores for θ. i.e.*

$$I_{\theta\theta} = E\left[\left(S - E\left(S\right)\right)\left(S - E\left(S\right)\right)'\right] = E\left(SS'\right).$$

$$I_{\theta\theta} = \sum_{t=1}^{T} E\left[\frac{\partial}{\partial\theta}\ln f\left(y_t;\theta\right)\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right].$$

Proof. $I_{\theta\theta} = E\left(SS'\right) = E\left[\left(\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\ln f\left(y_t;\theta\right)\right)\left(\sum_{t=1}^{T}\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right)\right]$
$= E\left[\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\ln f\left(y_t;\theta\right)\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right) + \sum_{i=1}^{T}\sum_{j\neq i}\frac{\partial}{\partial\theta}\ln f\left(y_i;\theta\right)\frac{\partial}{\partial\theta'}\ln f\left(y_j;\theta\right)\right]$
$= \sum_{t=1}^{T} E\left[\frac{\partial}{\partial\theta}\ln f\left(y_t;\theta\right)\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right] + \sum_{i=1}^{T}\sum_{j\neq i} E\left(\frac{\partial}{\partial\theta}\ln f\left(y_i;\theta\right)\right) E\left(\frac{\partial}{\partial\theta'}\ln f\left(y_j;\theta\right)\right)$
by independence of y_i and y_j for $i \neq j$.

Note that

$$\begin{aligned}
E\left[\frac{\partial}{\partial\theta}\ln f\left(y_i;\theta\right)\right] &= E\left[\frac{1}{f\left(y_i;\theta\right)}\frac{\partial}{\partial\theta}f\left(y_i;\theta\right)\right] \\
&= \int_{-\infty}^{\infty}\left[\frac{1}{f\left(y_i;\theta\right)}\frac{\partial}{\partial\theta}f\left(y_i;\theta\right)\right] f\left(y_i;\theta\right) dy_i \\
&= \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f\left(y_i;\theta\right) dy_i \\
&= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f\left(y_i;\theta\right) dy_i \\
&= \frac{\partial}{\partial\theta}\left[1\right] = 0.
\end{aligned}$$

Thus,

$$I_{\theta\theta} = \sum_{t=1}^{T} E\left[\frac{\partial}{\partial\theta}\ln f\left(y_t;\theta\right)\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right]. \blacksquare$$

Example 117 *For the Poisson random variable in the previous example,*

$$I_{\theta\theta} = E\left(\left(S - E\left(S\right)\right)^2\right) = E\left(S^2\right) = E\left(\frac{\sum_{t=1}^{10} y_t}{\theta} - 10\right)^2 = E\left(\sum_{t=1}^{10}\left(\frac{y_t}{\theta} - 1\right)\right)^2$$
$$= E\left[\sum_{t=1}^{10}\left(\frac{y_t}{\theta} - 1\right)^2 + \sum_{i=1}^{10}\sum_{j\neq i}\left(\frac{y_i}{\theta} - 1\right)\left(\frac{y_j}{\theta} - 1\right)\right]$$
$$= \sum_{t=1}^{10} E\left(\left(\frac{y_t}{\theta} - 1\right)^2\right) + \sum_{i=1}^{10}\sum_{j\neq i} E\left(\frac{y_i}{\theta} - 1\right) E\left(\frac{y_j}{\theta} - 1\right)$$

by independence of y_i and y_j for $i \neq j$.

$$= \sum_{t=1}^{10} E\left(\left(\frac{y_t}{\theta}-1\right)^2\right) + \sum_{i=1}^{10}\sum_{j\neq i}\left(\frac{E(y_i)}{\theta}-1\right)\left(\frac{E(y_j)}{\theta}-1\right)$$

$$= \sum_{t=1}^{10} E\left(\left(\frac{y_t}{\theta}-1\right)^2\right) + \sum_{i=1}^{10}\sum_{j\neq i}\left(\frac{\theta}{\theta}-1\right)\left(\frac{\theta}{\theta}-1\right)$$

$$= \sum_{t=1}^{10} E\left(\left(\frac{y_t}{\theta}-1\right)^2\right) = \frac{1}{\theta^2}\sum_{t=1}^{10} E\left((y_t-\theta)^2\right)$$

$$= \frac{1}{\theta^2}\sum_{t=1}^{10} Var(y_t) = \frac{1}{\theta^2}\sum_{t=1}^{10}\theta = \frac{10}{\theta}.$$

$$R.H.S. = \sum_{t=1}^{10} E\left[\left(\frac{\partial}{\partial\theta}\ln f(y_t;\theta)\right)\frac{\partial}{\partial\theta'}\ln f(y_t;\theta)\right]$$

$$= \sum_{t=1}^{10} E\left[\left(\frac{\partial}{\partial\theta}\ln f(y_t;\theta)\right)^2\right] = \sum_{t=1}^{10} E\left[\left(\frac{\partial}{\partial\theta}\ln\frac{\theta^{y_t}\exp(-\theta)}{y_t!}\right)^2\right]$$

$$= \sum_{t=1}^{10} E\left(\left(\frac{y_t}{\theta}-1\right)^2\right) = \frac{10}{\theta}.$$

Theorem 118 *The Fisher's Information Matrix can also be written as*

$$I_{\theta\theta} = -E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L(y;\theta)\right).$$

Proof. Since $E(S) = E\left(\frac{\partial}{\partial\theta}\ln L(y;\theta)\right)$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln L(y;\theta)\right) f(y_1,y_2,...,y_T;\theta)\,dy_1\cdots dy_T = 0.$$

Differentiating both sides with respect to θ',

$$\frac{\partial}{\partial\theta'}E(S) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta'}(SL(y;\theta))\,dy_1\cdots dy_T = 0.$$

This implies

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(S\frac{\partial}{\partial\theta'}L(y;\theta) + L(y;\theta)\frac{\partial}{\partial\theta'}S\right)dy_1\cdots dy_T = 0$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}S\frac{\partial}{\partial\theta'}L(y;\theta)\,dy_1\cdots dy_T = -\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L(y;\theta)\right)L(y;\theta)\,dy_1\cdots dy_T$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}S\frac{1}{L(y;\theta)}\left(\frac{\partial}{\partial\theta'}L(y;\theta)\right)L(y;\theta)\,dy_1\cdots dy_T = -E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L(y;\theta)\right)$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} S\left(\frac{\partial}{\partial\theta'}\ln L\left(y;\theta\right)\right) L\left(y;\theta\right) dy_1\cdots dy_T = -E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right)$$

$$E\left[SS'\right] = -E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right)$$

$$I_{\theta\theta} = -E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right).$$

Example 119 *For the Poisson random variable in the previous example,*

$$\begin{aligned}
-E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right) &= -E\left(\frac{\partial^2}{\partial\theta^2}\ln L\left(y;\theta\right)\right) \\
&= -E\left(\frac{\partial^2}{\partial\theta^2}\left(\left(\sum_{t=1}^{10} y_t\right)\ln\theta - 10\theta - \ln\left(\prod_{t=1}^{10} y_t!\right)\right)\right) \\
&= -E\left(\frac{\partial}{\partial\theta}\left(\frac{\sum_{t=1}^{10} y_t}{\theta} - 10\right)\right) = E\left(\frac{\sum_{t=1}^{10} y_t}{\theta^2}\right) = \frac{10}{\theta} \\
&= \sum_{t=1}^{T} E\left[\left(\frac{\partial}{\partial\theta}\ln f\left(y_t;\theta\right)\right)\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right].
\end{aligned}$$

Theorem 120 *If $y_t \sim$ i.i.d. with density $f\left(y_t;\theta\right)$, $\widehat{\theta}$ is any **unbiased** estimator of θ, the minimum variance of $\widehat{\theta}$ that can be attained is $I_{\theta\theta}^{-1}$.*

Proof. Note that since $\widehat{\theta}$ is unbiased, we have

$$E\left(\widehat{\theta}\right) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta} f\left(y_1, y_2, ..., y_T;\theta\right) dy_1\cdots dy_T = \theta.$$

Differentiating both sides with respect to θ',

$$\frac{\partial}{\partial\theta'}E\left(\widehat{\theta}\right) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\frac{\partial}{\partial\theta'}\prod_{t=1}^{T} f\left(y_t;\theta\right) dy_1\cdots dy_T = I$$

This implies

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\sum_{t=1}^{T}\left(\Pi_{j\neq t}f\left(y_j;\theta\right)\frac{\partial}{\partial\theta'}f\left(y_t;\theta\right)\right)dy_1\cdots dy_T = I$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\sum_{t=1}^{T}\left(\Pi_{j\neq t}f\left(y_j;\theta\right)f\left(y_t;\theta\right)\frac{1}{f\left(y_t;\theta\right)}\frac{\partial}{\partial\theta'}f\left(y_t;\theta\right)\right)dy_1\cdots dy_T = I$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\sum_{t=1}^{T}\left(\Pi_{j=1}^{T}f\left(y_j;\theta\right)\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right)dy_1\cdots dy_T = I$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\Pi_{j=1}^{T}f\left(y_j;\theta\right)\left(\sum_{t=1}^{T}\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right)dy_1\cdots dy_T = I$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}S'\Pi_{j=1}^{T}f\left(y_j;\theta\right)dy_1\cdots dy_T = I$$

Thus

$$E\left(\widehat{\theta}S'\right) = I$$

Using the fact that

$$E\left(\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)'\right)E\left(SS'\right) - \left[E\left(\left(\widehat{\theta}-\theta\right)S'\right)\right]^2$$

is a positive semi-definite matrix, we have

$$E\left(\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)'\right)E\left(SS'\right) - \left[E\left(\widehat{\theta}S'\right)\right]^2$$

is a positive semi-definite matrix. Hence

$$E\left(\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)'\right)I_{\theta\theta} - I$$

is a positive semi-definite matrix. Therefore

$$VarCov\left(\widehat{\theta}\right) - I_{\theta\theta}^{-1}$$

is a positive semi-definite matrix. ■

We call $I_{\theta\theta}^{-1}$ the **Cramér-Rao lower bound** of an unbiased estimator $\widehat{\theta}$.

Properties of Maximum Likelihood Estimators

1. Consistency:

$$p\lim \widehat{\theta}_{ML} = \theta.$$

2. Asymptotic Normality:

$$\sqrt{T}\left(\widehat{\theta}_{ML} - \theta\right) \xrightarrow{d} N\left(0, \lim_{T\to\infty} T\left(I_{\theta\theta}^{-1}\right)\right).$$

3. Asymptotic efficiency: $\widehat{\theta}_{ML}$ is asymptotically efficient and achieves the Cramér-Rao lower bound.

$$AsyVar\left(\widehat{\theta}_{ML}\right) = I_{\theta\theta}^{-1}.$$

4. Invariance: If $g(\cdot)$ is a continuous function, then the ML estimator for $g(\theta)$ is $g\left(\widehat{\theta}_{ML}\right)$

Exercise 0.73 *In examples 98-102, suppose* $X_t \sim i.i.d.(\mu, \sigma^2)$,

i) Which estimators are unbiased?

ii) Rank the efficiency of the estimators in the five examples.

Exercise 0.74 *Construct an estimator which is biased, consistent and less efficient than the simple average* $\overline{X}$.

Exercise 0.75 *Suppose the span of human life follows an i.i.d. distribution with an unknown upper bound* $c < \infty$. *Suppose we have a sample of* T *observations* $X_1, X_2, ..., X_T$ *on people's life span, construct a consistent estimator for* c *and explain why your estimator is consistent.*

Exercise 0.76 *Suppose we have a sample of* 3 *independent observations* X_1, X_2 *and* X_3 *drawn from a distribution with mean* μ *and variance* σ^2. *Which of the following estimators is/are unbiased? Which one is more efficient? Explain.*

$$\widehat{X}_a = \frac{X_1 + 2X_2 + X_3}{4}, \qquad \widehat{X}_b = \frac{X_1 + X_2 + X_3}{3}.$$

Exercise 0.77 *True or false?*

a. An estimator is unbiased if it is consistent.

b. An estimator is efficient if it is unbiased.

c. The Maximum Likelihood estimator is unbiased.

d. The Maximum Likelihood estimator is asymptotically more efficient than any estimators.

Exercise 0.78 *Consider a random sample of 10 observations from a Normal distribution* $y_1, y_2, ..., y_{10}$. *The density of* y_t *is*

$$f(y_t; \theta_1, \theta_2) = \sqrt{\frac{\theta_2}{2\pi}} \exp\left(-\frac{\theta_2}{2}\left(y_t - \frac{1}{\theta_1}\right)^2\right)$$

where θ_1, θ_2 are unknown parameters.

(a) Find the log-likelihood function.

(b) Find the score functions.

(c) Now let the observations be

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
-5	-4	-3	-2	-1	1	2	3	4	5

(i) Find the values of ML estimates for θ_1 and θ_2.

(ii) Evaluate the Fisher's Information Matrix.

Exercise 0.79 *Consider a random sample of 10 observations from a Normal distribution* $y_1, y_2, ..., y_{10}$. *The density of* y_t *is*

$$f\left(y_t; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \mu)^2}{2\sigma^2}\right)$$

where μ and σ^2 are unknown mean and variance of the population respectively.

(a) Find the log-likelihood function.

(b) Find the score functions.

(c) Find the ML estimators for μ and σ^2.

(d) Find the Fisher's Information Matrix.

Exercise 0.80 *Consider a simple linear regression model:*

$$Y_t = \beta_0 + \beta_1 X_t + u_t, \qquad t = 1, 2, ..., T.$$

i) Write down the OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

ii) Given $Cov\left(\overline{u}, \widehat{\beta}_1\right) = 0$, show that $Cov\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = -\overline{X}Var\left(\widehat{\beta}_1\right)$.

Explain intuitively why this covariance depends on $\overline{X}$, discuss cases where $\overline{X} > 0$, $\overline{X} = 0$, and $\overline{X} < 0$. (Hint: Use the fact that the estimated regression line must pass through the point $\left(\overline{X}, \overline{Y}\right)$, and see how the intercept and slope vary as this regression line rotates about the point $\left(\overline{X}, \overline{Y}\right)$.)

iii) If $E(u_t) = -2$, will $\widehat{\beta}_0$ and $\widehat{\beta}_1$ be biased? Explain your answers.

Exercise 0.81 *Consider the model:* $Y_t = \beta_0 + \beta_1 X_t + u_t, \qquad t = 1, 2, ..., T$

a) Suppose we have four observations of (X_t, Y_t), $t = 1, 2, 3, 4$.

X_t	0	1	c	$1-c$
Y_t	0	1	1	0

Find the followings in term of c:

i) $\widehat{\beta}_0$, $\widehat{\beta}_1$

ii) $\widehat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 X_t$ for $t = 1, 2, 3, 4$

iii) $ESS = \sum_{t=1}^{4} \left(Y_t - \widehat{Y}_t\right)^2$

iv) $TSS = \sum_{t=1}^{4} \left(Y_t - \overline{Y}\right)^2$

v) $R^2 = 1 - \dfrac{ESS}{TSS}$

b) For what value(s) of c will the $\widehat{\beta}_1$ equal 1?

c) For what value(s) of c will the R^2 be maximized? For what value(s) of c will the R^2 be minimized?

Exercise 0.82 *Consider the following density function of a random variable* X.

$$\begin{aligned} f(x;\theta) &= 1 \qquad \text{for } \theta < x < \theta + 1; \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

i) Find the moment generating function of x.

ii) Sketch the graph of $f(x;1)$, $f(x;2)$ and $f(x;3)$.

Let X_1 and X_2 constitute a random sample of size 2 from the above population.

iii) Find the joint density of X_1 and X_2.

iv) Find the likelihood function $L(x;\theta)$ and the log-likelihood function $\ln L(x;\theta)$.

v) Find the score $S = \dfrac{\partial}{\partial\theta} \ln L(x;\theta)$.

Exercise 0.83 *Suppose the random variable* $y_t \sim N(\mu, \sigma^2)$. *Let* $\theta = (\mu, \sigma^2)$

a) Derive the log-Likelihood function $\ln L(y;\theta)$, Scores function S and the Fisher's Information Matrix $I_{\theta\theta}$.

b) Derive the ML estimator $\widehat{\theta}$.

Exercise 0.84 *Suppose the random variable* $y_t \sim N(\exp(\theta), 1)$, $t = 1, 2, ..., 100$, y_i *and* y_j *are independent for all* $i \neq j$. *Thus*

$$f(y_t;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(y_t - e^{\theta}\right)^2}{2}}.$$

a) Derive the log-Likelihood function $\ln L(y;\theta)$ and the scores function S.

b) Derive the ML estimator $\widehat{\theta}$.

c) Show that the Fisher's Information Matrix is

$$I_{\theta\theta} = 100e^{2\theta}.$$

Exercise 0.85 *Consider a random sample of 10 observations from a Poisson distribution* $y_1, y_2, ..., y_{10}$.

The probability of observing the value y_t is

$$f(y_t; \theta) = \frac{\theta^{y_t} \exp(-\theta)}{y_t!}.$$

Thus the likelihood function is

$$L(y; \theta) = \frac{\theta^{y_1+y_2...+y_{10}} \exp(-10\theta)}{\prod_{t=1}^{10} y_t!}.$$

(a) Find the log-likelihood function.

(b) Find the score function.

(c) Find the ML estimator.

(d) Find the Fisher's Information Matrix.

(e) Now let the observations be

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
1	2	2	3	3	3	4	4	4	4

Perform a Wald test for the following hypothesis at $\alpha = 5\%$

$$\begin{aligned} H_0 &: \theta = 3, \\ H_1 &: \theta > 3. \end{aligned}$$

(From the Chi-square table with one degree of freedom and $\alpha = 5\%$, the critical value is $\chi_1^2(\alpha) = 3.84146$).

Exercise 0.86 *Given the data* $x = (x_1, x_2, ..., x_T)'$. x_t *is an i.i.d. random variable with density function*

$$f(x_t;\theta) = \frac{1}{\theta} e^{-\frac{x_t}{\theta}} \qquad 0 < x_t < \infty$$

a) Find the likelihood function $L(x;\theta)$ and the log-likelihood function $\ln L(x;\theta)$.

b) Find the score $S = \frac{\partial}{\partial \theta} \ln L(x;\theta)$.

c) Find the Fisher's Information Matrix

$$I_{\theta\theta} = \sum_{t=1}^{T} E\left(\left[\frac{\partial}{\partial \theta} \ln f(x_t;\theta) \right]^2 \right)$$

d) Show that $I_{\theta\theta}$ can be written as

$$I_{\theta\theta} = -E\left(\frac{\partial^2}{\partial \theta^2} \ln L(x;\theta) \right)$$

e) Find the ML estimator of θ and obtain the asymptotic distribution of this estimator.

Now suppose we observe the data $y = (y_1, y_2, ..., y_T)'$ and $y_t = -\ln x_t$.

f) Show that the density function of y_t is given by

$$f(y_t;\theta) = \frac{1}{\theta} e^{-y_t - \frac{1}{\theta} e^{-y_t}} \qquad -\infty < y_t < \infty$$

g) Find the likelihood function $L(y;\theta)$ and the log-likelihood function $\ln L(y;\theta)$.

h) Find the score $S = \frac{\partial}{\partial \theta} \ln L(y;\theta)$.

i) Find the Fisher's Information Matrix

$$I_{\theta\theta} = \sum_{t=1}^{T} E\left(\left[\frac{\partial}{\partial \theta} \ln f(y_t;\theta) \right]^2 \right)$$

j) Show that $I_{\theta\theta}$ can be written as

$$I_{\theta\theta} = -E\left(\frac{\partial^2}{\partial \theta^2} \ln L(y;\theta) \right)$$

k) Find the ML estimator of θ and obtain the asymptotic distribution of this estimator.

l) If $y = g(x)$ is a monotonic, continuous and differentiable function of x. Suppose we only observe the data of y and we only know the density of x. Show the MLE derived by the following two procedures are equivalent.

i) Transform the density of x into the density of y by $f_y(y;\theta) = f_x(x;\theta)\left|\frac{dx}{dy}\right|$, and maximize $\ln L(y;\theta)$.

ii) Transform the data of y into $x = g^{-1}(y)$, and maximize $\ln L(x;\theta)$.

Exercise 0.87 *Given the data $x = (x_1, x_2, ..., x_T)'$. x_t is an i.i.d. random variable with density function*

$$f(x_t; a, b, c) = a + bx_t. \qquad c < x_t < c + 1.$$

a) Find the likelihood function $L(x; a, b, c)$ and the log-likelihood function $\ln L(x; a, b, c)$.

b) Find the score $S = \frac{\partial}{\partial\theta}\ln L(x;\theta), \theta = (a, b, c)'$.

c) Find the Fisher's Information Matrix

$$I_{\theta\theta} = \sum_{t=1}^{T} E\left(\frac{\partial}{\partial\theta}\ln f(x_t;\theta)\frac{\partial}{\partial\theta'}\ln f(x_t;\theta)\right)$$

d) Derive the ML estimator $\widehat{\theta}$.

Exercise 0.88 *Consider the following density function of a random variable X.*

$$\begin{aligned} f(x;\theta) &= \theta x \qquad \text{for } 0 \le x \le \sqrt{\frac{2}{\theta}}; \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

i) Sketch the graph of $f(x;1)$, $f(x;2)$ and $f(x;3)$.

Let $X_1, X_2, ..., X_T$ constitute a random sample of size T from the above population.

ii) Find the joint density of $X_1, X_2, ..., X_T$.

iii) Find the likelihood function $L(x;\theta)$ and the log-likelihood function $\ln L(x;\theta)$.

iv) Find the score $S = \dfrac{\partial}{\partial\theta} \ln L(x;\theta)$, does the score have zero expectation?

v) Find the ML estimator $\widehat{\theta}$. Is your estimator consistent? Explain.

vi) Find the Fisher's information matrix $I(\theta)$ using

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2} \ln L(x;\theta)\right).$$

ECO5120: Econometric Theory and Application, Fall 2006
Prof. T.L. Chong
HANDOUT 5
HYPOTHESIS TESTING

When we observe a phenomenon, we would like to explain it by a hypothesis. We usually post a null hypothesis and an alternative hypothesis.

For example, when we observe that the death toll in winter is higher than those in other seasons, we may post a null hypothesis that the death toll is negatively related to temperature. The alternative hypothesis would be that the death toll has nothing to do with or that it is positively related to temperature.

A hypothesis is not a theorem. A theorem is always true when certain assumptions are held, whereas a hypothesis is just a guess. Thus, we have to test how likely our hypothesis is going to be correct. In testing a hypothesis, we cannot be 100 percent sure that it is correct, otherwise it becomes a theorem. Thus, we may commit errors when concluding a hypothesis. There are two possible types of errors.

Definition 121 *Rejection of the null hypothesis when it is true is called a* ***Type I Error****; the probability of committing a type I error is denoted by* α.

Definition 122 *Acceptance of the null hypothesis when it is false is called a* ***Type II Error****; the probability of committing a type II error is denoted by* β.

We want to reduce both Type I and Type II errors as much as we can. However, as there is no free lunch, we cannot reduce both errors. Reducing the chance of committing Type I Error will increase the chance of committing Type II Error and vice versa.

Example 123 *Suppose a random variable* X *comes from either* $U(0,1)$ *or* $U(0.5,1.5)$*, but we do not know which one is the true population. Thus we test the hypothesis that:*

$$\begin{aligned} H_0 &: X \sim U(0,1); \\ H_1 &: X \sim U(0.5, 1.5) \end{aligned}$$

Suppose we use a single observation X_1 to test the hypothesis. It is obvious that if $X_1 < 0.5$, then we accept the null and reject the alternative hypothesis. If $X_1 > 1$, then we reject the null and accept the alternative. However, if $0.5 \leq X_1 \leq 1$, then we have a problem in judging which hypothesis is true. What we can do is to set a rule. Suppose we only want a 5% probability of committing type I error, i.e., $\alpha = 5\%$, then our rule is to reject H_0 if $X_1 \geq 0.95$. By doing so, even if H_1 is true, we will accept H_0 as far as $X_1 < 0.95$, thus the probability of committing Type II error is

$$\beta = \Pr(X_1 < 0.95 | X_1 \sim U(0.5, 1.5)) = 0.45.$$

Testing Statistical Hypothesis

The current framework of testing statistical hypothesis is largely due to the work of Neyman and Pearson in the late 1920s, early 30s, complementing Fisher's work on estimation.

Definition 124 *A **statistical hypothesis** is an assertion about the distribution of one or more random variables. If the statistical hypothesis completely specifies the distribution, it is called a simple hypothesis; if it does not, it is called a composite hypothesis. In general, we can write $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$.*

For example, $H_0 : \theta \leq 0.5$ is a composite statistical hypothesis since it does not completely specify the distribution. If the null hypothesis is $H_0 : \theta = 0.5$, then it is a simple statistical hypothesis.

Definition 125 *A **test** of a statistical hypothesis is a rule which, when the experimental sample values have been obtained, leads to a decision to accept or to reject the hypothesis under consideration.*

Definition 126 *Let C be that subset of the sample space which, in accordance with a prescribed test, leads to the rejection of the hypothesis under consideration. Then C is called the* ***critical region*** *of the test.*

Definition 127 *The probability of rejecting H_0 when H_0 is false at some point $\theta_1 \in \Theta_1$, i.e. $\Pr(x \in C; \theta = \theta_1)$ is called the* ***power*** *of the test at $\theta = \theta_1$.*

Definition 128 *$P(\theta) = \Pr(x \in C; \theta \in \Theta_0 \cup \Theta_1)$ is called the* ***power function*** *of the test defined by the rejection region C.*

Example 129 *Suppose the random variable X has a density function of the form*

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right) \qquad 0 < x < \infty; \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Suppose we want to test

$$\begin{aligned} H_0 &: \theta = 2; \\ H_1 &: \theta > 2. \end{aligned}$$

Thus $\Theta_0 = \{2\}$ and $\Theta_1 = (2, \infty)$.

A random sample of size 2 is used. The rejection region C is set at

$$C = \left\{ \left(x_1, x_2; \overline{x} = \frac{x_1 + x_2}{2} > 4.75 \right) \right\}.$$

$$\begin{aligned} P(\theta) &= \text{Power at } \theta \\ &= \Pr\left((X_1, X_2) \in C; \theta \geq 2\right) \\ &= \Pr\left(x_1 + x_2 \geq 9.5; \theta \geq 2\right) \\ &= 1 - \Pr\left(x_1 + x_2 < 9.5; \theta \geq 2\right) \\ &= 1 - \int_0^{9.5} \int_0^{9.5-x_2} \frac{1}{\theta^2} \exp\left(-\frac{x_1 + x_2}{\theta}\right) dx_1 dx_2 \\ &= \left(\frac{\theta + 9.5}{\theta}\right) \exp\left(-\frac{9.5}{\theta}\right). \end{aligned}$$

If H_0 is true, the joint density function of X_1 and X_2 is

$$\begin{aligned} f(x_1; 2) f(x_1; 2) &= \frac{1}{4} \exp\left(-\frac{x_1 + x_2}{2}\right) \qquad 0 < x_1, x_2 < \infty; \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

$$\begin{aligned} P(2) &= \text{Power at } (\theta = 2) \\ &= \Pr\left(x_1 + x_2 \geq 9.5; \theta = 2\right) \\ &= \left(\frac{2 + 9.5}{2}\right) \exp\left(-\frac{9.5}{2}\right) \\ &\simeq 0.05. \end{aligned}$$

$$\begin{aligned} P(4) &= \text{Power at } (\theta = 4) \\ &= \left(\frac{4 + 9.5}{4}\right) \exp\left(-\frac{9.5}{4}\right) \\ &\simeq .314. \end{aligned}$$

$$\begin{aligned} P(9.5) &= \text{Power at } (\theta = 9.5) \\ &= 2 \exp(-1) \\ &\simeq .736. \end{aligned}$$

$$\begin{aligned} P(\infty) &= \text{Power at } (\theta = \infty) \\ &= 1. \end{aligned}$$

Definition 130 *The **significance level** of the test or (the **size** of the critical region) is the maximum value of the power function of the test when H_0 is true.*

$$\alpha = \max_{\theta \in \Theta_0} P(\theta).$$

If the set Θ_0 contains only one element θ_0, then $\alpha = P(\theta_0)$. The maximum operator is called for when Θ_0 is an interval or contains more than one single point.

Definition 131 *A test is **unbiased** if its power is greater than or equal to its size for all values of parameters.*

If a test is biased, then the power is less than the size, which means

$$\begin{aligned} \Pr(\text{Reject } H_0 | H_0 \text{ is false}) &< \Pr(\text{Reject } H_0 | H_0 \text{ is true}) \\ 1 - \Pr(\text{Accept } H_0 | H_0 \text{ is false}) &< 1 - \Pr(\text{Accept } H_0 | H_0 \text{ is true}) \\ \Pr(\text{Accept } H_0 | H_0 \text{ is true}) &< \Pr(\text{Accept } H_0 | H_0 \text{ is false}). \end{aligned}$$

In other words, if a test is biased, then for some values of the parameter, we are more likely to accept the null when it is false than when it is true.

Definition 132 *A test is **consistent** if its power goes to one as the sample size grows to infinity.*

The Normal Test

Consider a random sample X_1, X_2,...,X_T drawn from a **normal** distribution with unknown mean μ and a **known variance** σ^2. We would like to test whether μ equals a particular value μ_0. i.e.,

$$H_0 : \mu = \mu_0;$$

μ_0 is a pre-specified value, e.g. $\mu_0 = 0$. We construct a test statistic Z, where

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{T}}.$$

Under $H_0 : \mu = \mu_0$, $X_t \sim N(\mu_0, \sigma^2)$. Since the sum of normal random variable is also normal, as a result, $\overline{X}$ is also normally distributed for all sample size T, no matter T is small or large. Thus

$$\overline{X} = \frac{1}{T}(X_1 + X_2 + ... + X_T) \sim N\left(\mu_0, \frac{\sigma^2}{T}\right).$$

Hence

$$Z \sim N(0, 1).$$

In the two-sided case (i.e. $H_1 : \mu \neq \mu_0$), we reject H_0 at a significance level α, if $|Z| > Z_{\frac{\alpha}{2}}$. For example $Z_{0.025} = 1.96$.

In the one-sided case (i.e. $H_1 : \mu > (<)\mu_0$), we reject H_0 at a significance level α if $Z > Z_\alpha$ $(Z < -Z_\alpha)$.

A $100(1-\alpha)\%$ **confidence interval** for μ is

$$\left(\overline{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{T}}, \overline{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{T}}\right).$$

If μ_0 does not fall into this interval, we reject H_0 at the significance level α.

The normal test is of limited use since we have two very strong assumptions that the observations X_t come from the normal distribution and that the variance is known. A more commonly used test is the t-test, which is called for when the population variance is unknown and the sample size is small.

The t-Test

Consider a random sample X_1, X_2,...X_T drawn from a **normal** distribution with unknown mean μ and **unknown variance** σ^2. We would like to test whether μ equals a particular value μ_0.

$$H_0 : \mu = \mu_0.$$

We construct a test statistic, defined as

$$t_{obs} = \frac{\overline{X} - \mu_0}{\widehat{\sigma}/\sqrt{T}},$$

where t_{obs} stands for the observed value of the statistic under the null hypothesis that $\mu = \mu_0$.

What is the distribution of t_{obs}? Recall that

$$\widehat{\sigma} = \sqrt{\frac{\sum_{t=1}^{T} \left(X_t - \overline{X}\right)^2}{T-1}}.$$

Note that

$$t_{obs} = \frac{\overline{X} - \mu_0}{\widehat{\sigma}/\sqrt{T}} = \frac{\frac{\overline{X}-\mu_0}{\sigma/\sqrt{T}}}{\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(\frac{X_t-\overline{X}}{\sigma}\right)^2}}.$$

Under $H_0 : \mu = \mu_0$, $X_t \sim N(\mu_0, \sigma^2)$, thus $\overline{X} = \frac{1}{T}(X_1 + X_2 + ... + X_T) \sim N\left(\mu_0, \frac{\sigma^2}{T}\right)$, and

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{T}} \sim N(0,1).$$

Further, it can be shown that (very difficult)

$$\sum_{t=1}^{T}\left(\frac{X_t - \overline{X}}{\sigma}\right)^2$$

has a Chi-square distribution with degrees of freedom $(T-1)$, and that (also very difficult)

$$\frac{\overline{X}-\mu_0}{\sigma/\sqrt{T}}$$

and

$$\sum_{t=1}^{T}\left(\frac{X_t-\overline{X}}{\sigma}\right)^2$$

are independent.

Recall the definition of a t-distribution,

$$t_{obs}=\frac{\overline{X}-\mu_0}{\widehat{\sigma}/\sqrt{T}}=\frac{\frac{\overline{X}-\mu_0}{\sigma/\sqrt{T}}}{\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(\frac{X_t-\overline{X}}{\sigma}\right)^2}}=\frac{N(0,1)}{\sqrt{\chi^2_{T-1}/(T-1)}}$$

will have a t-distribution with degrees of freedom $(T-1)$.

In the two-sided case (i.e. $H_1:\mu\neq\mu_0$), we reject H_0 at a significance level α if $|t|>t_{\frac{\alpha}{2},T-1}$. For example, $t_{0.025,9}=2.262$.

In the one-sided case (i.e. $H_1:\mu>(<)\mu_0$), we reject H_0 at a significance level α if $t>t_{\alpha,T-1}$ $(t<-t_{\alpha,T-1})$.

A $100(1-\alpha)\%$ **confidence interval** for μ is

$$\left(\overline{X}-t_{\frac{\alpha}{2},T-1}\frac{\widehat{\sigma}}{\sqrt{T}},\overline{X}+t_{\frac{\alpha}{2},T-1}\frac{\widehat{\sigma}}{\sqrt{T}}\right).$$

If μ_0 does not fall into this interval, we reject H_0 at the significance level α.

Example 133 *Suppose the height of the population of Hong Kong is normally distributed $N(\mu,\sigma^2)$. Suppose we want to test a hypothesis that the mean height of the population of Hong Kong at a certain time is μ =160cm. We test this based on a sample of 10 people of Hong Kong, the sample mean being $\overline{X}$ =165cm and the standard error (note that standard error is the*

square root of the sample variance while standard deviation is the square root of the population variance) is $\widehat{\sigma}$ *=5cm.*

Thus we test

$$\begin{aligned} H_0 &: \mu = 160 \\ H_1 &: \mu \neq 160 \end{aligned}$$

Since the sample size is small and σ^2 is unknown, we use the t-test, the observed t-value is calculated by

$$t_{obs} = \frac{\overline{X} - \mu_0}{\widehat{\sigma}/\sqrt{T}} = \frac{165 - 160}{5/\sqrt{10}} = 3.163.$$

t_{obs} will have a t-distribution with degrees of freedom equal $T - 1$.

In the two-sided case, we reject H_0 at a significance level α if $|t_{obs}| > t_{\frac{\alpha}{2},T-1}$.

Now, let $\alpha = 5\%$, then

$$t_{0.025,9} = 2.262.$$

Since $|t_{obs}| > t_{0.025,9}$, we reject H_0 at $\alpha = 5\%$. This means we are 95% sure that the population mean is not equal to 160cm.

A 95% **confidence interval** for μ is

$$\overline{X} \mp t_{0.025,9} \left(\frac{\widehat{\sigma}}{\sqrt{10}} \right) = 165 \mp 2.262 \left(\frac{5}{\sqrt{10}} \right) = (161.4, 168.6) .$$

Since 160 does not fall into this interval, we reject H_0 at $\alpha = 5\%$.

Note that the conclusion depends on the value of α that we set, if we set $\alpha = 1\%$, then

$$t_{0.01,9} = 3.25.$$

Since $|t_{obs}| < t_{0.01,9}$, we do not reject H_0 at $\alpha = 1\%$. This means we cannot be 99% sure that the population mean is not equal to 160cm.

What if $\mathbf{X}_t$ are not Normally Distributed?

Thus far, we assume that the observations are normally distributed. What if this assumption does not hold?

Consider a random sample X_1, X_2,...X_T drawn from **any** distribution with unknown finite mean μ and a finite **unknown variance** σ^2. We would like to test whether μ equal a particular value μ_0.

$$H_0 : \mu = \mu_0.$$

If the sample size is small, say if $T < 30$, then we cannot test the hypothesis since we do not know what the behavior of the sample mean $\overline{X}$ and sample variance $\widehat{\sigma}^2$ if X_t is not normally distributed.

However, if the sample size is large, say $T > 30$, we can apply the Central Limited Theorem that $\overline{X}$ is normally distributed and the Law of Large Number that $\widehat{\sigma}^2$ will converge to the population variance σ^2.

Then the test statistic

$$Z = \frac{\overline{X} - \mu_0}{\widehat{\sigma}/\sqrt{T}}$$

will be approximately normally distributed as $N(0,1)$.

In the two-sided case(i.e., $H_1 : \mu \neq \mu_0$), we reject H_0 at a significance level α, if $|Z| > Z_{\frac{\alpha}{2}}$. For example $Z_{0.025} = 1.96$.

In the one-sided case(i.e., $H_1 : \mu > (<)\mu_0$), we reject H_0 at a significance level α if $Z > Z_\alpha$ $(Z < -Z_\alpha)$.

A $100(1-\alpha)\%$ **confidence interval** for μ is

$$\overline{X} \mp Z_{\frac{\alpha}{2}} \frac{\widehat{\sigma}}{\sqrt{T}}.$$

If μ_0 does not fall into this interval, we reject H_0 at the significance level α.

Thus if the observations X_t are not normal, we need a large sample to carry out the test.

Hypothesis testing on β's in a simple regression model

We run a linear regression for the model

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

because we want to see whether Y is linearly depending on X, i.e., we want to see whether β_1 equals zero.

After the estimation, we would like to test some hypotheses. Suppose we find $\widehat{\beta}_1 = 0.34$ from the sample, we may want to test whether the true parameter β_1 equals zero or not. That is, we may want to test $H_0 : \beta_1 = 0$. We must perform this test because if we cannot reject H_0, that implies X cannot explain Y and the regression model is useless. When we test this hypothesis, we have to form a test statistic and find the distribution of this test statistic. Usually, we will look up the t-table, thus we may use a test statistic which follow a t-distribution. As I mentioned previously, when we use the t-distribution, we have to assume that the observations are coming from a normal distribution. In the case of regression model, the random elements are u_t. Therefore we have to make the assumption that $u_t \sim N(0, \sigma^2)$.

This assumption is not necessary as far as estimation is concerned. It is called for when we want to perform hypothesis testing on β's. Suppose we perform a two-sided test on β_1:

$$\begin{aligned} H_0 &: \beta_1 = 0; \\ H_1 &: \beta_1 \neq 0. \end{aligned}$$

A standard way to test the hypothesis is to form a test statistic

$$M = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{Var\left(\widehat{\beta}_1\right)}},$$

where $\widehat{\beta}_1$ is the OLS estimator for the unknown parameter β_1 and

$$Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}.$$

Note that since $\widehat{\beta}_1$ is unbiased,

$$E\left(M\right) = \frac{E\left(\widehat{\beta}_1 - \beta_1\right)}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}} = 0,$$

and

$$Var\left(M\right) = Var\left(\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}}\right) = \frac{Var\left(\widehat{\beta}_1\right)}{\frac{\sigma^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}} = 1.$$

Thus, the test statistic will have a distribution with mean zero and variance 1, but what is its exact distribution? This depends on whether σ^2 is known or not. Note that

$$M = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}} = \frac{\frac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}}$$

$$= \frac{X_1 - \overline{X}}{\sqrt{\sigma^2\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}u_1 + \frac{X_1 - \overline{X}}{\sqrt{\sigma^2\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}u_1 + ... + \frac{X_T - \overline{X}}{\sqrt{\sigma^2\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}u_T,$$

which is a linear combination of u_t. Since u_t has a normal distribution, if σ^2 is known, by the property that normal plus normal is still normal, the test statistic M will have a $N(0,1)$ distribution.

The problem again, is that σ^2 is unknown in the real world, so we will have to estimate it. Recall that σ^2 is the variance of u_t in the true model:

$$Y_t = \beta_0 + \beta_1 X_t + u_t.$$

After we get the OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$, the estimated residual is

$$\widehat{u}_t = Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_t$$

and we define

$$\widehat{\sigma}^2 = \frac{\sum_{t=1}^{T}\widehat{u}_t^2}{T-2}.$$

We use $\widehat{\sigma}^2$ to estimate σ^2.

Two questions here. First, why $\sum_{t=1}^{T}\widehat{u}_t^2$ but not $\sum_{t=1}^{T}\left(\widehat{u}_t - \overline{\widehat{u}}\right)^2$? Second, why we have to use $(T-2)$, but not T?

The answer to the first question is $\sum_{t=1}^{T}\widehat{u}_t = \sum_{t=1}^{T}\left(Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_t\right) = 0$ by the first normal equation (*). Thus $\overline{\widehat{u}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{u}_t = 0$.

The reason why we have to use $(T-2)$ is because we want $\widehat{\sigma}^2$ to be an unbiased estimator of σ^2. This number should be equal to the number of

$\beta's$ in the regression. If we have a multiple regression with k $\beta's$, then it should be $(T-k)$ at the bottom. It is the same reason why we usually put $(T-1)$ at the bottom when forming a sample variance of a random variable, all because we want to get an unbiased estimator of σ^2. Now,

$$\begin{aligned} M &= \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\widehat{\sigma}^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}}\sqrt{\frac{\sigma^2}{\widehat{\sigma}^2}} \\ &= N(0,1) \times \sqrt{\frac{\sigma^2}{\widehat{\sigma}^2}} = \frac{N(0,1)}{\sqrt{\frac{\widehat{\sigma}^2}{\sigma^2}}} \\ &= \frac{N(0,1)}{\sqrt{\frac{\frac{1}{T-2}\sum_{t=1}^{T}\widehat{u}_t^2}{\sigma^2}}} = \frac{N(0,1)}{\sqrt{\frac{\sum_{t=1}^{T}\left(\frac{\widehat{u}_t}{\sigma}\right)^2}{T-2}}}. \end{aligned}$$

It can be shown that (very difficult) $\sum_{t=1}^{T}\left(\frac{\widehat{u}_t}{\sigma}\right)^2$ has a Chi-square distribution with degree of freedom $(T-2)$, and that $\sum_{t=1}^{T}\left(\frac{\widehat{u}_t}{\sigma}\right)^2$ is independent of $\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}}$, thus the test statistic $M = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{T-2}}{T-2}}}$ will have a t-distribution with degrees of freedom $(T-2)$. This explains why we have to use the t-table for hypothesis testing in regression models.

Asymptotic Distribution of a Nonlinear Function

Theorem 134 *If* $\sqrt{T}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} N(0, \sigma^2)$, *and if* $g(\cdot)$ *is a continuous differentiable function, then*

$$\sqrt{T}\left(g\left(\widehat{\theta}\right) - g(\theta)\right) \xrightarrow{d} N\left(0, \left\{g'(\theta)\right\}^2 \sigma^2\right).$$

Proof. Let ξ lies between $\widehat{\theta}$ and θ, taking Taylor expansion around θ, we have:

$$\begin{aligned}
g\left(\widehat{\theta}\right) &= g(\theta) + g'(\theta)\left(\widehat{\theta}-\theta\right) + \frac{1}{2}g''(\xi)\left(\widehat{\theta}-\theta\right)^2 \\
\sqrt{T}\left(g\left(\widehat{\theta}\right)-g(\theta)\right) &= \sqrt{T}g'(\theta)\left(\widehat{\theta}-\theta\right) + \frac{1}{2}g''(\xi)\left[\sqrt{T}\left(\widehat{\theta}-\theta\right)\right]\left(\widehat{\theta}-\theta\right) \\
&= \sqrt{T}g'(\theta)\left(\widehat{\theta}-\theta\right) + \frac{1}{2}O(1)O_p(1)o_p(1) \\
&= \sqrt{T}g'(\theta)\left(\widehat{\theta}-\theta\right) + o_p(1) \\
&\overset{d}{=} g'(\theta)\left[\sqrt{T}\left(\widehat{\theta}-\theta\right)\right] \\
&\overset{d}{\to} g'(\theta)N\left(0,\sigma^2\right) \\
&\overset{d}{=} N\left(0,\left\{g'(\theta)\right\}^2\sigma^2\right). \blacksquare
\end{aligned}$$

Three Asymptotic Tests

Thus far, we have only discussed tests for a single parameter that based on finite sample. We now discuss three tests which can handle complicated restrictions for multiple parameters. The three tests are the Wald, Likelihood Ratio and Lagrange Multiplier tests. Asymptotically, the three tests are equivalent and have a Chi-square distribution. The tests are based on the maximum likelihood estimation and use the asymptotic normality of the ML estimators.

Wald Test

Now consider how to test the linear restriction of the form

$$H_0 : R\theta = r.$$

For example, if θ is a scalar and if we want to test $H_0 : \theta = 0$, then $R = 1$ and $r = 0$.

If we want to test $H_0 : \theta_1 = 3$ and $\theta_2 = 4$, then

$$R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \qquad r = \begin{pmatrix} 3 \\ 4 \end{pmatrix}.$$

Let $\widehat{\theta}$ be the ML estimate, if the restriction is valid, $R\widehat{\theta} - r$ should be close to zero. Hence we directly test whether $R\widehat{\theta} - r = 0$. Note that

$$R\widehat{\theta} - r \xrightarrow{d} N\left(0, RVar\left(\widehat{\theta}\right)R'\right).$$

The Wald test statistic is defined as

$$W = \left(R\widehat{\theta} - r\right)'\left[R\widehat{Var}\left(\widehat{\theta}\right)R'\right]^{-1}\left(R\widehat{\theta} - r\right) \xrightarrow{d} \chi^2_k,$$

where k is the number of restrictions imposed, $\widehat{Var}\left(\widehat{\theta}\right) = I^{-1}\left(\widehat{\theta}\right)$ is the inverse of the Fisher's Information Matrix evaluated at the ML estimate $\widehat{\theta}$.

What is important here is that $\widehat{\theta}$ maximizes the likelihood function ignoring the restrictions and so it is called the unrestricted estimator. We never actually find an estimate of θ that is compatible with the restriction.

For nonlinear restriction, we write

$$H_0 : \psi(\theta) = 0.$$

For example, suppose we want to test $H_0 : \theta_1^2 + \theta_2 - 3 = 0$, then $\psi(\theta) = \theta_1^2 + \theta_2 - 3$.

The Wald test becomes

$$W = \psi\left(\widehat{\theta}\right)'\widehat{Var}\left(\psi\left(\widehat{\theta}\right)\right)^{-1}\psi\left(\widehat{\theta}\right) \xrightarrow{d} \chi^2_k.$$

Likelihood Ratio Test

Consider the restriction $H_0 : \psi(\theta) = 0$. Let $\widehat{\theta}_R$ be the restricted estimate under H_0. If the restriction is valid, $\widehat{\theta}_R$ should be close to the unrestricted ML estimate $\widehat{\theta}$. Thus $\ln L\left(y;\widehat{\theta}_R\right) - \ln L\left(y;\widehat{\theta}\right)$ should be close to zero. Note that $L\left(y;\widehat{\theta}_R\right)$ is the likelihood under restriction, $L\left(y;\widehat{\theta}\right)$ is the likelihood without restriction, with the possibility that the restriction may be wrong, it is clear that

$$L\left(y;\widehat{\theta}_R\right) \leq L\left(y;\widehat{\theta}\right).$$

The Likelihood Ratio is defined as

$$\frac{L\left(y;\widehat{\theta}_R\right)}{L\left(y;\widehat{\theta}\right)} \leq 1$$

and the LR test is defined as

$$LR = -2\ln\frac{L\left(y;\widehat{\theta}_R\right)}{L\left(y;\widehat{\theta}\right)} = 2\left(\ln L\left(y;\widehat{\theta}\right) - \ln L\left(y;\widehat{\theta}_R\right)\right) \xrightarrow{d} \chi^2_k.$$

LR is a non-negative number and has an asymptotic Chi-square distribution with degrees of freedom equal the number of restrictions imposed.

Lagrange Multiplier Test

Consider the restriction $H_0 : \psi(\theta) = 0$. If the restriction is valid, the restricted estimator should be near the point that maximizes the log likelihood. Therefore the slope of the log-likelihood should be near zeros at the restricted estimator.

The restricted estimator $\widehat{\theta}_R$ for θ is obtained by solving the following Lagrangian function

$$\ln L^*(y;\theta) = \ln L(y;\theta) + \psi(\theta)\lambda.$$

Differentiating $\ln L^*(y;\theta)$ with respect to θ and λ, we have the following necessary conditions:

$$\frac{\partial \ln^* L(y;\theta)}{\partial\theta} = \frac{\partial \ln L(y;\theta)}{\partial\theta} + \psi'(\theta)\lambda = 0$$

and

$$\frac{\partial \ln^* L(y;\theta)}{\partial\lambda} = \psi(\theta) = 0.$$

If the restrictions are valid, $\widehat{\theta}_R = \widehat{\theta}$, where $\widehat{\theta}$ is the unrestricted estimator solving $\dfrac{\partial \ln L(y;\theta)}{\partial\theta} = 0$.

$$\frac{\partial \ln L\left(y;\widehat{\theta}_R\right)}{\partial\theta} = \frac{\partial \ln L\left(y;\widehat{\theta}\right)}{\partial\theta} = 0.$$

Thus testing $\psi(\theta) = 0$ can be reduced to testing $\dfrac{\partial \ln L\left(y;\widehat{\theta}_R\right)}{\partial\theta} = 0$.

Define

$$LM = \left(\frac{\partial \ln L\left(y;\widehat{\theta}_R\right)}{\partial\theta}\right)' \left[I\left(\widehat{\theta}_R\right)\right]^{-1} \left(\frac{\partial \ln L\left(y;\widehat{\theta}_R\right)}{\partial\theta}\right) \xrightarrow{d} \chi^2_k,$$

where $I\left(\widehat{\theta}_R\right)$ is Fisher's Information Matrix evaluated at $\widehat{\theta}_R$.

LM has an asymptotic Chi-square distribution with degrees of freedom equal the number of restrictions imposed.

Example 135 *Consider a random sample of 10 observations from a Poisson distribution $y_1, y_2, ..., y_{10}$ in Handout 3. The density of each observation is*

$$f(y_t;\theta) = \frac{\theta^{y_t}\exp(-\theta)}{y_t!},$$

$$L(y;\theta) = \frac{\theta^{y_1+y_2...+y_{10}}\exp(-10\theta)}{\prod_{t=1}^{10} y_t!},$$

$$\ln L(y;\theta) = \left(\sum_{t=1}^{10} y_t\right)\ln\theta - 10\theta - \ln\left(\prod_{t=1}^{10} y_t!\right).$$

Let the observations be

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
3	2	5	6	2	1	7	3	4	3

then

$$\widehat{\theta} = \frac{\sum_{t=1}^{10} y_t}{10} = 3.6,$$

$$S = \frac{\partial \ln L(y;\theta)}{\partial \theta} = \frac{\sum_{t=1}^{10} y_t}{\theta} - 10.$$

The Fisher's Information Matrix is

$$I(\theta) = \frac{10}{\theta}.$$

Suppose we want to test

$$\begin{aligned} H_0 &: \theta = 3, \\ H_1 &: \theta > 3. \end{aligned}$$

Now, we have

$$R = 1, \qquad r = 3,$$

$$\widehat{Var}\left(\widehat{\theta}\right) = I\left(\widehat{\theta}\right)^{-1} = \frac{\widehat{\theta}}{10} = .36,$$

$$\widehat{\theta} = 3.6,$$

$$W = \left(R\widehat{\theta} - r\right)' \left[R\widehat{Var}\left(\widehat{\theta}\right) R'\right]^{-1} \left(R\widehat{\theta} - r\right) = \frac{\left(\widehat{\theta} - 3\right)^2}{I^{-1}\left(\widehat{\theta}\right)} = \frac{(3.6-3)^2}{.36} = 1.$$

We can also rewrite the hypothesis as

$$\begin{aligned} H_0 &: \psi(\theta) = 0, \\ H_1 &: \psi(\theta) > 0. \end{aligned}$$

where

$$\psi(\theta) = \theta - 3,$$

$$\widehat{\theta}_R = 3,$$

$$\begin{aligned} LR &= -2\ln\left(\frac{L\left(y;\widehat{\theta}_R\right)}{L\left(y;\widehat{\theta}\right)}\right) = 2\left(\ln L\left(y;\widehat{\theta}\right) - \ln L\left(y;\widehat{\theta}_R\right)\right) \\ &= 2\left(\left(\sum_{t=1}^{10} y_t\right)\ln\widehat{\theta} - 10\widehat{\theta} - \ln\left(\prod_{t=1}^{10} y_t!\right) - \left(\sum_{t=1}^{10} y_t\right)\ln\widehat{\theta}_R + 10\widehat{\theta}_R + \ln\left(\prod_{t=1}^{10} y_t!\right)\right) \\ &= 2\left((36)\ln(3.6) - 10(3.6) - (36)\ln(3) + 10(3)\right) \\ &= 1.1271521. \end{aligned}$$

$$\begin{aligned} LM &= \left(\frac{\partial \ln L\left(y;\widehat{\theta}_R\right)}{\partial\theta}\right)' \left[I\left(\widehat{\theta}_R\right)\right]^{-1} \left(\frac{\partial \ln L\left(y;\widehat{\theta}_R\right)}{\partial\theta}\right) \\ &= \frac{\left(\frac{\sum_{t=1}^{10} y_t}{\widehat{\theta}_R} - 10\right)^2}{\frac{10}{\widehat{\theta}_R}} = \frac{\left(\frac{36}{3} - 10\right)^2}{\frac{10}{3}} = 1.2. \end{aligned}$$

Since we have only one restriction that $\theta - 3 = 0$, we look up the Chi-square table with one degree of freedom. Let $\alpha = 5\%$, since $\chi_1^2(\alpha) = 3.84146$, we do not reject H_0 at $\alpha = 5\%$ in all the 3 tests.

Remark 2 *Note that $\widehat{\theta}_R$ is the restricted estimate under H_0. For example, if the null is $H_0 : \theta = 3$, then imposing H_0 gives $\widehat{\theta}_R = 3$. However, if the null is $H_0 : \theta > 0$, the we have to go through the ML estimation procedure and find an estimator which is positive and maximizes the likelihood.*

Exercise 0.89 *The following table is the Labour Force Participation Rates for **male**, using age group from **20 to 59**, for the year **1994**. The table is adopted from Hong Kong Annual Digest of Statistics 1996 Edition, page 13, Table 2.1.*

X (middle of the age group)	Y (%)
22	80.2
27	97.8
32	98.3
37	98.6
42	98.6
47	97.3
52	92.4
57	78.3

where

Y=Labour force participation rate;

X=Middle age in each age group.

i) Plot (X,Y).

ii) Run the following regression model

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

Find the values of $\widehat{\beta}_0$, $\widehat{\beta}_1$. What is the meaning of $\widehat{\beta}_0$ in this case? Interpret $\widehat{\beta}_1$.

iii) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$. Is the labour force participation rate stable for men? If not, is it increasing or decreasing with age?

vi) Repeat part i) to iii) using the labour force participation rate for female in the same year.

Exercise 0.90 *In a judicial trial, suppose the null hypothesis is that "the defendant is not guilty".*

(a) How to state the alternative hypothesis?

(b) What is the Type I Error in this case?

(c) What is the Type II Error in this case?

(d) How can you fully eliminate Type I Error in this case? How will this affect the chance of committing Type II Error?

(e) How can you fully eliminate Type II Error in this case? How will this affect the chance of committing Type I Error?

(f) How can you fully eliminate both Errors in this case?

(g) Suppose the defendant is charged with the murder of first degree, whose penalty is the capital punishment (death). From your point of view, which type of error has a more serious consequence?

Exercise 0.91 *A random sample of size* $T = 12$ *from a normal population has the sample mean* $\overline{X} = 28$ *and sample variance* $\widehat{\sigma}^2 = 3$.

(a) Construct a 95% confidence interval for the population mean μ.

(b) Test the hypothesis $H_0 : \mu = 30$ against $H_1 : \mu \neq 30$ at $\alpha = 5\%$.

Exercise 0.92 *Let* X_t *be the monthly total number of births in Hong Kong. Assume that* $X_t \sim N(\mu, \sigma^2)$. *Consider a sample of* X_t *from April 1998 to September 2003.*

(a) Find $\overline{X}$ and $\widehat{\sigma}^2$

(b) Use t-test to test the hypothesis $H_0 : \mu = 4500$ against $H_1 : \mu \neq 4500$ at $\alpha = 5\%$.

(c) Construct a 95% confidence interval for the population mean μ.

Exercise 0.93 *Let* X_t *be the monthly total number of deaths in Hong Kong. Assume that* $X_t \sim N(\mu, \sigma^2)$. *Consider a sample of* X_t *from April 1999 to September 2004.*

(a) Find $\overline{X}$ and $\widehat{\sigma}^2$

(b) Use t-test to test the hypothesis $H_0 : \mu = 3000$ against $H_1 : \mu < 3000$ at $\alpha = 5\%$.

Exercise 0.94 *Let* X_t *be the monthly total number of marriages in Hong Kong. Assume that* $X_t \sim N(\mu, \sigma^2)$. *Consider a sample of* X_t *from April 1999 to September 2004.*

(a) Find $\overline{X}$ and $\widehat{\sigma}^2$

(b) Use t-test to test the hypothesis $H_0 : \mu = 3000$ against $H_1 : \mu > 3000$ at $\alpha = 5\%$.

Exercise 0.95 *A random sample of size* $T = 100$ *from a population has the sample mean* $\overline{X} = 28$ *and sample variance* $\widehat{\sigma}^2 = 3$.

(a) Construct a 95% confidence interval for the population mean μ.

(b) Test the hypothesis $H_0 : \mu = 30$ against $H_1 : \mu \neq 30$ at $\alpha = 5\%$.

(Note that we cannot apply the t-test as we do not assume that the observations come from a normal distribution.)

Exercise 0.96 *Let*

Y=private consumption expenditure at constant (2000) market price;

X=Expenditure-based GDP at constant (2000) market price.

i) Find the values of X and Y for the period 1999-2005

ii) Plot (X,Y).

iii) Run the following regression model

$$Y_t = \beta_0 + \beta_1 X_t + u_t.$$

Find the values of $\widehat{\beta}_0$, $\widehat{\beta}_1$. What is the meaning of $\widehat{\beta}_0$ in this case? Interpret $\widehat{\beta}_1$.

iv) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$.

v) Using the estimated model to predict the value of Y in 2006 using X in 2006.

Exercise 0.97 *From the Hong Kong Annual Digest of Statistics 2005 Edition, find the Statistics of Results of Hong Kong Certificate of Education Examination 2004.*

Let

Y= % of student getting A.

X=Number sat.

i) If a student want to get 10 straight A in HKCEE, which 10 subjects will you recommend him/her to take?

ii) If a student want to fail 10 subjects in HKCEE, which 10 subjects will you recommend him/her to take?

iii) Run the following regression model

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

Find the values of $\widehat{\beta}_0$, $\widehat{\beta}_1$. What is the meaning of $\widehat{\beta}_0$ in this case? Interpret $\widehat{\beta}_1$.

iv) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$. Does the chance of getting an A depend on the number of candidates in the exam? If so, in which direction?

Exercise 0.98 *Use 9/99 to 9/06 Hang Seng Index, End of Month, closing price data to run a regression HSI on TIME, where HSI is the value of the Hang Seng index, and TIME=1 for September 1999, 2 for October 1999, and so on. Is the slope coefficient significantly different from 0 at* $\alpha = 5\%$*? Predict the value of Hang Seng index for End of October 2006.*

Now use the natural logarithm of Hang Seng index ln(HSI) as the dependant variable, run the regression ln(HSI) on TIME. Is the slope coefficient significantly different from 0 at $\alpha = 5\%$? Predict the value of ln(HSI) for October 2006, and take the exponential of this predicted value, i.e. calculate $e^{\widehat{\ln(HSI)}}$ and use it as the predicted value for HSI.

Finally, get the actual value of HSI at the end of October 2006, and compare your predicted values above with this actual value. Which one is closer to the true value, and why?

Exercise 0.99 *Let* $X_1, X_2, ..., X_T$ *be independent random variables come from* $N(\theta, 1)$ *distribution. Suppose* $\overline{X} = 3.5$*, perform the Wald, LM and LR tests on* $H_0 : \theta = 0$ *versus* $H_1 : \theta > 0$.

Exercise 0.100 *If* X_1 *and* X_2 *are independent* $N(0, 1)$ *random variables.*

a) What is the distribution of $X_1^2 + X_2^2$?

b) Let $\overline{X} = \dfrac{X_1 + X_2}{2}$, $Z_1 = X_1 - \overline{X}$, $Z_2 = X_2 - \overline{X}$, show that $Z_2 = -Z_1$.

c) What is the distribution of $\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2$?

Exercise 0.101 *If* X_t $(t = 1, 2, ..., T)$ *are i.i.d. normal random variables with* $E(X_t) = \mu$ *and* $Var(X_t) = \sigma^2$. *Show that* $\overline{X}$ *and* $S^2 = \dfrac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}{T}$ *are independent.*

Exercise 0.102 *Let* X_1 *and* X_2 *constitute a random sample of size 2 from the population*

$$\begin{aligned} f(x;\theta) &= \theta x^{\theta-1} \qquad \text{for } 0 < x < 1; \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

i) Sketch the graph of $f(x;1)$, $f(x;2)$ and $f(x;3)$.

ii) Find the joint density of X_1 and X_2.

iii) If the critical region $x_1 + x_2 \geq \frac{3}{2}$ is used to test $H_0 : \theta = 1$ against $H_1 : \theta > 1$. Show that the power function is

$$P(\theta) = \theta \int_{.5}^{1} x^{\theta-1}\left(1 - \left(\frac{3}{2} - x\right)^{\theta}\right) dx$$

iv) Find $P(1)$, $P(2)$, $P(3)$.

Exercise 0.103 *Consider the following density function of a random variable* X.

$$\begin{aligned} f(x;\theta) &= 1 \qquad \text{for } \theta < x < \theta + 1; \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

i) Consider testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Suppose we reject H_0 when $X_1 + X_2 > \dfrac{3}{2}$. Find the power function $P(\theta)$. Plot $P(1)$, $P(2)$, $P(3)$ and $P(\infty)$.

Exercise 0.104 *True/False.*

a. Rejection of the null hypothesis when it is true is called the Type I Error.

b. A test is unbiased if it is consistent.

c. The Wald, LR and LM tests will give the same conclusion in finite samples.

d. The LR test has no power when the log-liklelihood function is a flat line.

e. The LM test does not apply when the log-likelihood function is always increasing in parameters.

Exercise 0.105 *Suppose we draw observations y_t independently from two uniform distributions, $U(0,\theta_1)$ and $U(0,\theta_2)$ respectively, with $\theta_2 > \theta_1 > 0$. However, we do not know which distribution an observation belongs to. Let p be the chance that an observation is coming from $U(0,\theta_1)$. We would like to estimate the three parameters θ_1, θ_2 and p.*

(a) Find the likelihood and log-likelihood functions.
(b) Find the score functions for θ_1, θ_2 and p.
(c) Can we estimate the three parameters?
(d) Can we estimate θ_1 and θ_2 if we know $p = 0.5$?
(e) Can we estimate θ_1 and θ_2 if we know $p = 0.5$ and $\theta_1 = 1 - \theta_2$?
(f) Show how to perform a LR test for (i) $p = 0.5$; (ii) $\theta_1 = 1 - \theta_2$.

Exercise 0.106 *Consider a random sample of 10 observations from a Normal distribution $y_1, y_2, ..., y_{10}$. The density of y_t is*

$$f\left(y_t;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(y_t-\mu\right)^2}{2\sigma^2}\right)$$

where μ and σ^2 are unknown mean and variance of the population respectively.

(a) Find the log-likelihood function.
(b) Find the score functions.
(c) Find the ML estimators for μ and σ^2.

(d) Find the Fisher's Information Matrix.

(e) Now let the observations be

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
0	1	2	2	2	2	3	3	3	4

Perform the Wald, Likelihood Ratio and Lagrange Multiplier tests for the following hypothesis at $\alpha = 5\%$:

$$\begin{aligned} H_0 &: \mu = \sigma^2, \\ H_1 &: \mu > \sigma^2. \end{aligned}$$

(From the Chi-square table with one degree of freedom and $\alpha = 5\%$, the critical value is $\chi_1^2(\alpha) = 3.84146$).

Exercise 0.107 *Let X have a Poisson distribution with mean θ, i.e.*

$$\Pr(X = x) = \frac{\theta^x e^{-\theta}}{x!}$$

i) Show that $\sum_{x=0}^{\infty} \frac{\theta^x}{x!} = e^{\theta}$.

ii) Show that the moment generating function of X is equal to

$$M(t) = E(\exp(tX)) = e^{\theta\left(e^t - 1\right)}$$

and find $E(X)$ and $Var(X)$.

iii) Let $X_1, ..., X_{12}$ denote a random sample of size 12 from this distribution. Find the moment generating function of $Z = X_1 + ... + X_{12}$. Show that Z is also a Poisson random variable with mean 12θ.

iv) Consider the simple hypothesis $H_0 : \theta = \frac{1}{2}$ and the alternative composite hypothesis $H_1 : \theta > \frac{1}{2}$. We reject H_0 if and only if $\overline{X} = \frac{X_1 + ... + X_{12}}{12} > \frac{3}{4}$. Let $P(\theta)$ be the power function of the test. Find $P\left(\frac{1}{2}\right), P\left(\frac{3}{4}\right), P(\infty)$.

v) Sketch the graph of $P(\theta)$.

vi) What is the significance level of the test?

Exercise 0.108 *Let X have a Poisson distribution with mean θ. Consider the simple hypothesis $H_0 : \theta = \frac{1}{2}$ and the alternative composite hypothesis $H_1 : \theta < \frac{1}{2}$. Thus $\Theta = \left\{\theta : 0 < \theta \leq \frac{1}{2}\right\}$. Let $X_1, ..., X_{12}$ denote a random sample of size 12 from this distribution. We reject H_0 if and only if $\overline{X} = \frac{X_1 + ... + X_{12}}{12} \leq \frac{1}{6}$. Let $P(\theta)$ be the power function of the test. Find $P\left(\frac{1}{2}\right)$, $P\left(\frac{1}{3}\right), P\left(\frac{1}{4}\right), P\left(\frac{1}{6}\right), P\left(\frac{1}{12}\right)$. Sketch the graph of $P(\theta)$. What is the significance level of the test?*

Exercise 0.109 *Suppose the random variable $y_t \sim N(\exp(\theta), 1)$, $t = 1, 2, ..., 100$, y_i and y_j are independent for all $i \neq j$. Thus*

$$f(y_t; \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(y_t - e^{\theta}\right)^2}{2}}.$$

a) Derive the log-Likelihood function $\ln L(y; \theta)$ and the scores function S.

b) Derive the ML estimator $\widehat{\theta}$.

c) Show that the Fisher's Information Matrix is

$$I(\theta) = 100e^{2\theta}.$$

d) Suppose we would like to test $H_0 : \theta = 0$ versus $H_1 : \theta < 0$.

Define a Wald test

$$W\left(\widehat{\theta}\right) = \frac{\widehat{\theta}^2}{\widehat{Var}\left(\widehat{\theta}\right)} = \frac{\widehat{\theta}^2}{I^{-1}\left(\widehat{\theta}\right)} = \widehat{\theta}^2 I\left(\widehat{\theta}\right),$$

where $I\left(\widehat{\theta}\right)$ is the Fisher's Information Matrix evaluated at $\widehat{\theta}$.

Since we have only one restriction, $W\left(\widehat{\theta}\right)$ has an asymptotic chi-square distribution with 1 degree of freedom. Thus at $\alpha = 5\%$, we reject H_0 when $W\left(\widehat{\theta}\right) > 3.84146$;

i) Suppose H_0 is rejected when $\widehat{\theta} = -2$ at $\alpha = 5\%$. Intuitively, when $\widehat{\theta} = -3$, do you expect to reject or not to reject H_0? Explain.

ii) Plot $W\left(\widehat{\theta}\right)$ at $\widehat{\theta} = 0, -1, -2, -3, -4, -\infty$;

iii) Determine whether we should reject H_0 at $\alpha = 5\%$ when $\widehat{\theta} = 0, -1, -2, -3, -4, -\infty$.

iv) Derive the power function of this test.

Exercise 0.110 *Consider the following density function of a random variable X.*

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta} \quad \text{for } 0 \leq x \leq \theta; \\ &= 0 \quad \text{elsewhere.} \end{aligned}$$

i) Find the moment generating function of x.

ii) Sketch the graph of $f(x;1)$, $f(x;2)$ and $f(x;3)$.

Let $X_1, X_2, ..., X_T$ constitute a random sample of size T from the above population.

iii) Find the joint density of $X_1, X_2, ..., X_T$.

iv) Find the likelihood function $L(x;\theta)$ and the log-likelihood function $\ln L(x;\theta)$.

v) Find the score $S = \dfrac{\partial}{\partial\theta} \ln L(x;\theta)$, does the score have zero expectation?

vi) Find the ML estimator $\widehat{\theta}$. Is your estimator consistent? Explain.

vii) Find the Fisher's information matrix using $I(\theta)$ using

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ln L(x;\theta)\right).$$

viii) Suppose we would like to test $H_0 : \theta = 1$ versus $H_1 : \theta > 1$.

Define a Wald test

$$W\left(\widehat{\theta}\right) = \left(\widehat{\theta} - 1\right)^2 I\left(\widehat{\theta}\right),$$

where $I\left(\widehat{\theta}\right)$ is the Fisher's Information Matrix evaluated at $\widehat{\theta}$.

Since we have only one restriction, $W\left(\widehat{\theta}\right)$ has an asymptotic chi-square distribution with 1 degree of freedom. Thus at $\alpha = 5\%$, we reject H_0 when $W\left(\widehat{\theta}\right) > 3.84146$;

Now consider the case where the sample size $T = 1$;

a) show that $\widehat{\theta} = X_1$.

b) if $\widehat{\theta} = \frac{1}{3}$, intuitively, should we reject or not reject H_0? Now compute $W\left(\widehat{\theta}\right)$ at $\widehat{\theta} = \frac{1}{3}$. Is H_0 rejected at $\alpha = 5\%$?

c) if H_0 is true, can $\widehat{\theta} = 2$? Intuitively, should we reject or not reject H_0 if $\widehat{\theta} = 2$? Now compute $W\left(\widehat{\theta}\right)$ at $\widehat{\theta} = 2$. Is H_0 rejected at $\alpha = 5\%$?

d) plot $W\left(\widehat{\theta}\right)$ at $\widehat{\theta} = 1, 2, 3, 4, \infty$.

ix) For $\theta \geq 1$, plot the power functions of this test for $T = 1, 2, 3, 4, \infty$.

x) Explain why the test is not properly behaved. Design a test for above hypothesis.

Exercise 0.111 *Consider the following model:*

$$y_t = y_{t-1}^{\theta} + u_t, \qquad t = 1, 2, ..., T,$$

Suppose we estimate θ via the MLE method by assuming u_t follows independent $N(0,1)$.

(a) Derive the log-Likelihood function $\ln L(y;\theta)$ and the scores function S.

(b) Describe how to get the ML estimator $\widehat{\theta}$.

(c) Describe how to get the Information Matrix.

(d) Describe how to form a Wald test for $\theta = 1$.

Exercise 0.112 *Consider Section 17.5.4 in Greene 5th edition.*

Perform the Wald, LM and LR tests on
a) $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.
b) $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

Exercise 0.113 *Greene, Chapter 17, Exercises 1-4.*

ECO5120: Econometric Theory and Application, Fall 2006

Prof. T.L. Chong

HANDOUT 6

NONLINEAR LEAST SQUARES ESTIMATION

Definition 136 *A sequence of function $S_T(\theta)$ converge to $S(\theta)$* ***pointwise*** *in Θ, if for any given $\theta \in \Theta$,*

$$|S_T(\theta) - S(\theta)| = o(1).$$

Definition 137 *A sequence of function $S_T(\theta)$ converge to $S(\theta)$* ***uniformly*** *in Θ if*

$$\sup_{\theta \in \Theta} |S_T(\theta) - S(\theta)| = o(1).$$

Uniform convergence implies pointwise convergence, but the not the other way around.

Example 138 $\Theta = [0, 2]$,

$$\begin{aligned} S_T(\theta) &= T\theta \qquad \theta \in [0, \frac{1}{2T}] \\ &= 1 - T\theta \qquad \theta \in (\frac{1}{2T}, \frac{1}{T}] \\ &= 0 \qquad \theta \in (\frac{1}{T}, 1] \\ &= \frac{T}{T+1}(\theta - 1) \qquad \theta \in (1, 1.5] \\ &= \frac{T}{T+1}(2 - \theta) \qquad \theta \in (1.5, 2] \end{aligned}$$

$$\begin{aligned} S(\theta) &= 0 \qquad \theta \in [0, 1] \\ &= \theta - 1 \qquad \theta \in (1, 1.5] \\ &= 2 - \theta \qquad \theta \in (1.5, 2] \end{aligned}$$

For any given $\theta \in (0, 2]$,

$$|S_T(\theta) - S(\theta)| = o(1),$$

but

$$\sup_{\theta \in \Theta} |S_T(\theta) - S(\theta)| = \frac{1}{2} \neq o(1).$$

Definition 139 *A sequence of random variable $S_T(\theta)$ converge to $S(\theta)$ in probability* ***pointwise*** *if for any given $\theta \in \Theta$,*

$$|S_T(\theta) - S(\theta)| = o_p(1).$$

Definition 140 *A sequence of random variable $S_T(\theta)$ converge to $S(\theta)$ in probability* ***uniformly*** *if*

$$\sup_{\theta \in \Theta} |S_T(\theta) - S(\theta)| = o_p(1).$$

Uniform convergence in probability implies pointwise convergence in probability, but not the other way around.

Example 141 $\Theta = (0, \infty)$,

$$\begin{aligned} S_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t + ZT\theta \qquad \text{for } \theta \in (0, \frac{1}{2T}] \\ &= \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t + Z(1 - T\theta) \qquad \text{for } \theta \in (\frac{1}{2T}, \frac{1}{T}] \\ &= 0 \qquad \text{for } \theta \in \left(\frac{1}{T}, \infty\right) \end{aligned}$$

where $\{\varepsilon_t\}_{t=1}^{T}$ is an i.i.d. zero-mean, finite variance stochastic sequence, Z is a binary random variable which takes -1 with probability 0.5 and 1 with probability 0.5.

$$S(\theta) = 0 \qquad \text{for } \theta \in \Theta$$

For any given $\theta \in \Theta$,

$$|S_T(\theta) - S(\theta)| = o_p(1)$$

However

$$\sup_{\theta \in \Theta} |S_T(\theta) - S(\theta)| = \left| \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t + \frac{Z}{2} \right| \neq o_p(1)$$

since

$$\Pr\left(\sup_{\theta \in \Theta} |S_T(\theta) - S(\theta)| \geq \epsilon \right) = \Pr\left(\left| \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t + \frac{Z}{2} \right| \geq \epsilon \right) \to \Pr\left(\left| \frac{Z}{2} \right| \geq \epsilon \right) = 1.$$

Nonlinear Least Squares Estimation

Suppose the true relationship between y and x is

$$y_t = f(x_t; \theta_0) + u_t,$$

where θ_0 is a vector of true parameters which are unknown. For a given sample of size T, we want to construct an estimator to estimate θ_0. We use the least squares estimation method which minimizes the sum of squared errors, i.e.,

$$\min_{\theta} \sum_{t=1}^{T} u_t^2.$$

Or equivalently, we minimize $\frac{1}{T} \sum_{t=1}^{T} u_t^2$. Let

$$S_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} u_t^2 = \frac{1}{T} \sum_{t=1}^{T} (y_t - f(x_t; \theta))^2.$$

Definition 142 *The **nonlinear least squares estimator** is defined as*

$$\widehat{\theta}_T = Arg\min_{\theta} S_T(\theta),$$

where

$$S_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}(y_t - f(x_t;\theta))^2.$$

Definition 143 *A **nonlinear regression model** is one for which the first-order conditions for least squares estimation of the parameters are nonlinear functions of the parameters.*

The first order condition is

$$\frac{\partial S_T(\theta)}{\partial \theta} = 0,$$

or

$$\frac{2}{T}\sum_{t=1}^{T}(y_t - f(x_t;\theta))\frac{\partial f(x_t;\theta)}{\partial \theta} = 0.$$

and this is a nonlinear function in θ.

Example 144 *Consider the model*

$$y_t = \alpha + \beta x_t + u_t, \qquad t = 1, 2, ..., T.$$

This is a linear model, a special case of nonlinear model.

$$f(x_t;\theta_0) = \alpha_0 + \beta_0 x_t.$$

$$\theta_0 = \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}.$$

$$S_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2.$$

First-order conditions

$$\frac{\partial S_T(\theta)}{\partial \alpha} = -\frac{2}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t) = 0$$

$$\frac{\partial S_T(\theta)}{\partial \beta} = -\frac{2}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)\, x_t = 0$$

Thus we have

$$\widehat{\beta} = \frac{\sum_{t=1}^{T}(x_t - \overline{x})\, y_t}{\sum_{t=1}^{T}(x_t - \overline{x})^2},$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}.$$

Example 145 *Consider the model*

$$y_t = \alpha x_t^{\beta} + u_t, \qquad t = 1, 2, ..., T.$$

$$f(x_t; \theta_0) = \alpha_0 x_t^{\beta_0}.$$

$$\theta_0 = \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}.$$

$$S_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \alpha x_t^{\beta}\right)^2.$$

First-order conditions

$$\frac{\partial S_T(\theta)}{\partial \alpha} = -\frac{2}{T}\sum_{t=1}^{T}\left(y_t - \alpha x_t^{\beta}\right) x_t^{\beta} = 0$$

$$\frac{\partial S_T(\theta)}{\partial \beta} = -\frac{2}{T}\sum_{t=1}^{T}\left(y_t - \alpha x_t^{\beta}\right) x_t^{\beta} \ln x_t = 0$$

From the first equation,

$$\widehat{\alpha} = \frac{\sum_{t=1}^{T} y_t x_t^{\beta}}{\sum_{t=1}^{T} x_t^{2\beta}}.$$

Putting this into the second equation gives

$$\sum_{t=1}^{T} \left[y_t - \frac{\sum_{t=1}^{T} y_t x_t^{\beta}}{\sum_{t=1}^{T} x_t^{2\beta}} x_t^{\beta} \right] x_t^{\beta} \ln x_t = 0.$$

Thus $\widehat{\beta}$ solves

$$\sum_{t=1}^{T} y_t x_t^{\beta} \ln x_t - \frac{\sum_{t=1}^{T} y_t x_t^{\beta} \sum_{t=1}^{T} x_t^{2\beta} \ln x_t}{\sum_{t=1}^{T} x_t^{2\beta}} = 0.$$

Consistency of NLS Estimators

Definition 146 *A set in R^K is compact if and only if it is closed and bounded.*

Since the nonlinear estimator usually does not have a closed form solution, we need to verify the following assumptions for it to be consistent.

Assumptions:

(A1) The parameter space Θ is a compact subset of R^K.

(B1) $S_T(\theta)$ is continuous in $\theta \in \Theta$.

(C1) $S_T(\theta)$ converges to a non-stochastic function $S(\theta)$ in probability **uniformly** in $\theta \in \Theta$ as $T \to \infty$, and $S(\theta)$ attains a **unique** global minimum at θ_0. i.e.

$$\sup_{\theta \in \Theta} |S_T(\theta) - S(\theta)| = o_p(1)$$

and

$$\theta_0 \overset{def}{=} Arg\min_{\theta\in\Theta} S(\theta) \qquad \text{is unique.}$$

The first and second assumptions guarantee the existence of a minimum of $S_T(\theta)$. The last assumption is needed for the solution to be unique.

If all the three assumptions are satisfied, then the nonlinear least squares estimator will be consistent. If any one of the assumptions is violated, it does not mean the estimator is not consistent. It may still be consistent, but we are not confident about it. If a nonlinear LS estimator is inconsistent, then at least one of the assumptions must be violated.

Example 147 *The following model satisfies assumptions* $(A1)$ *to* $(C1)$*:*

$$y_t = x_t^{\theta} + u_t, \qquad (t = 1, 2, ..., T.)$$

where
x_t are distributed uniformly in $(0, 1)$;
$E(x^r) = \int_0^1 x^r dx = \dfrac{1}{1+r}$;
$u_t \sim i.i.d.(0, \sigma^2)$, $\sigma^2 < \infty$;
u and x are independent;
$\theta \in \Theta = [0, 2]$;
$\theta_0 = 1$.

Obvious $\Theta = [0, 2]$ a compact subset of R. Thus, $(A1)$ is satisfied.

$$f(x_t; \theta) = x_t^{\theta}.$$

$$S_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - x_t^{\theta}\right)^2.$$

For any given sample $\{x_t, y_t\}_{t=1}^T$, $S_T(\theta)$ is continuous in θ, since $|S_T(\theta+\delta) - S_T(\theta)| \to 0$ as $\delta \to 0$ for all $\theta \in \Theta$. Thus, $(B1)$ is satisfied. Note that

$$\begin{aligned} S_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T}\left(x_t^{\theta_0}+u_t-x_t^{\theta}\right)^2 \\ &\xrightarrow{p} \sigma^2+\sum_{t=1}^{T}\left(x_t^2+x_t^{2\theta}-2x_t^{\theta+\theta_0}\right) \\ &= \sigma^2+E\left(x^2\right)+E\left(x^{2\theta}\right)-2E\left(x^{\theta+1}\right) \\ &= \sigma^2+\frac{1}{3}+\frac{1}{1+2\theta}-\frac{2}{2+\theta}. \end{aligned}$$

Therefore, we let

$$S(\theta)=\sigma^2+\frac{1}{3}+\frac{1}{1+2\theta}-\frac{2}{2+\theta}.$$

It is easily shown than $S(\theta)$ attains a unique global minimum at $\theta=\theta_0=1$.

Now, consider

$$\begin{aligned} &\sup_{\theta\in[0,2]}\left|S_T(\theta)-S(\theta)\right| \\ &=\sup_{\theta\in[0,2]}\left|\frac{1}{T}\sum_{t=1}^{T}\left(y_t-x_t^{\theta}\right)^2-S(\theta)\right| \\ &=\sup_{\theta\in[0,2]}\left|\frac{1}{T}\sum_{t=1}^{T}\left(x_t+u_t-x_t^{\theta}\right)^2-S(\theta)\right| \\ &=\sup_{\theta\in[0,2]}\left|\frac{1}{T}\sum_{t=1}^{T}u_t^2+\frac{1}{T}\sum_{t=1}^{T}\left(x_t-x_t^{\theta}\right)^2+\frac{2}{T}\sum_{t=1}^{T}u_t\left(x_t-x_t^{\theta}\right)-S(\theta)\right| \\ &=\sup_{\theta\in[0,2]}\left|\begin{array}{c}\frac{1}{T}\sum_{t=1}^{T}u_t^2+\frac{1}{T}\sum_{t=1}^{T}x_t^2+\frac{1}{T}\sum_{t=1}^{T}x_t^{2\theta}-\frac{2}{T}\sum_{t=1}^{T}x_t^{1+\theta} \\ +\frac{2}{T}\sum_{t=1}^{T}u_t\left(x_t-x_t^{\theta}\right)-\sigma^2-\frac{1}{3}-\frac{1}{1+2\theta}+\frac{2}{1+\theta+\theta_0}\end{array}\right| \\ &\leq \sup_{\theta\in[0,2]}\left\{\left|\frac{1}{T}\sum_{t=1}^{T}u_t^2-\sigma^2\right|+\left|\frac{1}{T}\sum_{t=1}^{T}x_t^2-\frac{1}{3}\right|+\left|\frac{1}{T}\sum_{t=1}^{T}x_t^{2\theta}-\frac{1}{1+2\theta}\right|\right. \\ &\left.+\left|\frac{2}{T}\sum_{t=1}^{T}x_t^{1+\theta}-\frac{2}{2+\theta}\right|+\left|\frac{2}{T}\sum_{t=1}^{T}u_t\left(x_t-x_t^{\theta}\right)\right|\right\} \\ &\leq \left|\frac{1}{T}\sum_{t=1}^{T}u_t^2-\sigma^2\right|+\left|\frac{1}{T}\sum_{t=1}^{T}x_t^2-\frac{1}{3}\right|+\sup_{\theta\in[0,2]}\left|\frac{1}{T}\sum_{t=1}^{T}x_t^{2\theta}-\frac{1}{1+2\theta}\right| \\ &+\sup_{\theta\in[0,2]}\left|\frac{2}{T}\sum_{t=1}^{T}x_t^{1+\theta}-\frac{2}{2+\theta}\right|+\sup_{\theta\in[0,2]}\left|\frac{2}{T}\sum_{t=1}^{T}u_t\left(x_t-x_t^{\theta}\right)\right| \\ &=o_p(1)+o_p(1)+o_p(1)+o_p(1)+o_p(1) \\ &=o_p(1). \end{aligned}$$

So $(C1)$ is satisfied.

Theorem 148 *Under assumption (A1) to (C1), we have*

$$\widehat{\theta}_T \xrightarrow{p} \theta_0.$$

Proof. Let N be an open neighborhood in R^K containing θ_0, then $N^c \cap \Theta$ is compact. Therefore $\min_{\theta \in N^c \cap \Theta} S_T(\theta)$ exists. Denote

$$\epsilon = \min_{\theta \in N^c \cap \Theta} S(\theta) - S(\theta_0) > 0.$$

Let A_T be the event

$$-\frac{\epsilon}{2} < S_T(\theta) - S(\theta) < \frac{\epsilon}{2} \qquad \forall \theta \in \Theta.$$

Since A_T holds for all θ, which implies it holds for $\theta = \theta_0$ and $\theta = \widehat{\theta}_T$. Thus A_T implies

$$-\frac{\epsilon}{2} \leq S_T(\theta_0) - S(\theta_0) < \frac{\epsilon}{2}$$

and

$$-\frac{\epsilon}{2} \leq S_T\left(\widehat{\theta}_T\right) - S\left(\widehat{\theta}_T\right) < \frac{\epsilon}{2}.$$

The above inequalities also imply

$$S_T(\theta_0) - S(\theta_0) < \frac{\epsilon}{2}$$

and

$$-\frac{\epsilon}{2} \leq S_T\left(\widehat{\theta}_T\right) - S\left(\widehat{\theta}_T\right).$$

Summing up of the above inequalities gives

$$\begin{aligned} S_T(\theta_0) - S(\theta_0) - \frac{\epsilon}{2} &\leq \frac{\epsilon}{2} + S_T\left(\widehat{\theta}_T\right) - S\left(\widehat{\theta}_T\right) \\ S\left(\widehat{\theta}_T\right) - S(\theta_0) &< \epsilon + S_T\left(\widehat{\theta}_T\right) - S_T(\theta_0) \\ S\left(\widehat{\theta}_T\right) - S(\theta_0) &< \epsilon - \delta \end{aligned}$$

where

$$\delta = S_T(\theta_0) - S_T\left(\widehat{\theta}_T\right) > 0$$

due to the fact that $S_T(\theta)$ is minimized at $\widehat{\theta}_T$.
Thus,

$$S\left(\widehat{\theta}_T\right) - S(\theta_0) < \epsilon - \delta < \epsilon = \min_{\theta \in N^c \cap \Theta} S(\theta) - S(\theta_0).$$

This implies

$$S\left(\widehat{\theta}_T\right) < \min_{\theta \in N^c \cap \Theta} S(\theta),$$

which in turn implies

$$\widehat{\theta}_T \notin N^c \cap \Theta$$

or

$$\widehat{\theta}_T \in N.$$

Thus,

$$A_T \Rightarrow \widehat{\theta}_T \in N,$$

$$\Pr(A_T) \leq \Pr\left(\widehat{\theta}_T \in N\right).$$

Taking limit,

$$1 = \lim_{T \to \infty} \Pr(A_T) \leq \lim_{T \to \infty} \Pr\left(\widehat{\theta}_T \in N\right) \leq 1.$$

Thus,

$$\lim_{T \to \infty} \Pr\left(\widehat{\theta}_T \in N\right) = 1.$$

Thus, $\widehat{\theta}_T$ converges to θ_0 in probability. ■

Asymptotic Normality of NLS Estimators

An nonlinear estimator usually does not have a closed form solution. Even it does, it is often highly complicated. How do we know that such a complicated expression will have a normal distribution asymptotically? For a nonlinear LS estimator to be asymptotically normal, we need to verified the following assumptions.

Assumptions:

(A2) The parameter space Θ is an open subset of R^K. θ_0 belongs to the interior of Θ.

(B2) $S_T(\theta)$ is continuous in an open neighborhood $N_1(\theta_0)$ of the true parameter θ_0.

(C2) $S_T(\theta)$ converges to a non-stochastic function $S(\theta)$ in probability uniformly in an open neighborhood $N_2(\theta_0)$ of θ_0 as $T \to \infty$, and $S(\theta)$ attains a strict local minimum at θ_0.

(D2) $\dfrac{\partial S_T(\theta)}{\partial \theta}$ exists and is continuous in an open neighborhood $N_1(\theta_0)$ of θ_0.

(E2) $\dfrac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}$ exists and is continuous in an open, convex neighborhood of θ_0.

(F2) $\left(\dfrac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}\right)_{\theta_T^*} \xrightarrow{p} A(\theta_0)$ for any sequence $\theta_T^* \xrightarrow{p} \theta_0$, where $A(\theta_0)$ is a finite non-singular matrix defined as

$$A(\theta_0) = \lim_{T\to\infty} E\left(\frac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}\right)_{\theta_0}.$$

(G2) $\sqrt{T}\left(\dfrac{\partial S_T(\theta)}{\partial \theta}\right)_{\theta_0} \xrightarrow{d} N(0, B(\theta_0))$, where

$$B(\theta_0) = \lim_{T\to\infty} TE\left[\left(\frac{\partial S_T(\theta)}{\partial\theta}\right)_{\theta_0}\left(\frac{\partial S_T(\theta)}{\partial\theta'}\right)_{\theta_0}\right].$$

It is sometimes very difficult to go through all these assumptions. To have some practices, consider a linear model first, since a linear model is a special case of a nonlinear model.

Example 149 *The following model satisfies assumption* $(A2)$ *to* $(G2)$*:*

$$y_t = \theta x_t + u_t, \qquad (t = 1, 2, ..., T.)$$

where
x_t are distributed uniformly in $(0,1)$;
$E(x^r) = \int_0^1 x^r dx = \frac{1}{1+r}$;
$u_t \sim i.i.d.(0, \sigma^2)$, $\sigma^2 < \infty$;
u and x are independent;
$\theta \in \Theta = (0,2)$;
$\theta_0 = 1$.

The parameter space $\Theta = (0,2)$ is an open subset of R^K. $\theta_0 = 1$ belongs to the interior of $(0,2)$.

$$f(x_t;\theta) = \theta x_t.$$

$$S_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \theta x_t)^2.$$

For any given set of sample $\{x_t, y_t\}_{t=1}^T$, $S_T(\theta)$ is continuous in θ in an open neighborhood of θ_0 since

$$\begin{aligned} S_T(\theta+\delta) - S_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T}\left((y_t - (\theta+\delta)x_t)^2 - (y_t - \theta x_t)^2\right) \\ &= \delta\frac{1}{T}\sum_{t=1}^{T}(x_t\delta + 2y_t + 2\theta x_t)x_t, \end{aligned}$$

which tends to zero as $\delta \to 0$ for all θ in an open neighborhood of θ_0. Thus, $(B2)$ is satisfied.

Note that

$$S_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}(\theta_0 x_t + u_t - \theta x_t)^2 \xrightarrow{p} \sigma^2 + (\theta_0 - \theta)^2 E\left(x^2\right) = \sigma^2 + \frac{(1-\theta)^2}{3}.$$

Therefore, we let

$$S(\theta) = \sigma^2 + \frac{(1-\theta)^2}{3}.$$

It is easily shown that $S(\theta)$ attains a unique global maximum at $\theta = \theta_0 = 1$. Also, since

$$\begin{aligned}
&\sup_{\theta\in N_2(\theta_0)} |S_T(\theta) - S(\theta)| \\
&= \sup_{\theta\in N_2(\theta_0)} \left|\frac{1}{T}\sum_{t=1}^{T}(y_t - \theta x_t)^2 - S(\theta)\right| \\
&= \sup_{\theta\in N_2(\theta_0)} \left|\frac{1}{T}\sum_{t=1}^{T}(x_t + u_t - \theta x_t)^2 - S(\theta)\right| \\
&= \sup_{\theta\in N_2(\theta_0)} \left|\frac{1}{T}\sum_{t=1}^{T}u_t^2 + \frac{1}{T}(1-\theta)^2\sum_{t=1}^{T}x_t^2 + \frac{2(1-\theta)}{T}\sum_{t=1}^{T}u_t x_t - S(\theta)\right| \\
&\le \sup_{\theta\in N_2(\theta_0)} \left|\frac{1}{T}\sum_{t=1}^{T}u_t^2 - \sigma^2\right| + (1-\theta)^2 \sup_{\theta\in N_2(\theta_0)} \left|\frac{1}{T}\sum_{t=1}^{T}x_t^2 - \frac{1}{3}\right| + (1-\theta)\sup_{\theta\in N_2(\theta_0)} \left|\frac{2}{T}\sum_{t=1}^{T}u_t x_t\right| \\
&= \left|\frac{1}{T}\sum_{t=1}^{T}u_t^2 - \sigma^2\right| + (1-\theta)^2\left|\frac{1}{T}\sum_{t=1}^{T}x_t^2 - \frac{1}{3}\right| + (1-\theta)\left|\frac{2}{T}\sum_{t=1}^{T}u_t x_t\right| \\
&= o_p(1) + o_p(1) + o_p(1) \\
&= o_p(1).
\end{aligned}$$

So $(C2)$ is satisfied.

For any given sample $\{x_t, y_t\}_{t=1}^{T}$,

$$\frac{\partial S_T(\theta)}{\partial \theta} = -\frac{2}{T}\sum_{t=1}^{T}(y_t - \theta x_t)x_t = -\frac{2}{T}\sum_{t=1}^{T}y_t x_t + \left(\frac{2}{T}\sum_{t=1}^{T}x_t^2\right)\theta$$

which is a linear function of θ. Hence it exists and is continuous in an open neighborhood $N_1(\theta_0)$ of θ_0. So $(D2)$ is satisfied.

For any given sample $\{x_t, y_t\}_{t=1}^T$,

$$\frac{\partial S_T^2(\theta)}{\partial \theta^2} = \frac{2}{T}\sum_{t=1}^T x_t^2$$

which exists. Also, it is independent of θ and so it is always continuous in an open, convex neighborhood of θ_0. Thus, $(E2)$ is satisfied.

$$\left.\frac{\partial S_T^2(\theta)}{\partial \theta^2}\right|_{\theta_T^*} = \frac{2}{T}\sum_{t=1}^T x_t^2 \xrightarrow{p} E\left(x^2\right) = \frac{1}{3}$$

for any sequence $\theta_T^* \xrightarrow{p} \theta_0$. So $(F2)$ is satisfied. Now, note that

$$\sqrt{T}\left(\frac{\partial S_T(\theta)}{\partial \theta}\right)_{\theta_0} = \sqrt{T}\left[-\frac{2}{T}\sum_{t=1}^T (y_t - \theta_0 x_t) x_t\right] = -\frac{2}{\sqrt{T}}\sum_{t=1}^T u_t x_t \xrightarrow{d} N(0, B(\theta_0))$$

by the central Limit Theorem, where

$$\begin{aligned} B(\theta_0) &= \lim_{T\to\infty} TE\left[\left(-\frac{2}{T}\sum_{t=1}^T u_t x_t\right)^2\right] = \lim_{T\to\infty} T\left[\frac{4}{T^2}\sum_{t=1}^T E\left(u_t^2 x_t^2\right)\right] \\ &= \lim_{T\to\infty}\frac{4}{T}\sum_{t=1}^T E\left(u_t^2\right)E\left(x_t^2\right) = \lim_{T\to\infty}\frac{4}{T}\sum_{t=1}^T\left(\frac{1}{3}\right) = \frac{4\sigma^2}{3}. \end{aligned}$$

Thus, $(G2)$ is satisfied.

Theorem 150 *Let Θ_T be the set of roots of the equation $\dfrac{\partial S_T(\theta)}{\partial \theta} = 0$ corresponding to the local minima, and $\left\{\widehat{\theta}_T\right\}$ a sequence obtained by choosing one element from Θ_T. Then under assumptions (A2) to (G2), we have*

$$\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) \xrightarrow{d} N\left(0, A(\theta_0)^{-1} B(\theta_0) A(\theta_0)^{-1}\right).$$

Proof. Taking a Taylor expansion, we have

$$\begin{aligned}
\left(\frac{\partial S_T(\theta)}{\partial \theta}\right)_{\widehat{\theta}_T} &= \left(\frac{\partial S_T(\theta)}{\partial \theta}\right)_{\theta_0} + \left(\frac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}\right)_{\theta_T^*} \left(\widehat{\theta}_T - \theta_0\right) \\
0 &= \left(\frac{\partial S_T(\theta)}{\partial \theta}\right)_{\theta_0} + \left(\frac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}\right)_{\theta_T^*} \left(\widehat{\theta}_T - \theta_0\right) \\
\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) &= -\left(\frac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}\right)_{\theta_T^*}^{-1} \left(\sqrt{T}\frac{\partial S_T(\theta)}{\partial \theta}\right)_{\theta_0}.
\end{aligned}$$

Using assumption (F2), (G2) and Theorem 91(ii), we prove the above theorem. ■

Exercise 0.114 *Repeat the exercise of section 9.3.1 on page 171 of Greene 5^{th} edition. You can use the nonlinear least squares estimation procedure in Microfit window version for this problem.*

Exercise 0.115 *Consider the following data set on production*

L_t	K_t	Q_t
0.228	0.802	0.257
0.258	0.249	0.184
0.821	0.771	1.213
0.767	0.511	0.523
0.495	0.758	0.848
0.487	0.425	0.763
0.678	0.452	0.623
0.748	0.817	1.031
0.727	0.845	0.569
0.695	0.958	0.882
0.458	0.084	0.108
0.981	0.021	0.026
0.002	0.295	0.004
0.429	0.277	0.046

a) Estimate the Cobb-Douglas production function

$$Q_t = \alpha L_t^{\beta} K_t^{\gamma} + u_t.$$

b) Estimate the CES production function

$$Q_t = \alpha \left(\beta L_t^{\gamma} + (1-\beta) K_t^{\gamma}\right)^{\delta} + u_t.$$

c) Estimate the following production function

$$Q_t = \alpha^{\beta L_t K_t} + u_t.$$

You can use the nonlinear least squares estimation procedure in Microfit window version for this problem.

Exercise 0.116 *Consider the following model:*

$$y_t = x_t^{\beta} + u_t, \qquad t = 1, 2, ..., T.$$

a) Let $T = 10$, $x_t \sim U(0,1)$, $u_t \sim N(0,1)$, x and u are independent. Let the true parameter be $\beta_0 = 2$. Write a GAUSS program to generate y_t. Print the values of $y_1, ..., y_{10}$.

b) Write a GAUSS program to estimate the parameter β_0 for $T = 10$. Print the estimated value $\widehat{\beta}$.

c) Now repeat the program for $T = 20, 40, 60, ..., 200$. Plot the trend of the estimators for $T = 20, 40, 60, ..., 200$. Is the nonlinear least squares estimator consistent?

d) Use GAUSS to simulate the sampling distribution of the nonlinear least squares estimator for $T = 10, 100$ and 500, using 1000 replications. Is the nonlinear least squares estimators asymptotically normal?

Exercise 0.117 *True or False? Explain:*

(a) If a sequence of functions $S_T(\theta)$ converge to $S(\theta)$ uniformly in the parameter space Θ, then $S_T(\theta)$ converge to $S(\theta)$ pointwise in Θ.

(b) In the nonlinear least squares estimation, if the parameter space is $\Theta = [0, 1]$ and the true parameter is $\theta = 1$, then the nonlinear least squares estimator is asymptotically normally distributed.

(c) The nonlinear least squares estimation relies on the distribution of the error term.

Exercise 0.118 *Consider the model*

$$y_t = (\alpha x_t)^{\beta} + u_t, \qquad t = 1, 2, ..., T.$$

$$f(x_t; \theta_0) = (\alpha_0 x_t)^{\beta_0}.$$

$$\theta_0 = \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}.$$

$$S_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - (\alpha x_t)^{\beta}\right)^2.$$

(a) Derive the First-order conditions

$$\frac{\partial S_T(\theta)}{\partial \alpha} = 0$$

and

$$\frac{\partial S_T(\theta)}{\partial \beta} = 0.$$

(b) Is the model solvable? explained.

Exercise 0.119 *Consider the following model:*

$$y_t = \alpha x_t^{\beta} + u_t, \qquad t = 1, 2, ..., T.$$

a) Let $x_t \sim U(0,1)$, $u_t \sim N(0,1)$, x and u are independent. Let the true parameters be $\alpha_0 = 2$, $\beta_0 = 2$. Write a GAUSS program to estimate the parameters α_0, β_0 for $T = 20$. Print the estimated values $\widehat{\alpha}$ and $\widehat{\beta}$.

b) Plot the trend of the estimators for $T = 20, 40, 60, ..., 200$. Are the nonlinear least squares estimators consistent?

c) Use GAUSS to simulate the sampling distribution of the nonlinear least squares estimators for $T = 10, 100$ and 500, using 1000 replications. Are the nonlinear least squares estimators asymptotically normal?

d) Repeat a) to c) if $\alpha_0 = 0$.
e) Repeat a) to c) if $\beta_0 = 0$.

Exercise 0.120 *Consider the following model:*

$$y_t = \alpha^{\beta x_t} + u_t, \qquad t = 1, 2, ..., T.$$

a) Derive and simplify the first order conditions for the NLS estimator.
b) Let $x_t \sim N(0,1)$, $u_t \sim N(0,1)$, x and u are independent. Let the true parameters be $\alpha_0 = 2$, $\beta_0 = 2$. Write a GAUSS program to estimate the parameters α_0, β_0 for $T = 20$. Print the estimated values $\widehat{\alpha}$ and $\widehat{\beta}$.
c) Now repeat the program for $T = 20, 40, 60, ..., 200$. Plot the trend of the estimators for $T = 20, 40, 60, ..., 200$. Are the nonlinear least squares estimators consistent? Why?
d) Use GAUSS to simulate the sampling distribution of the nonlinear least squares estimators for $T = 10, 100$ and 500, using 1000 replications. Are the nonlinear least squares estimators asymptotically normal? Why?

Exercise 0.121 *(Difficult) Consider the following model:*

$$y_t = y_{t-1}^{\theta} + u_t, \qquad t = 1, 2, ..., T,$$

(a) Describe how to get the nonlinear least-squares estimator for θ.

(b) If the true value of $\theta = 1$, what will be the asymptotic distribution of the nonlinear least square estimator?

Now suppose we estimate θ via the MLE method by assuming u_t follows independent $N(0,1)$.

(c) Derive the log-Likelihood function $\ln L(y;\theta)$ and the scores function S.

(d) Describe how to get the ML estimator $\widehat{\theta}$.

(e) Describe how to get the Information Matrix.

(f) Describe how to form a Wald test for $\theta = 1$.

Exercise 0.122 *Consider the model*

$$y_t = \frac{x_t}{\theta} + u_t, \qquad t = 1, 2, ..., T.$$

$$S_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \frac{x_t}{\theta}\right)^2.$$

(a) Derive the first-order condition

$$\frac{\partial S_T(\theta)}{\partial \theta} = 0.$$

Is this a linear or nonlinear model?

(b) Solve the least square estimator for θ.

(c) Suppose we want to test the null hypothesis that y_t does not depends on x_t, write down the null hypothesis in terms of θ.

(d) Suppose we construct a Wald test under the null $H_0 : \theta = \theta_0$,

$$Wald = \frac{\left(\widehat{\theta} - \theta_0\right)^2}{\widehat{Var}\left(\widehat{\theta}\right)}$$

Discuss the problem associated with the test in this case, and suggest some remedies for the test.

Exercise 0.123 *The following GAUSS can not be executed, find out 5 mistakes in the program.*

```
output file=5120.out reset;
"Simulate the distribution of nonlinear least square estimator";
T=500;
N=1000;
"N=";N;
"T=";T;
beta0=2;
"beta0=";beta0;
betahat=zeros(N,1);
range=4;
increment=0.1;
"increment=";increment;
start=0;
"beta begins from";start;
"beta ends at";start+range;
y=zeros(T,1);
j=1;
do until j<N;
beta=zeros(range/increment,1);
RSS=9^ 99.*ones(range/increment,1);
x=rndu(T,1);
u=rndn(T,1);
y=x^ beta0+u;
```

```
m=1;
do until m>range*increment;
beta[m,1]=start+m*increment;
e=y-x^beta[m,1];
RSS[m,1]=e'e;
endo;
mstar=minindc(RSS);
betahat[j,1]=beta[mstar,1];
j=j+1;
endo;
library pgraph;
graphset;
begwind;
title("T=500, N=1000, beta0=2");
xlabel("T^.5*(betahat-beta0)");
ylabel("frequency");
hist(T^.5*(betahat-beta0),10)
endwind;
```

ECO5120: Econometric Theory and Application, Fall 2006
Prof. T.L. Chong
HANDOUT 7
DISCRETE AND LIMITED DEPENDENT VARIABLE MODELS

Linear Probability Model

In empirical studies, we often encounter variables which are qualitative rather than quantitative. For example, we may be interested in whether people participate in the labor force or not; whether people get married or not; whether people buy a car or not, etc., all these yes-no decisions are not easily quantifiable. In the case where the dependent variable is qualitative, we normally use the technique that, if the dependent variable falls into a certain category, we give it a value of 1, and assign a value of 0 to it if it falls into another category.

Suppose Y is a binary variable, consider a simple regression model

$$Y_t = \beta_0 + \beta_1 X_t + u_t.$$

Note very carefully that we cannot simply assume u_t to be $i.i.d.\,(0, \sigma^2)$, as Y_t cannot be treated as a predicted value in a regression line plus an arbitrary residual. This is because Y_t only takes either 0 or 1, so the residuals also take only two possible values for a given value of X_t.

First, note that

$$E(Y_t) = 1 \times \Pr(Y_t = 1) + 0 \times \Pr(Y_t = 0) = \Pr(Y_t = 1).$$

Further, if $Y_t = 1$, then $u_t = 1 - \beta_0 - \beta_1 X_t$, and if $Y_t = 0$, $u_t = -\beta_0 - \beta_1 X_t$.

$$\begin{aligned} E(u_t) &= (1-\beta_0-\beta_1 X_t)\Pr(Y_t=1)+(-\beta_0-\beta_1 X_t)\Pr(Y_t=0) \\ &= (1-\beta_0-\beta_1 X_t)\Pr(Y_t=1)+(-\beta_0-\beta_1 X_t)(1-\Pr(Y_t=1)) \\ &= \Pr(Y_t=1)-\beta_0-\beta_1 X_t. \end{aligned}$$

We can still assume $E(u_t)=0$ in order to get an unbiased estimator. This will imply

$$\Pr(Y_t=1)-\beta_0-\beta_1 X_t=0,$$

or

$$\Pr(Y_t=1)=\beta_0+\beta_1 X_t.$$

We call this a linear probability model, and β_1 is interpreted as the marginal effect of X_t on the probability of getting $Y_t=1$. To give a concrete example, suppose we have data on two groups of people, one group purchase sport car while the other purchase family car.

We define $Y_t=1$ if a family car is purchased and $Y_t=0$ if a sport car is purchased. Suppose X_t is the family size. Then β_1 is interpreted as: if there is one more member in the family, by how much will the chance of buying a family car increase?

The advantage of using the linear probability model is that it is simple, just run a regression and you will get the parameters of interest. However, there are a lot of problems associated with the linear probability model.

Heteroskedasticity

The first problem is that we cannot assume $Var(u_t)$ to be a constant in this framework. To see why, note that

$$
\begin{aligned}
Var\left(u_t\right) &= E\left(u_t^2\right) - E^2\left(u_t\right) = E\left(u_t^2\right) \\
&= \left(1-\beta_0-\beta_1 X_t\right)^2 \Pr\left(Y_t=1\right) + \left(-\beta_0-\beta_1 X_t\right)^2 \Pr\left(Y_t=0\right) \\
&= \left(1-\beta_0-\beta_1 X_t\right)^2 \Pr\left(Y_t=1\right) + \left(\beta_0+\beta_1 X_t\right)^2 \Pr\left(Y_t=0\right) \\
&= \left(1-\beta_0-\beta_1 X_t\right)\left(\beta_0+\beta_1 X_t\right),
\end{aligned}
$$

which is not a constant and will vary with X_t. Further, it may even be negative. Thus, we have the problem of heteroskedasticity, and the estimators will be inefficient.

Now since the disturbance is heteroskedastic, the OLS estimator will be inefficient, therefore we may use Generalized Least Squares to obtain efficient estimates.

If $0 < \widehat{Y}_t < 1$ for all t, we can get GLS estimators by dividing all the observations by $\sqrt{\left(1-\widehat{\beta}_0-\widehat{\beta}_1 X_t\right)\left(\widehat{\beta}_0+\widehat{\beta}_1 X_t\right)} = \sqrt{\left(1-\widehat{Y}_t\right)\widehat{Y}_t}$.

Non-normality of the disturbances

An additional problem is that the error distribution is not normal. This is because given the value of X_t, the disturbance u_t only takes 2 values, namely, $u_t = 1-\beta_0-\beta_1 X_t$ or $u_t = -\beta_0-\beta_1 X_t$. Thus, u_t actually follows the binomial distribution.

We cannot apply the classical statistical tests to the estimated parameters when the sample is small, since the tests depend on the normality of the errors. However, as sample size increases indefinitely, it can be shown that the OLS estimators tend to be normally distributed. Therefore, in large samples, statistical inference of the LPM can be carried out as usual.

Questionable value of R^2 as a measure of goodness of fit

The conventional R^2 is of limited use in the dichotomous response models. Since all the Y values will either lie along the X axis or along the line corresponding to 1, no linear regression line will fit the data well. As a

result, the conventional R^2 is likely to be much lower than 1 for such models. In most cases, the R^2 ranges from 0.2 to 0.6.

Nonfulfillment of 0 <$\widehat{\Pr(Y_t = 1)}$< 1.

The other problem is on prediction. Since

$$\widehat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 X_t = \widehat{\Pr(Y_t = 1)}$$

is the predicted probability of Y_t being equal to 1 given X_t, which must be bounded between 0 and 1 theoretically. However, the predicted value here is unbounded as we do not impose any restrictions on the values of X_t. An obvious solution for this problem is to set extreme predictions equal to 1 or 0, thereby constraining predicted probabilities within the zero-one interval.

This solution is not very satisfying either, as it suggests that we might have a predicted probability of 1 when it is entirely possible that an event may not occur, or we might have a predicted probability 0 when an event may actually occur. While the estimation procedure might yield unbiased estimates, the predictions obtained from the estimation process are clearly biased.

An alternative approach is to re-estimate the parameters subject to the constraint that the predicted value is bounded between zero and one. Since predicted value is the value in a regression curve, we must find a function $\widehat{Y}_t = g(X_t, \beta)$ such that $0 \leq g(X_t, \beta) \leq 1$ for all β and X_t. Clearly $g(X_t, \beta)$ cannot be linear in either β or X, i.e. $g(X_t, \beta) = \beta_0 + \beta_1 X_t$ will not work.

If we can find a function which is bounded between zero and one, then we can solve the problem of unrealistic prediction. What kind of function will be bounded between zero and one? Actually there are a lot of such functions, one of them is the cumulative distribution function. For example, a normal distribution has an increasing, S-shaped CDF bounded between zero and one. Another example is

$$g(X_t, \beta) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_t)]}.$$

Note that as $\beta_1 X_t \to -\infty$, $g(X_t, \beta) \to 0$, and as $\beta_1 X_t \to \infty$, $g(X_t, \beta) \to 1$. Since $g(X_t, \beta)$ is not linear in β, we cannot use the linear least squares method. Instead, the non-linear least squares or Maximum Likelihood estimation methods should be used.

Random Utility Model

Suppose you have to make a decision on two alternatives, say, whether to buy a sport car or family car. Given the characteristics X_i of individual i , for example, his/her family size, income, etc. Let

$$U_{i1} = \alpha_0 + \alpha_1 X_i + \varepsilon_{i1},$$

$$U_{i2} = \gamma_0 + \gamma_1 X_i + \varepsilon_{i2},$$

where U_{i1} is the utility derived from a family car, and U_{i2} is the utility derived from a sport car. The individual will buy a family car if $U_{i1} > U_{i2}$, or $U_{i1} - U_{i2} > 0$. Subtracting the second equation from the first equation gives

$$U_{i1} - U_{i2} = \alpha_0 - \gamma_0 + (\alpha_1 - \gamma_1) X_i + \varepsilon_{i1} - \varepsilon_{i2}.$$

Suppose we define $Y_i^* = U_{i1} - U_{i2}$, $\beta_0 = \alpha_0 - \gamma_0$, $\beta_1 = \alpha_1 - \gamma_1$, $u_i = \varepsilon_{i1} - \varepsilon_{i2}$. We can rewrite the model as

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i.$$

However, we cannot observe the exact value of Y_i^*, what we observe is whether the individual buy a family car or not. That is, we only observe whether $Y_i^* > 0$ or $Y_i^* < 0$. If $Y_i^* > 0$, the individual will buy a family car, we assign a value $Y_i = 1$ for this observation, and assign $Y_i = 0$ otherwise.

Denote the density function and distribution function of u_i by $f(\cdot)$ and $F(\cdot)$ respectively, and suppose it is symmetric about zero, i.e., $f(u_i) = f(-u_i)$, and $F(u_i) = 1 - F(-u_i)$. We then have:

$$\begin{aligned}
\Pr(Y_i = 1) &= \Pr(Y_i^* > 0) \\
&= \Pr(\beta_0 + \beta_1 X_i + u_i > 0) \\
&= \Pr(-u_i < \beta_0 + \beta_1 X_i) \\
&= \Pr(u_i < \beta_0 + \beta_1 X_i) \qquad \text{since } u_i \text{ is symmetrically distributed about zero,} \\
&= F(\beta_0 + \beta_1 X_i),
\end{aligned}$$

and

$$\Pr(Y_i = 0) = 1 - \Pr(Y_i = 1) = 1 - F(\beta_0 + \beta_1 X_i).$$

Note that the marginal effects of an increase in X_i in the probability is nonlinear in $\beta's$, in particular,

$$\frac{\partial \Pr(Y = 0)}{\partial X_i} = -f(\beta_0 + \beta_1 X_i)\beta_1,$$

$$\frac{\partial \Pr(Y_i = 1)}{\partial X_i} = f(\beta_0 + \beta_1 X_i)\beta_1.$$

Consider the case where $\beta_1 > 0$, since $f(\cdot) > 0$, we have

$$\begin{aligned}
\frac{\partial \Pr(Y_i = 0)}{\partial X_i} &< 0 \\
\frac{\partial \Pr(Y_i = 1)}{\partial X_i} &> 0.
\end{aligned}$$

Maximum Likelihood Estimation (MLE) of the Probit and Logit Models

Let $L(y_1, y_2, ..., y_T; \beta)$ be the joint probability density of the sample observations when the true parameter is β. This is a function of $y_1, y_2, ..., y_T$ and β. As a function of the sample observation it is called a joint probability density function of $y_1, y_2, ..., y_T$. As a function of the parameter β it is called

the **likelihood function** for β**.** The MLE method is to choose a value of β which maximizes $L(y_1, y_2, ..., y_T; \beta)$.

Intuitively speaking, if you are faced with several values of β, each of which might be the true value, your best guess is the value which would have made the sample actually observed have the highest probability.

Suppose we have T observations of Y and X, where Y takes the value zero or one. The probability of getting such observations is

$$\begin{aligned} L &= \Pr(Y_1 = y_1, Y_2 = y_2, ..., Y_T = y_T) \\ &= \Pr(Y_1 = y_1) \Pr(Y_2 = y_2) ... \Pr(Y_T = y_T) \end{aligned}$$

by the independence of u_t.

Since y_t only takes either zero or one, we can group them into two groups.

$$\begin{aligned} L &= \prod_{y_t=1} \Pr(Y_t = 1) \prod_{y_t=0} \Pr(Y_t = 0) \\ &= \prod_{y_t=1} F(\beta_0 + \beta_1 X_t) \prod_{y_t=0} [1 - F(\beta_0 + \beta_1 X_t)] \\ &= \prod_{t=1}^{T} [F(\beta_0 + \beta_1 X_t)]^{Y_t} [1 - F(\beta_0 + \beta_1 X_t)]^{1-Y_t}. \end{aligned}$$

$$\begin{aligned} \ln L &= \ln \left\{ \prod_{t=1}^{T} [F(\beta_0 + \beta_1 X_t)]^{Y_t} [1 - F(\beta_0 + \beta_1 X_t)]^{1-Y_t} \right\} \\ &= \sum_{t=1}^{T} \ln \left\{ [F(\beta_0 + \beta_1 X_t)]^{Y_t} [1 - F(\beta_0 + \beta_1 X_t)]^{1-Y_t} \right\} \\ &= \sum_{t=1}^{T} Y_t \ln F(\beta_0 + \beta_1 X_t) + \sum_{t=1}^{T} (1 - Y_t) \ln [1 - F(\beta_0 + \beta_1 X_t)]. \end{aligned}$$

We want to maximize L, or equivalently, maximize $\ln L$ since $\ln(\cdot)$ is a monotonic increasing function. The first order conditions are

$$\frac{\partial \ln L}{\partial \beta_0} = \sum_{t=1}^{T} Y_t \frac{f\left(\beta_0 + \beta_1 X_t\right)}{F\left(\beta_0 + \beta_1 X_t\right)} - \sum_{t=1}^{T} \left(1 - Y_t\right) \frac{f\left(\beta_0 + \beta_1 X_t\right)}{1 - F\left(\beta_0 + \beta_1 X_t\right)} = 0,$$

$$\frac{\partial \ln L}{\partial \beta_1} = \sum_{t=1}^{T} Y_t X_t \frac{f\left(\beta_0 + \beta_1 X_t\right)}{F\left(\beta_0 + \beta_1 X_t\right)} - \sum_{t=1}^{T} \left(1 - Y_t\right) X_t \frac{f\left(\beta_0 + \beta_1 X_t\right)}{1 - F\left(\beta_0 + \beta_1 X_t\right)} = 0.$$

These two equations can be solved to obtain estimators for $\beta' s$. However, as $\ln L$ is a highly nonlinear function of $\beta' s$, we cannot easily get the estimator of $\beta' s$ by simple substitution. We may use the grid-search method and a computer algorithm to solve the problem.

The MLE procedure has a number of desirable properties. When sample size is large, all estimators are consistent and also efficient if there is no misspecification on the probability distribution. In addition, all parameters are known to be normally distributed when sample size is large.

If we **assume** u_t to be **normally** distributed $N\left(0, \sigma^2\right)$, i.e.,

$$f\left(\beta_0 + \beta_1 X_t\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\beta_0 + \beta_1 X_t\right)^2}{2\sigma^2}\right),$$

$$F\left(\beta_0 + \beta_1 X_t\right) = \int_{-\infty}^{\beta_0 + \beta_1 X_t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt,$$

then we have the **Probit Model**.

The first order condition can be simplified to

$$\frac{\partial \ln L}{\partial \beta_0} = \sum_{Y_t=1} \frac{\exp\left(-\frac{\left(\beta_0 + \beta_1 X_t\right)^2}{2\sigma^2}\right)}{\int_{-\infty}^{\beta_0 + \beta_1 X_t} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt} - \sum_{Y_t=0} \frac{\exp\left(-\frac{\left(\beta_0 + \beta_1 X_t\right)^2}{2\sigma^2}\right)}{\int_{\beta_0 + \beta_1 X_t}^{\infty} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt} = 0,$$

$$\frac{\partial \ln L}{\partial \beta_1} = \sum_{Y_t=1} \frac{X_t \exp\left(-\frac{(\beta_0+\beta_1 X_t)^2}{2\sigma^2}\right)}{\int_{-\infty}^{\beta_0+\beta_1 X_t} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt} - \sum_{Y_t=0} \frac{X_t \exp\left(-\frac{(\beta_0+\beta_1 X_t)^2}{2\sigma^2}\right)}{\int_{\beta_0+\beta_1 X_t}^{\infty} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt} = 0.$$

Although the normal distribution is a commonly used distribution, its distribution function is not a closed form function of u_t. As the two first order conditions above involve the integration operator, the computational cost will be tremendous. For mathematical convenience, the **logistic distribution** is proposed:

$$\begin{aligned} f(\beta_0+\beta_1 X_t) &= \frac{\exp(\beta_0+\beta_1 X_t)}{(1+\exp(\beta_0+\beta_1 X_t))^2}, \\ F(\beta_0+\beta_1 X_t) &= \frac{\exp(\beta_0+\beta_1 X_t)}{1+\exp(\beta_0+\beta_1 X_t)}. \end{aligned}$$

If we assume u_t to have a logistic distribution, then we have the **Logit Model**. The first order condition can be simplified to

$$\frac{\partial \ln L}{\partial \beta_0} = \sum_{Y_t=1} \frac{1}{1+\exp(\beta_0+\beta_1 X_t)} - \sum_{Y_t=0} \frac{1}{1+\exp(-\beta_0-\beta_1 X_t)} = 0,$$

$$\frac{\partial \ln L}{\partial \beta_1} = \sum_{Y_t=1} \frac{X_t}{1+\exp(\beta_0+\beta_1 X_t)} - \sum_{Y_t=0} \frac{X_t}{1+\exp(-\beta_0-\beta_1 X_t)} = 0.$$

Polychotomous Variables with Unordered Data: The Multinomial Logit Model

Suppose there are n individuals and J categories, e.g., Occupational choice. Define $Y_{ij} = 1$ if individual i choose category j, and $Y_{ij} = 0$ otherwise. Thus, $\sum_{j=1}^{J} Y_{ij} = 1$ for all i.

For a simple analysis, we assume $J = 3$. Assume that an individual i whose utilities associated with three alternatives are given by

$$U_{ij} = X_i'\beta_j + \varepsilon_{ij}, \qquad j = 1, 2, 3.$$

where X and β are vectors.

Assume that ε_{ij} are independent and identically distributed, each with the *extreme value* distribution

$$F(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij})),$$

$$f(\varepsilon_{ij}) = \exp(-\varepsilon_{ij}) \exp(-\exp(-\varepsilon_{ij})).$$

Now, suppose there are 3 categories, category 1, 2 and 3. The probability that individual i will choose category 2 is

$$
\begin{aligned}
& \Pr\left(Y_{i2}=1\right) \\
= \; & \Pr\left(U_{i2}>U_{i1} \text{ and } U_{i2}>U_{i3}\right) \\
= \; & \Pr\left(X_i'\beta_2+\varepsilon_{i2}>X_i'\beta_1+\varepsilon_{i1} \text{ and } X_i'\beta_2+\varepsilon_{i2}>X_i'\beta_3+\varepsilon_{i3}\right) \\
= \; & \Pr\left(\varepsilon_{i1}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_1\right) \text{ and } \varepsilon_{i3}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_3\right)\right) \\
= \; & \int_{-\infty}^{\infty} f\left(\varepsilon_{i2}\right)\Pr\left(\varepsilon_{i1}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_1\right) \text{ and } \varepsilon_{i3}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_3\right)|\varepsilon_{i2}\right)d\varepsilon_{i2} \\
= \; & \int_{-\infty}^{\infty} f\left(\varepsilon_{i2}\right)\Pr\left(\varepsilon_{i1}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_1\right)|\varepsilon_{i2}\right)\Pr\left(\varepsilon_{i3}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_3\right)|\varepsilon_{i2}\right)d\varepsilon_{i2} \\
= \; & \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\varepsilon_{i2}+X_i'(\beta_2-\beta_1)} f\left(\varepsilon_{i1}\right)d\varepsilon_{i1}\right]\left[\int_{-\infty}^{\varepsilon_{i2}+X_i'(\beta_2-\beta_3)} f\left(\varepsilon_{i3}\right)d\varepsilon_{i3}\right]dF\left(\varepsilon_{i2}\right) \\
= \; & \int_{-\infty}^{\infty}\left[\exp\left(-\exp\left(-\varepsilon_{i2}-X_i'\left(\beta_2-\beta_1\right)\right)\right)\right]\left[\exp\left(-\exp\left(-\varepsilon_{i2}-X_i'\left(\beta_2-\beta_3\right)\right)\right)\right]dF\left(\varepsilon_{i2}\right) \\
= \; & \int_{-\infty}^{\infty}\exp\left[-\exp\left(-\varepsilon_{i2}\right)\exp\left(-X_i'\left(\beta_2-\beta_1\right)\right)\right]\exp\left[-\exp\left(-\varepsilon_{i2}\right)\exp\left(-X_i'\left(\beta_2-\beta_3\right)\right)\right]dF\left(\varepsilon_{i2}\right) \\
= \; & \int_{-\infty}^{\infty} F\left(\varepsilon_{i2}\right)^{\exp\left(-X_i'(\beta_2-\beta_1)\right)} F\left(\varepsilon_{i2}\right)^{\exp\left(-X_i'(\beta_2-\beta_3)\right)} dF\left(\varepsilon_{i2}\right) \\
= \; & \int_{-\infty}^{\infty} F\left(\varepsilon_{i2}\right)^{\exp\left(-X_i'(\beta_2-\beta_1)\right)+\exp\left(-X_i'(\beta_2-\beta_3)\right)} dF\left(\varepsilon_{i2}\right) \\
= \; & \left[\frac{F\left(\varepsilon_{i2}\right)^{1+\exp\left(-X_i'(\beta_2-\beta_1)\right)+\exp\left(-X_i'(\beta_2-\beta_3)\right)}}{1+\exp\left(-X_i'\left(\beta_2-\beta_1\right)\right)+\exp\left(-X_i'\left(\beta_2-\beta_3\right)\right)}\right]_{-\infty}^{\infty} \\
= \; & \frac{1}{1+\exp\left(-X_i'\left(\beta_2-\beta_1\right)\right)+\exp\left(-X_i'\left(\beta_2-\beta_3\right)\right)} \\
= \; & \frac{\exp\left(X_i'\beta_2\right)}{\exp\left(X_i'\beta_1\right)+\exp\left(X_i'\beta_2\right)+\exp\left(X_i'\beta_3\right)}.
\end{aligned}
$$

Therefore, if there are J categories, the probability that individual i will choose the jth category will be

$$\Pr\left(Y_{ij}=1\right)=\frac{\exp\left(X_i'\beta_j\right)}{\sum_{k=1}^{J}\exp\left(X_i'\beta_k\right)}.$$

One problem arises here, the β_j here cannot be identified as if we change all the β to $\beta+c$, where c is a vector of any constant, $\Pr\left(Y_{ij}=1\right)$ will still be the same since

$$\frac{\exp\left(X_i'\left(\beta_j+c\right)\right)}{\sum_{k=1}^{J}\exp\left(X_i'\left(\beta_k+c\right)\right)}=\frac{\exp\left(X_i'c\right)\exp\left(X_i'\left(\beta_j+c\right)\right)}{\exp\left(X_i'c\right)\sum_{k=1}^{J}\exp\left(X_i'\left(\beta_k+c\right)\right)}=\frac{\exp\left(X_i'\beta_j\right)}{\sum_{k=1}^{J}\exp\left(X_i'\beta_k\right)}.$$

Therefore, for the parameter to be identified, we must impose some restrictions on β. We can simply let $\beta_1=0$, so that

$$\Pr\left(Y_{i1}=1\right)=\frac{1}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)},$$

$$\Pr\left(Y_{ij}=1\right)=\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}\qquad j=2,3,...,J.$$

So the likelihood function is

$$L=\prod_{i=1}^{n}\prod_{j=1}^{J}\Pr\left(Y_{ij}=1\right)^{Y_{ij}}=\prod_{i=1}^{n}\prod_{j=1}^{J}\left[\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}\right]^{Y_{ij}}$$

By using the conditions that $\beta_1=0$ and $\sum_{j=1}^{J}Y_{ij}=1$, we have

$$\begin{aligned}\ln L &= \sum_{i=1}^{n}\sum_{j=1}^{J}Y_{ij}\ln\left(\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}Y_{ij}\left(X_i'\beta_j-\ln\left[1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)\right]\right)\\ &= \sum_{i=1}^{n}\left(\sum_{j=1}^{J}Y_{ij}X_i'\beta_j-\left(\sum_{j=1}^{J}Y_{ij}\right)\ln\left[1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)\right]\right)\\ &= \sum_{i=1}^{n}\left(\sum_{j=2}^{J}Y_{ij}X_i'\beta_j-\ln\left[1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)\right]\right).\end{aligned}$$

$$\frac{\partial\ln L}{\partial\beta_j}=\sum_{i=1}^{n}\left(Y_{ij}X_i'-\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}X_i'\right)=\sum_{i=1}^{n}\left(Y_{ij}-\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}\right)X_i'.$$

Ordered Data

Some multinomial-choice variables are inherently ordered, e.g., Bond ratings, opinion surveys, employment (unemployed, part time, or full time). Consider the model

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i.$$

where Y_i^* is unobserved. What we observe is

$$\begin{aligned} Y_i &= 1 \quad \text{if } \mu_0 < Y_i^* \leq \mu_1, \\ &= 2 \quad \text{if } \mu_1 < Y_i^* \leq \mu_2, \\ &= 3 \quad \text{if } \mu_2 < Y_i^* \leq \mu_3, \\ &\vdots \\ &= J \quad \text{if } \mu_{J-1} < Y_i^* \leq \mu_J, \end{aligned}$$

where $\mu_0 = -\infty$ and $\mu_J = \infty$. Other $\mu' s$ are unknown parameters to be estimated with $\beta' s$.

$$\begin{aligned} \Pr(Y_i = j) &= \Pr\left(\mu_{j-1} < Y_i^* \leq \mu_j\right) \\ &= \Pr\left(\mu_{j-1} < \beta_0 + \beta_1 X_i + u_i \leq \mu_j\right) \\ &= \Pr\left(u_i \leq \mu_j - \beta_0 - \beta_1 X_i\right) - \Pr\left(u_i \leq \mu_{j-1} - \beta_0 - \beta_1 X_i\right) \\ &= F\left(\mu_j - \beta_0 - \beta_1 X_i\right) - F\left(\mu_{j-1} - \beta_0 - \beta_1 X_i\right), \end{aligned}$$

We can either assume u_i is normally distributed, or has a logistic distribution.

Suppose we have n observations of Y and X, where Y takes the value $1, 2, ..., J$. The probability of getting such observations is

$$L = \Pr(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) = \Pr(Y_1 = y_1) \Pr(Y_2 = y_2) ... \Pr(Y_n = y_n)$$

by the independence of u_i.

The likelihood function is

$$\begin{aligned} L &= \prod_{y_i=1} \Pr(Y_i = 1) \prod_{y_i=2} \Pr(Y_i = 2) ... \prod_{y_i=J} \Pr(Y_i = J). \\ &= \prod_{i=1}^{n} \prod_{j=1}^{J} \left[F\left(\mu_j - \beta_0 - \beta_1 X_i\right) - F\left(\mu_{j-1} - \beta_0 - \beta_1 X_i\right) \right]^{d_j} \end{aligned}$$

where $d_j = 1$ if $Y_i = j$ and $d_j = 0$ otherwise.

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{J} d_j \ln \left\{ \left[F\left(\mu_j - \beta_0 - \beta_1 X_i\right) - F\left(\mu_{j-1} - \beta_0 - \beta_1 X_i\right) \right] \right\}.$$

Consider a simple example:

Example 151 *Suppose there are only 3 ordered categories, then*

$$\begin{aligned} \Pr(Y_i = 1) &= F(\mu_1 - \beta_0 - \beta_1 X_i), \\ \Pr(Y_i = 2) &= F(\mu_2 - \beta_0 - \beta_1 X_i) - F(\mu_1 - \beta_0 - \beta_1 X_i), \\ \Pr(Y_i = 3) &= 1 - F(\mu_2 - \beta_0 - \beta_1 X_i). \end{aligned}$$

Consider the case where $\beta_1 > 0$. For the three probabilities, the marginal effects of changes in the regressors are

$$\begin{aligned} \frac{\partial \Pr(Y_i = 1)}{\partial X_i} &= -f(\mu_1 - \beta_0 - \beta_1 X_i)\beta_1 < 0, \\ \frac{\partial \Pr(Y_i = 2)}{\partial X_i} &= [f(\mu_2 - \beta_0 - \beta_1 X_i) - f(\mu_1 - \beta_0 - \beta_1 X_i)]\beta_1 = ?, \\ \frac{\partial \Pr(Y_i = 3)}{\partial X_i} &= f(\mu_2 - \beta_0 - \beta_1 X_i)\beta_1 > 0. \end{aligned}$$

Thus, in the general case, given the signs of the coefficients, only the signs of the changes in $\Pr(Y_i = 1)$ and $\Pr(Y_i = J)$ are unambiguous. What happens to the middle cell is ambiguous.

Truncation of data

Sometimes we cannot perfectly observe the actual value of the dependent variable. In the previous section, when decisions are dichotomous (yes-no decision), we only observe the sign of the dependent variable. If we only observe a subpopulation such as individuals with income above a certain level, then the data is said to be lower-truncated, in the sense that we can never observe people with income below that level.

Let Y be a random variable which takes values between $-\infty$ and ∞, with $f(Y) \geq 0$ and $\int_{-\infty}^{\infty} f(Y)\, dY = 1$. Suppose Y is being lower-truncated at $Y = a$, and we can only observe those Y that are bigger than a. Now since we only observe $Y > a$, $\Pr(Y > a) = \int_a^{\infty} f(Y) < 1$, so we have to change the unconditional density function $f(Y)$ into a conditional density function $f(Y|Y > a)$ such that $\int_a^{\infty} f(Y|Y > a)\, dY = 1$. Recall the definition of conditional probability that $\Pr(A|B) = \dfrac{Pr(A \cap B)}{P(B)}$. Let A be the event that $Y < c$, and B be the event that $Y > a$.

$$\Pr(Y < c|Y > a) = \frac{\Pr(Y < c \cap Y > a)}{P(Y > a)} = \frac{\int_a^c f(Y)\, dY}{\int_a^{\infty} f(Y)\, dY},$$

$$f(Y = c|Y > a) = \frac{d \Pr(Y < c|Y > a)}{dc} = \frac{f(c)}{\int_a^{\infty} f(Y)\, dY}.$$

Example 152 *Suppose Y is uniformly distributed in the $[0, 1]$ interval, we know that $f(Y) = 1$ and $F(Y) = Y$. Thus, it is easy to find the unconditional probability $\Pr(Y > 3/4) = 1/4$. But suppose now we know that Y must be greater than $1/2$, how will this re-adjust our prediction for $\Pr(Y > 3/4)$?*

Solution: Using the above rule

$$\Pr\left(Y > \frac{3}{4} \middle| Y > \frac{1}{2}\right) = \frac{\Pr\left(Y > \frac{3}{4} \cap Y > \frac{1}{2}\right)}{\Pr\left(Y > \frac{1}{2}\right)} = \frac{\Pr\left(Y > \frac{3}{4}\right)}{\Pr\left(Y > \frac{1}{2}\right)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

Moments of Truncated Distributions

Note that $E(Y)$ is a weighted average of $E(Y|Y < a)$ and $E(Y|Y > a)$ since

$$\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} Y f(Y)\, dY \\
&= \int_{-\infty}^{a} Y f(Y)\, dY + \int_{a}^{\infty} Y f(Y)\, dY \\
&= \int_{-\infty}^{a} Y \frac{f(Y)}{\Pr(Y < a)} dY \Pr(Y < a) + \int_{a}^{\infty} Y \frac{f(Y)}{\Pr(Y > a)} dY \Pr(Y > a) \\
&= \int_{-\infty}^{a} Y f(Y|Y < a)\, dY \Pr(Y < a) + \int_{a}^{\infty} Y f(Y|Y > a)\, dY \Pr(Y > a) \\
&= E(Y|Y < a) \Pr(Y < a) + E(Y|Y > a) \Pr(Y > a).
\end{aligned}$$

This implies

$$\min\{E(Y|Y < a),\, E(Y|Y > a)\} < E(Y) < \max\{E(Y|Y < a),\, E(Y|Y > a)\}.$$

Since $E(Y|Y < a) < E(Y|Y > a)$, we have

$$\begin{aligned}
E(Y|Y \geq a) &= \int_{a}^{\infty} Y f(Y|Y \geq a)\, dY \geq E(Y), \\
E(Y|Y < a) &= \int_{-\infty}^{a} Y f(Y|Y < a)\, dY \leq E(Y).
\end{aligned}$$

If the truncation is from below, the mean of the truncated variable is greater than the mean of the original one. If the truncation is from above, the mean of the truncated variable is smaller than the mean of the original one.

Example 153 *Find* $E(u|u>1)$ *and* $Var(u|u>1)$ *if* $f(u)=\exp(-u)$, $u>0$, *and compare them to their unconditional means and variances.*

Solution:

$$\begin{aligned}
E(u \mid u>1) &= \int_1^\infty uf(u \mid u>1)\,du \\
&= \frac{1}{1-F(1)}\int_1^\infty uf(u)\,du \\
&= \frac{1}{e^{-1}}\int_1^\infty u\exp(-u)\,du \\
&= \frac{1}{e^{-1}}\left\{[-u\exp(-u)]_1^\infty + \int_1^\infty \exp(-u)\,du\right\} \\
&= \frac{e^{-1}}{e^{-1}} + \frac{1-F(1)}{1-F(1)} \\
&= 2 \\
&> E(u) = 1.
\end{aligned}$$

$$\begin{aligned}
Var(u \mid u>1) &= E\left(u^2 \mid u>1\right) - [E(u \mid u>1)]^2 \\
&= \int_1^\infty u^2 f(u \mid u>1)\,du - 4 \\
&= \frac{1}{1-F(1)}\int_1^\infty u^2 f(u)\,du - 4 \\
&= e\int_1^\infty u^2 f(u)\,du - 4 \\
&= e\int_1^\infty u^2\exp(-u)\,du - 4 \\
&= e\left[\left[-u^2\exp(-u)\right]_1^\infty + 2\int_1^\infty u\exp(-u)\,du\right] - 4 \\
&= e\left[e^{-1} + 2\times 2e^{-1}\right] - 4 \\
&= 1 = Var(u).\ \blacksquare
\end{aligned}$$

Maximum Likelihood Estimation of the Truncated Model

Consider the simple model

$$Y_t = \beta_0 + \beta_1 X_t + u_t > a.$$

$$\begin{aligned} \Pr(Y_t > a) &= \Pr(\beta_0 + \beta_1 X_t + u_t > a) \\ &= \Pr(u_t > a - \beta_0 - \beta_1 X_t) \\ &= 1 - F(a - \beta_0 - \beta_1 X_t). \end{aligned}$$

The Likelihood function is

$$\begin{aligned} L &= f(Y_1 = y_1, Y_2 = y_2, ..., Y_T = y_T | Y_1 > a, Y_2 > a, ..., Y_T > a) \\ &= f(y_1 - \beta_0 - \beta_1 X_1 | Y_1 > a) f(y_2 - \beta_0 - \beta_1 X_2 | Y_2 > a) ... f(y_T - \beta_0 - \beta_1 X_T | Y_T > a) \end{aligned}$$

The Log-likelihood function is

$$\begin{aligned} \ln L &= \ln [f(y_1 - \beta_0 - \beta_1 X_1 | Y_1 > a) f(y_2 - \beta_0 - \beta_1 X_2 | Y_2 > a) ... f(y_T - \beta_0 - \beta_1 X_T | Y_T > a)] \\ &= \sum_{t=1}^{T} \ln f(y_t - \beta_0 - \beta_1 X_t | Y_t > a) = \sum_{t=1}^{T} \ln \frac{f(y_t - \beta_0 - \beta_1 X_t)}{\Pr(Y_t > a)} \\ &= \sum_{t=1}^{T} \ln f(y_t - \beta_0 - \beta_1 X_t) - \sum_{t=1}^{T} \ln [1 - F(a - \beta_0 - \beta_1 X_t)]. \end{aligned}$$

First order conditions:

$$\frac{\partial \ln L}{\partial \beta_0} = -\sum_{t=1}^{T} \frac{f'(y_t - \beta_0 - \beta_1 X_t)}{f(y_t - \beta_0 - \beta_1 X_t)} - \sum_{t=1}^{T} \frac{f(a - \beta_0 - \beta_1 X_t)}{1 - F(a - \beta_0 - \beta_1 X_t)} = 0,$$

$$\frac{\partial \ln L}{\partial \beta_1} = -\sum_{t=1}^{T} X_t \frac{f'(y_t - \beta_0 - \beta_1 X_t)}{f(y_t - \beta_0 - \beta_1 X_t)} - \sum_{t=1}^{T} X_t \frac{f(a - \beta_0 - \beta_1 X_t)}{1 - F(a - \beta_0 - \beta_1 X_t)} = 0.$$

Maximum Likelihood Estimation of the Tobit Model

Sometimes data are **censored** rather than truncated. When the dependent variable is censored, values in a certain range are all transformed to a single value. Suppose we are interested in the demand for a certain hotel's accommodation, if the demand is higher than the hotel's capacity, we will never know the value of actual demand, and all this over-demand values are reported as the total number of rooms in this hotel. We may also observe people either work for a certain hour or not work at all. If people do not work at all, their optimal working hour may be negative. But we will never observe a negative working hour, we observe zero working hour instead. Suppose the data is lower-censored at zero.

$$\begin{aligned} Y_t^* &= \beta_0 + \beta_1 X_t + u_t, \\ Y_t &= 0 \quad \text{if } Y_t^* \leq 0, \\ Y_t &= Y_t^* \quad \text{if } Y_t{}^* > 0. \end{aligned}$$

Y_t^* is not observable, we can only observe Y_t and X_t. To fully utilize the information, if the observation is not censored, we calculate the density value at that point of observation $f\left(Y_t - \beta_0 - \beta_1 X_t\right)$. If the observation is censored, we use the probability of observing a censored value $\Pr\left(Y_t = 0\right)$. Note that:

$$\begin{aligned} \Pr\left(Y_t = 0\right) &= \Pr\left(\beta_0 + \beta_1 X_t + u_t \leq 0\right) \\ &= \Pr\left(u_t \leq -\beta_0 - \beta_1 X_t\right) \\ &= 1 - F\left(\beta_0 + \beta_1 X_t\right). \end{aligned}$$

The Likelihood function is

$$L = \prod_{Y_t > 0} f\left(Y_t - \beta_0 - \beta_1 X_t\right) \prod_{Y_t = 0} \Pr\left(Y_t = 0\right).$$

The Log-likelihood function is

$$\begin{aligned}\ln L &= \ln\left[\prod_{Y_t>0} f\left(Y_t-\beta_0-\beta_1 X_t\right)\prod_{Y_t=0}\Pr\left(Y_t=0\right)\right] \\ &= \sum_{Y_t>0}\ln f\left(Y_t-\beta_0-\beta_1 X_t\right)+\sum_{Y_t=0}\ln\left[1-F\left(\beta_0+\beta_1 X_t\right)\right].\end{aligned}$$

First order condition:

$$\frac{\partial \ln L}{\partial \beta_0}=-\sum_{Y_t>0}\frac{f'\left(Y_t-\beta_0-\beta_1 X_t\right)}{f\left(Y_t-\beta_0-\beta_1 X_t\right)}-\sum_{Y_t=0}\frac{f\left(\beta_0+\beta_1 X_t\right)}{1-F\left(\beta_0+\beta_1 X_t\right)}=0,$$

$$\frac{\partial \ln L}{\partial \beta_1}=-\sum_{Y_t>0}X_t\frac{f'\left(Y_t-\beta_0-\beta_1 X_t\right)}{f\left(Y_t-\beta_0-\beta_1 X_t\right)}-\sum_{Y_t=0}X_t\frac{f\left(\beta_0+\beta_1 X_t\right)}{1-F\left(\beta_0+\beta_1 X_t\right)}=0.$$

If $u_t\sim N\left(0,\sigma^2\right)$, and let $\phi\left(\cdot\right)$ and $\Phi\left(\cdot\right)$ denote the density and distribution functions of an $N\left(0,1\right)$ respectively.

$$f\left(Y_t-\beta_0-\beta_1 X_t\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(Y_t-\beta_0-\beta_1 X_t\right)^2}{2\sigma^2}\right)=\frac{1}{\sigma}\phi\left(\frac{Y_t-\beta_0-\beta_1 X_t}{\sigma}\right).$$

$$f'\left(Y_t-\beta_0-\beta_1 X_t\right)=\frac{1}{\sigma^2}\phi'\left(\frac{Y_t-\beta_0-\beta_1 X_t}{\sigma}\right).$$

$$f\left(\beta_0+\beta_1 X_t\right)=\frac{1}{\sigma}\phi\left(\frac{\beta_0+\beta_1 X_t}{\sigma}\right).$$

$$F\left(\beta_0+\beta_1 X_t\right)=\Phi\left(\frac{\beta_0+\beta_1 X_t}{\sigma}\right).$$

Then the log-likelihood can be rewritten as

$$\ln L=\sum_{Y_t>0}\ln\frac{1}{\sigma}\phi\left(\frac{Y_t-\beta_0-\beta_1 X_t}{\sigma}\right)+\sum_{Y_t=0}\ln\left[1-\Phi\left(\frac{\beta_0+\beta_1 X_t}{\sigma}\right)\right].$$

Example 154 *Consider the model* $Y_t = \beta_0 + \beta_1 X_t + u_t$. *If the dependent variable is upper-truncated at* c_1 *and lower-censored at* c_2, *for any 2 constants* $c_2 < c_1 < \infty$. *Derive the log-likelihood function of such a model.*

Solution: The likelihood function is given by

$$\begin{aligned} L &= \prod_{Y_t > c_2} f\left(Y_t - \beta_0 - \beta_1 X_t \mid Y_t < c_1\right) \prod_{Y_t = c_2} \Pr\left(Y_t = c_2 \mid Y_t < c_1\right) \\ &= \prod_{Y_t > c_2} \frac{f\left(Y_t - \beta_0 - \beta_1 X_t\right)}{\Pr\left(Y_t < c_1\right)} \prod_{Y_t = c_2} \frac{\Pr\left(Y_t = c_2\right)}{\Pr\left(Y_t < c_1\right)}. \end{aligned}$$

where

$$\begin{aligned} \Pr\left(Y_t = c_2\right) &= \Pr\left(\beta_0 + \beta_1 X_t + u_t < c_2\right) \\ &= \Pr\left(u_t < c_2 - \beta_0 - \beta_1 X_t\right) \\ &= F\left(c_2 - \beta_0 - \beta_1 X_t\right) \\ \text{and } \Pr\left(Y_t < c_1\right) &= \Pr\left(\beta_0 + \beta_1 X_t + u_t < c_1\right) \\ &= F\left(c_1 - \beta_0 - \beta_1 X_t\right). \end{aligned}$$

The log-likelihood function is given by

$$\begin{aligned} \ln L &= \sum_{Y_t > c_2} \ln \frac{f\left(Y_t - \beta_0 - \beta_1 X_t\right)}{\Pr\left(Y_t < c_1\right)} + \sum_{Y_t = c_2} \ln \frac{\Pr\left(Y_t = c_2\right)}{\Pr\left(Y_t < c_1\right)} \\ &= \sum_{Y_t > c_2} \ln \frac{f\left(Y_t - \beta_0 - \beta_1 X_t\right)}{F\left(c_1 - \beta_0 - \beta_1 X_t\right)} + \sum_{Y_t = c_2} \ln \frac{F\left(c_2 - \beta_0 - \beta_1 X_t\right)}{F\left(c_1 - \beta_0 - \beta_1 X_t\right)}. \blacksquare \end{aligned}$$

Exercise 0.124 *Consider the truncated model*

$$Y_i = \beta_0 + \beta_1 X_i + u_i > a,$$

where u_i are i.i.d. with density function and distribution function

$$f\left(u_i\right) = \exp\left(-u_i\right)$$

and

$$F(u_i) = 1 - \exp(-u_i)$$

respectively.

(a) Show that

$$\Pr(Y_i > a) = \exp(\beta_0 + \beta_1 X_i - a)$$

(b) Suppose we have n observations of Y and X, find the log-likelihood function.

(c) Find $\dfrac{\partial \ln L}{\partial \beta_0}$ and $\dfrac{\partial \ln L}{\partial \beta_1}$. Discuss the identifiability of β_0 and β_1.

Exercise 0.125 *Consider the Probit model*

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i.$$

Suppose we can only observe the sign of Y_i^*. If $Y_i^* > 0$, we assign a value $Y_i = 1$ for this observation, and assign $Y_i = 0$ otherwise. Denote the density function and distribution function of u_i by $f(\cdot)$ and $F(\cdot)$ respectively, where

$$f(u_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{u_i^2}{2\sigma^2}\right),$$

$$F(u_i) = \int_{-\infty}^{u_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv,$$

(a) Show that

$$\Pr(Y_i = 1) = F(\beta_0 + \beta_1 X_i),$$

and

$$\Pr(Y_i = 0) = 1 - F(\beta_0 + \beta_1 X_i).$$

(b) Suppose we have T observations of Y and X, where Y takes the value zero or one. Assume u_i to be independent, show that the log-likelihood function can be simplified to

$$\ln L = \sum_{i=1}^{T} Y_i \ln \int_{-\infty}^{\beta_0+\beta_1 X_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv + \sum_{i=1}^{T} (1 - Y_i) \ln \left[\int_{\beta_0+\beta_1 X_i}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv\right.$$

(c) Let $w = \dfrac{v}{\sigma}$, show that

$$\ln L = \sum_{i=1}^{T} Y_i \ln \int_{-\infty}^{\frac{\beta_0}{\sigma}+\frac{\beta_1}{\sigma} X_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) dw + \sum_{i=1}^{T} (1 - Y_i) \ln \left[\int_{\frac{\beta_0}{\sigma}+\frac{\beta_1}{\sigma} X_i}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) dw\right].$$

(d) Given the data $\{X_i, Y_i\}_{i=1}^{T}$, suppose $\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}\right) = (1, 2, 3)$ maximizes the log-likelihood function, will $\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}\right) = (2, 4, 6)$ also maximize the log-likelihood function? Discuss the identifiability of β_0 and β_1.

Exercise 0.126 *Consider the following linear probability model:*

$$\begin{aligned} DIVORCE_i &= \beta_0 + \beta_1 INCOME_i + \beta_2 YEARMARRIED_i + \beta_3 AFFAIR_i \\ &\quad + \beta_4 CHILDREN_i + u_i, \end{aligned}$$

where

$DIVORCE_i = 1$ if couple i got divorce in the year of the survey, and $DIVORCE_i = 0$ if not.

$INCOME_i =$ family's monthly income of couple i (in dollars).

$YEARMARRIED_i =$ years of marriage of couple i.

$AFFAIR_i = 1$ if the husband or the wife (or both) has an extramarital affair, and $AFFAIR_i = 0$ if not.

$CHILDREN_i =$ number of children of couple i.

a) Show that $E\left(DIVORCE_i\right) = \Pr\left(DIVORCE_i = 1\right)$.

b) Interpret each of the above coefficients $\beta_0, ..., \beta_4$.

c) Show that $E\left(u_i\right) = 0$ implies

$$\begin{aligned}\Pr\left(DIVORCE_i = 1\right) &= \beta_0 + \beta_1 INCOME_i + \beta_2 YEARMARRIED_i + \beta_3 AFFAIR_t \\ &\quad + \beta_4 CHILDREN_i\end{aligned}$$

d) Show that $\text{Var}(u_i) = \Pr\left(DIVORCE_i = 1\right)\Pr\left(DIVORCE_i = 0\right)$.

e) Suppose the we estimate the model by OLS and get:

$$\begin{aligned}\widehat{DIVORCE}_i &= .5 - .0002 INCOME_i - .015 YEARMARRIED_i + .9 AFFAIR_i \\ &\quad - .03 CHILDREN_i.\end{aligned}$$

What is the chance of getting divorce for:

i) a couple with 6 years of marriage, 2 children, family's monthly income of 1000 dollars, and no extramarital affair.

ii) a couple with 1 year of marriage, no children, family's monthly income of 2000 dollars, and the husband has an extramarital affair.

iii) a couple with 30 years of marriage, 3 children, family's monthly income of 4000 dollars, and the wife has an extramarital affair.

f) State an advantage and a shortcoming of the linear probability model.

Exercise 0.127 *Consider the following linear probability model:*

$$\begin{aligned}AFFAIR_i &= \beta_0 + \beta_1 INCOME_i + \beta_2 SPOUSEINCOME_i + \beta_3 YEARMARRIED_i \\ &\quad + \beta_4 CHILDREN_i + \beta_5 HRTOGETHER_i + \beta_6 SEX_i + u_i\end{aligned}$$

where

$AFFAIR_i = 1$ if individual i has an extramarital affair, and $= 0$ if not,

$INCOME_i =$ monthly income of individual i (in dollars),

$SPOUSEINCOME_i =$ monthly income of the spouse of individual i,

$YEARMARRIED_i =$years of marriage of individual i,

$CHILDREN_i =$ number of children of individual i,

$HRTOGETHER_i =$number of hours per week that individual i spends with his/her spouse.

$SEX_i = 1$ if individual i is a male, and $= 0$ otherwise.

(a) Interpret each of the above coefficients $\beta_1, ..., \beta_6$, what are their expected signs? Explain.

(b) Show that $E(u_i) = 0$ implies

$$\begin{aligned}\Pr(AFFAIR_i = 1) &= \beta_0 + \beta_1 INCOME_i + \beta_2 SPOUSEINCOME_i \\ &\quad + \beta_3 YEARMARRIED_i + \beta_4 CHILDREN_i \\ &\quad + \beta_5 HRTOGETHER_i + \beta_6 SEX_i.\end{aligned}$$

(c) Show that $\text{Var}(u_i) = \Pr(AFFAIR_i = 1)\Pr(AFFAIR_i = 0)$.

(d) Suggest a method to fix the problem of heteroskedasticity in part (c). What is the advantage and shortcoming of your method?

(e) Suppose the we estimate the model by OLS and get:

$$\begin{aligned}\widehat{AFFAIR}_i &= .5 + .008 INCOME_i - .009 SPOUSEINCOME_i - .015 YEARMARRIED_i \\ &\quad - .03 CHILDREN_i - .004 HRTOGETHER_i + .007 SEX_i\end{aligned}$$

What is the chance of having an extramarital affair for:

i) a man with 6 years of marriage, 2 children, monthly income of 1000 dollars, wife's income is 800 and he spends 100 hours per week with his wife.

ii) a woman with 1 years of marriage, 1 child, monthly income of 1000 dollars, husband's income is 900 and she spends 56 hours per week with his husband.

iii) a man with 30 years of marriage, 3 children, monthly income of 700 dollars, wife's income is 500 and he spends 120 hours per week with his wife.

Exercise 0.128 *Using the data in Table 19.1 in Greene, repeat the calculation of Table 19.2 using*

(a) Linear probability model;
(b) Probit model;
(c) Logit model;
(b) Nonlinear regression model with

$$Y_t = \frac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 X_{1t} + ... + \beta_k X_{kt}\right)\right]} + u_t.$$

Exercise 0.129 *Find* $E(X)$ *and* $Var(X)$ *of the random variable* X *with*

$$F(x) = \exp\left(-\exp\left(-x\right)\right),$$

$$f(x) = \exp\left(-x\right)\exp\left(-\exp\left(-x\right)\right).$$

Exercise 0.130 *Find* $E(u|u>1)$ *and* $Var(u|u>1)$ *if* $u \sim N(0,1)$*, and compare them to their unconditional means and variances.*

Exercise 0.131 *True/False. Explain.*

(a) If we only observe a subpopulation such as individuals with income above a certain level, then the data is said to be lower-truncated.

(b) If we only observe a subpopulation, such as individuals with income above a certain level, then the data are said to be lower-censored.

(c) When the dependent variable is censored, values in a certain range are all transformed to a single value.

(d) When the dependent variable is truncated, values in a certain range are all transformed to a single value.

(e) If X is a random variable which has an extreme value distribution with density $f(x) = \exp(-x)\exp(-\exp(-x))$ for $-\infty < x < \infty$. Let $Y = \exp(-X)$, then $E(Y) = 1$.

Exercise 0.132 *A Probit model assumes that the error term has a uniform distribution. True/False.*

Exercise 0.133 *If we only observe a subpopulation, such as individuals with income above a certain level, then the data are said to be lower-censored. True/False.*

Exercise 0.134 $Var(X) \geq Var(X|X = a)$ *for any random variable* X *and constant* a*. True/False.*

Exercise 0.135 *Greene, 5th edition, Chapter 21, Exercises 1 & 3;.Chapter 22, Exercises 1 & 2.*

ECO5120: Econometric Theory and Application, Fall 2006
Prof. T.L. Chong
HANDOUT 8
AR(1) , I(1), NEARLY-I(1) PROCESSES and VAR Model

Definition 155 *A process y_t is said to be* ***weakly stationary (or covariance stationary)*** *if*

$$\begin{aligned} E(y_t) &= \mu < \infty \qquad \text{for all } t, \\ E\left(y_t^2\right) &< \infty \qquad \text{for all } t, \\ E(y_t - \mu)(y_{t-j} - \mu) &= \gamma_{|j|} < \infty \qquad \text{for all } t \text{ and any } j = \pm 1, \pm 2, ... \end{aligned}$$

Notice that if a process is covariance stationary, the covariance between y_t and y_{t-j} depends only on j, the length of time separating the observations.

Definition 156 *A process is said to be* ***strictly stationary*** *if the joint distribution of*

$(y_t,\ y_{t+j_1},\ y_{t+j_2},...,\ y_{t+j_n},)$ depends only on $(j_1,\ j_2, ...,\ j_n)$, for all t, and any j_1, $j_2, ..., j_n$, n. i.e. The joint density

$$f(y_t,\ y_{t+j_1},\ y_{t+j_2},...,\ y_{t+j_n},) = f(y_s,\ y_{s+j_1},\ y_{s+j_2},...,\ y_{s+j_n},)$$

for any s and t.

Weakly stationarity and strictly stationarity do not imply each other, a process can be strictly stationary but not weakly stationary. For example, if the process has a Cauchy distribution, then its moments do not exist, and it is not weakly stationary. However, as long as its distribution does not change over time, it is strongly stationary. It is also possible to find a process that is covariance-stationary but not strictly stationary, e.g., the mean and covariance are not functions of time, but perhaps higher moments such as $E(y_t^4)$ and $E(y_t^5)$ are.

If a process is strictly stationary *with finite second moments*, then it must be covariance-stationary.

AR(1) Process

Consider an autoregressive process of order 1,

$$\begin{aligned} y_t &= \beta y_{t-1} + u_t, \\ |\beta| &< 1, \\ u_t &\sim iid\left(0, \sigma^2\right). \end{aligned}$$

We are interested in finding the mean and variance of the process y_t. Assume that time starts from -∞, then by repeating substitution, we can show that

$$\begin{aligned} y_t &= \beta\left(\beta y_{t-2} + u_{t-1}\right) + u_t \\ &= \beta^2 y_{t-2} + \beta u_{t-1} + u_t \\ &= \beta^2\left(\beta y_{t-3} + u_{t-2}\right) + \beta u_{t-1} + u_t \\ &= ..... \\ &= u_t + \beta u_{t-1} + \beta^2 u_{t-2} + \beta^3 u_{t-3} + ... + \beta^{t-1} u_1 \\ &= \sum_{k=0}^{\infty} \beta^k u_{t-k}. \end{aligned}$$

$$E\left(y_t\right) = E\left(\sum_{k=0}^{\infty} \beta^k u_{t-k}\right) = \sum_{k=0}^{\infty} \beta^k E\left(u_{t-k}\right) = 0.$$

$$\begin{aligned} Var\left(y_t\right) &= Var\left(\sum_{k=0}^{\infty} \beta^k u_{t-k}\right) = \sum_{k=0}^{\infty} \beta^{2k} Var\left(u_{t-k}\right) = \sigma^2 \sum_{k=0}^{\infty} \beta^{2k} \\ &= \frac{\sigma^2}{1-\beta^2}. \end{aligned}$$

We are interested in the unknown parameter β. For a data set $\{y_t\}_{t=1}^T$, the OLS estimator for β is given by

$$\widehat{\beta}_T = \frac{\sum_{t=1}^{T} y_{t-1}y_t}{\sum_{t=1}^{T} y_{t-1}^2} = \beta + \frac{\sum_{t=1}^{T} y_{t-1}u_t}{\sum_{t=1}^{T} y_{t-1}^2}.$$

Theorem 157 *In an AR(1) model without intercept, if $|\beta| < 1$, and the error terms are i.i.d., then the OLS estimator is consistent with an asymptotic distribution given by*

$$\sqrt{T}\left(\widehat{\beta}_T - \beta\right) = \frac{\sum_{t=1}^{T} y_{t-1}u_t/\sqrt{T}}{\sum_{t=1}^{T} y_{t-1}^2/T} \xrightarrow{d} N\left(0, 1-\beta^2\right).$$

Proof. Exercise.

Thus, in a stationary AR process, the OLS estimators is asymptotically normally distributed.

Asymptotic Test for $\mathbf{H}_0 : \beta = \beta_0$, where $|\beta_0| < 1$.

Under the null $H_0 : \beta = \beta_0$, where $|\beta_0| < 1$, the process is stationary and as $T \to \infty$ the test statistic

$$t = \frac{\widehat{\beta}_T - \beta_0}{\sqrt{\widehat{Var}\left(\widehat{\beta}_T\right)}} \xrightarrow{d} N(0,1),$$

where

$$\widehat{Var}\left(\widehat{\beta}_T\right) = \frac{\widehat{\sigma}^2}{\sum_{t=1}^{T} y_{t-1}^2},$$

$$\widehat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(y_t - \widehat{\beta}_T y_{t-1}\right)^2 \xrightarrow{p} \sigma^2.$$

AR(1) process with AR(1) error term

Consider the process

$$\begin{aligned} y_t &= \beta y_{t-1} + u_t, \\ u_t &= \rho u_{t-1} + \varepsilon_t, \end{aligned}$$

where ε_t is i.i.d. and independent of u_{t-1}.

We know that the OLS estimator for β will be biased and inconsistent in this case. What will it converge to?

Theorem 158 *In an AR(1) model without intercept, if $|\beta| < 1$, and the error terms are also an AR(1) process, then the OLS estimator is inconsistent and*

$$\widehat{\beta} \xrightarrow{p} \beta + \frac{\rho\left(1-\beta^2\right)}{1+\beta\rho}.$$

Proof. Exercise.

I(1) Process

When $\beta = 1$, we call the process y_t an integrated process of order 1 , $I(1)$, or sometimes it is called the Unit-Root Process, Random Walk process, etc.. In the AR(1) model, as long as $|\beta| < 1$, the process y_t has a finite long run variance, and it is covariance stationary. However, when β equals 1, the variance of y_t explodes, and y_t is *nonstationary* as a result. The $I(1)$ process is widely used in Economics and Finance. For example, in the stock market, many people believe that the stock price is a random walk process. Given the information set Ω_t in period t, the prediction of tomorrow stock price is today's price, mathematically speaking,

$$\begin{aligned} P_{t+1} &= P_t + u_{t+1}, \\ E\left(P_{t+1}|\Omega_t\right) &= E\left(P_t|\Omega_t\right) + E\left(u_{t+1}|\Omega_t\right) = P_t + 0 = P_t. \end{aligned}$$

Consider the process

$$\begin{aligned} y_t &= y_{t-1} + u_t, \\ y_0 &= 0, \\ u_t &\sim i.i.d.\left(0, \sigma^2\right). \end{aligned}$$

$$\begin{aligned} y_t &= y_{t-2} + u_{t-1} + u_t \\ &= y_{t-3} + u_{t-2} + u_{t-1} + u_t \\ &= ..... \\ &= \sum_{k=0}^{t-1} u_{t-k}. \end{aligned}$$

$$E\left(y_t\right) = E\left(\sum_{k=0}^{t-1} u_{t-k}\right) = \sum_{k=0}^{t-1} E\left(u_{t-k}\right) = 0.$$

$$Var\left(y_t\right) = Var\left(\sum_{k=0}^{t-1} u_{t-k}\right) = \sum_{k=0}^{t-1} Var\left(u_{t-k}\right) = t\sigma^2.$$

Thus

$$\begin{aligned} y_t &\sim N\left(0, t\sigma^2\right), \\ \frac{y_t}{\sigma\sqrt{t}} &\sim N\left(0, 1\right). \end{aligned}$$

Note that the OLS estimator is given by

$$\widehat{\beta}_T = 1 + \frac{\sum_{t=1}^{T} y_{t-1} u_t}{\sum_{t=1}^{T} y_{t-1}^2}.$$

In an I(1) process, the term $\frac{\sum_{t=1}^{T} y_{t-1}u_t}{\sum_{t=1}^{T} y_{t-1}^2}$ behaves in a very strange way. Let us study the numerator first. Note that

$$y_t^2 = (y_{t-1} + u_t)^2 = y_{t-1}^2 + 2y_{t-1}u_t + u_t^2.$$

Thus,

$$\begin{aligned}
y_{t-1}u_t &= \frac{1}{2}\left(y_t^2 - y_{t-1}^2 - u_t^2\right) \\
\sum_{t=1}^{T} y_{t-1}u_t &= \sum_{t=1}^{T} \frac{1}{2}\left(y_t^2 - y_{t-1}^2 - u_t^2\right) \\
&= \frac{1}{2}\left(y_T^2 - y_0^2 - \sum_{t=1}^{T} u_t^2\right) \\
\frac{1}{T\sigma^2}\sum_{t=1}^{T} y_{t-1}u_t &= \frac{1}{2\sigma^2 T} y_T^2 - \frac{1}{2\sigma^2 T}\sum_{t=1}^{T} u_t^2 \\
&= \frac{1}{2}\left(\frac{y_T}{\sigma\sqrt{T}}\right)^2 - \frac{1}{2\sigma^2 T}\sum_{t=1}^{T} u_t^2 \\
&\xrightarrow{d} \frac{1}{2}\left(N(0,1)\right)^2 - \frac{1}{2} \\
&\overset{d}{=} \frac{1}{2}\left(\chi_1^2 - 1\right).
\end{aligned}$$

Therefore, the term in the numerator, after adjustment, will have a distribution related to a Chi-square random variable.

Definition 159 *A **Standard Brownian Motion** $B(r)$ is a continuous-time stochastic process, associating each date $r \in [0,1]$ such that:*

1) $B(0) = 0$;
2) For any dates $0 \leq r_1 < r_2 < ... < r_m \leq 1$,

$$B(r_1), B(r_2) - B(r_1), B(r_3) - B(r_2), ..., B(r_m) - B(r_{m-1}), B(1) - B(r_m)$$

are independent multivariate normal with $[B(r_i) - B(r_{i-1})] \sim N(0, r_i - r_{i-1})$.

Theorem 160 *The asymptotic distribution of the OLS estimator when* $\beta = 1$ *is given by*

$$T\left(\widehat{\beta}_T - 1\right) \xrightarrow{d} \frac{B(1)^2 - 1}{2\int_0^1 B(r)^2\, dr}.$$

Note that both the rate of convergence and the asymptotic distribution of $\widehat{\beta}_T$ under $\beta = 1$ are different from the case where $|\beta| < 1$.

Asymptotic Test for $\mathbf{H_0} : \beta = 1$.

Under the null $H_0 : \beta = 1$, the process is nonstationary, it will be shown that the test statistic is not asymptotically normal. To see this, note that

$$t = \frac{\widehat{\beta}_T - 1}{\sqrt{\widehat{Var}\left(\widehat{\beta}_T\right)}} = \frac{\frac{\sum_{t=1}^T y_{t-1}u_t}{\sum_{t=1}^T y_{t-1}^2}}{\sqrt{\frac{\widehat{\sigma}^2}{\sum_{t=1}^T y_{t-1}^2}}} = \frac{\frac{\sum_{t=1}^T y_{t-1}u_t/T}{\sqrt{\sum_{t=1}^T y_{t-1}^2/T^2}}}{\sqrt{\widehat{\sigma}^2}} \xrightarrow{d} \frac{\frac{1}{2}\sigma^2\left(B(1)^2 - 1\right)}{\sqrt{\sigma^2\int_0^1 B(r)^2\, dr}} = \frac{B(1)^2 - 1}{2\sqrt{\int_0^1 B(r)^2\, dr}}.$$

Theorem 161 *Under* $H_0 : \beta = 1$*, the asymptotic distribution of the t-statistic is given by*

$$\frac{B(1)^2 - 1}{2\sqrt{\int_0^1 B(r)^2\, dr}},$$

which is a Dickey-Fuller distribution.

Vector Autoregression

Suppose we like to study the relation between a set of k economic variables and their lags, we can run a p^{th} order vector autoregression (VAR(p)) model.

$$y_t = c + \Gamma_1 y_{t-1} + ... + \Gamma_p y_{t-p} + u_t,$$

where

$y_t = (y_{1t}, ..., y_{kt})'$ is a k by 1 random vector of variables of interest.

For example, $y_t = (GNP_t, M2_t, IR_t)$ represent the gross national income, money demand and interest rate at time t.

$u_t = (u_{1t}, ..., u_{kt})'$ is a k by 1 random vector of uncorrelated disturbances with zero mean and contemporaneous covariance matrix $E(u_t u_t') = \Omega$.

$c = (c_1, ..., c_k)'$ is a k by 1 fixed vector of intercept terms allowing for possibility of nonzero $E(y_t)$.

$$\Gamma_i = \begin{pmatrix} \Gamma_{i11} & \Gamma_{i12} & \cdots & \Gamma_{i1k} \\ \Gamma_{i21} & \Gamma_{i22} & \cdots & \Gamma_{i2k} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{ik1} & \Gamma_{ik2} & \cdots & \Gamma_{ikk} \end{pmatrix}$$

are fixed k by k coefficient matrices, $i = 1, 2, ..., p$.

Note that the system can be rewritten as

$$\underset{kp\times 1}{Y_t} = \underset{kp\times 1}{C} + \underset{kp\times kp}{\Gamma}\underset{kp\times 1}{Y_{t-1}} + \underset{kp\times 1}{U_t},$$

where

$$Y_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{pmatrix}, C = \begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$\Gamma = \begin{pmatrix} \Gamma_1 & \Gamma_2 & \cdots & \cdots & \Gamma_p \\ I_k & 0 & \cdots & 0 & 0 \\ 0 & I_k & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_k & 0 \end{pmatrix}, U_t = \begin{pmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

A stable VAR can be written in infinite order Vector Moving Average form. Note that

$$\begin{aligned} Y_t &= C + \Gamma Y_{t-1} + U_t \\ &= C + \Gamma \left(C + \Gamma Y_{t-2} + U_{t-1} \right) + U_t \\ &= ... \\ &= \mu + \sum_{i=0}^{\infty} \Gamma^i U_{t-i}, \end{aligned}$$

where

$$\mu = \sum_{i=0}^{\infty} \Gamma^i C.$$

The matrix Γ^i can be interpreted as

$$\underset{kp \times kp}{\frac{\partial Y_t}{\partial U'_{t-i}}} = \Gamma^i.$$

A plot of the element of $\frac{\partial Y_t}{\partial U'_{t-i}}$ as a function of i is called the **impulse-response function**.

More demanding materials

The Functional Central Limit Theorem

Define

$$X_T(r) = \frac{1}{T} \sum_{t=1}^{[Tr]} u_t, \qquad r \in [0,1],$$

where $[Tr]$ stands for the largest integer less than or equal to Tr. That is

$$\begin{aligned} X_T(r) &= 0 \qquad \text{for } 0 \leq r < \frac{1}{T} \\ &= \frac{y_1}{T} \qquad \text{for } \frac{1}{T} \leq r < \frac{2}{T} \\ &\vdots \\ &= \frac{y_T}{T} \qquad \text{for } r = 1 \end{aligned}$$

Note that $X_T(r)$ is a step function of r in $[0,1]$.

$$\sqrt{T}X_T(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} u_t = \sqrt{\frac{[Tr]}{T}}\left(\frac{1}{\sqrt{[Tr]}}\sum_{t=1}^{[Tr]} u_t\right)$$

It should be noted that, as $T \to \infty$,

$$\sqrt{\frac{[Tr]}{T}} \to \lim_{T\to\infty}\sqrt{\frac{[Tr]}{T}} = \sqrt{r}$$

and

$$\frac{1}{\sqrt{[Tr]}}\sum_{t=1}^{[Tr]} u_t \xrightarrow{d} N\left(0, \sigma^2\right).$$

Thus

$$\sqrt{T}X_T(r) \xrightarrow{d} N\left(0, r\sigma^2\right)$$

and

$$\sqrt{T}\frac{X_T(r)}{\sigma} \xrightarrow{d} N(0, r).$$

Note that, the asymptotic behavior of $\sqrt{T}\dfrac{X_T(\cdot)}{\sigma}$ can be described by a standard Brownian motion. To see this, note that,

1) $\sqrt{T}\dfrac{X_T(0)}{\sigma} \xrightarrow{p} 0$;

2) For any $0 \leq r_1 < r_2 < ... < r_m \leq 1$.

$$\left(\sqrt{T}\frac{X_T(r_1)}{\sigma} - \sqrt{T}\frac{X_T(0)}{\sigma}\right), \left(\sqrt{T}\frac{X_T(r_2)}{\sigma} - \sqrt{T}\frac{X_T(r_1)}{\sigma}\right), ..., \left(\sqrt{T}\frac{X_T(1)}{\sigma} - \sqrt{T}\frac{X_T(r_m)}{\sigma}\right)$$

are independent increments with $\left(\sqrt{T}\frac{X_T(r_i)}{\sigma} - \sqrt{T}\frac{X_T(r_{i-1})}{\sigma}\right) \xrightarrow{d} N(0, r_i - r_{i-1})$.

The functional central limit theorem states that

$$\sqrt{T}\frac{X_T(\cdot)}{\sigma} \xrightarrow{d} B(\cdot)$$

Note that $X_T(\cdot)$ denotes a random function while $X_T(r)$ denotes the value that the function assumes at date r, which is a random variable. Thus, when the function is evaluated at $r = 1$, the conventional central limit theorem is obtained, which is a special case of the functional central limit theorem, i.e.

$$\sqrt{T}\frac{X_T(1)}{\sigma} \xrightarrow{d} B(1) \sim N(0,1)$$

Theorem 162 *The **Continuous Mapping Theorem** states that, if a sequence of random continuous functions, $W_T : r \in [0,1] \to R$ with $W_T(\cdot) \xrightarrow{d} W(\cdot)$ where $W(\cdot)$ is a stochastic function and $g(\cdot)$ is a continuous functional, then $g(W_T(\cdot)) \xrightarrow{d} g(W(\cdot))$.*

Remark 3 *A functional is a function of function, for example the integral is a functional. Continuity of a functional $g(\cdot)$ means that for any $\varepsilon > 0$, there exists a $\delta > 0$ such that if $h(r)$ and $k(r)$ are any continuous bounded functions on $[0,1]$, such that $|h(r) - k(r)| < \delta$ for all $r \in [0,1]$, then $|g(h(\cdot)) - g(k(\cdot))| < \varepsilon$.*

Therefore by the continuous mapping theorem, we have

$$\int_0^1 \sqrt{T}X_T(r)\,dr \xrightarrow{d} \sigma \int_0^1 B(r)\,dr.$$

Now define

$$S_T(r) = \left[\sqrt{T}X_T(r)\right]^2.$$

That is

$$\begin{aligned} S_T(r) &= 0 \quad \text{for } 0 \leq r < \frac{1}{T} \\ &= \frac{y_1^2}{T} \quad \text{for } \frac{1}{T} \leq r < \frac{2}{T} \\ &\vdots \\ &= \frac{y_T^2}{T} \quad \text{for } r = 1 \end{aligned}$$

Note that $S_T(r)$ is a step function of r in $[0,1]$,

$$\int_0^1 S_T(r)\, dr = \sum_{t=1}^{T-1} \frac{y_t^2}{T} \left(\frac{1}{T}\right).$$

By the continuous mapping theorem, we also have

$$\int_0^1 S_T(r)\, dr = \int_0^1 \left[\sqrt{T} X_T(r)\right]^2 dr \xrightarrow{d} \sigma^2 \int_0^1 B(r)^2\, dr.$$

We have more results presented in the following proposition.

Proposition 163 *If y_t follows an I(1) process*

$$\begin{aligned} y_t &= y_{t-1} + u_t, \\ y_0 &= 0, \\ u_t &\sim i.i.d.\left(0, \sigma^2\right). \end{aligned}$$

Then

(a) $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} u_t \xrightarrow{d} \sigma B(1)$;
(b) $\frac{1}{T} \sum_{t=1}^{T} y_{t-1} u_t \xrightarrow{d} \frac{1}{2}\sigma^2 \left(B(1)^2 - 1\right)$;
(c) $\frac{1}{T^{3/2}} \sum_{t=1}^{T} t u_t \xrightarrow{d} \sigma B(1) - \sigma \int_0^1 B(r)\, dr$;
(d) $\frac{1}{T^{3/2}} \sum_{t=1}^{T} y_{t-1} \xrightarrow{d} \sigma \int_0^1 B(r)\, dr$;
(e) $\frac{1}{T^2} \sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{d} \sigma^2 \int_0^1 B(r)^2\, dr$;
(f) $\frac{1}{T^{5/2}} \sum_{t=1}^{T} t y_{t-1} \xrightarrow{d} \sigma \int_0^1 rB(r)\, dr$;

(g) $\frac{1}{T^3}\sum_{t=1}^{T} t y_{t-1}^2 \xrightarrow{d} \sigma^2 \int_0^1 rB(r)^2\, dr$;
(h) $\frac{1}{T^{n+1}}\sum_{t=1}^{T} t^n \to \frac{1}{1+n}$.

Proof. Exercise.

Nearly-I(1) Process

In an AR(1) model, when β is very close to 1 by a rate of $\frac{1}{T}$, we call the process y_t a near-integrated or near unit-root process. Consider the process

$$\begin{aligned} y_t &= \left(1 - \frac{c}{T}\right) y_{t-1} + u_t, \qquad t = 1, 2, ..., T, \\ y_0 &= 0, \\ c &> 0, \\ u_t &\sim i.i.d.\left(0, \sigma^2\right). \end{aligned}$$

$$
\begin{aligned}
y_t &= \left(1-\frac{c}{T}\right) y_{t-1} + u_t \\
&= \left(1-\frac{c}{T}\right)^2 y_{t-2} + \left(1-\frac{c}{T}\right) u_{t-1} + u_t \\
&= ..... \\
&= \left(1-\frac{c}{T}\right)^t y_0 + u_t + \left(1-\frac{c}{T}\right) u_{t-1} + \left(1-\frac{c}{T}\right)^2 u_{t-2} + ... + \left(1-\frac{c}{T}\right)^{t-1} u_1 \\
&= \sum_{k=0}^{t-1} \left(1-\frac{c}{T}\right)^k u_{t-k} \\
&= \sum_{k=0}^{t-1} u_{t-k} + \sum_{k=0}^{t-1} \left(\left(1-\frac{c}{T}\right)^k - 1 \right) u_{t-k} \\
&= \sum_{k=0}^{t-1} u_{t-k} + \sum_{k=1}^{t-1} \left(\left(1-\frac{c}{T}\right)^k - 1 \right) u_{t-k} \\
&= \sum_{k=0}^{t-1} u_{t-k} + \sum_{k=1}^{t-1} \left(\left(\left(1-\frac{c}{T}\right) - 1 \right) \sum_{i=0}^{k-1} \left(1-\frac{c}{T}\right)^i \right) u_{t-k} \\
&= \sum_{k=0}^{t-1} u_{t-k} - \frac{c}{T} \sum_{k=1}^{t-1} \sum_{i=0}^{k-1} \left(1-\frac{c}{T}\right)^i u_{t-k} \\
&= \sum_{k=0}^{t-1} u_{t-k} - \frac{c}{T} \sum_{i=1}^{t-1} \left(1-\frac{c}{T}\right)^{i-1} \sum_{k=1}^{t-i} u_k \\
&= \sum_{k=1}^{t} u_k - \frac{c}{T} \sum_{i=1}^{t-1} \left[\left(1-\frac{c}{T}\right)^T \right]^{(i-1)/T} \sum_{k=1}^{t-i} u_k .
\end{aligned}
$$

Therefore,

$$
\frac{1}{\sqrt{T}} y_t = \frac{1}{\sqrt{T}} \sum_{k=1}^{t} u_k - c \sum_{i=1}^{t-1} \left[\left(1-\frac{c}{T}\right)^T \right]^{(i-1)/T} \frac{1}{\sqrt{T}} \sum_{k=1}^{t-i} u_k \frac{1}{T}.
$$

Let

$$
\begin{aligned}
t &= [rT], \\
t-i &= [sT], \\
i &= [(r-s)T].
\end{aligned}
$$

$$\begin{aligned}\frac{1}{\sqrt{T}}y_{[rT]} &= \frac{1}{\sqrt{T}}\sum_{k=1}^{[rT]}u_k - c\sum_{i=1}^{[rT]-1}\left[\left(1-\frac{c}{T}\right)^T\right]^{([rT]-[sT]-1)/T}\left(\frac{1}{\sqrt{T}}\sum_{k=1}^{[sT]}u_k\right)\frac{1}{T} \\ &\Rightarrow \sigma\left[B(r) - c\int_0^r \exp\left(-c(r-s)\right)B(s)\,ds\right] \overset{def}{=} \sigma K_c(r).\end{aligned}$$

Proposition 164 *If y_t follows a near I(1) process*

$$\begin{aligned}y_t &= \left(1-\frac{c}{T}\right)y_{t-1} + u_t, \\ y_0 &= 0, \qquad c > 0, \\ u_t &\sim i.i.d.\left(0,\sigma^2\right).\end{aligned}$$

Then

(a) $\frac{1}{\sqrt{T}}y_{[rT]} \Rightarrow \sigma K_c(r)$;

(b) $\frac{1}{T}\sum_{t=1}^T y_{t-1}u_t \Rightarrow \sigma^2\int_0^1 K_c(r)\,dB(r)$;

(c) $\frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2 \Rightarrow \sigma^2\int_0^1 K_c(r)^2\,dr$.

Proof. Exercise.

Theorem 165 *Using (b) and (c) in the above proposition, the asymptotic distribution of the OLS estimator when $\beta = 1 - \frac{c}{T}$ is given by*

$$\begin{aligned}T\left(\widehat{\beta}_T - \beta\right) &= \frac{\sum_{t=1}^T y_{t-1}u_t/T}{\sum_{t=1}^T y_{t-1}^2/T^2} \xrightarrow{d} \mathfrak{L}(c), \\ \mathfrak{L}(c) &= \frac{\int_0^1 K_c(r)\,dB(r)}{\int_0^1 K_c(r)^2\,dr}.\end{aligned}$$

Proof. Exercise.

$$\mathfrak{L}(0) = \frac{\int_0^1 B(r)\, dB(r)}{\int_0^1 B(r)^2\, dr} = \frac{B(1)^2 - 1}{2\int_0^1 B(r)^2\, dr}.$$

Proof. Exercise.

Theorem 166 *As $c \to \infty$ and $T \to \infty$, $\frac{c}{T} < 1$, the asymptotic distribution of the OLS estimator when $\beta = 1 - \frac{c}{T}$ is given by*

$$\sqrt{\sum_{t=1}^{T} y_{t-1}^2}\left(\widehat{\beta}_T - \beta\right) = \frac{\sum_{t=1}^{T} y_{t-1}u_t/\sqrt{T}}{\sqrt{\sum_{t=1}^{T} y_{t-1}^2/T}} \xrightarrow{d} N\left(0, \sigma^2\right).$$

Proof. Exercise.

Exercise 0.136 *Explain why the long-run variance of an I(1) process does not exist.*

Exercise 0.137 *True/False. Explain.*

(a) If a process is covariance stationary, then the covariance between y_t and y_{t-j} depends only on t.

(b) If a process is covariance stationary, then the covariance between y_t and y_{t-j} depends only on j.

(c) Weak stationarity is implied by strict stationarity.

(d) In an AR(1) model without an intercept, if $|\beta| < 1$, and the error terms are i.i.d. with zero mean and finite variance, then $\sqrt{T}\left(\widehat{\beta}_T - \beta\right) \xrightarrow{d} N(0, 1-\beta)$.

(e) In an AR(1) model without an intercept, if $|\beta| < 1$, and the error terms are i.i.d. with zero mean and finite variance, then $\sqrt{T}\left(\widehat{\beta}_T - \beta\right) \xrightarrow{d} N\left(0, 1-\beta^2\right)$.

(f) In an AR(1) model without an intercept, if $\beta = 1$, and the error terms are i.i.d. with zero mean and finite variance, then the test statistic $\dfrac{\widehat{\beta}_T - 1}{\sqrt{\widehat{Var}\left(\widehat{\beta}_T\right)}} \xrightarrow{d} N(0,1)$.

Exercise 0.138 *Define* $X_t = u_t - u_{t-1}$, $\overline{X} = \dfrac{\sum_{t=1}^{T} X_t}{T}$. *Find* $E\left(\overline{X}\right), Var\left(\overline{X}\right)$ *and examine whether the Central Limit Theorem applies to* $\overline{X}$ *in the following cases:*

a) $u_t = u_{t-1} + \varepsilon_t$, where $\{\varepsilon_t\}_{t=0}^{T} \sim i.i.d.\left(0, \sigma_\varepsilon^2\right)$, $\sigma_\varepsilon^2 < \infty$.
b) $\{u_t\}_{t=0}^{T} \sim i.i.d.\left(0, \sigma_u^2\right)$, $\sigma_u^2 < \infty$.

Exercise 0.139 *Consider the following process*

$$\begin{aligned} y_t &= \beta y_{t-1} + u_t, \\ y_0 &= 0, \\ u_t &\sim i.i.d.N\left(0,1\right). \end{aligned}$$

a) Write a Gauss program to simulate the above process y_t for $\beta = 0.5$ and $\beta = 1$, using a sample size $T = 100$.

b) Use Gauss to simulate the distribution of $\sqrt{T}\left(\widehat{\beta}_T - 0.5\right)$ for $\beta = 0.5$, and $T\left(\widehat{\beta}_T - 1\right)$ for $\beta = 1$, where $\widehat{\beta}_T = \dfrac{\sum_{t=1}^{T} y_{t-1} y_t}{\sum_{t=1}^{T} y_{t-1}^2}$, using a sample size $T = 100$, with the number of replications $N = 10000$.

c) Use Gauss to simulate the distribution of the test statistic

$$t = \frac{\widehat{\beta}_T - 0.5}{\sqrt{\widehat{Var}\left(\widehat{\beta}_T\right)}} \qquad \text{for } \beta = 0.5,$$

and

$$t = \frac{\widehat{\beta}_T - 1}{\sqrt{\widehat{Var}\left(\widehat{\beta}_T\right)}} \qquad \text{for } \beta = 1,$$

where

$$\widehat{Var}\left(\widehat{\beta}_T\right) = \frac{\widehat{\sigma}^2}{\sum_{t=1}^{T} y_{t-1}^2},$$

$$\widehat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(y_t - \widehat{\beta}_T y_{t-1}\right)^2,$$

using a sample size $T = 25, 50, 100, 250, 500, 1000$, with the number of replications $N = 10000$. Calculate the 1%, 2.5%, 5%, 10%, 50%, 90%, 95%, 97.5% and 99% critical values. Compare your results to a t-table and a Dickey Fuller table.

Exercise 0.140 *Test whether the Hang Seng Index follows a unit root process, using year 2006* ***daily closing*** *price data from January to October.*

Exercise 0.141 *Suppose in GNP and Consumption are related as in the following VAR(2) model:*

$$\begin{pmatrix} GNP_t \\ C_t \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} + \begin{pmatrix} .7 & .1 \\ 0 & .4 \end{pmatrix}\begin{pmatrix} GNP_{t-1} \\ C_{t-1} \end{pmatrix} + \begin{pmatrix} -.2 & 0 \\ 0 & .1 \end{pmatrix}\begin{pmatrix} GNP_{t-2} \\ C_{t-2} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix},$$

(a) Write the process $y_t = \begin{pmatrix} GNP_t \\ C_t \end{pmatrix}$ in VAR(1) form of

$$\underset{4\times 1}{Y_t} = \underset{4\times 1}{C} + \underset{4\times 4}{\Gamma}\,\underset{4\times 1}{Y_{t-1}} + \underset{4\times 1}{U_t},$$

where $Y_t = \begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix}, C = \begin{pmatrix} c \\ 0 \end{pmatrix}, \Gamma = \begin{pmatrix} \Gamma_1 & \Gamma_2 \\ I_2 & 0 \end{pmatrix}, U_t = \begin{pmatrix} u_t \\ 0 \end{pmatrix}.$

(b) Calculate $\underset{4\times 4}{\Gamma}^{\,i}$ for $i = 2, 3$.

Exercise 0.142 *Suppose in China the growth rate of income and money demand and the interest rate are related as in the following VAR(2) model:*

$$\begin{pmatrix} GNP_t \\ M2_t \\ IR_t \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} .7 & .1 & 0 \\ 0 & .4 & .1 \\ .9 & 0 & .8 \end{pmatrix}\begin{pmatrix} GNP_{t-1} \\ M2_{t-1} \\ IR_{t-1} \end{pmatrix} + \begin{pmatrix} -.2 & 0 & 0 \\ 0 & .1 & .1 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} GNP_{t-2} \\ M2_{t-2} \\ IR_{t-2} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix},$$

$$\Omega = \begin{pmatrix} .26 & .03 & 0 \\ .03 & .09 & 0 \\ 0 & 0 & .81 \end{pmatrix}.$$

a) Write the process $y_t = \begin{pmatrix} GNP_t \\ M2_t \\ IR_t \end{pmatrix}$ in VAR(1) form.

b) Calculate

$$\underset{6\times 6}{\frac{\partial Y_t}{\partial U'_{t-i}}} = \Gamma^i$$

for i=0,1 and 2. What is the limit as $i \to \infty$?

Exercise 0.143 *Consider the process*

$$\begin{aligned} y_t &= \beta y_{t-1} + u_t, \\ y_0 &= 0, \\ u_t &\sim i.i.d.\left(0, \sigma^2\right). \end{aligned}$$

a) Find the asymptotic distribution of the OLS $\widehat{\beta}_T$ if the true value of β is

i) $\beta = 0.5$

ii) $\beta = 1$;

ii) $\beta = -1$;

iv) $\beta = 1 - \frac{1}{T}$.

b) Use Gauss to simulate the distribution of $\widehat{\beta}_T$ for the 2 cases in part a), with T=100,1000, with N=10000, $u_t \sim i.i.d.N\left(0, 1\right)$.

c) Redo parts a) and b) if the true process is

$$\begin{aligned} y_t &= \alpha + \beta y_{t-1} + u_t, \\ y_0 &= 0, \\ \alpha &\neq 0, \\ u_t &\sim i.i.d.\left(0, \sigma^2\right). \end{aligned}$$

d) Redo parts a) and b) if the true process is

$$\begin{aligned} y_t &= \alpha + \beta y_{t-1} + \gamma t + u_t, \\ y_0 &= 0, \\ \alpha &\neq 0, \\ \gamma &\neq 0, \\ u_t &\sim i.i.d.\left(0, \sigma^2\right). \end{aligned}$$

Exercise 0.144 *Consider the process*

$$\begin{aligned} y_t &= (-1)^t y_{t-1} + u_t, \\ y_0 &= 0, \\ u_t &\sim i.i.d.\left(0, \sigma^2\right). \end{aligned}$$

a) Show that

$y_1 = u_1,$

$y_2 = u_1 + u_2,$

$y_3 = -u_1 - u_2 + u_3,$

$y_4 = -u_1 - u_2 + u_3 + u_4,$

$y_5 = u_1 + u_2 - u_3 - u_4 + u_5.$

b) Let $\widehat{\beta}_T = \dfrac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2} y_{t-1}^2}$, show that

$$\widehat{\beta}_T = \frac{\sum_{t=even}^{T} y_{t-1}^2 - \sum_{t=odd}^{T} y_{t-1}^2 + \sum_{t=1}^{T} y_{t-1} u_t}{\sum_{t=1}^{T} y_{t-1}^2}.$$

c) Find $E\left(y_1^2 - y_2^2\right), E\left(y_3^2 - y_4^2\right), E\left(\sum_{t=even}^{T} y_{t-1}^2 - \sum_{t=odd}^{T} y_{t-1}^2\right)$.

d) Find plim $\widehat{\beta}_T$.

e) Write a Gauss program to simulate the distribution of $\widehat{\beta}_T$ for the T=100000, and N=10000, $u_t \sim i.i.d.N(0,1)$.

Exercise 0.145 *Consider the following model:*

$$\begin{aligned} y_t &= \beta_1 y_{t-1} + u_t, \qquad t = 1, 2, ..., k_0, \\ y_t &= \beta_2 y_{t-1} + u_t, \qquad t = k_0 + 1, k_0 + 2, ..., T. \end{aligned}$$

We let $\tau = \dfrac{k}{T}$ and make the following assumptions:
$(A1)$ $y_0 = 0$;
$(A2)$ $\varepsilon_t \sim i.i.d.(0, \sigma^2)$ $\forall t$, $0 < \sigma^2 < \infty$ and $E(\varepsilon_t^4) < \infty$;
$(A3)$ $\tau_0 = \dfrac{k_0}{T} \in [\underline{\tau}, \overline{\tau}] \subset (0,1)$.

For any given constant τ, the OLS estimators are given by:

$$\begin{aligned} \widehat{\beta}_1(\tau) &= \frac{\sum_{t=1}^{[\tau T]} y_t y_{t-1}}{\sum_{t=1}^{[\tau T]} y_{t-1}^2}, \\ \widehat{\beta}_2(\tau) &= \frac{\sum_{t=[\tau T]+1}^{T} y_t y_{t-1}}{\sum_{t=[\tau T]+1}^{T} y_{t-1}^2}. \end{aligned}$$

(a) Suppose $|\beta_1| < 1$ and $|\beta_2| < 1$. Show that:
For $0 < \tau \leq \tau_0$,
(i) $\widehat{\beta}_1(\tau) \xrightarrow{p} \beta_1$.
(ii) $\widehat{\beta}_2(\tau) \xrightarrow{p} \dfrac{(\tau_0 - \tau)\left(1 - \beta_2^2\right)\beta_1 + (1 - \tau_0)\left(1 - \beta_1^2\right)\beta_2}{(\tau_0 - \tau)\left(1 - \beta_2^2\right) + (1 - \tau_0)\left(1 - \beta_1^2\right)}$.
For $\tau_0 < \tau < 1$,

$(iii)\,\widehat{\beta}_1(\tau) \xrightarrow{p} \dfrac{\tau_0\left(1-\beta_2^2\right)\beta_1 + (\tau-\tau_0)\left(1-\beta_1^2\right)\beta_2}{\tau_0\left(1-\beta_2^2\right) + (\tau-\tau_0)\left(1-\beta_1^2\right)}$.
$(iv)\,\widehat{\beta}_2(\tau) \xrightarrow{p} \beta_2$.

(b) Suppose $|\beta_1| < 1$ and $\beta_2 = 1$. Show that:
For $0 < \tau \leq \tau_0$,
$(i)\ \widehat{\beta}_1(\tau) \xrightarrow{p} \beta_1$.
$(ii)\ \widehat{\beta}_2(\tau) \xrightarrow{p} 1$.
For $\tau_0 < \tau < 1$,
$(iii)\ \widehat{\beta}_1(\tau) \xrightarrow{p} 1$.
$(iv)\ \widehat{\beta}_2(\tau) \xrightarrow{p} 1$.

c) Suppose $\beta_1 = 1$ and $|\beta_2| < 1$. Show that:
For $0 < \tau \leq \tau_0$,
$(i)\ \widehat{\beta}_1(\tau) \xrightarrow{p} 1$.
$(ii)\ \widehat{\beta}_2(\tau) \xrightarrow{p} 1$.
For $\tau_0 < \tau < 1$,
$(iii)\,\widehat{\beta}_1(\tau) \xrightarrow{p} 1$.
$(iv)\ \widehat{\beta}_2(\tau) \xrightarrow{p} \beta_2$.

(d) Assume $|\beta_1| < 1$ and $\beta_2 = 1$. Find probability limit of $\dfrac{\sum_{t=1}^{[\tau_0 T]} y_{t-1}^2}{\sum_{t=[\tau_0 T]+1}^{T} y_{t-1}^2}$
as T goes to infinity.

Exercise 0.146 *Consider the process*

$$\begin{aligned} y_t^* &= \beta y_{t-1}^* + u_t. \\ y_0^* &= 0. \\ u_t &\sim i.i.d.\left(0, \sigma_u^2\right). \end{aligned}$$

Suppose y_t^* is not observable and we only observe y_t, where $y_t = y_t^* + \varepsilon_t$, $\varepsilon_t \sim i.i.d.\,(0, \sigma_\varepsilon^2)$. $\{y_t^*\}_{t=1}^T$, $\{\varepsilon_t\}_{t=1}^T$and $\{u_t\}_{t=1}^T$ are independent.

a) Suppose $|\beta| < 1$, show that as $t \to \infty$,

$$E\left(y_t^{*2}\right) \overset{def}{=} \sigma_*^2 = \frac{\sigma_u^2}{1-\beta^2}.$$

b) Let

$$A = \frac{\sigma_*^2}{\sigma_*^2 + \sigma_\varepsilon^2}$$

and

$$\widehat{\beta}_T = \frac{\sum_{t=2}^T y_t y_{t-1}}{\sum_{t=2}^T y_{t-1}^2}.$$

Show that

$$\widehat{\beta}_T \xrightarrow{p} \beta A.$$

c) Suppose $\beta = 1$, show that plim $\widehat{\beta}_T = 1$.

d) Write a Gauss program to simulate the distribution of $\widehat{\beta}_T$ for $T = 100000$, $N = 10000$, $u_t \sim i.i.d.N\,(0,1)$ and $\varepsilon_t \sim i.i.d.N\,(0,1)$

Exercise 0.147 *Consider the process*

$$\begin{aligned} y_t &= \beta y_{t-1} + u_t. \\ y_0 &= 0. \\ \beta &= 1. \end{aligned}$$

Suppose y_t takes values between 0 and 1. (e.g. if y_t is the unemployment rate at time t, then it takes values between zero and 1.)

(a) Given y_{t-1}, let a_t be the lower bound of u_t, and b_t be the upper bound of u_t. Find a_t and b_t.

(b) Given y_{t-1}, Let u_t be uniformly distributed in $[a_t, b_t]$, find $E(u_t|y_{t-1})$ and $Var(u_t|y_{t-1})$.

(c) Find the asymptotic distribution of

$$\widehat{\beta}_T = \frac{\sum_{t=1}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2}.$$

Exercise 0.148 *Consider the following model:*

$$y_t = y_{t-1}^{\theta} + u_t, \qquad t = 1, 2, ..., T,$$

(a) Describe how to get the nonlinear least-squares estimator for θ.

(b) If the true value of $\theta = 1$, what will be the asymptotic distribution of the nonlinear least square estimator?

Now suppose we estimate θ via the MLE method by assuming u_t follows independent $N(0, 1)$.

(c) Derive the log-Likelihood function $\ln L(y; \theta)$ and the scores function S.

(d) Describe how to get the ML estimator $\widehat{\theta}$.

(e) Describe how to get the Information Matrix.

(f) Describe how to form a Wald test for $\theta = 1$.

Exercise 0.149 *Let $y_t = y_{t-1} + u_t$ be an I(1) process with $y_0 = 0$ and $\{u_t\}_{t=1}^{T}$ follow an i.i.d.$N(0, 1)$ distribution with density function $f(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$. Suppose we run the following regression:*

$$\max\{y_t - y_{t-1}, y_{t-1} - y_{t-2}\} = \alpha + \beta \min\{y_t - y_{t-1}, y_{t-1} - y_{t-2}\} + \varepsilon_t, \qquad (t = 2, ...T.)$$

i.e., we regress $\max\{y_t - y_{t-1}, y_{t-1} - y_{t-2}\}$ on $\min\{y_t - y_{t-1}, y_{t-1} - y_{t-2}\}$ with an intercept. Let $\widehat{\beta}_T$ be the OLS estimator for the slope parameter β. Find

$$p\lim \frac{\widehat{\beta}_T + 1}{\widehat{\beta}_T}.$$

Exercise 0.150 *Consider the following time series model*

$$y_t = ((\beta y_{t-1})^m + \varepsilon_t^m)^{\frac{1}{m}} \qquad \text{for } t = 1, 2, ..., T, m = 1, 2, 3, ...$$

(a) Simplify the model for $m = 1$ and $m = \infty$.

Now consider the following model:

$$y_t = \max\{\beta y_{t-1}, \varepsilon_t\} \qquad \text{for } t = 1, 2, ..., T,$$

where $\beta > 0$, ε_t are constant mean and finite variance i.i.d. positive valued random variables.

Suppose the true $\beta = 1$ and $y_0 = 0$.. Define the estimator for β as

$$\widehat{\beta} = \min\left\{\frac{y_2}{y_1}, \frac{y_3}{y_2}, ..., \frac{y_T}{y_{T-1}}\right\}$$

(b) Show that the model can be rewritten as

$$y_t = \max\{\varepsilon_1, \varepsilon_2, ..., \varepsilon_t\} \qquad \text{for } t = 1, 2, ..., T.$$

(c) Show that $\Pr\left(\widehat{\beta} = 1\right) = 1 - \frac{1}{T!}$.

(d) Is $\widehat{\beta}$ a consistent estimator for β when $\beta = 1$?

Exercise 0.151 *Consider the following time series process:*

$$\begin{aligned} y_t &= A_t y_{t-1} + \varepsilon_t, \\ A_t &= a + u_t, \end{aligned}$$

where $\{\varepsilon_t\}_{t=0}^{\infty}$ and $\{u_t\}_{t=0}^{\infty}$ are two independent random variable with zero mean and finite variance σ_ε^2 and σ_u^2, i.i.d. random variables. We let $y_0 = 0$.

(a) Show, by successive substitution, that

$$y_t = \varepsilon_t + \sum_{j=1}^{t-1} \left(\prod_{i=0}^{j-1} A_{t-i} \right) \varepsilon_{t-j}.$$

(b) Find $E(y_t)$ and $Var(y_t)$.

(c) Under what condition will $Var(y_t)$ be finite? Under what condition will it be infinite as t goes to infinity?

(d) Let $\frac{1}{T}\sum_{t=1}^{T}(y_t - \overline{y})^2$ be the sample variance of y_t. As the sample size tends to infinity, compare the sample variance for the following three cases:
(i) $a = 1$, $\sigma_u^2 = 0$;
(ii) $a^2 = 0.5$, $\sigma_u^2 = 0.5$;
(iii) $a = -1$, $\sigma_u^2 = 0$.

Exercise 0.152 *Show that $\mathfrak{L}(c)$ can also be written as*

$$\begin{aligned} \mathfrak{L}(c) &= \frac{\int_0^1 (1+br)^{-1} B(r)\, dB(r)}{\int_0^1 (1+br)^{-2} B(r)^2\, dr}, \\ b &= (\exp(2c) - 1). \end{aligned}$$

Exercise 0.153 *(Difficult) Suppose the true process Y_t^* is an AR(1) process*

$$
\begin{aligned}
Y_t^* &= \beta Y_{t-1}^* + u_t, \\
Y_0^* &= 0, \\
u_t &\sim i.i.d.N\left(0,1\right).
\end{aligned}
$$

But we only observe

$$
\begin{aligned}
Y_t &= 0 \quad \text{if } Y_t^* \leq 0, \\
Y_t &= Y_t^* \quad \text{if } {Y_t}^* > 0.
\end{aligned}
$$

(a) Find the probability limit of the OLS estimator $\widehat{\beta}_T = \dfrac{\sum_{t=2}^{T} Y_{t-1} Y_t}{\sum_{t=2}^{T} Y_{t-1}^2}$.

(b) Describe how to use MLE to estimate β consistently based on the observed data $\{Y_t\}_{t=1}^T$.

ECO5120: Econometric Theory and Application, Fall 2006
Prof. T.L. Chong
HANDOUT 9
SOME TIME SERIES MODELS

Cointegration Model

An $(n \times 1)$ vector time series is said to be cointegrated if each of the series taken individually is I(1), while some linear combinations of the series is stationary. For example, consider the model

$$\begin{aligned} y_{1t} &= \gamma y_{2t} + u_{1t} \\ y_{2t} &= y_{2,t-1} + u_{2t}, \end{aligned}$$

while u_{1t} and u_{2t} are uncorrelated white noise processes. Thus we have

$$\Delta y_{2t} = u_{2t},$$

$$\begin{aligned} \Delta y_{1t} &= \gamma \Delta y_{2t} + \Delta u_{1t} \\ &= \gamma u_{2t} + u_{1t} - u_{1,t-1} \\ &= \varepsilon_t, \end{aligned}$$

with

$$\begin{aligned} E(\varepsilon_t) &= 0; \\ Var(\varepsilon_t) &= \gamma^2 Var(u_{2t}) + Var(u_{1t}) + Var(u_{1,t-1}) < \infty; \\ Cov(\varepsilon_t, \varepsilon_{t-k}) &= -Var(u_{1,t-k}) \qquad \text{for k}= \pm 1 \\ &= 0 \qquad \text{for k}= \pm 2, \pm 3, ... \end{aligned}$$

Thus, ε_t is a stationary process and hence y_{1t} is also an $I(1)$ process.

Hence, although both y_{1t} and y_{2t} are $I(1)$ processes, their linear combinations $(y_{1t} - \gamma y_{2t})$ can be an $I(0)$ process. In this case, we say y_{1t} and y_{2t} are cointegrated with the cointegrating vector $(1, -\gamma)$.

Structural Break Model

Consider the sequence of random variables

$$Y_t = \beta_1 + u_t, \qquad t = 1, 2, ..., k_0,$$

$$Y_t = \beta_2 + u_t, \qquad t = k_0 + 1, k_0 + 2, ..., T.$$

$u_t = i.i.d.(0, \sigma^2)$. β_1, β_2 and k_0 are unknown.

Estimation

For any given k, the pre- and post-shift estimators are

$$\widehat{\beta}_1(k) = \frac{\sum_{t=1}^{k} Y_t}{k},$$

$$\widehat{\beta}_2(k) = \frac{\sum_{t=k+1}^{T} Y_t}{T-k}.$$

The residual sum of squares at k is

$$RSS(k) = \sum_{t=1}^{k} \left(y_t - \widehat{\beta}_1(k)\right)^2 + \sum_{t=k+1}^{T} \left(y_t - \widehat{\beta}_2(k)\right)^2.$$

The least square estimator for k_0 is defined as

$$\widehat{k} = Argmin_k \; RSS(k).$$

Testing for structural break with unknown break date

We want to test if there is a break in coefficient, i.e.,

$$H_0 : \beta_1 = \beta_2,$$

or equivalently

$$H_0 : \beta_2 - \beta_1 = 0.$$

Note that under the null hypothesis the parameter k_0 does not exist.
For any given k, the Wald test statistic is

$$W(k) = \frac{\left(\widehat{\beta}_2(k) - \widehat{\beta}_1(k)\right)^2}{\widehat{Var}\left(\widehat{\beta}_2(k) - \widehat{\beta}_1(k)\right)},$$

where $\widehat{\beta}_1(k)$ and $\widehat{\beta}_2(k)$ are the least squares estimators for β_1 and β_2.

$$\begin{aligned}
Var\left(\widehat{\beta}_2(k) - \widehat{\beta}_1(k)\right) &= Var\left(\widehat{\beta}_1(k)\right) + Var\left(\widehat{\beta}_2(k)\right) - 2Cov\left(\widehat{\beta}_1(k), \widehat{\beta}_2(k)\right) \\
&= Var\left(\widehat{\beta}_1(k)\right) + Var\left(\widehat{\beta}_2(k)\right) \qquad \text{by the independence of } y_t \\
&= Var\left(\frac{1}{k}\sum_{t=1}^{k} y_t\right) + Var\left(\frac{1}{T-k}\sum_{t=k+1}^{T} y_t\right) \\
&= Var\left(\frac{1}{k}\sum_{t=1}^{k} u_t\right) + Var\left(\frac{1}{T-k}\sum_{t=k+1}^{T} u_t\right) \\
&= \frac{\sigma^2}{k} + \frac{\sigma^2}{T-k} \\
&= \frac{T\sigma^2}{k(T-k)}.
\end{aligned}$$

Thus

$$\widehat{Var}\left(\widehat{\beta}_2 - \widehat{\beta}_1\right) = \frac{T\widehat{\sigma}^2}{k(T-k)} = \frac{RSS(k)}{k(T-k)}$$

Hence, the Wald test becomes

$$W(k) = k(T-k)\frac{\left(\widehat{\beta}_2(k) - \widehat{\beta}_1(k)\right)^2}{RSS(k)}.$$

The Wald test above is only for a particular k. Under the null hypothesis, there is no break at all k. Thus, we have to form a test which can incorporate the case for all k. To avoid the boundary case, we restrict the set of k to be such that $\frac{k}{T} \in [a,b] \subset (0,1)$. We use the Sup-Wald test defined as

$$\sup_{\frac{k}{T}\in[a,b]\subset(0,1)} W(k) = \sup_{\frac{k}{T}\in[a,b]\subset(0,1)} k(T-k)\frac{\left(\widehat{\beta}_2(k) - \widehat{\beta}_1(k)\right)^2}{RSS(k)}.$$

$$W(k) \xrightarrow{d} \frac{(B(\tau) - \tau B(1))^2}{\tau(1-\tau)},$$

and

$$\sup_{\frac{k}{T}\in[a,b]\subset(0,1)} W(k) \xrightarrow{d} \sup_{\tau\in[a,b]\subset(0,1)} \frac{(B(\tau) - \tau B(1))^2}{\tau(1-\tau)},$$

where $B(\tau)$ is a standard Brownian motion on $[0,1]$.

Fractionally Integrated Model

Definition 167 *A time series process $\{y_t\}$ is said to be integrated of order d if $(1-L)^d y_t$ is stationary, where L is a lag operator such that $Ly_t = y_{t-1}$. If d is not an integer, then the process is said to be fractionally integrated.*

Consider the following model:

$$(1-L)^d y_t = u_t \qquad t = 1, 2, ..., T$$

where L is the lag operator and u_t is white noise.

$$y_t = (1-L)^{-d} u_t$$

$$\frac{\partial (1-L)^{-d}}{\partial L} = d(1-L)^{-d-1}$$

$$\frac{\partial^2 (1-L)^{-d}}{\partial L^2} = d(d+1)(1-L)^{-d-2}$$

and

$$\frac{\partial^j (1-L)^{-d}}{\partial L^j} = (d+j-1)(d+j-2)...(d+1)d(1-L)^{-d-j}$$

A power serious expansion for $(1-L)^{-d}$ around $L=0$ is given by

$$\begin{aligned}(1-L)^{-d} &= 1 + dL + \frac{1}{2!}(d+1)dL^2 + \frac{1}{3!}(d+2)(d+1)dL^3 + \cdots \\ &= \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} L^j,\end{aligned}$$

where $\Gamma(x)$ is the gamma function defined as

$$\begin{aligned}\Gamma(x) &= \int_0^{\infty} z^{x-1}\exp(-z)\,dz \qquad \text{for } x>0 \\ \Gamma(x) &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(x+k)k!} + \int_1^{\infty} z^{x-1}\exp(-z)\,dz \qquad \text{for } x<0, x\neq 0,-1,-2,-3,...\end{aligned}$$

Thus,

$$y_t = (1-L)^{-d} u_t = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} u_{t-j}.$$

The j^{th} autocorrelation of this $ARFIMA(0,d,0)$ process is given by

$$\rho_j(d) = \prod_{i=1}^{j} \frac{d+i-1}{i-d}.$$

Definition 168 *A time series process is said to have long memory if* $\sum_{j=-\infty}^{\infty} |\rho_j| = \infty$.

Fractionally integrated process has the following properties:

(i) A fractionally integrated process with $d > 0$ is a long memory process.

(ii) A fractionally integrated process with $d \geq 0.5$ is a nonstationary process.

Estimating d via the Autocorrelation Function

The parameter of interest in the model is d. There are a number of ways to estimate the parameter d. Tieslau, Schmidt and Baillie (1996) propose a minimum distance estimator of d defined to be

$$\widehat{d} = \underset{d\in(-.5,.25)}{Argmin} \left[\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\left(d\right)\right]' C^{-1} \left[\widehat{\rho} - \boldsymbol{\rho}\left(d\right)\right],$$

where

$\boldsymbol{\rho}\left(d\right)$ is a n by 1 vector with the j^{th} element $\rho_j\left(d\right)$.

$\widehat{\boldsymbol{\rho}}$ is a n by 1 vector with the j^{th} element $\widehat{\rho}_j$.

Since $E\left(y\right) = 0$, the sample autocorrelations can be defined as:

$$\widehat{\rho}_j = \frac{\sum_{t=j+1}^{T} y_t y_{t-j}}{\sum_{t=1}^{T} y_t^2}.$$

C is the asymptotic variance covariance matrix of $\widehat{\boldsymbol{\rho}}$

$c_{i,j}$ is given by

$$c_{i,j} = \sum_{s=1}^{\infty} \left(\rho_{s+i} + \rho_{s-i} - 2\rho_s\rho_i\right)\left(\rho_{s+j} + \rho_{s-j} - 2\rho_s\rho_j\right).$$

The Shortcoming of TSB's Estimator

In Tables 2,3 and 4 of TSB's paper, a substantial efficiency loss occurs when the first-order correlation is not used for the estimation of d. This

implies that the first-order autocorrelation carries most of the information needed for the estimation of d. This is due to the fact that the mapping between $\rho_n(d)$ and d is not one to one for all $n \geq 2$.

Note that when $n = 1$

$$\rho_1(d) = \frac{d}{1-d}.$$

In this instance, the mapping between d and $\rho_1(d)$ is one to one. However, for $n \geq 2$, different values of d may generate the same $\rho_n(d)$. Consider the values of d used in Table 2 of TSB's paper. Table A shows all others values of d which share the same n^{th} order autocorrelation for $n = 2, 3$.

Table A: Values of d which share the same n^{th} order autocorrelation for $n = 2, 3$.

d	$n=2$	$n=3$
$-.49$	$-.2576$	$-.1392, -2.6410$
$-.45$	$-.2895$	$-.1608, -2.6801$
$-.4$	$-.3333$	$-.1917, -2.7213$
$-.3$	$-.4375$	$-.2714, -2.7665$
$-.2$	$-.5714$	$-.3879, -2.7298$
$-.1$	$-.75$	$-.5772, -2.5430$
0	-1	$-1, -2$
$.1$	-1.375	$-1.3497 \pm .9213i$
$.2$	-2	$-1.0789 \pm 1.5197i$
$.24$	-2.3846	$-.9317 \pm 1.7192i$

For example, consider the case where $n = 2$, we have

$$\rho_2(d) = \prod_{i=1}^{2} \frac{d+i-1}{i-d}.$$

In this case, $\rho_2(-0.4) = \rho_2(-.3333) = -0.0714$.

Thus, if the true d is $d_0 = -0.4$, and if we estimate d_0 by using the second order autocorrelation only, the estimator converges to the set $\{-0.4, -.3333\}$.

Obviously, for $n = 1$, the criterion function is U-shape and thus it has a unique minimum. However, for $n \geq 2$, the shape changes with the true value of d.

The existence of multiple solution widens the variation of $\widehat{d}$. This will make the variance of $\sqrt{T}\left(\widehat{d} - d\right)$ diverge to infinity as the sample size increases.

Estimating d via the Partial Autocorrelation Function

Chong (2000) proposes another estimation method for d. The estimator differs from TSB's estimator in that it uses the sample partial correlation function to form the moment conditions. The n^{th} order partial autocorrelation function of a fractionally integrated process is:

$$\alpha_n(d) = \frac{d}{n-d}.$$

Since the mapping of $\alpha_n(d)$ and d is one to one for all n, we can either use a single $\alpha_n(d)$ or a combination of them to form an estimator of d.

To obtain the estimate of $\alpha_n(d)$. Let

$$X_n = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ y_1 & 0 & \cdots & 0 \\ y_2 & y_1 & \cdots & 0 \\ \vdots & y_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & y_1 \\ \vdots & \vdots & \cdots & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-n} \end{pmatrix},$$

$$Y = (y_1, y_2, ..., y_T)',$$

$$\widehat{\boldsymbol{\beta}}(n) = \begin{pmatrix} \widehat{\beta}_{n,1} & \widehat{\beta}_{n,2} & \cdots & \widehat{\beta}_{n,n-1} & \widehat{\beta}_{n,n} \end{pmatrix}' = (X_n' X_n)^{-1} X_n Y \xrightarrow{p} \Phi(n-1)^{-1} \boldsymbol{\rho}(n),$$

where $\Phi(n-1)$ is a $n \times n$ Toeplitz matrix defined as

$$\Phi(n-1) = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{pmatrix},$$

$$\boldsymbol{\rho}(n) = \begin{pmatrix} \rho_1 & \rho_2 & \cdots & \rho_{n-1} & \rho_n \end{pmatrix}'.$$

The element of $\widehat{\boldsymbol{\beta}}(n)$ will converge in probability to a function of d, in particular,

$$\begin{aligned} \widehat{\boldsymbol{\beta}}(2) \xrightarrow{p} \Phi(1)^{-1} \boldsymbol{\rho}(2) &= \begin{pmatrix} \rho_1 \frac{1-\rho_2}{1-\rho_1^2} & \frac{\rho_1^2-\rho_2}{\rho_1^2-1} \end{pmatrix}' = \begin{pmatrix} \frac{2d}{2-d} & \frac{d}{2-d} \end{pmatrix}', \\ \widehat{\boldsymbol{\beta}}(3) \xrightarrow{p} \Phi(2)^{-1} \boldsymbol{\rho}(3) &= \begin{pmatrix} \frac{3d}{3-d} & \frac{3d(1-d)}{(3-d)(2-d)} & \frac{d}{3-d} \end{pmatrix}', \\ \widehat{\boldsymbol{\beta}}(4) \xrightarrow{p} \Phi(3)^{-1} \boldsymbol{\rho}(4) &= \begin{pmatrix} \frac{4d}{4-d} & \frac{6d(1-d)}{(4-d)(3-d)} & \frac{4d(1-d)}{(4-d)(3-d)} & \frac{d}{4-d} \end{pmatrix}'. \end{aligned}$$

Thus,

$$\widehat{\beta}_{n,n} \xrightarrow{p} \frac{d}{n-d} = \alpha_n(d).$$

Hence, the n^{th} order sample partial autocorrelation can be obtained from the estimated coefficient of y_{t-n} in the regression of y_t on $y_{t-1}, y_{t-2}, ..., y_{t-n}$.

Chong's estimator of d is defined to be

$$\widehat{d} = \underset{d \in (-.5,.25)}{Argmin} S(d).$$

where

$$S(d) = \left[\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}(d)\right]' \Omega^{-1} \left[\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}(d)\right].$$

$\boldsymbol{\alpha}(d)$ is a n by 1 vector with the j^{th} element $\dfrac{d}{j-d}$.

$\widehat{\boldsymbol{\alpha}}$ is a n by 1 vector with the j^{th} element $\widehat{\beta}_{j,j}$.

The $(l,m)^{th}$ element of Ω is given by

$$\Omega_{l,m} = \lim_{T\to\infty} TCov\left(\widehat{\beta}_{l,l}, \widehat{\beta}_{m,m}\right) = \lim_{T\to\infty} T\left[L(l)\, E\left(\widehat{\boldsymbol{\beta}}(l) - \boldsymbol{\beta}(l)\right)\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}(m)\right)' L(m)'\right],$$

where

$$L(i) = \underbrace{(0\ 0\ ...\ 0\ 1)}_{i \text{ terms}}$$

To find $E\left(\widehat{\boldsymbol{\beta}}(l) - \boldsymbol{\beta}(l)\right)\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}(m)\right)'$. Note that

$$\begin{aligned}\widehat{\boldsymbol{\rho}}(n) - \boldsymbol{\rho}(n) &= \widehat{\Phi}(n-1)\,\widehat{\boldsymbol{\beta}}(n) - \Phi(n-1)\,\boldsymbol{\beta}(n) \\ &= \Phi(n-1)\left(\widehat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}(n)\right) + \left(\widehat{\Phi}(n-1) - \Phi(n-1)\right)\boldsymbol{\beta}(n) + O_p\left(T^{-1}\right),\end{aligned}$$

we have

$$\widehat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}(n) = \Phi(n-1)^{-1}\,\Delta(n),$$

where

$$\Delta(n) = (\widehat{\boldsymbol{\rho}}(n) - \boldsymbol{\rho}(n)) - \left(\widehat{\Phi}(n-1) - \Phi(n-1)\right)\boldsymbol{\beta}(n) + O_p\left(T^{-1}\right).$$

Hence, $\Omega_{l,m}$ is reduced to

$$\lim_{T\to\infty} T\left[L(l)\,\Phi(l-1)^{-1}\, E\left(\Delta(l)\,\Delta(m)'\right)\Phi(m-1)^{-1}\, L(m)'\right].$$

Note that

$$\begin{aligned}&\lim_{T\to\infty} TE\left(\Delta(l)\,\Delta(m)'\right) \\ =\ & C(l,m) - \lim_{T\to\infty} E\left(\widehat{\Phi}(l-1) - \Phi(l-1)\right)\boldsymbol{\beta}(l)\left(\widehat{\boldsymbol{\rho}}(m) - \boldsymbol{\rho}(m)\right)' \\ & - \lim_{T\to\infty} E\left(\widehat{\boldsymbol{\rho}}(l) - \boldsymbol{\rho}(l)\right)\boldsymbol{\beta}(m)'\left(\widehat{\Phi}(m-1) - \Phi(m-1)\right)' \\ & + \lim_{T\to\infty} E\left(\widehat{\Phi}(l-1) - \Phi(l-1)\right)\boldsymbol{\beta}(l)\,\boldsymbol{\beta}(m)'\left(\widehat{\Phi}(m-1) - \Phi(m-1)\right),\end{aligned}$$

where $C(l,m)$ is a l by m matrix with the $(i,j)^{th}$ element $c_{i,j}$.
The $(i,j)^{th}$ element of the matrix $\lim_{T\to\infty} TE\left(\Delta(l)\Delta(m)'\right)$ is given by

$\lim_{T\to\infty} TE\left(\Delta(l)\Delta(m)'\right)_{i,j} = c_{i,j} - \sum_{h=1,h\neq i}^{l} \beta_{l,h} c_{|i-h|,j} - \sum_{k=1,k\neq j}^{m} \beta_{m,k} c_{|j-k|,i} + \sum_{h=1,h\neq i}^{l} \sum_{k=1,k\neq j}^{m} c_{|i-h|,|k-j|} \beta_{l,h} \beta_{m,k}$.

Hence, the matrix Ω^{-1} can be constructed, and we can evaluate $S(d)$ at various values of d and find a d to minimize $S(d)$.

References:

1. Andrews D.W.K.(1993)"Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821-856.

2. Chong, T.T.L. (1995) "Partial Parameter Consistency in a Misspecified Structural Change Model," *Economics Letters*, 49, 351-357.

3. Chong, T.T.L. and C.S. Lui (1999)"Estimating the Fractionally Integrated Process in the Presence of Measurement Errors," *Economics Letters*, 63, 285-294.

4. Chong, T.T.L. (2000) "Estimating the Differencing Parameter via the Partial Autocorrelation Function," *Journal of Econometrics,* 97, 365-381.

5. Chong, T.T.L. (2001a) "Structural Change in AR(1) Models," *Econometric Theory*, 17, 87-155.

6. Chong, T.T.L. (2001b) "Estimating the Locations and Number of Change Points by the Sample-Splitting Method," *Statistical Papers*, 42, 53-79.

7. Chong, T.T.L. (2003) "Generic Consistency of the Break-Point Estimator under Specification Errors," *Econometrics Journal*, 6, 167-192.

8. Tieslau, M.A., P. Schmidt and R.T. Baillie (1996)"A Minimum Distance Estimator for Long-Memory Processes," *Journal of Econometrics,* 71, 249-64.

Exercise 0.154 *Verify the values in Table 1 of Andrews 1993, (Econometrica, pp.821-856). Use sample size T=100,1000,3600, number of repetitions N=10000.*

Exercise 0.155 *True/False?*

(a) If two variables X and Y are cointegrated, then they are both non-stationary.

(b) A Brownian motion has independent increments.

Exercise 0.156 *Consider the following model.*

$$y_t = \beta_1 x_t + u_t \qquad t = 1, 2, ..., k_0$$

$$y_t = \beta_2 x_t + u_t \qquad k_0 + 1, k_0 + 2, ..., T$$

$t = 1, 2, ..., T$, $x_t \sim U(0, 1)$, $u_t \sim N(0, 1)$, $\{x_t\}_{t=1}^T$ and $\{u_t\}_{t=1}^T$ are independent.

a) Derive the Least Squares estimators for $\widehat{\beta}_1$, $\widehat{\beta}_2$.

b) Construct a Sup-Wald Test for the hypothesis that $H_0 : \beta_1 = \beta_2$.

c) Derive the asymptotic distribution for the Sup-Wald test.

d) Under $H_0 : \beta_1 = \beta_2 (= 2, \text{say})$, use GAUSS to simulate the sampling distribution of $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{k}$ and SupWald for T=50, 100, 1000, using 20000 replications.

Exercise 0.157 *Consider the model:*

$$y_t = \beta_1 x_t + \varepsilon_t \qquad t \leq k_0$$

$$y_t = \beta_2 x_t + \varepsilon_t \qquad t > k_0$$

$(t = 1, 2, ..., T)$,

where β_1 and β_2 are the pre-shift and post-shift regression slope parameters respectively, and let $\frac{k_0}{T} = \tau_0$ be fixed.

(a) For any given k, find the OLS estimators of $\widehat{\beta}_1(k), \widehat{\beta}_2(k)$ and $\widehat{k}$.

(b) If the true x_t is misspecified as x_t^2, will the OLS estimators of $\widehat{\tau} = \frac{\widehat{k}}{T}$ be consistent? Will the OLS estimators of β_1, β_2 be consistent? Explain.

(c) If the true model has two breaks

$$y_t = \beta_1 x_t + \varepsilon_t \qquad t \leq k_1$$

$$y_t = \beta_2 x_t + \varepsilon_t \qquad k_1 < t \leq k_2$$

$$y_t = \beta_3 x_t + \varepsilon_t \qquad t > k_2$$

but we estimate a one-break model, plot a graph to approximate the behavior of $\frac{1}{T} RSS(\tau)$ when T is large, where $RSS(\tau)$ is the residual sum of squares for any given $\tau \in [0, 1]$. What will τ converge to? What factor(s) will affect the probability limit $\widehat{\tau}$?

Exercise 0.158 *Consider the following model.*

$$\begin{aligned} y_t &= \beta_1 x_t^* + u_t, \qquad t = 1, 2, ..., k_0, \\ y_t &= \beta_2 x_t^* + u_t, \qquad k_0 + 1, k_0 + 2, ..., T. \end{aligned}$$

Suppose x_t^* is not observable and we only observe x_t.

a) Let $\tau = \frac{k}{T}$, $k = 1, 2, ..., T$. For any given τ, derive the Least squares estimators $\widehat{\beta}_{1\tau}$, $\widehat{\beta}_{2\tau}$.

b) Suppose $x_t = x_t^* + \varepsilon_t$, where $\{x_t^*\}_{t=1}^T$, $\{\varepsilon_t\}_{t=1}^T$and $\{u_t\}_{t=1}^T$ are independent.

$$\begin{aligned} E\left(x_t^{*2}\right) &= \sigma_*^2, \\ \varepsilon_t &\sim i.i.d.\left(0, \sigma_\varepsilon^2\right). \end{aligned}$$

Let

$$\begin{aligned} A &= \frac{\sigma_*^2}{\sigma_*^2 + \sigma_\varepsilon^2}, \\ \Gamma_{1\tau} &= \beta_1 \frac{\tau_0}{\tau} + \beta_2 \frac{\tau - \tau_0}{\tau}, \\ \Gamma_{2\tau} &= \beta_1 \frac{\tau_0 - \tau}{1 - \tau} + \beta_2 \frac{1 - \tau_0}{1 - \tau}. \end{aligned}$$

Show that for $\tau \in [\underline{\tau}, \tau_0]$,

$$\begin{aligned} \widehat{\beta}_{1\tau} &\xrightarrow{p} \beta_1 A, \\ \widehat{\beta}_{2\tau} &\xrightarrow{p} \Gamma_{2\tau} A. \end{aligned}$$

and for $\tau \in (\tau_0, \overline{\tau}]$,

$$\begin{aligned} \widehat{\beta}_{1\tau} &\xrightarrow{p} \Gamma_{1\tau} A, \\ \widehat{\beta}_{2\tau} &\xrightarrow{p} \beta_2 A. \end{aligned}$$

c) Let

$$RSS_T(\tau) = \sum_{t=1}^{[\tau T]} \left(y_t - \widehat{\beta}_{1\tau} x_t\right)^2 + \sum_{t=[\tau T]+1}^{T} \left(y_t - \widehat{\beta}_{2\tau} x_t\right)^2.$$

Show that $\frac{1}{T} RSS_T(\tau) \xrightarrow{p} h(\tau)$ where for $\tau \in [\underline{\tau}, \tau_0)$,

$$\begin{aligned} h(\tau) &= \sigma_u^2 + (1 - \tau_0)\left(\beta_2^2 - \beta_1^2\right)\sigma_*^2 + \beta_1^2 (1 - A)\sigma_*^2 + (1 - \tau)\left(\beta_1^2 - \Gamma_{2\tau}^2\right) A\sigma_*^2, \\ \frac{\partial h(\tau)}{\partial \tau} &\leq 0, \\ \frac{\partial^2 h(\tau)}{\partial \tau^2} &\leq 0. \end{aligned}$$

For $\tau = \tau_0$,

$$h(\tau_0) = \sigma_u^2 + \left(\tau_0 \beta_1^2 + (1-\tau_0)\beta_2^2\right)(1-A)\sigma_*^2.$$

For $\tau \in (\tau_0, \overline{\tau}]$,

$$\begin{aligned} h(\tau) &= \sigma_u^2 + \tau_0\left(\beta_1^2 - \beta_2^2\right)\sigma_*^2 + \beta_2^2(1-A)\sigma_*^2 + \tau\left(\beta_2^2 - \Gamma_{1\tau}^2\right)A\sigma_*^2, \\ \frac{\partial h(\tau)}{\partial \tau} &\geq 0, \\ \frac{\partial^2 h(\tau)}{\partial \tau^2} &\leq 0. \end{aligned}$$

d) Construct a Sup-Wald Test for the hypothesis that $H_0 : \beta_1 = \beta_2$.

e) Under $H_0 : \beta_1 = \beta_2 (= 2, \text{say})$, write a GAUSS program to simulate the sampling distribution of the SupWald test for $T = 1000$, using 20000 replications, $\varepsilon_t \sim N(0,1)$, $x_t^* \sim U(0,1)$, $u_t \sim N(0,1)$, $\{x_t^*\}_{t=1}^T$, $\{\varepsilon_t\}_{t=1}^T$and $\{u_t\}_{t=1}^T$ are independent.

Exercise 0.159 *Find two text books which give the definition of the gamma function* $\Gamma(x)$ *for* $x < 0$.

Exercise 0.160 *Consider the following model:*

$$(1-L)^d y_t = u_t \qquad t = 1, 2, ..., T$$

where L is the lag operator and u_t follows an i.i.d.$(0, \sigma^2)$ process for $0 < \sigma^2 < \infty$.

a) Suppose

$$(1-L)^d = \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)} L^j.$$

Show that

i)

$$y_t = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} u_{t-j},$$

ii)

$$y_{t+n} = \sum_{j=-n}^{\infty} \frac{\Gamma(j+n+d)}{\Gamma(d)\Gamma(j+n+1)} u_{t-j},$$

iii)

$$E(y_{t+n}y_t) = \sigma^2 \sum_{j=0}^{\infty} \frac{\Gamma(j+d)\Gamma(j+n+d)}{(\Gamma(d))^2 \Gamma(j+1)\Gamma(j+n+1)},$$

iv)

$$E(y_t^2) = \sigma^2 \sum_{j=0}^{\infty} \left(\frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} \right)^2.$$

b) Using

$$\sum_{j=0}^{\infty} \frac{\Gamma(j+d)\Gamma(j+n+d)}{(\Gamma(d))^2 \Gamma(j+1)\Gamma(j+n+1)} = \frac{\Gamma(n+d)\Gamma(1-2d)}{\Gamma(d)\Gamma(1-d)\Gamma(n+1-d)}$$

and

$$\sum_{j=0}^{\infty} \left(\frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} \right)^2 = \frac{\Gamma(1-2d)}{\Gamma^2(1-d)}.$$

Show that the n^{th} autocorrelation of this process is given by

$$\rho_n = \frac{\Gamma(1-d)\Gamma(n+d)}{\Gamma(d)\Gamma(n+1-d)}.$$

c) Show that the ρ_n can also be written as

$$\rho_n = \prod_{i=1}^{n} \frac{d+i-1}{i-d}.$$

Now suppose we run a regression of y_t on $y_{t-1}, y_{t-2}, .., y_{t-n}$, the estimated coefficients are

$$\widehat{\boldsymbol{\beta}}(n) = \begin{pmatrix} \widehat{\beta}_{n,1} & \widehat{\beta}_{n,2} & \cdots & \widehat{\beta}_{n,n-1} & \widehat{\beta}_{n,n} \end{pmatrix}'.$$

d) Show that

$$\widehat{\boldsymbol{\beta}}(n) \xrightarrow{p} \Phi(n-1)^{-1} \boldsymbol{\rho}(n),$$

where

$$\Phi(n-1) = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{pmatrix},$$

$$\boldsymbol{\rho}(n) = \begin{pmatrix} \rho_1 & \rho_2 & \cdots & \rho_{n-1} & \rho_n \end{pmatrix}'.$$

e) Show that, as $T \to \infty$,

$$\widehat{\boldsymbol{\beta}}(2) \xrightarrow{p} \begin{pmatrix} \frac{2d}{2-d} & \frac{d}{2-d} \end{pmatrix}'.$$

f) Suppose we also have the results that

$$\widehat{\boldsymbol{\beta}}(3) \xrightarrow{p} \Phi(2)^{-1} \boldsymbol{\rho}(3) = \begin{pmatrix} \frac{3d}{3-d} & \frac{3d(1-d)}{(3-d)(2-d)} & \frac{d}{3-d} \end{pmatrix}',$$

$$\widehat{\boldsymbol{\beta}}(4) \xrightarrow{p} \Phi(3)^{-1} \boldsymbol{\rho}(4) = \begin{pmatrix} \frac{4d}{4-d} & \frac{6d(1-d)}{(4-d)(3-d)} & \frac{4d(1-d)}{(4-d)(3-d)} & \frac{d}{4-d} \end{pmatrix}',$$

and that, in general, as $T \to \infty$,

$$plim\widehat{\beta}_{n,1} = plim\left(n\widehat{\beta}_{n,n}\right),$$

for d belongs to $(-0.5, 0.25)$.

Discuss how to use the above information to derive a test for the hypothesis that

$H_0 : \exists d \in (-.5, .25)$ such that $y_t \sim ARFIMA(0, d, 0)$.
$H_1 : H_0$ is not true.

Exercise 0.161 *Using GAUSS to generate an* $ARFIMA(0, d, 0)$ *process with* $d = .5,\ -1,\ 0,\ -.1,\ -.25,\ -.5$.

Exercise 0.162 *Using GAUSS to recompute the values of Table 1 and Table 3 of TSB's paper, approximate $c_{i,j}$ by*

$$c_{i,j} = \sum_{s=1}^{T} \left(\rho_{s+i} + \rho_{s-i} - 2\rho_s\rho_i\right)\left(\rho_{s+j} + \rho_{s-j} - 2\rho_s\rho_j\right),$$

where $T = 100000$.

Show that there are some typos in TSB's paper. In their Table 1, for $d = 0$ and $n = 3$, the value should be 0.7347 instead of 0.7437. In the same table, for $d = -0.3$ and $n = 20$, the correct value should be 0.7426 instead of 0.8625. In Table 3, for $d = 0$ and $n = 2 - 6$, $n = 3 - 7$, the true values should be 2.035 and 3.8198 respectively instead of 1.5866 and 2.4957.

Exercise 0.163 *Suppose the true model is an $ARFIMA(1, d, 0)$ process of the form*

$$(1 - L)^d (1 - \phi L) y_t = \varepsilon_t.$$

Show that

$$y_t = (1 - \phi L)^{-1} (1 - L)^{-d} \varepsilon_t = \sum_{j=0}^{\infty} \mu(j) L^j \varepsilon_t,$$

where

$$\mu(j) = \sum_{i=0}^{j} \left(\frac{\Gamma(i + d)}{\Gamma(d)\Gamma(i + 1)} \phi^{j-i}\right),$$

and

$$\widehat{\rho}_n \xrightarrow{p} \frac{\sum_{j=n}^{\infty} (\mu(j)\mu(j - n))}{\sum_{j=0}^{\infty} (\mu(j))^2}.$$

If the true model is an $ARFIMA(0, d, 1)$ process of the form

$$(1 - L)^d y_t = (1 + \theta L) \varepsilon_t$$

Show that

$$y_t = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} L^j (1+\theta L)\varepsilon_t = \sum_{j=0}^{\infty} \lambda(j) L^j \varepsilon_t,$$

where

$$\begin{aligned} \lambda(0) &= 1, \\ \lambda(j) &= \frac{\Gamma(j-1+d)}{\Gamma(d)\Gamma(t)}\theta + \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} \qquad for\ j \geq 1. \end{aligned}$$

$$\begin{aligned} &\widehat{\rho}_n \xrightarrow{p} \frac{\sum_{j=n}^{\infty} (\lambda(j)\lambda(j-n))}{\sum_{j=0}^{\infty} (\lambda(j))^2} \\ = &\frac{\left(1+\theta^2\right)\Gamma(n+d)}{\Gamma(n+1-d)} + \frac{\theta\Gamma(n+1+d)}{\Gamma(n+2-d)} + \frac{\theta\Gamma(n-1+d)}{\Gamma(n-d)} \\ &\frac{\left(1+\theta^2\right)\Gamma(d)}{\Gamma(1-d)} + \frac{2\theta\Gamma(1+d)}{\Gamma(2-d)}. \end{aligned}$$

Exercise 0.164 *Consider the following model:*

$$(1-L)^d y_t^* = u_t, \qquad t = 1, 2, ..., T,$$

where L is a lag operator such that $Ly_t^* = y_{t-1}^*$.
Suppose the true values of $\{y_t^*\}_{t=1}^T$ are not observable. Instead, we observe

$$y_t = y_t^* + \varepsilon_t \qquad t = 1, 2, ..., T,$$

where $\{\varepsilon_t\}_{t=1}^T$ is the measurement error process.
(a) Show that

$$\widehat{\rho}_j = \frac{\sum_{t=1}^{T-j} (y_t - \overline{y})(y_{t+j} - \overline{y})}{\sum_{t=1}^{T} (y_t - \overline{y})^2} \xrightarrow{p} A\left(\rho_j + \frac{\gamma_j}{\sigma_*^2}\right)$$

where

$$A = \frac{\sigma_*^2}{\sigma_*^2 + \gamma_0},$$

$$\sigma_*^2 = Var\left(y_t^*\right) = \sigma_u^2 \frac{\Gamma\left(1-2d\right)}{\Gamma^2\left(1-d\right)},$$

$$\gamma_j = Cov\left(\varepsilon_t, \varepsilon_{t+j}\right).$$

(b) Suppose we run a regression of y_t on $y_{t-1}, y_{t-2}, ..., y_{t-n}$, show that the estimators converge in probability to:

$$\widehat{\boldsymbol{\beta}}\left(n\right) \xrightarrow{p} \boldsymbol{\Phi}\left(n\right)^{-1}\left(\boldsymbol{\rho}\left(n\right) + \sigma_*^{-2}\boldsymbol{\gamma}\left(n\right)\right),$$

where

$$\widehat{\boldsymbol{\beta}}\left(n\right) = \begin{pmatrix} \widehat{\beta}_{n,1} & \widehat{\beta}_{n,2} & \cdots & \widehat{\beta}_{n,n-1} & \widehat{\beta}_{n,n} \end{pmatrix}',$$

$$\boldsymbol{\rho}\left(n\right) = \begin{pmatrix} \rho_1 & \rho_2 & \cdots & \rho_{n-1} & \rho_n \end{pmatrix}',$$

$$\boldsymbol{\gamma}\left(n\right) = \begin{pmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_{n-1} & \gamma_n \end{pmatrix}',$$

$$\boldsymbol{\Phi}\left(n\right) = \begin{pmatrix} A^{-1} & \rho_1 + \frac{\gamma_1}{\sigma_*^2} & \cdots & \rho_{n-1} + \frac{\gamma_{n-1}}{\sigma_*^2} \\ \rho_1 + \frac{\gamma_1}{\sigma_*^2} & A^{-1} & \cdots & \rho_{n-2} + \frac{\gamma_{n-2}}{\sigma_*^2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} + \frac{\gamma_{n-1}}{\sigma_*^2} & \rho_{n-2} + \frac{\gamma_{n-2}}{\sigma_*^2} & \cdots & A^{-1} \end{pmatrix}.$$

Exercise 0.165 *Consider the following model:*

$$\begin{aligned} \left(1-L\right)^{d_1} y_t &= \varepsilon_t \qquad \text{for } t = 1, 2, ..., k_0, \\ \left(1-L\right)^{d_2} y_t &= \varepsilon_t \qquad \text{for } t = k_0+1, k_0+2, ...T. \end{aligned}$$

where

$$(1-L)^d = \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)\,\Gamma(j+1)} L^j,$$

$\Gamma(x)$ is the gamma function such that $\Gamma(x) = (x-1)\,\Gamma(x-1)$, and L is the lag operator such that $LZ_t = Z_{t-1}$ for any time series variable Z_t.

(a) Show that the model can be rewritten as

$$\begin{aligned} y_t &= \sum_{j=0}^{\infty} \frac{\Gamma(j+d_1)}{\Gamma(d_1)\,\Gamma(j+1)} \varepsilon_{t-j} \qquad \text{for } t \leq k_0, \\ y_t &= \sum_{j=0}^{\infty} \frac{\Gamma(j+d_2)}{\Gamma(d_2)\,\Gamma(j+1)} \varepsilon_{t-j} \qquad \text{for } t > k_0. \end{aligned}$$

Assume that
$(A1)$ $\varepsilon_t \sim i.i.d.\,(0, \sigma^2)$ $\forall t$, $0 < \sigma^2 < \infty$ and $E(\varepsilon_t^4) < \infty$;
$(A2)$ $\tau_0 = \dfrac{k_0}{T} \in [\underline{\tau}, \overline{\tau}] \subset (0,1)$;
$(A3)$ $(d_1, d_2) \in (-.5, .25) \times (-.5, .25)$.

If there is no structural change such that $d_1 = d_2 = d$, the h^{th} $(h = 1, 2, 3, ...)$ order autocovariance and autocorrelation of the process y_t are given by

$$\gamma_h(d) = Cov(y_t, y_{t-h}) = \rho_h(d)\,\gamma_0(d)$$

and

$$\rho_h(d) = \frac{Cov(y_t, y_{t-h})}{Var(y_t)\,Var(y_{t-h})} = \frac{\Gamma(h+d)\,\Gamma(1-d)}{\Gamma(h+1-d)\,\Gamma(d)}$$

respectively, where

$$\gamma_0(d) = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma^2(1-d)}.$$

(b) Show that $\rho_1(d) = \dfrac{d}{1-d}$.

(c) Since we have two regimes in our model, let $h = m - l$, show that for any $l \leq k_0 < m$, , we have

$$Cov\left(y_l, y_m\right) = \sigma^2 \frac{\Gamma\left(h + d_2\right)\Gamma\left(1 - d_1 - d_2\right)}{\Gamma\left(h + 1 - d_1\right)\Gamma\left(d_2\right)\Gamma\left(1 - d_2\right)}$$

and

$$Corr\left(y_l, y_m\right) = \frac{\Gamma\left(h + d_2\right)\Gamma\left(1 - d_1\right)\Gamma\left(1 - d_1 - d_2\right)}{\Gamma\left(h + 1 - d_1\right)\Gamma\left(d_2\right)\sqrt{\Gamma\left(1 - 2d_1\right)\Gamma\left(1 - 2d_2\right)}}.$$

Now let $k = [\tau T]$, where $[\cdot]$ is the greatest integer function. We define

$$\widehat{\tau} = \underset{\tau \in [\underline{\tau}, \overline{\tau}]}{Arg\max} M_T\left(\tau\right),$$

where

$$M_T\left(\tau\right) = \left[\widehat{\rho}_1\left(1, [\tau T]\right) - \widehat{\rho}_1\left([\tau T] + 1, T\right)\right]^2,$$

$$\widehat{\rho}_1\left(i, j\right) = \frac{\sum_{t=i+1}^{j} y_t y_{t-1}}{\sum_{t=i+1}^{j} y_{t-1}^2}.$$

Thus, the break-point estimate is defined to be the time when the difference between the two first-order autocorrelations is maximized.

(d) Show that for $\tau \leq \tau_0$, we have

$$M_T\left(\tau\right) \xrightarrow{p} \frac{\left(d_1 - d_2\right)^2}{\left(1 - d_1\right)^2\left(1 - d_2\right)^2} h_1^2\left(\tau\right),$$

where

$$h_1\left(\tau\right) = \frac{\left(1 - \tau_0\right)\Gamma\left(1 - 2d_2\right)\Gamma^2\left(1 - d_1\right)}{\left(\tau_0 - \tau\right)\Gamma\left(1 - 2d_1\right)\Gamma^2\left(1 - d_2\right) + \left(1 - \tau_0\right)\Gamma\left(1 - 2d_2\right)\Gamma^2\left(1 - d_1\right)}.$$

and for $\tau > \tau_0$,

$$M_T(\tau) \xrightarrow{p} \frac{(d_1 - d_2)^2}{(1-d_1)^2 (1-d_2)^2} h_2^2(\tau),$$

where

$$h_2(\tau) = \frac{\tau_0 \Gamma^2 (1-d_2) \Gamma(1-2d_1)}{\tau_0 \Gamma(1-2d_1) \Gamma^2(1-d_2) + (\tau - \tau_0) \Gamma(1-2d_2) \Gamma^2(1-d_1)}.$$

(e) Is $\widehat{\tau}$ is a consistent estimator for τ_0? Explain in words.

(f) After getting the estimate of the change point, write down the consistent estimators for d_1 and d_2.

Exercise 0.166 *In the threshold model, verify that*

(i) $E(\xi_i) = 3$;
(ii) The moment generating function of ξ_i $(i = 1, 2)$ is $\dfrac{1}{(1-t)(1-2t)}$;
(iii) $E(\xi) = 6$;
(iv) The m.g.f. of ξ is $\left(\dfrac{1}{(1-t)(1-2t)}\right)^2$.

ECO5120: Econometric Theory and Application, Fall 99
Prof. T.L. Chong
HANDOUT 8
FRACTIONALLY INTEGRATED PROCESSES

A time series process $\{y_t\}$ is said to be integrated of order d if $(1-L)^d y_t$ is stationary, where L is a lag operator such that $Ly_t = y_{t-1}$. If d is not an integer, then the process is said to be fractionally integrated. Consider the following model:

$$(1-L)^d y_t = u_t \qquad t = 1, 2, ..., T$$

where L is the lag operator and u_t is white noise.

$$y_t = (1-L)^{-d} u_t$$

$$\frac{\partial (1-L)^{-d}}{\partial L} = d(1-L)^{-d-1}$$

$$\frac{\partial^2 (1-L)^{-d}}{\partial L^2} = d(d+1)(1-L)^{-d-2}$$

and

$$\frac{\partial^j (1-L)^{-d}}{\partial L^j} = (d+j-1)(d+j-2)...(d+1)d(1-L)^{-d-j}$$

A power series expansion for $(1-L)^{-d}$ around $L=0$ is given by

$$\begin{aligned}(1-L)^{-d} &= 1 + dL + \frac{1}{2!}(d+1)dL^2 + \frac{1}{3!}(d+2)(d+1)dL^3 + \cdots \\ &= \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} L^j,\end{aligned}$$

where $\Gamma(x)$ is the gamma function defined as

$$\Gamma(x) = \int_0^\infty z^{x-1}\exp(-z)\,dz \qquad \text{for } x>0$$

$$\Gamma(x) = \sum_{k=0}^{\infty}\frac{(-1)^k}{(x+k)\,k!} + \int_1^\infty z^{x-1}\exp(-z)\,dz \qquad \text{for } x<0, x\neq 0,-1,-2,-3,...$$

Thus,

$$y_t = (1-L)^{-d}u_t = y_t = \sum_{j=0}^{\infty}\frac{\Gamma(j+d)}{\Gamma(d)\,\Gamma(j+1)}u_{t-j}.$$

Estimating d via the Autocorrelation Function

In a recent study, Tieslau, Schmidt and Baillie (1996) propose a minimum distance estimator of d defined to be

$$\widehat{d} = \underset{d\in(-.5,.25)}{Argmin}\left[\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}(d)\right]' C^{-1}\left[\widehat{\rho} - \boldsymbol{\rho}(d)\right],$$

where

$\boldsymbol{\rho}(d)$ is a n by 1 vector with the j^{th} element $\rho_j(d)$.

The j^{th} autocorrelation of this $ARFIMA(0,d,0)$ process is given by

$$\rho_j(d) = \prod_{i=1}^{j}\frac{d+i-1}{i-d},$$

$\widehat{\boldsymbol{\rho}}$ is a n by 1 vector with the j^{th} element $\widehat{\rho}_j$.

Since $E(y)=0$, the sample autocorrelations can be defined as:

$$\widehat{\rho}_j = \frac{\sum_{t=j+1}^{T} y_t y_{t-j}}{\sum_{t=1}^{T} y_t^2}.$$

C is the asymptotic variance covariance matrix of $\widehat{\boldsymbol{\rho}}$

$c_{i,j}$ is given by

$$c_{i,j} = \sum_{s=1}^{\infty}\left(\rho_{s+i} + \rho_{s-i} - 2\rho_s\rho_i\right)\left(\rho_{s+j} + \rho_{s-j} - 2\rho_s\rho_j\right).$$

The Shortcoming of TSB's Estimator

In Tables 2,3 and 4 of TSB's paper, a substantial efficiency loss occurs when the first-order correlation is not used for the estimation of d. This implies that the first-order autocorrelation carries most of the information needed for the estimation of d. We argue that their findings are due to the fact that the mapping between $\rho_n(d)$ and d is not one to one for all $n \geq 2$.

Note that when $n = 1$

$$\rho_1(d) = \frac{d}{1-d}.$$

In this instance, the mapping between d and $\rho_1(d)$ is one to one. However, for $n \geq 2$, different values of d may generate the same $\rho_n(d)$. Consider the values of d used in Table 2 of TSB's paper. Table A shows all others values of d which share the same n^{th} order autocorrelation for $n = 2, 3$.

Table A: Values of d which share the same n^{th} order autocorrelation for $n = 2, 3$.

d	$n=2$	$n=3$
$-.49$	$-.2576$	$-.1392, -2.6410$
$-.45$	$-.2895$	$-.1608, -2.6801$
$-.4$	$-.3333$	$-.1917, -2.7213$
$-.3$	$-.4375$	$-.2714, -2.7665$
$-.2$	$-.5714$	$-.3879, -2.7298$
$-.1$	$-.75$	$-.5772, -2.5430$
0	-1	$-1, -2$
$.1$	-1.375	$-1.3497 \pm .9213i$
$.2$	-2	$-1.0789 \pm 1.5197i$
$.24$	-2.3846	$-.9317 \pm 1.7192i$

For example, consider the case where $n = 2$, we have

$$\rho_2(d) = \prod_{i=1}^{2} \frac{d+i-1}{i-d}.$$

In this case, $\rho_2(-0.4) = \rho_2(-.3333) = -0.0714$.

Thus if the true d is $d_0 = -0.4$, and if we estimate d_0 by using the second order autocorrelation only, the estimator converges to the set $\{-0.4, -.3333\}$.

Obviously, for $n = 1$, the criterion function is U-shape and thus it has a unique minimum. However, for $n \geq 2$, the shape changes with the true value of d.

The existence of multiple solution widens the variation of $\widehat{d}$. This will make the variance of $\sqrt{T}\left(\widehat{d} - d\right)$ diverge to infinity as the sample size increases.

Estimating d via the Partial Autocorrelation Function

The following results are based on my recent study. I propose another estimation method for d. My estimator differs from TSB's estimator in that I use the sample partial correlation function to form the moment conditions. The n^{th} order partial autocorrelation function of a fractionally integrated process is:

$$\alpha_n(d) = \frac{d}{n-d}.$$

The expression can be found in Brockwell and Davis (1991, p. 522, Eq.13.2.10). A salient feature of $\alpha_n(d)$ is that its relationship with d is one to one for all n. Hence we can either use a single $\alpha_n(d)$ or a combination of them to form an estimator of d.

Next, let's discuss how to obtain an estimate of $\alpha_n(d)$. Let

$$X_n = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ y_1 & 0 & \cdots & 0 \\ y_2 & y_1 & \cdots & 0 \\ \vdots & y_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & y_1 \\ \vdots & \vdots & \cdots & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-n} \end{pmatrix},$$

$$Y = (y_1, y_2, ..., y_T)',$$

$$\boldsymbol{\rho}(n) = \begin{pmatrix} \rho_1 & \rho_2 & \cdots & \rho_{n-1} & \rho_n \end{pmatrix}',$$

$$\widehat{\boldsymbol{\beta}}(n) = \begin{pmatrix} \widehat{\beta}_{n,1} & \widehat{\beta}_{n,2} & \cdots & \widehat{\beta}_{n,n-1} & \widehat{\beta}_{n,n} \end{pmatrix}' = (X_n' X_n)^{-1} X_n Y$$

$$= \begin{pmatrix} \sum_{t=2}^{T} y_{t-1}^2 & \sum_{t=3}^{T} y_{t-1}y_{t-2} & \cdots & \sum_{t=n+1}^{T} y_{t-1}y_{t-n} \\ \sum_{t=3}^{T} y_{t-1}y_{t-2} & \sum_{t=3}^{T} y_{t-2}^2 & \cdots & \sum_{t=n+1}^{T} y_{t-2}y_{t-n} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=n+1}^{T} y_{t-1}y_{t-n} & \sum_{t=n+1}^{T} y_{t-2}y_{t-n} & \cdots & \sum_{t=n+1}^{T} y_{t-n}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=2}^{T} y_t y_{t-1} \\ \sum_{t=3}^{T} y_t y_{t-2} \\ \vdots \\ \sum_{t=n+1}^{T} y_t y_{t-n} \end{pmatrix}.$$

Dividing each element by $\sum_{t=1}^{T} y_t^2$ and take probability limit, we have:

$$\widehat{\boldsymbol{\beta}}(n) \xrightarrow{p} \Phi(n-1)^{-1} \boldsymbol{\rho}(n).$$

Since $E(y_t)$ is assumed to be 0 for all t, if we divide each element in the above matrix by $\sum_{t=2}^{T} y_{t-1}^2$ and take probability limit, we have:

$$\widehat{\boldsymbol{\beta}}(n) \xrightarrow{p} \Phi(n-1)^{-1} \boldsymbol{\rho}(n),$$

where $\Phi(n-1)$ is a $n \times n$ Toeplitz matrix defined as

$$\Phi(n-1)=\begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{pmatrix}.$$

The element of $\widehat{\boldsymbol{\beta}}(n)$ will converge in probability to a function of d, in particular,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}(2) \xrightarrow{p} \Phi(1)^{-1}\boldsymbol{\rho}(2) &= \begin{pmatrix} \rho_1\frac{1-\rho_2}{1-\rho_1^2} & \frac{\rho_1^2-\rho_2}{\rho_1^2-1} \end{pmatrix}' = \begin{pmatrix} \frac{2d}{2-d} & \frac{d}{2-d} \end{pmatrix}', \\
\widehat{\boldsymbol{\beta}}(3) \xrightarrow{p} \Phi(2)^{-1}\boldsymbol{\rho}(3) &= \begin{pmatrix} \frac{3d}{3-d} & \frac{3d(1-d)}{(3-d)(2-d)} & \frac{d}{3-d} \end{pmatrix}', \\
\widehat{\boldsymbol{\beta}}(4) \xrightarrow{p} \Phi(3)^{-1}\boldsymbol{\rho}(4) &= \begin{pmatrix} \frac{4d}{4-d} & \frac{6d(1-d)}{(4-d)(3-d)} & \frac{4d(1-d)}{(4-d)(3-d)} & \frac{d}{4-d} \end{pmatrix}'.
\end{aligned}$$

Thus,

$$\widehat{\beta}_{n,n} \xrightarrow{p} \frac{d}{n-d} = \alpha_n(d).$$

Hence, the n^{th} order sample partial autocorrelation can be obtained from the estimated coefficient of y_{t-n} in the regression of y_t on $y_{t-1}, y_{t-2}, ..., y_{t-n}$.

Our estimator of d is defined to be

$$\widehat{d} = \underset{d\in(-.5,.25)}{Argmin}\, S(d).$$

where

$$S(d) = [\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}(d)]' W [\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}(d)].$$

$\boldsymbol{\alpha}(d)$ is a n by 1 vector with the j^{th} element $\dfrac{d}{j-d}$.
$\widehat{\boldsymbol{\alpha}}$ is a n by 1 vector with the j^{th} element $\widehat{\beta}_{j,j}$.
W is a symmetric, positive-definite weighting matrix.
Let

$$D = \frac{\partial \boldsymbol{\alpha}(d)}{\partial d} = \begin{pmatrix} \frac{1}{(1-d)^2} & \frac{2}{(2-d)^2} & \cdots & \frac{n}{(n-d)^2} \end{pmatrix}'.$$

Note that

$$\frac{\partial S(d)}{\partial d} = -2D'W\left[\widehat{\alpha} - \alpha(d)\right],$$

$$\frac{\partial^2 S(d)}{\partial d^2} = 2D'W^{-1}D + o_p(1),$$

$$\sqrt{T}\left[\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}(d)\right] \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = \lim_{T\to\infty} T \begin{pmatrix} Var\left(\widehat{\beta}_{1,1}\right) & Cov\left(\widehat{\beta}_{1,1}, \widehat{\beta}_{2,2}\right) & \cdots & Cov\left(\widehat{\beta}_{1,1}, \widehat{\beta}_{n,n}\right) \\ Cov\left(\widehat{\beta}_{2,2}, \widehat{\beta}_{1,1}\right) & Var\left(\widehat{\beta}_{2,2}\right) & \cdots & Cov\left(\widehat{\beta}_{2,2}, \widehat{\beta}_{n,n}\right) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\left(\widehat{\beta}_{n,n}, \widehat{\beta}_{1,1}\right) & Cov\left(\widehat{\beta}_{n,n}, \widehat{\beta}_{2,2}\right) & \cdots & Var\left(\widehat{\beta}_{n,n}\right) \end{pmatrix}.$$

Note also that

$$\frac{\partial S(d)}{\partial \widehat{d}} = \frac{\partial S(d)}{\partial d} + \frac{\partial^2 S(d)}{\partial d_*^2}\left(\widehat{d} - d\right) = 0,$$

$$\widehat{d} - d = -\left[\frac{\partial^2 S(d)}{\partial d_*^2}\right]^{-1} \frac{\partial S(d)}{\partial d},$$

$$\sqrt{T}\frac{\partial S(d)}{\partial d} \xrightarrow{d} N\left(0, 4D'W\Omega WD\right),$$

$$\sqrt{T}\left(\widehat{d} - d\right) \xrightarrow{d} N\left(0, \left[D'WD\right]^{-1} D'W\Omega WD\left[D'WD\right]^{-1}\right).$$

Thus, the optimal weighting matrix is

$$W = \Omega^{-1,}$$

and

$$\sqrt{T}\left(\widehat{d} - d\right) \xrightarrow{d} N\left(0, \left[D'\Omega^{-1}D\right]^{-1}\right).$$

Note that the $(l,m)^{th}$ element of the variance-covariance matrix Ω is given by

$$\Omega_{l,m} = \lim_{T\to\infty} T\left[L(l)\, E\left(\widehat{\boldsymbol{\beta}}(l) - \boldsymbol{\beta}(l)\right)\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}(m)\right)' L(m)'\right],$$

where

$$L(i) = \underbrace{(0\ 0\ ...\ 0\ 1)}_{i \text{ terms}}$$

The remaining conundrum is to find $E\left(\widehat{\boldsymbol{\beta}}(l) - \boldsymbol{\beta}(l)\right)\left(\widehat{\boldsymbol{\beta}}(m) - \boldsymbol{\beta}(m)\right)'$. Since

$$\begin{aligned}
\widehat{\boldsymbol{\rho}}(n) - \boldsymbol{\rho}(n) &= \widehat{\Phi}(n-1)\,\widehat{\boldsymbol{\beta}}(n) - \Phi(n-1)\,\boldsymbol{\beta}(n) \\
&= \Phi(n-1)\left(\widehat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}(n)\right) + \left(\widehat{\Phi}(n-1) - \Phi(n-1)\right)\boldsymbol{\beta}(n) \\
&\quad + \left(\widehat{\Phi}(n-1) - \Phi(n-1)\right)\left(\widehat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}(n)\right) \\
&= \Phi(n-1)\left(\widehat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}(n)\right) + \left(\widehat{\Phi}(n-1) - \Phi(n-1)\right)\boldsymbol{\beta}(n) + O_p\left(T^{-1}\right),
\end{aligned}$$

we have

$$\widehat{\boldsymbol{\beta}}(n) - \boldsymbol{\beta}(n) = \Phi(n-1)^{-1}\,\Delta(n),$$

where

$$\Delta(n) = \left(\widehat{\boldsymbol{\rho}}(n) - \boldsymbol{\rho}(n)\right) - \left(\widehat{\Phi}(n-1) - \Phi(n-1)\right)\boldsymbol{\beta}(n) + O_p\left(T^{-1}\right).$$

Hence, $\Omega_{l,m}$ is reduced to

$$\lim_{T\to\infty} T\left[L(l)\,\Phi(l-1)^{-1}\, E\left(\Delta(l)\,\Delta(m)'\right)\Phi(m-1)^{-1} L(m)'\right].$$

Note that

$$
\begin{aligned}
& \lim_{T\to\infty} TE\left(\Delta(l)\,\Delta(m)'\right) \\
= \; & C(l,m) - \lim_{T\to\infty} E\left(\widehat{\Phi}(l-1) - \Phi(l-1)\right)\boldsymbol{\beta}(l)\left(\widehat{\boldsymbol{\rho}}(m) - \boldsymbol{\rho}(m)\right)' \\
& - \lim_{T\to\infty} E\left(\widehat{\boldsymbol{\rho}}(l) - \boldsymbol{\rho}(l)\right)\boldsymbol{\beta}(m)'\left(\widehat{\Phi}(m-1) - \Phi(m-1)\right)' \\
& + \lim_{T\to\infty} E\left(\widehat{\Phi}(l-1) - \Phi(l-1)\right)\boldsymbol{\beta}(l)\,\boldsymbol{\beta}(m)'\left(\widehat{\Phi}(m-1) - \Phi(m-1)\right),
\end{aligned}
$$

where $C(l,m)$ is a l by m matrix with the $(i,j)^{th}$ element $c_{i,j}$.

Lastly, the $(i,j)^{th}$ element of the matrix $\lim_{T\to\infty} TE\left(\Delta(l)\,\Delta(m)'\right)$ is given by

$$
\begin{aligned}
& \lim_{T\to\infty} TE\left(\Delta(l)\,\Delta(m)'\right)_{i,j} \\
& = c_{i,j} - \sum_{h=1,h\neq i}^{l} \beta_{l,h} c_{|i-h|,j} - \sum_{k=1,k\neq j}^{m} \beta_{m,k} c_{|j-k|,i} \\
& + \sum_{h=1,h\neq i}^{l}\sum_{k=1,k\neq j}^{m} E\left(\widehat{\Phi}(l-1) - \Phi(l-1)\right)_{i,h}\left(\widehat{\Phi}(m-1) - \Phi(m-1)\right)_{k,j}\beta_{l,h}\beta_{m,k} \\
& = c_{i,j} - \sum_{h=1,h\neq i}^{l} \beta_{l,h} c_{|i-h|,j} - \sum_{k=1,k\neq j}^{m} \beta_{m,k} c_{|j-k|,i} + \sum_{h=1,h\neq i}^{l}\sum_{k=1,k\neq j}^{m} c_{|i-h|,|k-j|}\beta_{l,h}\beta_{m,k}
\end{aligned}
$$

Thus, all the elements of the variance-covariance matrix Ω are uncovered. Hence, the matrix $W\,(=\Omega^{-1})$ can be constructed, and we can now evaluate $S(d)$ at various values of d.

Comparison of Asymptotic variance of $\sqrt{T}\left(\widehat{d}-d\right)$

We now compare the efficiency of our estimator to the TSB's estimator. In all the tables below, $c_{i,j}$ is approximated by

$$
c_{i,j} = \sum_{s=1}^{T}\left(\rho_{s+i} + \rho_{s-i} - 2\rho_s\rho_i\right)\left(\rho_{s+j} + \rho_{s-j} - 2\rho_s\rho_j\right)
$$

where $T = 5\times 10^6$.

Tables 1 to 5 below are the counterpart of tables 1 to 5 in TSB's paper respectively. We compare the performance of the asymptotic variance of $\sqrt{T}\left(\widehat{d}-d\right)$ of our estimator using partial autocorrelation with that using pure autocorrelation in TSB's paper.

Table 1: Asymptotic variance of $\sqrt{T}\left(\widehat{d}-d\right)$, using partial autocorrelation $1, ..., n$.

d	$n=1$	$n=2$	$n=3$	$n=5$	$n=10$	$n=20$
$-.49$	3.4012	2.0663	1.6214	1.2305	0.9776	0.8226
$-.45$	3.1351	1.9356	1.5338	1.2063	0.9481	0.8058
$-.4$	2.8205	1.7789	1.4277	1.1399	0.9115	0.7848
$-.3$	2.2506	1.4873	1.2273	1.0122	0.8395	0.7426
$-.2$	1.7578	1.2255	1.0431	0.8917	0.7697	0.7013
$-.1$	1.3405	0.9948	0.8772	0.7803	0.7034	0.6614
0	1.0000	0.8000	0.7347	0.6832	0.6453	0.6265
.1	0.7492	0.6590	0.6344	0.6187	0.6108	0.6089
.2	0.7181	0.7107	0.7105	0.7060	0.6908	0.6709
.24	1.0765	1.0535	1.0103	0.9373	0.8372	0.7582

For $n = 1$, the partial correlation function and the autocorrelation are identical. Thus the values in this column should not be much different from that of TSB's table 1.

The asymptotic variance of $\sqrt{T}\left(\widehat{d}-d\right)$ generated by our estimator does differ from that of TSB's but not in a uniform fashion. Table 1 makes it clear that for negative values of d, the asymptotic variance using partial autocorrelation is smaller than that using autocorrelation. However, for positive d, partial autocorrelation yields a larger asymptotic variance.

Table 2: Asymptotic variance of $\sqrt{T}\left(\widehat{d}-d\right)$, using n^{th} partial autocorrelation only

d	$n=1$	$n=2$	$n=3$	$n=5$	$n=10$
$-.49$	3.4012	7.9649	14.562	33.775	116.83
$-.45$	3.1351	7.5739	14.039	32.986	115.37
$-.4$	2.8205	7.1015	13.402	32.015	113.56
$-.3$	2.2506	6.2111	12.181	30.127	110.00
$-.2$	1.7578	5.3939	11.034	28.315	106.52
$-.1$	1.3405	4.6530	9.9660	26.593	103.16
0	1.0000	4.0000	9.0000	25.000	100.00
.1	0.7492	3.4821	8.2162	23.685	97.359
.2	0.7181	3.4821	8.3006	23.903	97.897
.24	1.0765	4.5826	10.111	27.129	104.44

The asymptotic variance can just be read off Table 2. Roughly speaking, the asymptotic variance is about n^2 for any given value of d. For $d < 0$, a switch from TSB's estimator to our estimator reduces the asymptotic variance by as large as a factor of 350. Also, the asymptotic variance grows monotonically and stably with n, a feature which the TSB's estimator lacks. Thus, the statistical inference based on partial autocorrelation function estimator should be more reliable.

Table 3: Asymptotic variance of $\sqrt{T}\left(\widehat{d}-d\right)$, using five partial autocorrelations

$d\backslash n$	$1-5$	$2-6$	$3-7$	$5-9$	$10-14$
$-.49$	1.2305	2.6125	4.4028	9.2274	28.378
$-.45$	1.2063	2.5530	4.3376	9.1503	28.270
$-.4$	1.1399	2.4806	4.2583	9.0569	28.139
$-.3$	1.0122	2.3425	4.1088	8.8829	27.897
$-.2$	0.8917	2.2159	3.9756	8.7328	27.694
$-.1$	0.7803	2.1075	3.8693	8.6247	27.564
0	0.6832	2.0350	3.8198	8.6091	27.603
.1	0.6187	2.0669	3.9412	8.8813	28.192
.2	0.7060	2.7407	5.1597	11.044	32.426
.24	0.9373	4.2363	7.8733	15.785	41.386

Notice, too, that for $n \geq 2$, the increases in d stimulate the decreases in the asymptotic variance, up to the point where $d = 0$. For positive d, partial autocorrelation yields a larger asymptotic variance.

Table 4: Asymptotic variance of $\sqrt{T}\left(\widehat{d} - d\right)$, using ten partial autocorrelations

$d\backslash n$	$1-10$	$2-11$	$3-12$	$5-14$	$10-19$
$-.49$	0.9776	1.9036	3.0681	6.0495	17.134
$-.45$	0.9481	1.8827	3.0563	6.0566	17.190
$-.4$	0.9115	1.8581	3.0440	6.0696	17.266
$-.3$	0.8395	1.8146	3.0294	6.1128	17.448
$-.2$	0.7697	1.7818	3.0337	6.1880	17.685
$-.1$	0.7034	1.7674	3.0704	6.3177	18.016
0	0.6453	1.7920	3.1748	6.5623	18.554
.1	0.6108	1.9309	3.1436	6.3321	19.733
.2	0.6908	2.7321	5.0047	9.9176	25.045
.24	0.8372	4.1292	7.8532	15.264	35.234

Misspecification

Thus far, we have assumed a fractionally integrated white noise process. We would like to examine how robust is our estimator to serial correlation. We will discuss the asymptotic bias in our estimator of d obtained from the $(0, d, 0)$ model caused by ignoring short-run dynamics. For comparison, we will consider estimators of d based on a single partial autocorrelation α_k. d^* is the value of d that generates the same value of α_k for $ARFIMA\,(0, d, 0)$ model.

We consider the case where the true model is an $ARFIMA\,(1, d, 0)$ process of the form

$$(1 - L)^d\,(1 - \phi L)\,y_t = \varepsilon_t$$

as well as the case where the true model is an $ARFIMA\,(0, d, 1)$ process of the form

$$(1 - L)^d\,y_t = (1 + \theta L)\,\varepsilon_t$$

Table 5 gives values of d^* for $d = 0.1$, 0.2, and 0.24, for the $(1, d, 0)$ and $(0, d, 1)$ models. For the $(1, d, 0)$ model we consider $\phi = 0.4$ and 0.8, whereas for the $(0, d, 1)$ model we consider $\theta = 0.4$ and 0.8.

Table 5: Asymptotic bias $|d^* - d|$ for the MDE from the $(0, d, 0)$ model

True model is ARFIMA $(1, d, 0)$ model							
d	ϕ	Lag 1	Lag 2	Lag 5	Lag 10	Lag 20	Lag 50
.1	0.4	0.238	0.068	0.024	0.013	0.009	0.008
.2	0.4	0.185	0.145	0.052	0.033	0.027	0.034
.24	0.4	0.162	0.180	0.065	0.044	0.039	0.052
.1	0.8	0.364	0.186	0.082	0.052	0.032	0.019
.2	0.8	0.278	0.412	0.168	0.108	0.073	0.056
.24	0.8	0.242	0.517	0.203	0.133	0.094	0.078
True model is ARFIMA $(0, d, 1)$ model							
d	θ	Lag 1	Lag 2	Lag 5	Lag 10	Lag 20	Lag 50
.1	0.4	0.208	0.343	0.035	0.006	0.003	0.001
.2	0.4	0.158	0.374	0.027	0.012	0.006	0.002
.24	0.4	0.137	0.387	0.024	0.014	0.007	0.003
.1	0.8	0.263	0.915	0.541	0.415	0.089	0.002
.2	0.8	0.197	0.922	0.521	0.426	0.094	0.004
.24	0.8	0.171	0.925	0.512	0.430	0.096	0.004

If the true model is an $ARFIMA\,(1, d, 0)$ model, our estimator yields a smaller bias for the number of lags more than or equal 5. If the true model is $ARFIMA\,(0, d, 1)$ model, our estimator has a larger bias for $\theta = 0.8$ and for the number of lags less than or equal 20. Generally speaking, the TSB's estimator performs better if the true model is $ARFIMA\,(0, d, 1)$ while our estimator is better if the true model is $ARFIMA\,(1, d, 0)$. Therefore our estimator is more robust to misspecification in the autoregressive component whereas the TSB's estimator is more robust to misspecification in the moving average component.

Reference:

1. Brockwell, P.J. and R.A. Davis (1991), *Time series: Theory and methods*, 2nd ed., Springer-Verlag.

2. Granger, C.W.J and R. Joyeus (1980)"An Introduction to the Long-Memory Time Series Models and Fractional Differencing." *Journal of Time Series Analysis,* 1, 15-29.

3. Tieslau, M.A., P. Schmidt and R.T. Baillie (1996)"A Minimum Distance Estimator for Long-Memory Processes," *Journal of Econometrics,* 71, 249-64.

ECO5120: Econometric Theory and Application, Fall 99
Prof. T.L. Chong
HANDOUT 9

ECO5120: Econometric Theory and Application, Fall 99
Prof. T.L. Chong
HANDOUT 10
MEASUREMENT ERRORS AND MISSCLASSIFICATION

Most studies in economics implicitly assume that the variables of interest are perfectly observed and measured. Little effort, however, is devoted to the issue of measurement errors. Many economic variables may be of imperfect quality, either because the true variables of interest are simply not observable(e.g. ability), or the observable data are suffered from a wide variety of errors including those resulting from poor sampling techniques. The consequences of errors in variables have been well established. It has been shown that measurement error in the dependent variable will not affect the consistency of the OLS estimator. However, measurement error in the independent variable usually introduces complications into the analysis. If the explanatory variables are suffered from errors, the parameters cannot be consistently estimated and will be biased toward zero. Various methods, such as finding a bound for the estimate and using instrumental variables have been suggested to fix the problem.

The problem of measurement error (or misclassification) occurs due to two major reasons. The first is that the person who conducts the survey commits a systematic mistake. The second reason is that the respondents of the survey have an incentive to lie. In order to make the model interpretable, we should know what kind of information we can still get in the presence of measurement errors.

The earliest studies on the consequences of measurement errors were done by Adcock (1877, 1878) and Madansky (1959), who considered the problem of fitting a straight line when both variables are subject to errors. Levi (1973) showed that in a simple linear regression model without intercept, if the explanatory variable is suffered from errors, the structural parameter will

be biased toward zero. Nelson (1995) obtained a similar result for the case where more than one independent variable are measured with errors. The phenomenon that the impact of the regressor on the dependent variable is diluted by measurement errors is called attenuation bias. The attenuation bias is usually linear, in the sense that the probability limit of the estimator is the true parameter multiplied by a positive constant less than 1. Chong and Lui (1998) show that the attenuation bias is nonlinear when measurement errors exist in a fractionally integrated model. Other studies on measurement errors include Stefanski (2000), Lee and Sepanski (1995), Nowak (1992), Hausman, Newey, Powell and Ichimura (1991), Whittemore and Keller (1988), Schafer (1986), Hausman (1977, 1995, 2001), Hausman and Griliches (1986).

Inconsistency of the OLS Estimator

As a simple exposition, suppose the true model is

$$y_t = \beta x_t^* + u_t \qquad t = 1, 2, ..., T.$$

Because the true values of x_t^* is not observable, it is proxied by an observable x_t where

$$x_t = x_t^* + \varepsilon_t.$$

If $\{u_t\}_{t=1}^{\infty}$ and $\{\varepsilon_t\}_{t=1}^{\infty}$ are i.i.d. zero mean finite variance random variables, and if ε_t and u_t are independent, the least squares estimator of β based on the observable $\{x_t, y_t\}_{t=1}^{T}$ is given by

$$\widehat{\beta} = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} x_t^2} \xrightarrow{p} \beta \frac{\sigma_*^2}{\sigma_*^2 + \sigma_\varepsilon^2},$$

where $\sigma_*^2 = E\left(x_t^{*2}\right)$, $\sigma_\varepsilon^2 = E\left(\varepsilon_t^2\right)$. Thus the OLS estimate will be biased towards zero. The estimator for the variance of u_t is inconsistent too since

$$\widehat{\sigma}_u^2 = \frac{\sum_{t=1}^{T} \widehat{u}_t^2}{T-1} \xrightarrow{p} \sigma_u^2 + \frac{\beta^2 \sigma_\varepsilon^2 \sigma_*^2}{\sigma_*^2 + \sigma_\varepsilon^2}.$$

Thus, the true parameters cannot be consistently estimated, unless there is no measurement error ($\sigma_\varepsilon^2 = 0$), or unless we have additional information on the "signal to noise ratio" $\frac{\sigma_*^2}{\sigma_\varepsilon^2}$.

Although measurement errors usually result in inconsistent estimates, there are exceptional cases where the parameters can still be consistently estimated. We will study two of these cases. The first one is the consistency of the break-point estimator in a structural-break model. The second is the unit-root model.

Missclassification of the dependent variable

Here study a regression model with the regressor being a dummy variable measured with errors.

Suppose the true model is

$$y_t = \alpha\left(1 - x_t^*\right) + \gamma x_t^* + u_t,$$

$t = 1, 2, ..., T$.

x_t^* is a zero-one dummy variable, $(1 - x_t^*)$ is the dummy variable of another category.

(α, γ) are true structural parameters.

Model (1) is more easily explained and understood in an alternative representation:

$$y_t = \alpha + \beta x_t^* + u_t,$$

where

$$\beta = \gamma - \alpha.$$

Note that β denotes the difference between the coefficients of the two groups.

Now, suppose the true value of x_t^* is not perfectly measured and is approximated by an observable x_t where

$$x_t = x_t^* + \varepsilon_t$$

and ε_t is the measurement error.

We regress y_t on x_t with an intercept. In the conventional case where x_t^* and ε_t are continuous variables, they are usually assumed to be independent of each other. However, an interesting feature of our model when x_t^* is a dummy variable is that x_t^* and ε_t will not be independent anymore.

We define

$$p = \Pr\left(x_t = 0 | x_t^* = 1\right),$$

and

$$q = \Pr\left(x_t = 1 | x_t^* = 0\right).$$

We assume the following:

$(A1)$ $u_t \sim i.i.d.\,(0, \sigma_u^2)$, $\sigma_u^2 < \infty$.

$(A2)$ $x_t^* \sim i.i.d.$ which takes 1 with probability a and 0 with $1-a$, where $0 \leq a \leq 1$.

$(A3)$ ε_t is dependent on x_t^* and has following distribution:

$$\begin{aligned} \text{If } x_t^* &= 1, \text{ then } \varepsilon_t = -1 \text{ with probability } p \text{ and } \varepsilon_t = 0 \text{ otherwise,} \\ \text{if } x_t^* &= 0, \text{ then } \varepsilon_t = 1 \text{ with probability } q \text{ and } \varepsilon_t = 0 \text{ otherwise.} \end{aligned}$$

$(A4)$ x_t^* and ε_t are independent of u_t.

Assumptions $(A1) - (A4)$ describe the nature of the dummy variable, disturbance term and measurement errors.

Lemma 169 *Under assumptions* $(A1)$ *to* $(A4)$*, we have*

$$E(x_t^*) = a,$$

$$Var\left(x_t^*\right) = a\left(1-a\right),$$

$$E(\varepsilon_t) = -ap + (1-a)q,$$

$$Var\left(\varepsilon_t\right) = 2ap - \left(ap - q\left(1-a\right)\right)\left(a\left(p+q\right) + 1 - q\right),$$

$$Cov\left(x_t^*, \varepsilon_t\right) = -a\left(p+q\right)\left(1-a\right),$$

$$E(x_t) = a\left(1-p\right) + (1-a)q,$$

$$E\left(x_t\varepsilon_t\right) = -ap + ap + (1-a)q = (1-a)q,$$

$$Var\left(x_t\right) = \left(ap + \left(1-a\right)\left(1-q\right)\right)\left(a\left(1-p\right) + \left(1-a\right)q\right).$$

Proof. Exercise.

It is worth noting that, unlike the case of continuous variable, the measurement error and the latent variable are not independent but negatively correlated in the dummy variable case.

Asymptotic Behavior of the Least-Squares Estimators

As already mentioned, we cannot treat the problem like the conventional case of measurement error without an intercept. It is because if we allow only one dummy regressor, we must include an intercept for the model to be identifiable. Also, the latent dummy variable and the measurement error are not independent in this case as shown in Lemma 1. The OLS estimators are defined as

$$\widehat{\beta} = \frac{\sum_{t=1}^{T}\left(x_t - \overline{x}\right)y_t}{\sum_{t=1}^{T}\left(x_t - \overline{x}\right)x_t},$$

$$\widehat{\alpha} = \frac{1}{T}\left(\sum_{t=1}^{T} y_t - \widehat{\beta}\sum_{t=1}^{T} x_t\right)$$

and

$$\widehat{\gamma} = \widehat{\alpha} + \widehat{\beta}.$$

The followings are the asymptotic properties of the OLS estimators. The proofs are provided in the Appendix.

$$\widehat{\alpha} \xrightarrow{p} \alpha\left(1 - \Phi_1\right) + \gamma\Phi_1,$$

$$\widehat{\gamma} \xrightarrow{p} \alpha\left(1 - \Phi_2\right) + \gamma\Phi_2,$$

where

$$\begin{aligned}\Phi_0 &= \frac{Cov\left(x_t, x_t^*\right)}{Var\left(x_t\right)} = \frac{Var\left(x_t^*\right) + Cov\left(\varepsilon_t, x_t^*\right)}{Var\left(x_t\right)} \\ &= \frac{a\left(1-a\right)\left(1-p-q\right)}{\left(ap + \left(1-a\right)\left(1-q\right)\right)\left(a\left(1-p\right) + \left(1-a\right)q\right)},\end{aligned}$$

$$0 \leq \Phi_1 = \frac{ap}{ap + \left(1-a\right)\left(1-q\right)} \leq 1,$$

$$0 \leq \Phi_2 = \Phi_0 + \Phi_1 = \frac{a\left(1-p\right)}{a\left(1-p\right) + \left(1-a\right)q} \leq 1.$$

Thus, $\widehat{\alpha}$ will be consistent if $\Phi_1 = 0$, i.e., $a = 0$ or $p = 0$. For $a = 0$, x_t^* always equals 0. For $p = 0$, there is no measurement error in the case of $x_t^* = 1$.

$\widehat{\gamma}$ will be consistent if $\Phi_2 = 1$, i.e., $a = 1$ or $q = 0$. For $a = 1$, x_t^* always equals 1. For $q = 0$, there is no measurement error in the case of $x_t^* = 0$.

Note that Φ_0 can be negative if $p + q > 1$, but will be less than 1 in absolute value, whereas Φ_1 and Φ_2 are values between zero and one.

Theorem 170 *If assumptions* $(A1) - (A4)$ *hold, then as* $T \to \infty$*, we have:*

$$\widehat{\alpha} \xrightarrow{p} \alpha\left(1-\Phi_1\right)+\gamma\Phi_1 \tag{1}$$

and

$$\widehat{\gamma} \xrightarrow{p} \alpha\left(1-\Phi_2\right)+\gamma\Phi_2. \tag{2}$$

Proof. Exercise.

The Theorem states that the structural estimators converge in probability to some convex combinations of the coefficients of the two dummy variables. In general, $\widehat{\alpha}$ will be consistent if $\Phi_1 = 0$. $\widehat{\gamma}$ will be consistent if $q = 0$.

To understand the implication of this result, suppose $x_t^* = 1$ if the respondent is a man, and $x_t^* = 0$ if the respondent is a woman. The Theorem implies that the coefficient for the group of women will be identified either if all the respondents are women, or if the group of men has no missclassification. Similarly, the coefficient for the group of men will be identified either if all the respondents are men, or if the group of women has no missclassification.

Inspection of the Theorem shows that without additional information about the measurement errors, it is not possible to recover the true pre- and post-shift parameters. There are several ways to extract the information of the true parameters under measurement errors. The most common one is to use instrumental variables, which is widely recognized as an important method for the analysis of linear measurement error model.

In our case, when Φ_1 and Φ_2 are known, in other words, when we know a, p and q, then we can identify the true pre- and post-shift parameters. The consistent estimators for the structural parameters are given in the following Theorem.

If p, q and a are known, then we have

$$\widetilde{\alpha} = \frac{\Phi_2\widehat{\alpha} - \Phi_1\widehat{\gamma}}{\Phi_0} \xrightarrow{p} \alpha$$

and

$$\widetilde{\gamma} = \frac{\widehat{\gamma}\left(1-\Phi_1\right) - \widehat{\alpha}\left(1-\Phi_2\right)}{\Phi_0} \xrightarrow{p} \gamma.$$

Note that the validity of the reparameterizations depends on the assumption that the value of Φ_0 is non-zero. In the cases where $\Phi_0 = 0$, i.e., $a = 0$, $a = 1$, or $p + q = 1$, we can never recover the true coefficients.

Some Special Cases

We have derived the general result in the previous section. Now, let us look at some interesting cases.

Case 1: Observationally Equivalent Groups $(p + q = 1)$

When $p = 1 - q$, the two different groups are measured with errors in a way that the statistical properties of the two observed groups are identical. For example, a group of men with 30 percent reported as women will be observationally equivalent to a group of women with 70 percent reported as men. In such a case, we have $\Phi_0 = 0$ and $\Phi_1 = \Phi_2 = a$.

Corollary 171 *If assumptions* $(A1) - (A4)$ *hold and* $p + q = 1$*, then as* $T \to \infty$*, we have:*

$$\widehat{\alpha} \xrightarrow{p} \alpha (1 - a) + \gamma a \tag{3}$$

and

$$\widehat{\gamma} \xrightarrow{p} \alpha (1 - a) + \gamma a. \tag{4}$$

The Corollary states that the OLS estimators converge to the same convex combination of the true parameters of the two groups.

In this case, even if the coefficients of the two groups are different, we cannot observe this due to the similarity of the statistical properties of the two observed groups. Further, since $\Phi_0 = 0$, $\widetilde{\alpha}$ and $\widetilde{\gamma}$ will all be undefined, the true parameters will *never* be identified.

Case 2: Same Coefficient for the Two Groups $(\alpha = \gamma)$

When the two groups share the same coefficient, we have $\alpha = \gamma$. This implies that $\beta = 0$.

Corollary 2: If assumptions $(A1) - (A4)$ hold, and if $\alpha = \gamma$, then as $T \to \infty$, we have:

$$\widehat{\alpha} \xrightarrow{p} \alpha \tag{5}$$

and

$$\widehat{\gamma} \xrightarrow{p} \gamma. \tag{6}$$

Thus, all the estimators will be consistent and converge to the true parameters, despite the fact that there are measurement errors.

Therefore, if the coefficients for the two groups are the same, measurement errors will have no effect on estimation at all.

Case 3: The Existence of One Group only ($a = 0$ *or* $a = 1$)

This happens if the survey method is not random enough so that only one group of people is being surveyed. However, even if there is only one category, we will still observe two categories due to measurement errors. We study the case where $a = 0$. The case where $a = 1$ will have an opposite interpretation and is therefore skipped. When $a = 0$, only the group defined to be zero exists, we have

$$\Phi_0 = \Phi_1 = \Phi_2 = 0.$$

Corollary 3: If assumptions $(A1)-(A4)$ hold and $a = 0$, then as $T \to \infty$, we have:

$$\widehat{\alpha} \xrightarrow{p} \alpha \tag{7}$$

and

$$\widehat{\gamma} \xrightarrow{p} \alpha. \tag{8}$$

In general, the probability limits of the estimators for both groups in the existence of only one group will be the true parameters of the existing group. The true coefficients of the non-existing group can never be identified even if we know the values of p and q.

Case 4: Measurement Error in One Group only ($p = 0$ *or* $q = 0$)

This happens when the members of a certain group have no incentive to misreport themselves, but some members in another group misreport themselves as the members of the opposite group. For example, suppose in a survey an individual is asked if he/she is homosexual, those who are not will most likely reveal the truth, but there are strong reasons to believe that some of the homosexuals may not tell the truth due to social pressure. Another example is that in a court trial of a criminal offense with death penalty. Excluding very special cases, those who did not commit the crime are unlikely to confess. However, some criminals may not confess even if they did commit the crime.

Now, we consider the case where $p = 0$, i.e., when the true dummy is 1, we measure it perfectly. This implies that $\Phi_1 = 0$ and

$$\Phi_0 = \Phi_2 = \frac{a}{a + (1 - a)\, q}.$$

Note that Φ_0 is a value between zero and one.

Corollary 4: If assumptions $(A1)-(A4)$ hold and $p = 0$, then as $T \to \infty$, we have:

$$\widehat{\alpha} \xrightarrow{p} \alpha \tag{9}$$

and

$$\widehat{\gamma} \xrightarrow{p} \alpha\,(1 - \Phi_2) + \gamma\Phi_2. \tag{10}$$

When a and q are known, we can even identify γ. The consistent estimator is

$$\widetilde{\gamma} = \frac{\widehat{\gamma} - \widehat{\alpha}\,(1 - \Phi_2)}{\Phi_2} \xrightarrow{p} \gamma. \tag{11}$$

Corollary 4 states that the structural parameters for the group of $(1 - x_t^*)$ can be identified despite the presence of measurement errors. However, the structural estimator for the group of x_t^* is biased towards a convex combination of the coefficients of the two dummy variables. Further, if we have information about q and a, then all the parameters can be identified. The case for $q = 0$ has an opposite interpretation and is therefore skipped.

Case 5: Perfect Measurement Error in One Group ($p = 1$ *or* $q = 1$)

When $p = 1$, i.e., when the true dummy is 1, we always measure it incorrectly. This happens when all the members in one group lie to the survey conductor. Then, we have $\Phi_2 = 0$ and

$$\Phi_0 = -\frac{a}{a + (1 - a)(1 - q)},$$

$$\Phi_1 = -\Phi_0.$$

Note that Φ_0 is negative and is less than 1 in absolute value.

Corollary 5: If assumptions $(A1)-(A4)$ hold and $p = 1$, then as $T \to \infty$, we have:

$$\widehat{\alpha} \xrightarrow{p} \alpha(1 - \Phi_1) + \gamma\Phi_1, \tag{12}$$

and

$$\widehat{\gamma} \xrightarrow{p} \alpha. \tag{13}$$

If we have the information about p, q and a, we can construct the consistent estimators as

$$\widetilde{\alpha} = \widehat{\gamma} \xrightarrow{p} \alpha \tag{14}$$

and

$$\widetilde{\gamma} = \frac{\widehat{\alpha} - \widehat{\gamma}(1 - \Phi_1)}{\Phi_1} \xrightarrow{p} \gamma. \tag{15}$$

Corollary 5 states that the structural estimators are inconsistent. The estimator for the group with perfect measurement error will converge in probability to the true coefficient of another group. The probability limit of the estimator of another group will be a convex combination of the true coefficients of the two groups. Thus in a large sample, the group with a higher value of the true coefficient will turn out to have a smaller value of estimated coefficient, and vice versa.

If we have information about p, q and a, then all the parameters can be retrieved. The case for $q = 1$ has an opposite interpretation and is therefore skipped.

Case 6: Perfect Measurement Error $(p = q = 1)$

When $p = q = 1$, we always measure the dummy variable incorrectly. This may be an imaginary scenario, but in a small survey with particular type of respondents, it may happen. Suppose there are only two types of people in a survey, one type is smart but humble, the other type is dull but arrogant. When they are asked if they think they are smart, those who are really smart will be humble enough to report that they are not that smart, but those who are dull will not consider themselves as dull and will report that they are smart.

In this case, we have $\Phi_0 = -1$, $\Phi_1 = 1$ and $\Phi_2 = 0$. Corollary 6 below states that the estimator for one group will converge to the coefficient of another group.

Corollary 6: If assumptions $(A1) - (A4)$ hold and $p = q = 1$, then as $T \to \infty$, we have:

$$\widehat{\alpha} \xrightarrow{p} \gamma \tag{16}$$

and

$$\widehat{\gamma} \xrightarrow{p} \alpha. \tag{17}$$

It follows that the effect of perfect measurement errors is to interchange the probability limits of the estimators. The estimator for one group will

converge to the true coefficient of another group. Consequently, we can define

$$\widetilde{\alpha} = \widehat{\gamma} \xrightarrow{p} \alpha, \tag{18}$$

$$\widetilde{\gamma} = \widehat{\alpha} \xrightarrow{p} \gamma. \tag{19}$$

Case 7: $p + q > 1$

When $p + q > 1$, the probability of committing measurement errors in at least one of the groups is higher than 0.5. In such a case, $\Phi_0 < 0$. Corollary 7 states that the relative importance of the marginal effect of the two categories will be misinterpreted if the measurement error is serious.

Corollary 7: If assumptions $(A1) - (A4)$ hold and $p+q > 1$, and assume that $\gamma > \alpha$ then as $T \to \infty$, we have:

$$p\lim \widehat{\gamma} < p\lim \widehat{\alpha}. \tag{20}$$

Proof. $p\lim \widehat{\gamma} - p\lim \widehat{\alpha} = \alpha(1 - \Phi_1) + \gamma\Phi_1 + \alpha(1 - \Phi_2) + \gamma\Phi_2 = (\gamma - \alpha)\Phi_0 < 0$.

Thus, even $\gamma > \alpha$, we will observe that $p\lim \widehat{\gamma} < p\lim \widehat{\alpha}$.

Monte Carlo Experiments

This experiment verifies the above Theorems and Corollaries. Consider the model

$$y_t = \alpha(1 - x_t^*) + \gamma x_t^* + u_t, \qquad (t = 1, 2, ..., T).$$

We perform the following experiment:

Let

$T = 50, 100, 1000, 10000.$

$u_t \sim nid(0, 1)$,

$x_t^* \sim i.i.d.\ Bernoulli((1,a),(0,1-a))$,
$x_t = x_t^* + \varepsilon_t$.

If $x_t^* = 1$, then $\varepsilon_t = -1$ with probability p and $\varepsilon_t = 0$ with probability $(1-p)$;
if $x_t^* = 0$, then $\varepsilon_t = 1$ with probability q and $\varepsilon_t = 0$ with probability $(1-q)$.

x_t^* and ε_t are independent of u_t. Φ_0, Φ_1 and Φ_2 are defined as in Section 3.

The simulations were programmed in GAUSS. For each value of a, p and q, we perform n replications, where $n = 1$ and 1000. Let $\overline{\widehat{\alpha}}_{(T,n)}$ and $\overline{\widehat{\gamma}}_{(T,n)}$ be the average values of the OLS estimators for a sample of size T in these n replications. Displayed in Table 1 are Monte Carlo simulation results for the seven special cases. The first column corresponds to case one, and so on.

Table 1: Estimation results for the 7 cases

$Case$	$p+q=1$	$\alpha=\gamma$	$a=0$	$p=0$	$p=1$	$p=q=1$	$p+q>1$
a,p,q	$.5,.3,.7$	$.5,.2,.2$	$0,.2,.2$	$.5,0,.4$	$.5,1,.3$	$.5,1,1$	$.5,.7,.7$
α,γ	$1,2$	$2,2$	$1,2$	$1,2$	$1,2$	$1,2$	$1,2$
$plim\widehat{\alpha},plim\widehat{\gamma}$	$1.5,1.5$	$2,2$	$1,1$	$1,1.714$	$1.588,1$	$2,1$	$1.7,1.3$
$\widehat{\widehat{\alpha}}_{(50,1)}$	1.475	2.000	.819	.799	1.506	2.255	1.594
$\widehat{\widehat{\alpha}}_{(50,1000)}$	1.492	2.011	.997	.990	1.594	2.007	1.709
$\widehat{\widehat{\alpha}}_{(100,1)}$	1.740	1.940	.854	.987	1.502	2.244	1.909
$\widehat{\widehat{\alpha}}_{(100,1000)}$	1.490	2.001	1.001	1.013	1.592	2.001	1.705
$\widehat{\widehat{\alpha}}_{(1000,1)}$	1.566	2.025	.986	1.091	1.612	2.020	1.626
$\widehat{\widehat{\alpha}}_{(1000,1000)}$	1.501	1.997	1.002	1.000	1.590	2.000	1.703
$\widehat{\widehat{\alpha}}_{(10000,1)}$	1.512	2.003	.999	1.008	1.583	1.997	1.698
$\widehat{\widehat{\alpha}}_{(10000,1000)}$	1.500	2.000	1.000	1.000	1.588	2.000	1.700
$\widehat{\widehat{\gamma}}_{(50,1)}$	1.420	2.437	.671	2.039	.888	.671	1.179
$\widehat{\widehat{\gamma}}_{(50,1000)}$	1.498	2.004	.989	1.718	.994	1.003	1.296
$\widehat{\widehat{\gamma}}_{(100,1)}$	1.410	1.952	1.042	1.620	.895	.963	1.461
$\widehat{\widehat{\gamma}}_{(100,1000)}$	1.501	1.999	.998	1.717	.992	1.009	1.304
$\widehat{\widehat{\gamma}}_{(1000,1)}$	1.529	2.007	.935	1.726	1.041	1.002	1.256
$\widehat{\widehat{\gamma}}_{(1000,1000)}$	1.498	2.000	1.001	1.717	1.001	1.001	1.300
$\widehat{\widehat{\gamma}}_{(10000,1)}$	1.498	1.992	1.009	1.716	.971	1.030	1.296
$\widehat{\widehat{\gamma}}_{(10000,1000)}$	1.500	2.000	1.000	1.715	.999	1.000	1.300

The simulated results in Table 1 largely conform to our theory that the OLS estimators converge to some convex combinations of the true parameters.

***

MEASUREMENT ERRORS IN STRUCTURAL-BREAK MODELS

It will be shown that in a structural-break model, the break point can still be consistently estimated in the presence of measurement errors.

The Consistency of $\widehat{\tau}$

Suppose the true model is

$$y_t = \beta_1 x_t^* 1\{t \leq k_0\} + \beta_2 x_t^* 1\{t > k_0\} + u_t \qquad t = 1, 2, ..., T.$$

where $1\{\cdot\}$ is an indicator function that equals 1 when the statement inside the bracket is true and equals 0 otherwise, β_1 and β_2 are true structural parameters for $0 < t \leq k_0$ and $k_0 < t \leq T$ respectively. Let $k = [\tau T]$, where $[\cdot]$ is the greatest integer function, $\tau \in [\underline{\tau}, \overline{\tau}]$ the break fraction. Now, suppose the true value of x_t^* is not perfectly measured and is approximated by an observable x_t where

$$x_t = x_t^* + \varepsilon_t$$

and ε_t is the measurement error. We shall assume for ease of exposition that:

$(A1)$ $\tau_0 \in [\underline{\tau}, \overline{\tau}] \subset (0, 1)$.

$(A2)$ $u_t \sim iid(0, \sigma_u^2)$, $\sigma_u^2 < \infty$.

$(A3)$ $\varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon^2 < \infty$, $E(\varepsilon_t^4) < \infty$.

$(A4)$ x_t^*, ε_t, and u_t are independent of one another.

$(A5)$ $E(x_t^{*4}) < \infty$.

$(A6)$

$$S_{**}(\tau) \overset{def}{=} \frac{1}{T}\sum_{t=1}^{[\tau T]} x_t^{*2} \xrightarrow{p} \tau\sigma_*^2$$

$$S_{\varepsilon\varepsilon}(\tau) \overset{def}{=} \frac{1}{T}\sum_{t=1}^{[\tau T]} \varepsilon_t^2 \xrightarrow{p} \tau\sigma_\varepsilon^2$$

uniformly for $\tau \in [\underline{\tau}, \overline{\tau}]$.

Assumptions $(A2)$ to $(A6)$ imply the followings:

$$\begin{aligned}
S_{*u}(\tau) &\overset{def}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^{[\tau T]} x_t^* u_t \quad \Rightarrow \quad B_{*u}(\tau) \\
S_{*\varepsilon}(\tau) &\overset{def}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^{[\tau T]} x_t^* \varepsilon_t \quad \Rightarrow \quad B_{*\varepsilon}(\tau) \\
S_{u\varepsilon}(\tau) &\overset{def}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^{[\tau T]} u_t \varepsilon_t \quad \Rightarrow \quad B_{u\varepsilon}(\tau) \\
\sqrt{T}\left(S_{**}(\tau) - \tau\sigma_*^2\right) &\Rightarrow \quad B_{**}(\tau) \\
\sqrt{T}\left(S_{\varepsilon\varepsilon}(\tau) - \tau\sigma_\varepsilon^2\right) &\Rightarrow \quad B_{\varepsilon\varepsilon}(\tau)
\end{aligned}$$

where " $\Rightarrow$ " denotes the weak convergence of a stochastic process, and $B_{*u}(\tau)$, $B_{*\varepsilon}(\tau)$, $B_{u\varepsilon}(\tau)$, $B_{**}(\tau)$, and $B_{\varepsilon\varepsilon}(\tau)$ are Gaussian processes with zero mean and variances $\tau\sigma_u^2\sigma_*^2$, $\tau\sigma_\varepsilon^2\sigma_*^2$, $\tau\sigma_u^2\sigma_\varepsilon^2$, $\tau\left(E\left(x_t^{*4}\right) - \sigma_*^4\right)$, $\tau\left(E\left(\varepsilon_t^4\right) - \sigma_\varepsilon^4\right)$ respectively.

For any given k, our estimated model is

$$\widehat{y}_t = \widehat{\beta}_{1\tau} x_t 1\{t \leq k\} + \widehat{\beta}_{2\tau} x_t 1\{t > k\}$$

The least-squares estimators of β_1 and β_2 based on the observable $\{x_t, y_t\}_{t=1}^T$ is given by:

$$\begin{aligned}
\widehat{\beta}_{1\tau} &= \frac{S_{xy}(\tau)}{S_{xx}(\tau)} \\
\widehat{\beta}_{2\tau} &= \frac{S_{xy}(1) - S_{xy}(\tau)}{S_{xx}(1) - S_{xx}(\tau)}
\end{aligned}$$

where

$$\begin{aligned}
S_{xx}(\tau) &\overset{def}{=} \frac{1}{T} \sum_{t=1}^{[\tau T]} x_t^2 \xrightarrow{p} \tau\left(\sigma_*^2 + \sigma_\varepsilon^2\right) \\
S_{xy}(\tau) &\overset{def}{=} \frac{1}{T} \sum_{t=1}^{[\tau T]} x_t y_t
\end{aligned}$$

We define the break-point estimator as:

$$\widehat{\tau} = Arg \min_{\tau \in (0,1)} \frac{1}{T} RSS_T(\tau)$$

where

$$RSS_T(\tau) = \sum_{t=1}^{[\tau T]} \left(y_t - \widehat{\beta}_{1\tau} x_t \right)^2 + \sum_{t=[\tau T]+1}^{T} \left(y_t - \widehat{\beta}_{2\tau} x_t \right)^2$$

For $\tau \in [\underline{\tau}, \overline{\tau}]$,

$$\begin{aligned} \widehat{\beta}_{1\tau} &\xrightarrow{p} [\beta_1 1\{\tau \leq \tau_0\} + \Gamma_{1\tau} 1\{\tau > \tau_0\}] A \\ \widehat{\beta}_{2\tau} &\xrightarrow{p} [\Gamma_{2\tau} 1\{\tau \leq \tau_0\} + \beta_2 1\{\tau > \tau_0\}] A \end{aligned}$$

where

$$A = \frac{\sigma_*^2}{\sigma_*^2 + \sigma_\varepsilon^2}$$

$$\begin{aligned} \Gamma_{1\tau} &= \beta_1 \frac{\tau_0}{\tau} + \beta_2 \frac{\tau - \tau_0}{\tau} \\ \Gamma_{2\tau} &= \beta_1 \frac{\tau_0 - \tau}{1 - \tau} + \beta_2 \frac{1 - \tau_0}{1 - \tau} \end{aligned}$$

$$\sup_{\tau \in [\underline{\tau}, \overline{\tau}]} \left| \frac{1}{T} RSS_T(\tau) - h(\tau) \right| = o_p(1)$$

where for $\tau \in [\underline{\tau}, \tau_0)$,

$$h(\tau) = \sigma_u^2 + (1 - \tau_0)\left(\beta_2^2 - \beta_1^2\right)\sigma_*^2 + \beta_1^2 (1 - A)\sigma_*^2 + (1 - \tau)\left(\beta_1^2 - \Gamma_{2\tau}^2\right) A \sigma_*^2$$

$$\frac{\partial h(\tau)}{\partial \tau} = -A(\beta_2 - \beta_1)^2 \sigma_*^2 \left(\frac{1 - \tau_0}{1 - \tau} \right)^2 \leq 0$$

$$\frac{\partial^2 h(\tau)}{\partial \tau^2} = -2A(\beta_2 - \beta_1)^2 \sigma_*^2 \frac{(1 - \tau_0)^2}{(1 - \tau)^3} \leq 0$$

Therefore $\frac{1}{T}RSS_T(\tau)$ converges uniformly to a non-increasing and concave function of τ for $\tau \in [\underline{\tau}, \tau_0)$.

For $\tau = \tau_0$,

$$h(\tau_0) = \sigma_u^2 + \left(\tau_0 \beta_1^2 + (1-\tau_0)\beta_2^2\right)(1-A)\sigma_*^2$$

Thus even if the change point can be consistently estimated, the variance of the regression error u_t will be over-estimated in general unless there is no measurement errors (i.e. when $A = 1$).

For $\tau \in (\tau_0, \overline{\tau}]$,

$$h(\tau) = \sigma_u^2 + \tau_0\left(\beta_1^2 - \beta_2^2\right)\sigma_*^2 + \beta_2^2(1-A)\sigma_*^2 + \tau\left(\beta_2^2 - \Gamma_{1\tau}^2\right)A\sigma_*^2$$

$$\frac{\partial h(\tau)}{\partial \tau} = A(\beta_2 - \beta_1)^2 \sigma_*^2 \left(\frac{\tau_0}{\tau}\right)^2 \geq 0$$

$$\frac{\partial^2 h(\tau)}{\partial \tau^2} = -2A(\beta_2 - \beta_1)^2 \sigma_*^2 \frac{(\tau_0)^2}{\tau^3} \leq 0$$

Thus for $\tau \in (\tau_0, \overline{\tau}]$, $h(\tau)$ is non-decreasing and concave.

To summarize, the criterion function $\frac{1}{T}RSS_T(\tau)$ converges uniformly to a piecewise concave function $h(\tau)$ whose minimum takes place at the true change point. This implies the true change point can be consistently estimated despite of the presence of the measurement errors. However, for all $\tau \in [\underline{\tau}, \overline{\tau}]$, $\widehat{\beta}_{1\tau}$ and $\widehat{\beta}_{2\tau}$ are inconsistent estimates for β_1 and β_2 respectively. Thus the consistency of the change-point estimator does not depend on the consistency of the structural estimators.

Theorem 172 *If assumptions* $(A1) - (A6)$ *hold, then under* $H_1 : \beta_1 \neq \beta_2$, *as* $T \to \infty$, *we have:*

$$\widehat{\tau} \xrightarrow{p} \tau_0$$
$$\widehat{\beta}_{1\widehat{\tau}} \xrightarrow{p} \beta_1 A$$
$$\widehat{\beta}_{2\widehat{\tau}} \xrightarrow{p} \beta_2 A$$

where $0 \leq A = \dfrac{\sigma_*^2}{\sigma_*^2 + \sigma_\varepsilon^2} \leq 1$

Proof. Exercise.

This Theorem states that the change point can be estimated consistently despite the existence of measurement errors. However, the pre-shift and post-shift structural estimators are biased towards zero as long as there are measurement errors ($\sigma_\varepsilon^2 > 0$). Therefore the true magnitude of break should be bigger than the observed difference between the pre-shift and post-shift estimates.

The Wald Test

Under the null of no structural break, the probability limits of the pre-shift and post-shift parameters should be the same (not necessarily coincide with true structural parameters). While in the presence of break(s), their probability limits are different. Therefore a test statistic that based on the difference between the estimated pre-shift and post-shift parameters will be a consistent test. We define a Wald-type statistic as follows:

$$W_T(\tau) = \frac{T^2 \tau (1-\tau)}{RSS_T(\tau)} \left(\widehat{\beta}_{2\tau} - \widehat{\beta}_{1\tau}\right)^2 S_{xx}(1)$$

Let $\mathbf{S}$ be a set whose closure lies in $(0,1)$. A Sup-Wald statistic defined as $\sup_{\tau \in \mathbf{S}} W_T(\tau)$ is constructed by searching the supremum of $W_T(\tau)$ over a wide range of possible break points.

Under $H_0 : \beta_1 = \beta_2 = \beta$, if there is no measurement error, both the pre-shift and post-shift estimators are consistent, so $\left(\widehat{\beta}_{2\tau} - \widehat{\beta}_{1\tau}\right)$ is a stochastic term of order $o_p(1)$ and is independent of the true parameter β. As a result, the limiting distribution of $W_T(\tau)$ will be independent of true parameters in the absence of measurement errors. However, when there are measurement errors, the term $\left(\widehat{\beta}_{2\tau} - \widehat{\beta}_{1\tau}\right)$ will depend on the true parameters β, σ_*^2, σ_ε^2, and σ_u^2. It will be shown that measurement errors will affect the asymptotic null distribution of $W_T(\tau)$ by a scaling parameter.

Definition 173 *A Brownian Bridge motion $BB(\tau)$ is defined as*

$$BB(\tau) = B(\tau) - \tau B(1)$$

where $B(\tau)$ is a standard Brownian motion on $[0,1]$.

Note that a Brownian Bridge motion has the property that $BB(0) = BB(1) = 0$.

Theorem 174 *If assumptions* $(A2)-(A6)$ *hold, then under* $H_0 : \beta_1 = \beta_2 = \beta$, *as* $T \to \infty$, *we have:*

$$W_T(\tau) \Rightarrow C\left(\beta, \sigma_*^2, \sigma_\varepsilon^2, \sigma_u^2\right) \frac{(BB(\tau))^2}{\tau(1-\tau)}$$

and

$$\sup_{\tau \in \mathbf{S}} W_T(\tau) \xrightarrow{d} C\left(\beta, \sigma_*^2, \sigma_\varepsilon^2, \sigma_u^2\right) \sup_{\tau \in \mathbf{S}} \frac{(BB(\tau))^2}{\tau(1-\tau)}$$

where

$$0 \leq C\left(\beta, \sigma_*^2, \sigma_\varepsilon^2, \sigma_u^2\right) = 1 + \frac{\beta^2\left[\sigma_*^4 E(\varepsilon_t^4) + \sigma_\varepsilon^4 E(x_t^{*4}) - 6\sigma_\varepsilon^4\sigma_*^4\right]}{\left(\sigma_*^2 + \sigma_\varepsilon^2\right)^2\left(\beta^2\sigma_*^2\sigma_\varepsilon^2 + \sigma_u^2\sigma_*^2 + \sigma_u^2\sigma_\varepsilon^2\right)}$$

$$= 1 + \frac{\beta^2\left[\sigma_*^4\left(E(\varepsilon_t^4) - 3\sigma_\varepsilon^4\right) + \sigma_\varepsilon^4\left(E(x_t^{*4}) - 3\sigma_*^4\right)\right]}{\left(\sigma_*^2 + \sigma_\varepsilon^2\right)^2\left(\beta^2\sigma_*^2\sigma_\varepsilon^2 + \sigma_u^2\sigma_*^2 + \sigma_u^2\sigma_\varepsilon^2\right)}$$

$\mathbf{S}$ denotes a set whose closure lies in $(0,1)$, and $BB(\tau)$ is a Brownian Bridge motion on $[0,1]$.

Proof. Exercise.

If $C\left(\beta, \sigma_*^2, 0, \sigma_u^2\right) < 1$, using the conventional critical values will tend to over-accept the null of no break, and the null will be over rejected if $C\left(\beta, \sigma_*^2, 0, \sigma_u^2\right) > 1$. Note that if there is no measurement errors, we have $C\left(\beta, \sigma_*^2, 0, \sigma_u^2\right) = 1$, and the conventional null distribution applies. Further, if x_t^* and ε_t are normally distributed, we will have $E(\varepsilon_t^4) = 3\sigma_\varepsilon^4$ and $E(x_t^{*4}) = 3\sigma_*^4$, $C\left(\beta, \sigma_*^2, \sigma_\varepsilon^2, \sigma_u^2\right) = 1$ again.

THE UNIT ROOT MODEL WITH MEASUREMENT ERRORS

Macroeconomists have nowadays recognized that unit root may exist in many variables of economic interest. Standard theorems in the unit-root literature, however, rule out the possibility of measurement errors in these economic time series. This leaves open the question of the validity of those previous findings if measurement errors exist.

The consequences of measurement errors in a simple unit root process will be discussed here. It will be proved that the parameter can still be consistently estimated, despite the presence of measurement errors. The convergence rate is still T. The limiting distribution of the OLS estimator, however, will be distorted. Its exact form depends on the distributional properties of the measurement error process.

Definition 175 *A sequence* $\{X_t\}_{t=-\infty}^{\infty}$ *is said to be strong mixing (*α*-mixing) if* $lim_{m\to\infty}\alpha_m = 0$ *where* $\alpha_m \equiv \sup_\tau \sup_{\left\{A\in\Im_{-\infty}^{\tau}, B\in\Im_{\tau+m}^{\infty}\right\}} |P(A\cap B) - P(A)P(B)|$ *and* $\Im_\tau^t \equiv \sigma(X_\tau, X_{\tau+1}, ..., X_t)$.

Many macroeconomic variables appear to be generated from unit root processes. Suppose the true data-generating process is a unit root process without drift:

$$y_t^* = y_{t-1}^* + u_t = y_0^* + \sum_{i=1}^{t} u_t = y_0^* + S_t \qquad t = 1, 2, ..., T.$$

where y_0^* is assumed to be drawn from a stationary processes with zero mean and a finite second moment, and u_t is a random variable satisfying the following assumptions:

Assumption 1:

$(a)\, E(u_t) = 0$ for all t.

$(b)\, \limsup_t \|u_t\|_\gamma < \infty$, for some $\gamma > 2$.

$(c)\, \sigma^2 \overset{def}{=} \lim_{T\to\infty} E\left(\frac{1}{T}S_T^2\right)$ exists and $\sigma^2 > 0$.

$(d)\, \{u_t\}_{t=1}^{\infty}$ are strong mixing (α-mixing) with mixing coefficients α_m that satisfy $\sum_{m=1}^{\infty} \alpha_m^{1-2/\gamma} < \infty$.

where $\|X\|_p$ is the L_p norm of X defined as $\|X\|_p = (E|X|^p)^{\frac{1}{p}}$ for $p \in [1, \infty)$ and $\|X\|_p = ess\sup|X|$ for $p = \infty$.

Assumption 1 remains in effect throughout the rest of this handout. The following lemma illustrates its role:

Lemma 1: Let $B(r)$ be the standard Brownian motion in $[0,1]$, if $\{u_t\}_{t=1}^{\infty}$ satisfies Assumption 1, then as $T \to \infty$,

(a) $B_T(r) \overset{def}{=} \dfrac{1}{\sqrt{T}\sigma} S_{[Tr]} \Rightarrow B(r)$,

(b) $\dfrac{1}{T}\sum_1^T y_{t-1}^* u_t \Rightarrow \dfrac{\sigma^2}{2}\left(B^2(1) - \dfrac{\sigma_u^2}{\sigma^2}\right)$, where $\sigma_u^2 = \lim_{T\to\infty} \dfrac{1}{T}\sum_{t=1}^T E(u_t^2)$.

(c) $\dfrac{1}{T^2}\sum_1^T y_{t-1}^{*2} \Rightarrow \sigma^2 \int_0^1 B^2(r)\, dr$.

Proof. Exercise.

Most macroeconomic data are aggregated and are therefore inevitably suffered from measurement and aggregation errors. In addition, the current value of an economic variable may greatly depend on its lagged values, so measurement errors may appear in both sides of a dynamic model. Therefore it would be of intrinsic interest to investigate the properties of the estimator in the presence of these kinds of errors. Suppose the true value of y_t^* cannot be accurately observed and is proxied by an observable y_t where

$$y_t = y_t^* + \varepsilon_t = \left(y_0^* + \sum_{i=1}^t u_i\right) + \varepsilon_t.$$

With a little rearrangement, the observed process y_t is generated by

$$y_t = y_{t-1} + v_t = y_0 + \sum_{i=1}^t v_t = y_0 + S_{vt}$$

where

$$v_t = u_t + \varepsilon_t - \varepsilon_{t-1}.$$

The initial measurement error ε_0 is assumed to be drawn from a stationary process with zero mean and a finite second moment. It is worth noting that even if the processes $\{u_t\}_{t=1}^{\infty}$ and $\{\varepsilon_t\}_{t=1}^{\infty}$ are i.i.d., the observed process $\{y_t\}_{t=1}^{\infty}$ still has serially correlated innovations. Thus, the existence of measurement errors in the unobserved unit root process is equivalent to the case where the observed process has a unit root with serially correlated disturbances.

The least-squares estimator for β based on the observations $\{y_t\}_{t=1}^{T}$ is given by

$$\widehat{\beta} = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} = 1 + \frac{\sum_{t=2}^{T} v_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}.$$

Case 1: $\varepsilon_t \sim I(0)$

We make the following assumptions for the process $\{\varepsilon_t\}_{t=1}^{\infty}$:

Assumption 2:

(a) $E(\varepsilon_t) = 0$ for all t.

(b) $\limsup_t \|\varepsilon_t\|_{\gamma_\varepsilon} < \infty$, for some $\gamma_\varepsilon > 2$.

(c) $\{\varepsilon_t\}_{t=1}^{\infty}$ are strong mixing with mixing coefficients $\alpha_{\varepsilon m}$ that satisfy $\sum_{t=1}^{\infty} \alpha_{\varepsilon m}^{1-2/\gamma_\varepsilon} < \infty$.

Lemma 2: Under Assumptions 1 and 2, we have:

(a) $E(v_t) = 0$ for all t.

(b) $\limsup_t \|v_t\|_{\gamma_v} < \infty$, for some $\gamma_v > 2$.

(c) $\{v_t\}_{t=1}^{\infty}$ are strong mixing with mixing coefficients α_{vm} that satisfy $\sum_{t=1}^{\infty} \alpha_{vm}^{1-2/\gamma_v} < \infty$.

(d) $\sigma_v^2 \overset{def}{=} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E(v_t^2) = \sigma_u^2 + 2\sigma_\varepsilon^2 - \lim_{T\to\infty} \frac{2}{T}\sum_{t=1}^{T} E(\varepsilon_t \varepsilon_{t-1})$.

where $\sigma_\varepsilon^2 \overset{def}{=} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E(\varepsilon_t^2)$ and σ_u^2 was defined in Lemma 1(b).

$(e)\,\sigma_V^2 \overset{def}{=} \lim_{T\to\infty} E\left(\frac{1}{T}S_{vT}^2\right) = \sigma^2.$

(f)As $T \to \infty$, $B_{vT}(r) \overset{def}{=} \frac{1}{\sqrt{T}\sigma_V} S_{v[Tr]} \Rightarrow B(r)$.

Proof. Exercise.

The properties of $\{v_t\}_{t=1}^{\infty}$ in Lemma 2 suffice to establish the consistency of $\widehat{\beta}$ in the following theorem.

Theorem 176 *Under Assumptions 1 and 2, as $T \to \infty$:*

$$T\left(\widehat{\beta}-1\right) \Rightarrow \frac{B^2(1) - \frac{\sigma_v^2}{\sigma^2}}{2\int_0^1 B^2(r)\,dr}$$

where σ^2 is defined in assumption 1(c), and σ_v^2 in Lemma 2(d).

Proof. Exercise.

Note that $\widehat{\beta}$ no longer has a standard Dickey-Fuller distribution, and its shape depends on the variance of the increment of measurement errors. Note that when there is no measurement error and when $\{u_t\}_{t=1}^{\infty}$ are i.i.d., i.e. $\sigma_\varepsilon^2 = 0$ and $\sigma_u^2 = \sigma^2$, the OLS estimator will have a standard Dickey-Fuller distribution.

Case 2: $\varepsilon_t \sim I(1)$

The previous case where the measurement error process is stationary suggests that $\frac{Var(\varepsilon_t)}{Var(y_t^*)}$ becomes negligible as the sample size grows. This implies the effect of measurement error will die out asymptotically, so that the consistency of the OLS estimator will be preserved. Empirically speaking, it is more likely for the magnitude of measurement errors to commensurate with the magnitude of the true process y_t^*.

To see how this affects our previous result, suppose the measurement errors follow a first order integrated process such that

$$\varepsilon_t = \varepsilon_{t-1} + \omega_t = \varepsilon_0 + \sum_{i=1}^{t} \omega_i.$$

The observed process y_t is generated by

$$y_t = y_0 + \varepsilon_t + \sum_{i=1}^{t} u_i = y_0 + \varepsilon_0 + \sum_{i=1}^{t} (u_i + \omega_i)$$

or

$$y_t = y_{t-1} + v_t$$

where

$$v_t = u_t + \omega_t$$

and ε_0 is assumed to be drawn from a stationary process with zero mean and a finite second moment. The increments of measurement errors satisfy the following assumptions:

Assumption 3:

(a) $E(\omega_t) = 0$ for all t.

(b) $\limsup_t \|\omega_t\|_{\gamma_\omega} < \infty$, for some $\gamma_\omega > 2$.

(c) $\{\omega_t\}_{t=1}^{\infty}$ are strong mixing with mixing coefficient $\alpha_{\omega m}$ that satisfy $\sum_{m=1}^{\infty} \alpha_{\omega m}^{1-2/\gamma_\omega} < \infty$.

(d) $\sigma_\Omega^2 \stackrel{def}{=} \lim_{T\to\infty} E\left(\frac{1}{T} S_{\omega T}^2\right)$ exists and is strictly positive.

(e) $\{\omega_t\}_{t=1}^{\infty}$ and $\{u_t\}_{t=1}^{\infty}$ are independent.

The observed innovations $\{v_t\}_{t=1}^{\infty}$ have the following properties:

Lemma 3: Under Assumptions 1 and 3, we have:

(a) $E(v_t) = 0$ for all t.

(b) $\limsup_t \|v_t\|_{\gamma_v} < \infty$, for some $\gamma_v > 2$.

$(c)\,\{v_t\}_{t=1}^{\infty}$ are strong mixing with mixing coefficients α_{vm} that satisfy $\sum_{m=1}^{\infty}\alpha_{vm}^{1-2/\gamma_v}<\infty$.

$(d)\,\sigma_v^2 \overset{def}{=} \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E\left(v_t^2\right)=\sigma_u^2+\sigma_\omega^2$,

where $\sigma_\omega^2 \overset{def}{=} \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}E\left(\omega_t^2\right)$ and σ_u^2 was defined in Lemma 1(b).

$(e)\lim_{T\to\infty}E\left(\frac{1}{T}S_{vT}^2\right)$ exists and is strictly positive.

(f)As $T\to\infty$, $B_{vT}(r)\Rightarrow B_v(r)$,

where $B_{vT}(r)$ is defined in Lemma 2(f) and $B_v(r)$ is a standard Brownian Motion on $[0,1]$.

Proof. Exercise.

Theorem 177 *Suppose $\varepsilon_t=\varepsilon_{t-1}+\omega_t$, then under Assumptions 1 and 3, as $T\to\infty$,*

$$T\left(\widehat{\beta}-1\right)\Rightarrow\frac{B_v^2(1)-\frac{\sigma_v^2}{\sigma_V^2}}{2\int_0^1 B_v^2(r)\,dr}.$$

Similar to the previous Theorem, $\widehat{\beta}$ will have a standard Dickey-Fuller distribution when $\sigma_\omega^2=0$ and $\sigma_u^2=\sigma^2$.

Proof. Exercise.

The robustness of the consistency of the OLS estimator in the presence of measurement error is surprising. It is more surprising that this consistency result is insensitive to whether these measurement errors are $I(0)$ or $I(1)$. The basic idea involved is that, if the true process is a unit root process with weakly dependent increments, then the observed process will be a unit root process with increments exhibiting another form of dependence.

Under fairly general circumstances, the OLS estimator approaches its asymptotic distribution at the rate of T, as rapid as that in the case of no measurement error. However, the limiting distribution will no longer be a

standard Dickey Fuller type distribution, its shape depends on the long-run variance and the order of the measurement error process.

Case 3: $\varepsilon_t = -\varepsilon_{t-1} + \omega_t$

We have shown previously that the consistency of OLS estimator is preserved under both $I(0)$ and $I(1)$ measurement errors. It seems that this consistency result is invariant to the type of measurement errors from which the true unit root process is suffering. This conjecture, however, is questionable. We will show in the following case that the OLS estimator is inconsistent under certain type of measurement errors. Suppose the measurement errors are generated by the following process:

$$\begin{aligned} \varepsilon_t &= -\varepsilon_{t-1} + \omega_t = -\left(-\varepsilon_{t-2} + \omega_{t-1}\right) + \omega_t \\ &= (-1)^t \varepsilon_0 + \sum_{i=1}^{t} \eta(i,t) \end{aligned}$$

where

$$\eta(i,t) = (-1)^{t-i} \omega_i.$$

Without loss of generality, suppose $\varepsilon_0 = 0$, then the observed process y_t is generated by

$$y_t = y_0 + \varepsilon_t + \sum_{i=1}^{t} u_i = y_0 + \sum_{i=1}^{t} \left(u_i + \eta(i,t)\right).$$

Lemma 4: If ω_i and u_j are independent for all i, j, then under Assumption 3,

(a) $E(\eta(i,t)) = 0$ for all i, t.

(b) $\limsup_t \|\eta(i,t)\|_{\gamma_\eta} < \infty$, for some $\gamma_\eta > 2$.

(c) $\{\eta(i,t)\}_{t=1}^{\infty}$ are strong mixing with mixing coefficients $\alpha_{\eta m}$ that satisfy $\sum_{m=1}^{\infty} \alpha_{\eta m}^{1-2/\gamma_\eta} < \infty$.

(d) $\sigma_\eta^2 \overset{def}{=} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} E(\eta_t^2) = \sigma_\omega^2$.

$(e)\,\sigma_\Lambda^2 \overset{def}{=} \lim_{T\to\infty} E\left(\frac{1}{T}S_{\eta T}^2\right)$ exists and is strictly positive.

(f)As $T\to\infty$, $B_{\eta T}(r) \Rightarrow B_\eta(r)$, where $B_\eta(r)$ is a standard Brownian Motion on $[0,1]$.

Proof. Exercise.

Theorem 178 *Suppose $\varepsilon_t = -\varepsilon_{t-1} + \omega_t$, then under Assumptions 1 and 3, the OLS estimator $\widehat{\beta}$ is inconsistent, and as $T\to\infty$,*

$$\widehat{\beta} \Rightarrow 1 - \frac{2\int_0^1 B_\eta^2(r)\,dr}{\int_0^1 \left(B_\eta(r) + \frac{\sigma}{\sigma_\Lambda}B(r)\right)^2 dr},$$

where $B(r)$ is defined in Lemma 1(a), $B_\eta(r)$ in Lemma 4(f) , and $B_\eta(r)$ is independent of $B(r)$.

Proof. Exercise.

Note that $\widehat{\beta} \xrightarrow{p} 1$ as $\frac{\sigma}{\sigma_\Lambda}\to\infty$, and $\widehat{\beta} \xrightarrow{p} -1$ as $\frac{\sigma}{\sigma_\Lambda}\to 0$. In other words, the OLS estimator will be closer to its true value as the measurement error process becomes less volatile.

MONTE CARLO EXPERIMENTS

Experiment 1: This experiment verifies Theorem 52 and 53. We have the following setup:

$y_t^* = y_{t-1}^* + u_t$, $u_t \sim nid(0,1)$, $t = 1,2,...,T$.

y_0 is drawn from a $nid(0,1)$ independent of $\{u_t\}_{t=1}^T$.

The observed process

$y_t = y_{t-1}^* + \varepsilon_t$, $\varepsilon_t = \rho\varepsilon_{t-1} + \omega_t \qquad \rho = -0.5,\ 0,\ 0.5,\ 1.$

ω_t is set to be a standard normal i.i.d. random variable independent of y_0 and $\{u_t\}_{t=1}^T$. i.e., $\omega_t \sim nid(0,1)$.

To get the limiting distribution of $\widehat{\beta}$ with high precision, we perform simulations at $T = 20000$ (sample size) and $N = 20000$ (number of replications). Let $\widehat{\beta}_N$ be the average of $\widehat{\beta}$ in these 20000 replications.

Table 1: Percentile of $T\left(\widehat{\beta}-1\right)$

	$\widehat{\beta}_N$	99%	95%	90%	50%	10%	5%	1%
$\rho=-.5$	.9993	.3452	$-.4204$	-1.053	-7.860	-32.31	-44.03	-72.43
$\rho=0$	.9996	.8842	.2839	$-.1363$	-4.198	-18.73	-25.75	-42.28
$\rho=.5$	.9997	1.110	.5623	.1813	-3.128	-14.44	-19.66	-32.65
$\rho=1$	.9999	2.024	1.269	.9175	$-.8389$	-5.756	-8.114	-14.06
$\sigma_\varepsilon^2=0$	.9999	2.062	1.293	.9285	$-.8412$	-5.580	-7.809	-14.00

Note from Table 1 that $\widehat{\beta}_N$ are very close to 1 under various values of ρ. And as ρ increases, the limiting distribution of $\widehat{\beta}$ shifts to the right. The case without measurement is tabulated in the last row of Table 1 for comparison purposes.

Note that the limiting distribution of $\widehat{\beta}$ in the absence of measurement error is closest to the case where $\rho=1$. This implies that if the measurement process is also a unit root process, its effects on the limiting distribution of $\widehat{\beta}$ will vanish.

Experiment 2: This experiment verifies Theorem 54.
$y_t^*=y_{t-1}^*+u_t$, $u_t\sim nid(0,\sigma^2)$, $t=1,2,...,T$.
y_0 is drawn from a $nid(0,1)$ independent of $\{u_t\}_{t=1}^T$.
The observed process
$y_t=y_{t-1}^*+\varepsilon_t, \qquad \varepsilon_t=-\varepsilon_{t-1}+\omega_t, \qquad \omega_t\sim nid(0,\sigma_\Lambda^2)$.
$T=20000, N=20000$
The results are summarized in Table 2.

Table 2: Percentile of $\widehat{\beta}$

	$\widehat{\beta}_N$	99%	95%	90%	50%	10%	5%	1%
$\frac{\sigma}{\sigma_\Lambda}=10$	.9501	.9993	.9982	.9971	.9798	.8715	.7985	.5776
$\frac{\sigma}{\sigma_\Lambda}=1$	$-.0002$	.9287	.8362	.7444	.0005	$-.7476$	$-.8389$	$-.9293$
$\frac{\sigma}{\sigma_\Lambda}=0.1$	$-.9506$	$-.5759$	$-.8003$	$-.8759$	$-.9803$	$-.9970$	$-.9982$	$-.9993$

Note that as predicted in Theorem 54, $\widehat{\beta} \xrightarrow{p} 1$ as $\frac{\sigma}{\sigma_\Lambda} \to \infty$, and $\widehat{\beta} \xrightarrow{p} -1$ as $\frac{\sigma}{\sigma_\Lambda} \to 0$, and $\widehat{\beta}$ converges in distribution to a random variable symmetrically distributed about zero when $\frac{\sigma}{\sigma_\Lambda} = 1$.

Exercise 0.167 *Use GAUSS to run the following experiment to verify Theorem 50:*

$y_t = -10x_t^* + u_t \qquad t = 1, 2, ..., 10000$
$y_t = 10x_t^* + u_t \qquad t = 10001, 10002, ..., 20000$
$x_t = x_t^* + \varepsilon_t$
$u_t \sim nid\,(0, 1)$
$x_t^* \sim nid\,(1, 1)$
$\varepsilon_t \sim nid\,(0, \sigma_\varepsilon^2) \qquad \sigma_\varepsilon^2 = 0, 1, 2, ..., 8.$
x_t^*, ε_t, and u_t are independent of one another.
$A = \frac{\sigma_*^2}{\sigma_*^2 + \sigma_\varepsilon^2} = \frac{2}{2 + \sigma_\varepsilon^2}$
$\frac{1}{T} RSS_T\,(\tau_0) \xrightarrow{p} h\,(\tau_0) = \sigma_u^2 + \left(\tau_0 \beta_1^2 + (1 - \tau_0)\,\beta_2^2\right)(1 - A)\,\sigma_*^2 = 1 + 200\,(1 - A)$

Fill up the Table below. For each value of σ_ε^2, perform 1 replication.

σ_ε^2	0	1	2	3	4	5	6	7	8
$\beta_1 A$									
$\beta_2 A$									
$h\,(\tau_0)$									
$\widehat{\tau}$									
$\widehat{\beta}_{1\widehat{\tau}}$									
$\widehat{\beta}_{2\widehat{\tau}}$									
$\frac{1}{T} RSS_T\,(\widehat{\tau})$									

Do your results support the findings in Theorem 50? Is $\frac{1}{T} RSS_T\,(\widehat{\tau})$ is much greater than the true value of $\sigma_u^2\,(= 1)$ when there are measurement errors? Does the change-point estimator coincide with the true change point

in all cases? Do the structural estimators converge to the true parameters multiplied by A?

Exercise 0.168 *Consider the process*

$$\begin{aligned} y_t^* &= \beta y_{t-1}^* + u_t. \\ y_0^* &= 0. \\ u_t &\sim i.i.d.\left(0, \sigma_u^2\right). \end{aligned}$$

Suppose y_t^* is not observable and we only observe y_t, where $y_t = y_t^* + \varepsilon_t$, $\varepsilon_t \sim i.i.d.\,(0, \sigma_\varepsilon^2)$. $\{y_t^*\}_{t=1}^T$, $\{\varepsilon_t\}_{t=1}^T$and $\{u_t\}_{t=1}^T$ are independent.

a) Suppose $|\beta| < 1$, show that as $t \to \infty$,

$$E\left(y_t^{*2}\right) \overset{def}{=} \sigma_*^2 = \frac{\sigma_u^2}{1-\beta^2}.$$

b) Let

$$A = \frac{\sigma_*^2}{\sigma_*^2 + \sigma_\varepsilon^2}$$

and

$$\widehat{\beta}_T = \frac{\sum_{t=2}^T y_t y_{t-1}}{\sum_{t=2}^T y_{t-1}^2}.$$

Show that

$$\widehat{\beta}_T \xrightarrow{p} \beta A.$$

c) Suppose $\beta = 1$, show that plim $\widehat{\beta}_T = 1$.

d) Write a Gauss program to simulate the distribution of $\widehat{\beta}_T$ for $T = 100000$, $N = 10000$, $u_t \sim i.i.d.N\,(0,1)$ and $\varepsilon_t \sim i.i.d.N\,(0,1)$

7. Use GAUSS to run the following experiment to verify Theorem 51, which states that measurement errors affect the limiting distribution of the Sup-Wald statistic.

True Model:

$$y_t = \beta x_t^* + u_t$$

Estimated Model:

$$y_t = \beta x_t + u_t$$

$t = 1, 2, ..., 1000$
$x_t = x_t^* + \varepsilon_t$
$x_t^* \sim n.i.d.\,(0, 1)$
$u_t \sim n.i.d.\,(0, 1)$
$\{x_t\}_{t=1}^T$ and $\{u_t\}_{t=1}^T$ are independent of each other.
$N = 10000$

Let b be the critical value such that $\Pr\left(\sup_{\tau\in(.15,.85)} W_T(\tau) > b\right) = \alpha$.

a) If $\varepsilon_t = 0$ for all t, i.e. there is no measurement error. Fill up the following table:

Table a:

$\underline{\beta = 0}$			$\underline{\beta = 1}$			$\underline{\beta = 2}$			$\underline{\beta = 3}$		
$C(0,1,0,1) = 1$			$C(1,1,0,1) = 1$			$C(2,1,0,1) = 1$			$C(3,1,0,1) = 1$		
α	.1	.05 .01		.1	.05 .01		.1	.05 .01		.1	.05 .01
b											

b) If $\varepsilon_t \sim U\left(-\sqrt{3}, \sqrt{3}\right)$, show that $\sigma_\varepsilon^2 = 1$, $E(\varepsilon_t^4) = \dfrac{9}{5}$ and $C(\beta, 1, 1, 1) = \dfrac{7\beta^2 + 20}{10\beta^2 + 20}$. Fill up the following table:

Table b:

$\underline{\beta = 0}$			$\underline{\beta = 1}$			$\underline{\beta = 2}$			$\underline{\beta = 3}$		
$C(0,1,1,1) = 1$			$C(1,1,1,1) = 0.9$			$C(2,1,1,1) = 0.8$			$C(3,1,1,1) = .75455$		
α	.1	.05 .01		.1	.05 .01		.1	.05 .01		.1	.05 .01
b											

c) If $\varepsilon_t \sim N(0,1)$, $\sigma_\varepsilon^2 = 1$, show that $E(\varepsilon_t^4) = 3$, $C(\beta,1,1,1) = 1$. Fill up the following table:

Table c:

$\underline{\beta = 0}$			$\underline{\beta = 1}$			$\underline{\beta = 2}$			$\underline{\beta = 3}$		
$C(0,1,1,1) = 1$			$C(1,1,1,1) = 1$			$C(2,1,1,1) = 1$			$C(3,1,1,1) = 1$		
α	.1	.05 .01		.1	.05 .01		.1	.05 .01		.1	.05 .01
b											

Exercise 0.169 *Repeat experiments 1 and 2.*

Proof of Theorem 50.

For $\tau \in [\underline{\tau}, \tau_0]$,

$$\widehat{\beta}_{1\tau} = \frac{S_{xy}(\tau)}{S_{xx}(\tau)} = \beta_1 \frac{S_{**}(\tau)}{S_{xx}(\tau)} + o_p(1) \xrightarrow{p} \beta_1 A$$

$$\widehat{\beta}_{2\tau} = \frac{S_{xy}(1) - S_{xy}(\tau)}{S_{xx}(1) - S_{xx}(\tau)} = \beta_1 \frac{S_{**}(\tau_0) - S_{**}(\tau)}{S_{xx}(1) - S_{xx}(\tau)} + \beta_2 \frac{S_{**}(1) - S_{**}(\tau_0)}{S_{xx}(1) - S_{xx}(\tau)} + o_p(1)$$

$$\xrightarrow{p} \left(\beta_1 \frac{\tau_0 - \tau}{1 - \tau} + \beta_2 \frac{1 - \tau_0}{1 - \tau} \right) A = \Gamma_{2\tau} A$$

Using

$$\frac{2}{T} \sum_{t=1}^{[\tau T]} \left(\beta_1 x_t^* - \widehat{\beta}_{1\tau} (x_t^* + \varepsilon_t) \right) u_t = o_p(1)$$

$$\frac{2}{T}\sum_{t=[\tau T]+1}^{k_0}\left(\beta_1 x_t^* - \widehat{\beta}_{2\tau}\left(x_t^*+\varepsilon_t\right)\right)u_t = o_p(1)$$

$$\frac{2}{T}\sum_{t=k_0+1}^{T}\left(\beta_2 x_t^* - \widehat{\beta}_{2\tau}\left(x_t^*+\varepsilon_t\right)\right)u_t = o_p(1)$$

and $(1-A)\sigma_*^2 = A\sigma_\varepsilon^2$, we have:

$$\begin{aligned}
&\tfrac{1}{T}RSS_T(\tau)\\
&= \tfrac{1}{T}\sum_{t=1}^{[\tau T]}\left(\beta_1 x_t^* + u_t - \widehat{\beta}_{1\tau}\left(x_t^*+\varepsilon_t\right)\right)^2 + \tfrac{1}{T}\sum_{t=[\tau T]+1}^{k_0}\left(\beta_1 x_t^* + u_t - \widehat{\beta}_{2\tau}\left(x_t^*+\varepsilon_t\right)\right)^2\\
&+ \tfrac{1}{T}\sum_{t=k_0+1}^{T}\left(\beta_2 x_t^* + u_t - \widehat{\beta}_{2\tau}\left(x_t^*+\varepsilon_t\right)\right)^2\\
&= \tfrac{1}{T}\sum_{t=1}^{T}u_t^2 + \widehat{\beta}_{1\tau}^2\tfrac{1}{T}\sum_{t=1}^{[\tau T]}\varepsilon_t^2 + \left(\beta_1 - \widehat{\beta}_{1\tau}\right)^2\tfrac{1}{T}\sum_{t=1}^{[\tau T]}x_t^{*2}\\
&+ \left(\beta_1 - \widehat{\beta}_{2\tau}\right)^2\tfrac{1}{T}\sum_{t=[\tau T]+1}^{k_0}x_t^{*2} + \left(\beta_2 - \widehat{\beta}_{2\tau}\right)^2\tfrac{1}{T}\sum_{t=k_0+1}^{T}x_t^{*2}\\
&+ \widehat{\beta}_{2\tau}^2\tfrac{1}{T}\sum_{t=[\tau T]+1}^{T}\varepsilon_t^2 + o_p(1)\\
&\xrightarrow{p} \sigma_u^2 + \beta_1^2A^2\tau\sigma_\varepsilon^2 + \left(\beta_1 - \beta_1 A\right)^2\tau\sigma_*^2 + \left(\beta_1 - \Gamma_{2\tau}A\right)^2(\tau_0-\tau)\sigma_*^2\\
&+ \left(\beta_2 - \Gamma_{2\tau}A\right)^2(1-\tau_0)\sigma_*^2 + \Gamma_{2\tau}^2A^2(1-\tau)\sigma_\varepsilon^2\\
&= \sigma_u^2 + (1-A)\sigma_*^2\tau\beta_1^2 + (\tau_0-\tau)\beta_1^2\sigma_*^2 + (1-\tau_0)\beta_2^2\sigma_*^2\\
&+ \left((\tau_0-\tau)A + (1-A)(1-\tau) + (1-\tau_0)A\right)A\Gamma_{2\tau}^2\sigma_*^2 - 2(1-\tau)\Gamma_{2\tau}^2A\sigma_*^2\\
&= \sigma_u^2 + (1-\tau_0)\left(\beta_2^2-\beta_1^2\right)\sigma_*^2 + (1-A)\sigma_*^2\beta_1^2 + (1-\tau)\left(\beta_1^2-\Gamma_{2\tau}^2\right)A\sigma_*^2\\
&\overset{def}{=} h(\tau)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial h(\tau)}{\partial\tau} &= -\beta_1^2A\sigma_*^2 - 2\Gamma_{2\tau}(\beta_2-\beta_1)\frac{1-\tau_0}{1-\tau}A\sigma_*^2 + \Gamma_{2\tau}^2A\sigma_*^2\\
&= -\beta_1^2A\sigma_*^2 + 2\Gamma_{2\tau}(\beta_1-\Gamma_{2\tau})A\sigma_*^2 + \Gamma_{2\tau}^2A\sigma_*^2 = -A\left(\beta_1-\Gamma_{2\tau}\right)^2\sigma_*^2 \le 0\\
\frac{\partial^2h(\tau)}{\partial\tau^2} &= -2A\left(\beta_1-\beta_2\right)^2\sigma_*^2\frac{(1-\tau_0)^2}{(1-\tau)^3} \le 0
\end{aligned}$$

For $\tau\in(\tau_0,\overline{\tau}]$

$$\begin{aligned}
\widehat{\beta}_{1\tau} &= \beta_1\frac{S_{**}(\tau_0)}{S_{xx}(\tau)} + \beta_2\frac{S_{**}(\tau)-S_{**}(\tau_0)}{S_{xx}(\tau)} + o_p(1)\\
&\xrightarrow{p}\left(\beta_1\frac{\tau_0}{\tau} + \beta_2\frac{\tau-\tau_0}{\tau}\right)A = \Gamma_{1\tau}A\\
\widehat{\beta}_{2\tau} &= \beta_2\frac{S_{**}(1)-S_{**}(\tau)}{S_{xx}(1)-S_{xx}(\tau)} + o_p(1) \xrightarrow{p} \beta_2A
\end{aligned}$$

Using

$$\frac{2}{T}\sum_{t=1}^{k_0}\left(\beta_1 x_t^* - \widehat{\beta}_{1\tau}\left(x_t^* + \varepsilon_t\right)\right)u_t = o_p\left(1\right)$$
$$\frac{2}{T}\sum_{t=k_0+1}^{[\tau T]}\left(\beta_2 x_t^* - \widehat{\beta}_{1\tau}\left(x_t^* + \varepsilon_t\right)\right)u_t = o_p\left(1\right)$$
$$\frac{2}{T}\sum_{t=[\tau T]+1}^{T}\left(\beta_2 x_t^* - \widehat{\beta}_{2\tau}\left(x_t^* + \varepsilon_t\right)\right)u_t = o_p\left(1\right)$$
and $\left(1-A\right)\sigma_*^2 = A\sigma_\varepsilon^2$, we have:

$$\begin{aligned}
&\tfrac{1}{T}RSS_T\left(\tau\right)\\
&= \tfrac{1}{T}\sum_{t=1}^{k_0}\left(\beta_1 x_t^* + u_t - \widehat{\beta}_{1\tau}\left(x_t^* + \varepsilon_t\right)\right)^2 + \tfrac{1}{T}\sum_{t=k_0+1}^{[\tau T]}\left(\beta_2 x_t^* + u_t - \widehat{\beta}_{1\tau}\left(x_t^* + \varepsilon_t\right)\right)^2\\
&+ \tfrac{1}{T}\sum_{t=[\tau T]+1}^{T}\left(\beta_2 x_t^* + u_t - \widehat{\beta}_{2\tau}\left(x_t^* + \varepsilon_t\right)\right)^2\\
&= \tfrac{1}{T}\sum_{t=1}^{T}u_t^2 + \widehat{\beta}_{1\tau}^2\tfrac{1}{T}\sum_{t=1}^{[\tau T]}\varepsilon_t^2 + \left(\beta_1 - \widehat{\beta}_{1\tau}\right)^2\tfrac{1}{T}\sum_{t=1}^{k_0}x_t^{*2}\\
&+ \left(\beta_2 - \widehat{\beta}_{1\tau}\right)^2\tfrac{1}{T}\sum_{t=k_0+1}^{[\tau T]}x_t^{*2} + \left(\beta_2 - \widehat{\beta}_{2\tau}\right)^2\tfrac{1}{T}\sum_{t=[\tau T]+1}^{T}x_t^{*2}\\
&+ \widehat{\beta}_{2\tau}^2\tfrac{1}{T}\sum_{t=[\tau T]+1}^{T}\varepsilon_t^2 + o_p\left(1\right)\\
&\xrightarrow{p} \sigma_u^2 + \Gamma_{1\tau}^2A^2\tau\sigma_\varepsilon^2 + \left(\beta_1 - \Gamma_{1\tau}A\right)^2\tau_0\sigma_*^2 + \left(\beta_2 - \Gamma_{1\tau}A\right)^2\left(\tau - \tau_0\right)\sigma_*^2\\
&+ \beta_2^2\left(1-A\right)^2\left(1-\tau\right)\sigma_*^2 + \beta_2^2A^2\left(1-\tau\right)\sigma_\varepsilon^2\\
&= \sigma_u^2 + \tau\Gamma_{1\tau}^2A\left(1-A\right)\sigma_*^2 + \left(1-\tau\right)\beta_2^2\left(1-A\right)\sigma_*^2 + \tau_0\beta_1^2\sigma_*^2 + \tau_0\Gamma_{1\tau}^2A^2\sigma_*^2\\
&+ \left(\tau - \tau_0\right)\beta_2^2\sigma_*^2 + \left(\tau - \tau_0\right)\Gamma_{1\tau}^2A^2\sigma_*^2 - 2\tau_0\beta_1\Gamma_{1\tau}A\sigma_*^2 - 2\left(\tau - \tau_0\right)\beta_2\Gamma_{1\tau}A\sigma_*^2\\
&= \sigma_u^2 + \tau_0\left(\beta_1^2 - \beta_2^2\right)\sigma_*^2 + \beta_2^2\left(1-A\right)\sigma_*^2 + \tau\left(\beta_2^2 - \Gamma_{1\tau}^2\right)A\sigma_*^2 \overset{def}{=} h\left(\tau\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial h\left(\tau\right)}{\partial\tau} &= \beta_2^2\sigma_*^2A - 2\Gamma_{1\tau}\left(\beta_2 - \beta_1\right)\frac{\tau_0}{\tau}A\sigma_*^2 - \Gamma_{1\tau}^2A\sigma_*^2\\
&= \beta_2^2\sigma_*^2A - 2\Gamma_{1\tau}\left(\beta_2 - \Gamma_{1\tau}\right)A\sigma_*^2 - \Gamma_{1\tau}^2A\sigma_*^2 = A\left(\beta_2 - \Gamma_{1\tau}\right)^2\sigma_*^2\\
&= A\left(\beta_2 - \beta_1\right)^2\sigma_*^2\left(\frac{\tau_0}{\tau}\right)^2 \geq 0
\end{aligned}$$

$$\frac{\partial^2 h\left(\tau\right)}{\partial\tau^2} = -2A\left(\beta_2 - \beta_1\right)^2\sigma_*^2\frac{\left(\tau_0\right)^2}{\tau^3} \leq 0$$

Thus $\frac{1}{T}RSS_T\left(\tau\right)$ converges in probability to a piecewise concave function of τ. Under the i.i.d. assumptions $(A2)$ to $(A6)$, it is not difficult to show that $\frac{1}{T}RSS_T\left(\tau\right)$ converges uniformly to $h\left(\tau\right)$. Thus its minimum should take place at the true change point, which implies:

$$\begin{aligned}
&\widehat{\tau} \xrightarrow{p} \tau_0 \\
\widehat{\beta}_{1\widehat{\tau}} &= \widehat{\beta}_{1\tau_0} + o_p(1) \xrightarrow{p} \beta_1 A \\
\widehat{\beta}_{2\widehat{\tau}} &= \widehat{\beta}_{2\tau_0} + o_p(1) \xrightarrow{p} \beta_2 A
\end{aligned}$$

■

Proof of Theorem 51.

$\widehat{\beta}_{1\tau} = \beta - \beta\frac{S_{\varepsilon\varepsilon}(\tau)}{S_{xx}(\tau)} + \frac{1}{\sqrt{T}}\frac{S(\tau;\beta)}{S_{xx}(\tau)}$

where $S(\tau;\beta) = -\beta S_{*\varepsilon}(\tau) + S_{*u}(\tau) + S_{u\varepsilon}(\tau)$

$\widehat{\beta}_{2\tau} = \beta - \beta\frac{S_{\varepsilon\varepsilon}(1) - S_{\varepsilon\varepsilon}(\tau)}{S_{xx}(1) - S_{xx}(\tau)} + \frac{1}{\sqrt{T}}\frac{S(1;\beta) - S(\tau;\beta)}{S_{xx}(1) - S_{xx}(\tau)}$

$$\begin{aligned}
&\sqrt{T}\left(\widehat{\beta}_{2\tau} - \widehat{\beta}_{1\tau}\right) \\
&= -\beta\sqrt{T}\frac{\left(S_{\varepsilon\varepsilon}(1) - \sigma_\varepsilon^2\right)S_{xx}(\tau) - \left(S_{\varepsilon\varepsilon}(\tau) - \tau\sigma_\varepsilon^2\right)S_{xx}(1) + \sigma_\varepsilon^2\left(S_{xx}(\tau) - \tau S_{xx}(1)\right)}{S_{xx}(\tau)\left[S_{xx}(1) - S_{xx}(\tau)\right]} \\
&+ \frac{S(1;\beta)S_{xx}(\tau) - S(\tau;\beta)S_{xx}(1)}{S_{xx}(\tau)\left[S_{xx}(1) - S_{xx}(\tau)\right]} \\
&= -\beta\sqrt{T}\frac{\left(S_{\varepsilon\varepsilon}(1) - \sigma_\varepsilon^2\right)\left(S_{xx}(\tau) - \tau\sigma_\varepsilon^2\right) - \left(S_{\varepsilon\varepsilon}(\tau) - \tau\sigma_\varepsilon^2\right)\left(S_{xx}(1) - \sigma_\varepsilon^2\right)}{S_{xx}(\tau)\left[S_{xx}(1) - S_{xx}(\tau)\right]} \\
&-\beta\sqrt{T}\frac{\sigma_\varepsilon^2\left(\left(S_{**}(\tau) - \tau\sigma_*^2\right) - \tau\left(S_{**}(1) - \sigma_*^2\right)\right)}{S_{xx}(\tau)\left[S_{xx}(1) - S_{xx}(\tau)\right]} - 2\beta\frac{\sigma_\varepsilon^2\left(S_{*\varepsilon}(\tau) - \tau S_{*\varepsilon}(1)\right)}{S_{xx}(\tau)\left[S_{xx}(1) - S_{xx}(\tau)\right]} \\
&+ \frac{S(1;\beta)S_{xx}(\tau) - S(\tau;\beta)S_{xx}(1)}{S_{xx}(\tau)\left[S_{xx}(1) - S_{xx}(\tau)\right]} \\
&\Rightarrow -\beta\sigma_*^2\frac{\tau B_{\varepsilon\varepsilon}(1) - B_{\varepsilon\varepsilon}(\tau)}{\tau(1-\tau)\left(\sigma_*^2 + \sigma_\varepsilon^2\right)^2} + \beta\sigma_\varepsilon^2\frac{\tau B_{**}(1) - B_{**}(\tau)}{\tau(1-\tau)\left(\sigma_*^2 + \sigma_\varepsilon^2\right)^2} + 2\beta\sigma_\varepsilon^2\frac{\tau B_{*\varepsilon}(1) - B_{*\varepsilon}(\tau)}{\tau(1-\tau)\left(\sigma_*^2 + \sigma_\varepsilon^2\right)^2} \\
&-\beta\frac{\tau B_{*\varepsilon}(1) - B_{*\varepsilon}(\tau)}{\tau(1-\tau)\left(\sigma_*^2 + \sigma_\varepsilon^2\right)} + \frac{\tau B_{*u}(1) - B_{*u}(\tau)}{\tau(1-\tau)\left(\sigma_*^2 + \sigma_\varepsilon^2\right)} + \frac{\tau B_{u\varepsilon}(1) - B_{u\varepsilon}(\tau)}{\tau(1-\tau)\left(\sigma_*^2 + \sigma_\varepsilon^2\right)} \\
&= \theta_0\sum_{i=1}^{5}\theta_i BB_i(\tau) \overset{d}{=} \eta BB(\tau)
\end{aligned}$$

where

$\theta_0 = \frac{1}{\tau(1-\tau)\left(\sigma_*^2 + \sigma_\varepsilon^2\right)^2}$, $\theta_1 = -\beta\sigma_*^2\sqrt{E\left(\varepsilon_t^4\right) - \sigma_\varepsilon^4}$, $\theta_2 = \beta\sigma_\varepsilon^2\sqrt{E\left(x_t^{*4}\right) - \sigma_*^4}$,

$\theta_3 = \beta\left(\sigma_\varepsilon^2 - \sigma_*^2\right)\sigma_*\sigma_\varepsilon$, $\theta_4 = \sigma_*\sigma_u\left(\sigma_*^2 + \sigma_\varepsilon^2\right)$, $\theta_5 = \sigma_u\sigma_\varepsilon\left(\sigma_*^2 + \sigma_\varepsilon^2\right)$

$$\begin{aligned}
\eta^2 &= \theta_0^2\sum_{i=1}^{5}\theta_i^2 \\
&= \frac{\beta^2\left(\sigma_*^4 E\left(\varepsilon_t^4\right) + \sigma_\varepsilon^4 E\left(x_t^{*4}\right) - 6\sigma_\varepsilon^4\sigma_*^4\right) + \left(\beta^2\sigma_*^2\sigma_\varepsilon^2 + \sigma_u^2\sigma_*^2 + \sigma_u^2\sigma_\varepsilon^2\right)\left(\sigma_*^2 + \sigma_\varepsilon^2\right)^2}{\tau^2(1-\tau)^2\left(\sigma_*^2 + \sigma_\varepsilon^2\right)^4}
\end{aligned}$$

$$BB_1(\tau) = \frac{\tau B_{\varepsilon\varepsilon}(1) - B_{\varepsilon\varepsilon}(\tau)}{\sqrt{E(\varepsilon_t^4) - \sigma_\varepsilon^4}},\ BB_2(\tau) = \frac{\tau B_{**}(1) - B_{**}(\tau)}{\sqrt{E(x_t^{*4}) - \sigma_*^4}},\ BB_3(\tau) = \frac{\tau B_{*\varepsilon}(1) - B_{*\varepsilon}(\tau)}{\sigma_*\sigma_\varepsilon},\ BB_4(\tau) = \frac{\tau B_{*u}(1) - B_{*u}(\tau)}{\sigma_*\sigma_u},\ BB_5(\tau) = \frac{\tau B_{u\varepsilon}(1) - B_{u\varepsilon}(\tau)}{\sigma_u\sigma_\varepsilon}$$

$BB_i(\tau)$ are Brownian Bridge motions on $[0,1]$ independent of one another, $i = 1,2,3,4,5$.

$$W_T(\tau) = \frac{\tau(1-\tau)}{\frac{1}{T}RSS_T(\tau)} T\left(\widehat{\beta}_{2\tau} - \widehat{\beta}_{1\tau}\right)^2 S_{xx}(1)$$

$$\Rightarrow \frac{\tau(1-\tau)}{\sigma_u^2 + \beta^2 \frac{\sigma_\varepsilon^2}{\sigma_*^2 + \sigma_\varepsilon^2}\sigma_*^2} \eta^2 BB(\tau)^2 (\sigma_*^2 + \sigma_\varepsilon^2) = C\left(\beta, \sigma_*^2, \sigma_\varepsilon^2, \sigma_u^2\right) \frac{BB(\tau)^2}{\tau(1-\tau)}$$

where

$$C\left(\beta, \sigma_*^2, \sigma_\varepsilon^2, \sigma_u^2\right) = 1 + \frac{\beta^2\left[\sigma_*^4 E(\varepsilon_t^4) + \sigma_\varepsilon^4 E(x_t^{*4}) - 6\sigma_\varepsilon^4\sigma_*^4\right]}{(\sigma_*^2 + \sigma_\varepsilon^2)^2\left(\beta^2\sigma_*^2\sigma_\varepsilon^2 + \sigma_u^2\sigma_*^2 + \sigma_u^2\sigma_\varepsilon^2\right)}$$

and $BB(\tau)$ is a Brownian Bridge motion on $[0,1]$.

By the Continuous Mapping Theorem, we have:

$$\sup_{\tau\in\mathbf{S}} W_T(\tau) \xrightarrow{d} C\left(\beta, \sigma_*^2, \sigma_\varepsilon^2, \sigma_u^2\right) \sup_{\tau\in\mathbf{S}} \frac{(BB(\tau))^2}{\tau(1-\tau)}$$

where $\mathbf{S}$ is a set whose closure lies in $(0,1)$. ■

Proof of Theorem 52:

$$\frac{1}{T^2}\sum_{t=1}^{T}y_{t-1}^2=\frac{1}{T^2}\sum_{t=1}^{T}\left(y_0+\varepsilon_{t-1}+\sum_{i=1}^{t-1}u_i\right)^2$$

$$=\frac{1}{T^2}\sum_{t=1}^{T}\left(S_{t-1}^2+2\left(y_0+\varepsilon_{t-1}\right)S_{t-1}+\left(y_0+\varepsilon_{t-1}\right)^2\right)$$

$$=\sigma^2\sum_{t=1}^{T}\int_{(t-1)/T}^{t/T}\frac{S_{[Tr]}^2}{T\sigma^2}dr+\frac{2\sigma}{\sqrt{T}}\sum_{t=1}^{T}\left(y_0+\varepsilon_{t-1}\right)\int_{(t-1)/T}^{t/T}\frac{S_{[Tr]}}{\sqrt{T}\sigma}dr$$

$$+\frac{1}{T}\left(y_0^2+\frac{2y_0}{T}\sum_{t=1}^{T}\varepsilon_{t-1}+\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{t-1}^2\right)$$

$$=\sigma^2\int_0^1 B_T^2\left(r\right)dr+\frac{2\sigma}{\sqrt{T}}\left(y_0\int_0^1 B_T\left(r\right)dr+\sum_{t=1}^{T}\varepsilon_{t-1}\int_{(t-1)/T}^{t/T}B_T\left(r\right)dr\right)$$

$$+\frac{1}{T}\left(y_0^2+\frac{2y_0}{T}\sum_{t=1}^{T}\varepsilon_{t-1}+\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{t-1}^2\right)$$

$$=\sigma^2\int_0^1 B_T^2\left(r\right)dr+o_p\left(1\right)\Rightarrow\sigma^2\int_0^1 B^2\left(r\right)dr$$

$$\frac{1}{T}\sum_{t=1}^{T}y_{t-1}\left(y_t-y_{t-1}\right)=\frac{1}{T}\sum_{t=1}^{T}\left(y_0+S_{v(t-1)}\right)v_t$$

$$=y_0\overline{v}+\frac{1}{2T}\sum_{t=1}^{T}\left(S_{vt}^2-S_{v(t-1)}^2-v_t^2\right)=y_0\overline{v}+\frac{1}{2T}S_{vT}^2-\frac{1}{2T}\sum_{t=1}^{T}v_t^2$$

$$=y_0\overline{v}+\frac{\sigma^2}{2}\left(B_T\left(1\right)+\frac{\varepsilon_T}{\sigma\sqrt{T}}\right)^2-\frac{1}{2T}\sum_{t=1}^{T}v_t^2$$

$$=y_0\overline{v}+\frac{\sigma^2}{2}B_T^2\left(1\right)-\frac{1}{2T}\sum_{t=1}^{T}v_t^2+o_p\left(1\right)\Rightarrow\frac{\sigma^2}{2}B^2\left(1\right)-\frac{\sigma_v^2}{2}$$

Thus, $T\left(\widehat{\beta}-1\right)=\dfrac{\dfrac{1}{T}\sum_{t=2}^{T}v_ty_{t-1}}{\dfrac{1}{T^2}\sum_{t=2}^{T}y_{t-1}^2}\Rightarrow\dfrac{B^2\left(1\right)-\dfrac{\sigma_v^2}{\sigma^2}}{2\displaystyle\int_0^1 B^2}$. ■

Proof of Theorem 53:

$$\frac{1}{T^2}\sum_{t=1}^{T}y_{t-1}^2=\frac{1}{T^2}\sum_{t=1}^{T}\left(y_0+\varepsilon_0+\sum_{i=1}^{t-1}v_i\right)^2$$

$$=\frac{1}{T^2}\sum_{t=1}^{T}\left(S_{v(t-1)}^2+2\left(y_0+\varepsilon_0\right)S_{v(t-1)}+\left(y_0+\varepsilon_0\right)^2\right)$$

$$=\sigma_V^2\sum_{t=1}^{T}\int_{(t-1)/T}^{t/T}\frac{S_{v[Tr]}^2}{T\sigma_V^2}dr+\frac{2}{\sqrt{T}}\sigma_V\left(y_0+\varepsilon_0\right)\sum_{t=1}^{T}\int_{(t-1)/T}^{t/T}\frac{S_{v[Tr]}}{\sqrt{T}\sigma_V}dr+\frac{\left(y_0+\varepsilon_0\right)^2}{T}$$

$$= \sigma_V^2 \int_0^1 B_{vT}^2(r)\,dr + \frac{2}{\sqrt{T}}\sigma_V (y_0+\varepsilon_0)\int_0^1 B_{vT}(r)\,dr + \frac{1}{T}(y_0+\varepsilon_0)^2$$
$$= \sigma_V^2 \int_0^1 B_{vT}^2(r)\,dr + o_p(1) \Rightarrow \sigma_V^2 \int_0^1 B_v^2(r)\,dr \text{ as } T\to\infty$$

$$\frac{1}{T}\sum_{t=1}^{T} y_{t-1}(y_t - y_{t-1}) = \frac{1}{T}\sum_{t=1}^{T}\left(S_{v(t-1)} + y_0 + \varepsilon_0\right) v_t$$
$$= (y_0+\varepsilon_0)\overline{v} + \frac{1}{2T}\sum_{t=1}^{T}\left(S_{vt}^2 - S_{v(t-1)}^2 - v_t^2\right)$$
$$= \frac{1}{2T}\sum_{t=1}^{T}\left(S_{vt}^2 - S_{v(t-1)}^2\right) - \frac{1}{2T}\sum_{t=1}^{T} v_t^2 + o_p(1)$$
$$= \frac{1}{2T}S_{vT}^2 - \frac{1}{2T}\sum_{t=1}^{T}(u_t+\omega_t)^2 + o_p(1)$$
$$= \frac{\sigma_V^2}{2}B_{vT}^2(1) - \frac{1}{2T}\sum_{t=1}^{T}u_t^2 - \frac{1}{2T}\sum_{t=1}^{T}\omega_t^2 + o_p(1)$$
$$\Rightarrow \frac{\sigma_V^2}{2}B_v^2(1) - \frac{\sigma_u^2}{2} - \frac{\sigma_\omega^2}{2}$$
$$\text{Thus, } T\left(\widehat{\beta}-1\right) = \frac{\frac{1}{T}\sum_{t=2}^{T} v_t y_{t-1}}{\frac{1}{T^2}\sum_{t=2}^{T} y_{t-1}^2} \Rightarrow \frac{B_v^2(1) - \frac{\sigma_u^2+\sigma_\omega^2}{\sigma_V^2}}{2\int_0^1 B_v^2}. \quad \blacksquare$$

Proof of Theorem 54:
$$\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}(y_t - y_{t-1}) = \frac{1}{T^2}\sum_{t=1}^{T}\left(y_0 + S_{v(t-1)}\right) v_t$$
$$= \tfrac{1}{T}y_0\overline{v} + \frac{1}{2T^2}\sum_{t=1}^{T}\left(S_{vt}^2 - S_{v(t-1)}^2 - v_t^2\right)$$
$$= o_p(1) + \frac{1}{2T^2}S_{vT}^2 - \frac{1}{2T^2}\sum_{t=1}^{T}(u_t + \varepsilon_t - \varepsilon_{t-1})^2$$
$$= -\frac{1}{2T^2}\sum_{t=1}^{T}(u_t + \omega_t - 2\varepsilon_{t-1})^2 + o_p(1)$$
$$= -\frac{2}{T^2}\sum_{t=1}^{T}\omega_t\varepsilon_{t-1} - \frac{2}{T^2}\sum_{t=1}^{T}\varepsilon_{t-1}^2 + o_p(1)$$
$$= -\frac{2}{T^2}\sum_{t=1}^{T}\varepsilon_{t-1}^2 + o_p(1) \Rightarrow -2\sigma_\Lambda^2\int_0^1 B_\eta^2(r)\,dr$$

$$\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2 = \frac{1}{T^2}\sum_{t=1}^{T}\left(y_0 + \sum_{i=1}^{t-1}(\eta(i,t) + u_i)\right)^2$$

$$= \frac{1}{T^2}\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}\left(\eta\left(i,t\right)+u_i\right)\right)^2 + o_p\left(1\right)$$

$$\Rightarrow \int_0^1 \left(\sigma_\Lambda B_\eta\left(r\right)+\sigma B\left(r\right)\right)^2 dr$$

$$\text{Thus, } \widehat{\beta} = 1 + \frac{\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}\left(y_t - y_{t-1}\right)}{\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2} \Rightarrow 1 - \frac{2\int_0^1 B_\eta^2\left(r\right)dr}{\int_0^1 \left(B_\eta\left(r\right)+\frac{\sigma}{\sigma_\Lambda}B\left(r\right)\right)^2 dr}.$$

■

Proof of Lemma 1:

(a)See Herrndorf (1984, Corollary 1, p.142).

(b) and (c), see Phillips (1987, Theorem 3.1(a),(b), p.282).

Proof of Lemma 2:

(a) Obvious.

(b) Let $\gamma_v = \min\left\{\gamma, \gamma_\varepsilon\right\} > 2$, then

$\limsup_t \left\|v_t\right\|_{\gamma_v} = \limsup_t \left\|u_t + \varepsilon_t - \varepsilon_{t-1}\right\|_{\gamma_v}$

$\leq \limsup_t \left(\left\|u_t\right\|_{\gamma_v} + \left\|\varepsilon_t\right\|_{\gamma_v} + \left\|\varepsilon_{t-1}\right\|_{\gamma_v}\right)$ by Minkowski's inequality

$\leq \limsup_t \left(\left\|u_t\right\|_{\gamma} + \left\|\varepsilon_t\right\|_{\gamma_\varepsilon} + \left\|\varepsilon_{t-1}\right\|_{\gamma_\varepsilon}\right)$ by Liapunov's inequality

$\leq \limsup_t \left\|u_t\right\|_{\gamma} + \limsup_t \left\|\varepsilon_t\right\|_{\gamma_\varepsilon} + \limsup_t \left\|\varepsilon_{t-1}\right\|_{\gamma_\varepsilon}$

$< \infty$ by Assumptions 1(b) and 2(b).

(c) See White (1984, Theorem 3.49, p.47).

$(d)\, \sigma_v^2 \overset{def}{=} \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} E\left(v_t^2\right) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} E\left(u_t+\varepsilon_t-\varepsilon_{t-1}\right)^2$

$= \sigma_u^2 + 2\sigma_\varepsilon^2 - \lim_{T\to\infty}\frac{2}{T}\sum_{t=1}^{T} E\left(\varepsilon_t\varepsilon_{t-1}\right)$

$(e)\, \sigma_V^2 = \lim_{T\to\infty} E\left(\frac{1}{T}S_{vT}^2\right) = \lim_{T\to\infty}\frac{1}{T}E\left(S_T+\varepsilon_T\right)^2$

$= \lim_{T\to\infty}\frac{1}{T}E\left(S_T^2\right) + \lim_{T\to\infty}\frac{2}{T}E\left(S_T\varepsilon_T\right) + \lim_{T\to\infty}\frac{1}{T}E\left(\varepsilon_T^2\right)$

$= \lim_{T\to\infty} E\left(\frac{1}{T}S_T^2\right) = \sigma^2.$

(f)As $T\to\infty$, $B_{vT}\left(r\right) \overset{def}{=} \frac{1}{\sqrt{T}\sigma_V}S_{v[Tr]} = \frac{\sigma}{\sigma_V}\frac{1}{\sqrt{T}\sigma}S_{[Tr]} + \frac{\varepsilon_{[Tr]}}{\sqrt{T}\sigma_V}$

$= \frac{\sigma}{\sigma_V}B_T\left(r\right) + o_p\left(1\right) \Rightarrow B\left(r\right).$

Proof of Lemma 3:

(a)Obvious.

(b) Let $\gamma_v = \min\{\gamma, \gamma_\omega\} > 2$, then $\limsup_t \|v_t\|_{\gamma_v} = \limsup_t \|u_t + \omega_t\|_{\gamma_v}$

$\leq \limsup_t \left(\|u_t\|_{\gamma_v} + \|\omega_t\|_{\gamma_v}\right)$ by Minkowski's inequality

$\leq \limsup_t \left(\|u_t\|_{\gamma} + \|\omega_t\|_{\gamma_\omega}\right)$ by Liapunov's inequality

$\leq \limsup_t \|u_t\|_{\gamma} + \limsup_t \|\omega_t\|_{\gamma_\omega} < \infty$ by Assumptions 1(b) and 3(b).

(c)See White (1984, Theorem 3.49, p.47).

$(d)\ \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E(v_t^2) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E(u_t + \omega_t)^2$

$= \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E(u_t^2) + \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E(\omega_t^2) + \lim_{T\to\infty} \frac{2}{T}\sum_{t=1}^{T} E(u_t)\, E(\omega_t)$

$= \sigma_u^2 + \sigma_\omega^2.$

$(e)\ \sigma_V^2 = \lim_{T\to\infty} E\left(\frac{1}{T} S_{vT}^2\right) = \lim_{T\to\infty} \frac{1}{T} E(S_T + \varepsilon_T)^2$

$= \lim_{T\to\infty} \frac{1}{T} E(S_T + S_{\omega T} + \varepsilon_0)^2 = \lim_{T\to\infty} \frac{1}{T} E(S_T^2 + S_{\omega T}^2)$

$= \sigma^2 + \sigma_\Omega^2$ which exists and is strictly positive.

(f)See Herrndorf (1984, Corollary 1, p.142).

Proof of Lemma 4:

$(a)\ E(\eta(i,t)) = (-1)^{t-i} E(\omega_i) = 0$ for all i, t.

$(b)\ \limsup_t \|\eta(i,t)\|_{\gamma_\eta} = \limsup_t \|\omega_t\|_{\gamma_\omega} < \infty$, for some $\gamma_\eta > 2$.

(c)See White (1984, Theorem 3.49, p.47).

$(d)\ \sigma_\eta^2 \overset{def}{=} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E(\eta_t^2) = \sigma_\omega^2.$

$(e)\ \sigma_\Lambda^2 \overset{def}{=} \lim_{T\to\infty} E\left(\frac{1}{T} S_{\eta T}^2\right)$

$= \lim_{T\to\infty} E\left(\frac{1}{T} S_{\omega T}^2\right)$ which exists and is strictly positive.

(f)See Herrndorf (1984, Corollary 1, p.142).

****

ECO5120, Econometrics II, HANDOUT 3

Prof. T.L. Chong

Fall 98

ASYMPTOTIC PROPERTIES OF EXTREMUM ESTIMATOR

Definition 179 *A sequence of function* $g_T(\theta)$ *converge to* $g(\theta)$ ***pointwise*** *in* Θ*, if for any given* $\theta \in \Theta$,

$$|g_T(\theta) - g(\theta)| = o(1)$$

Definition 180 *A sequence of function* $g_T(\theta)$ *converge to* $g(\theta)$ ***uniformly*** *in* Θ *if*

$$\sup_{\theta \in \Theta} |g_T(\theta) - g(\theta)| = o(1)$$

Uniform convergence implies pointwise convergence, but the not the other way around.

Example 181 $\Theta = [0, 2]$,

$$\begin{aligned} g_T(\theta) &= T\theta \qquad \theta \in \left[0, \frac{1}{2T}\right] \\ &= 1 - T\theta \qquad \theta \in (\frac{1}{2T}, \frac{1}{T}] \\ &= 0 \qquad \theta \in (\frac{1}{T}, 1] \\ &= \frac{T}{T+1}(\theta - 1) \qquad \theta \in (1, 1.5] \\ &= \frac{T}{T+1}(2 - \theta) \qquad \theta \in (1.5, 2] \end{aligned}$$

$$\begin{aligned} g(\theta) &= 0 \qquad \theta \in [0,1] \\ &= \frac{1}{2}(\theta - 1) \qquad \theta \in (1, 1.5] \\ &= \frac{1}{2}(2 - \theta) \qquad \theta \in (1.5, 2] \end{aligned}$$

For any given $\theta \in [0,2]$,

$$|g_T(\theta) - g(\theta)| = o(1)$$

But

$$\sup_{\theta \in \Theta} |g_T(\theta) - g(\theta)| = \frac{1}{2} \neq o(1).$$

Definition 182 *A sequence of random variable $g_T(\theta)$ converge to 0 in probability* ***pointwise*** *if for any given $\theta \in \Theta$,*

$$|g_T(\theta) - g(\theta)| = o_p(1)$$

Definition 183 *A sequence of random variable $g_T(\theta)$ converge to 0 in probability* ***uniformly*** *if*

$$\sup_{\theta \in \Theta} |g_T(\theta) - g(\theta)| = o_p(1)$$

Uniform convergence in probability implies pointwise convergence in probability, but the not the other way around.

Example 184 $\Theta = [0, \infty)$,

$$\begin{aligned} g_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t + ZT\theta \qquad \theta \in \left[0, \frac{1}{2T}\right] \\ &= \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t + Z(1 - T\theta) \qquad \theta \in (\frac{1}{2T}, \frac{1}{T}] \\ &= 0 \qquad \theta \in \left(\frac{1}{T}, \infty\right) \end{aligned}$$

where $\{\varepsilon_t\}_{t=1}^T$ is an i.i.d. zero-mean, finite variance stochastic sequence, Z is a binary random variable which takes -1 with probability 0.5 and 1 with probability 0.5.

$$g(\theta) = 0 \qquad \theta \in \Theta$$

For any given $\theta \in \Theta$,

$$|g_T(\theta) - g(\theta)| = o_p(1)$$

However

$$\sup_{\theta \in \Theta} |g_T(\theta) - g(\theta)| = \left| \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t + \frac{Z}{2} \right| \neq o_p(1)$$

since

$\Pr\left(\sup_{\theta \in \Theta} |g_T(\theta) - g(\theta)| \geq \epsilon\right) = \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t + \frac{Z}{2}\right| \geq \epsilon\right) \to \Pr\left(\left|\frac{Z}{2}\right| \geq \epsilon\right) =$ 1.

Definition 185 *Let $y = (y_1, y_{2...}, y_T)'$ be a T-vector of random variables and θ a K-vector of parameters. An extremum estimator $\widehat{\theta}_T$ is an estimator obtained by maximizing (or minimizing) a certain function $Q_T(y;\theta)$ defined over the parameter space Θ.*

$$\widehat{\theta}_T \overset{def}{=} Arg \max_{\theta \in \Theta} Q_T(y;\theta)$$

Such an estimator is sometimes referred to an M-estimator.

Consistency

Assumptions:

(A1) The parameter space Θ is a compact subset of R^K.

(B1) $Q_T(y;\theta)$ is continuous in $\theta \in \Theta$ for all y and is a measurable function of y for all $\theta \in \Theta$.

(C1) $\frac{1}{T} Q_T(y;\theta)$ converges to a non-stochastic function $Q(\theta)$ in probability uniformly in $\theta \in \Theta$ as $T \to \infty$, and $Q(\theta)$ attains a unique global maximum at θ_0. i.e.

$$\sup_{\theta\in\Theta}\left|\frac{1}{T}Q_T(y;\theta)-Q(\theta)\right|=o_p(1)$$

and

$$\theta_0 \overset{def}{=} Arg\max_{\theta\in\Theta} Q(\theta) \qquad \text{is unique}$$

Theorem 186 *Under assumption (A1) to (C1), we have*

$$\widehat{\theta}_T \xrightarrow{p} \theta_0$$

Proof. Let N be an open neighborhood in R^K containing θ_0. Then $N^c \cap \Theta$ is compact. Therefore $\max_{\theta\in N^c\cap\Theta} Q_T(y;\theta)$ exist. Denote

$$\epsilon = Q(\theta_0) - \max_{\theta\in N^c\cap\Theta} Q(\theta) > 0$$

Let A_T be the event

$$\left|\frac{1}{T}Q_T(y;\theta)-Q(\theta)\right|<\frac{\epsilon}{2} \qquad \forall\theta\in\Theta$$

Since $\widehat{\theta}_T$ and $\theta_0 \in \Theta$, we have

$$Q(\theta_0)-\frac{1}{T}Q_T(y;\theta_0)<\frac{\epsilon}{2}$$

and

$$\frac{1}{T}Q_T\left(y;\widehat{\theta}_T\right)-Q\left(\widehat{\theta}_T\right)<\frac{\epsilon}{2}$$

Summing up of the above inequalities gives

$$\frac{1}{T}Q_T\left(y;\widehat{\theta}_T\right)-Q\left(\widehat{\theta}_T\right)+Q(\theta_0)-\frac{1}{T}Q_T(y;\theta_0)<\epsilon$$

Since $Q_T\left(y;\widehat{\theta}_T\right)\geq Q_T(y;\theta_0)$, we have

$$Q(\theta_0)-Q\left(\widehat{\theta}_T\right)<\epsilon=Q(\theta_0)-\max_{\theta\in N^c\cap\Theta}Q(\theta)$$

This implies

$$Q\left(\widehat{\theta}_T\right) > \max_{\theta \in N^c \cap \Theta} Q\left(\theta\right)$$

which in turn implies

$$\widehat{\theta}_T \notin N^c \cap \Theta$$

or

$$\widehat{\theta}_T \in N$$

Thus, $A_T \Rightarrow \widehat{\theta}_T \in N$

$$\Pr\left(A_T\right) \leq \Pr\left(\widehat{\theta}_T \in N\right)$$

Taking limit and using assumption (C),

$$1 = \lim_{T \to \infty} \Pr\left(A_T\right) \leq \lim_{T \to \infty} \Pr\left(\widehat{\theta}_T \in N\right) \leq 1$$

Thus,

$$\lim_{T \to \infty} \Pr\left(\widehat{\theta}_T \in N\right) = 1$$

Thus, $\widehat{\theta}_T$ converges to θ_0 in probability. ■

Asymptotic Normality

Assumptions:

(A2) The parameter space Θ is an open subset of R^K. θ_0 belongs to the interior of Θ.

(B2) $Q_T\left(y;\theta\right)$ is continuous in an open neighborhood $N_1\left(\theta_0\right)$ of θ_0 for all y and is a measurable function of y for all $\theta \in \Theta$.

(C2) $\frac{1}{T}Q_T\left(y;\theta\right)$ converges to a non-stochastic function $Q\left(\theta\right)$ in probability uniformly in an open neighborhood $N_2\left(\theta_0\right)$ of θ_0 as $T \to \infty$, and $Q\left(\theta\right)$ attains a strict local maximum at θ_0.

(D2) $\dfrac{\partial Q_T\left(y,\theta\right)}{\partial \theta}$ exists and is continuous in an open neighborhood $N_1\left(\theta_0\right)$of θ_0.

(E2) $\dfrac{\partial^2 Q_T(y;\theta)}{\partial\theta\partial\theta'}$ exists and is continuous in an open, convex neighborhood of θ_0.

(F2) $\dfrac{1}{T}\left(\dfrac{\partial^2 Q_T(y;\theta)}{\partial\theta\partial\theta'}\right)_{\theta_T^*} \xrightarrow{p} A(\theta_0)$ for any sequence $\theta_T^* \xrightarrow{p} \theta_0$, where $A(\theta_0)$ is a finite non-singular matrix defined as

$$A(\theta_0) = \lim_{T\to\infty}\frac{1}{T}E\left(\frac{\partial^2 Q_T(y;\theta)}{\partial\theta\partial\theta'}\right)_{\theta_0}$$

(G2) $\dfrac{1}{\sqrt{T}}\left(\dfrac{\partial Q_T(y,\theta)}{\partial\theta}\right)_{\theta_0} \xrightarrow{d} N(0, B(\theta_0))$, where

$$B(\theta_0) = \lim_{T\to\infty}\frac{1}{T}E\left[\left(\frac{\partial Q_T(y,\theta)}{\partial\theta}\right)_{\theta_0}\left(\frac{\partial Q_T(y,\theta)}{\partial\theta'}\right)_{\theta_0}\right]$$

Theorem 187 *Let Θ_T be the set of roots of the equation $\dfrac{\partial Q_T(y,\theta)}{\partial\theta} = 0$ corresponding to the local maxima, and $\left\{\widehat{\theta}_T\right\}$ a sequence obtained by choosing one element from Θ_T. Then under assumptions (A2) to (G2), we have*

$$\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) \xrightarrow{d} N\left(0, A(\theta_0)^{-1} B(\theta_0) A(\theta_0)^{-1}\right)$$

Proof. Taking a Taylor expansion, we have

$$0 = \left(\frac{\partial Q_T(y,\theta)}{\partial\theta}\right)_{\widehat{\theta}_T} = \left(\frac{\partial Q_T(y,\theta)}{\partial\theta}\right)_{\theta_0} + \left(\frac{\partial^2 Q_T(y;\theta)}{\partial\theta\partial\theta'}\right)_{\theta_T^*}\left(\widehat{\theta}_T - \theta_0\right)$$

$$\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) = -\left(\frac{1}{T}\frac{\partial^2 Q_T(y;\theta)}{\partial\theta\partial\theta'}\right)_{\theta_T^*}^{-1}\left(\frac{1}{\sqrt{T}}\frac{\partial Q_T(y,\theta)}{\partial\theta}\right)_{\theta_0}$$

Using assumption (F2), (G2) and Theorem 27, we prove the above theorem. ■

Maximum Likelihood Estimation

Let $\{y_t\}_{i=1}^T$ be i.i.d. r.v. with joint density $f(y_1, y_{2...}, y_T;\theta)$, where $\theta = (\theta_1, \theta_2, ...\theta_K)'$. Since the sample values have been observed and therefore fixed number, we regard $f(y_t;\theta)$ as a function of θ.

Definition 188 *Let $y = (y_1, y_{2...}, y_T)'$, we defined the **likelihood function** as*

$$L(y;\theta) = f(y_1, y_2, ..., y_T;\theta) = \prod_{t=1}^{T} f(y_t;\theta).$$

Definition 189 *The scores S is a K by 1 vector defined as*

$$S = \frac{\partial}{\partial\theta} \ln L(y;\theta) = \sum_{t=1}^{T} \frac{\partial}{\partial\theta} \ln f(y_t;\theta)$$

Theorem 190 *The scores have zero expectation when the density for y_t is correctly specified.*

Proof.

$E(S) = E\left[\frac{\partial}{\partial\theta} \ln L(y;\theta)\right] = E\left[\frac{1}{L(y;\theta)} \frac{\partial}{\partial\theta} L(y;\theta)\right]$

$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\frac{1}{L(y;\theta)} \frac{\partial}{\partial\theta} L(y;\theta)\right] L(y;\theta)\, dy_1 \cdots dy_T$

$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} L(y;\theta)\, dy_1 \cdots dy_T$

$= \frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L(y;\theta)\, dy_1 \cdots dy_T$

$= \frac{\partial}{\partial\theta}[1] = 0.$ ■

Note that if the density is misspecified, say suppose the true joint density is $g(y_1, y_{2...}, y_T;\theta)$, then the expectation of score will not be zero in general since

$$E(S) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\frac{1}{L(y;\theta)} \frac{\partial}{\partial\theta} L(y;\theta)\right] g(y_1, y_{2...}, y_T;\theta)\, dy_1 \cdots dy_T \neq 0.$$

Definition 191 *Fisher's Information Matrix is the variance-covariance matrix of the scores for θ and it equals*

$$I_{\theta\theta} = \sum_{t=1}^{T} E\left[\left(\frac{\partial}{\partial\theta} \ln f(y_t;\theta)\right) \frac{\partial}{\partial\theta'} \ln f(y_t;\theta)\right]$$

To show this, note that

$I_{\theta\theta} = E(S - E(S))(S - E(S))' = E(SS')$

$= E\left[\left(\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\ln f\left(y_t;\theta\right)\right)\left(\sum_{t=1}^{T}\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right)\right]$

Since

$E\left[\left(\frac{\partial}{\partial\theta}\ln f\left(y_i;\theta\right)\right)\left(\frac{\partial}{\partial\theta'}\ln f\left(y_j;\theta\right)\right)\right]=0$

for all $i\neq j$ by independence.

We have

$I_{\theta\theta}=\sum_{t=1}^{T}E\left[\left(\frac{\partial}{\partial\theta}\ln f\left(y_t;\theta\right)\right)\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right].$

Theorem 192 *If $y_t\sim$ i.i.d. with density $f\left(y_t;\theta\right)$, $\widehat{\theta}$ is any unbiased estimator of θ, the minimum variance of $\widehat{\theta}$ that can be attained is $I_{\theta\theta}^{-1}$.*

Proof.

Note that since $\widehat{\theta}$ is unbiased, we have

$$E\left(\widehat{\theta}\right)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}f\left(y_1,y_2,...,y_T;\theta\right)dy_1\cdots dy_T=\theta.$$

Differentiating both sides w.r.t. θ',

$$\frac{\partial}{\partial\theta'}E\left(\widehat{\theta}\right)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\frac{\partial}{\partial\theta'}\prod_{t=1}^{T}f\left(y_t;\theta\right)dy_1\cdots dy_T=I$$

This implies

$$\begin{aligned}
\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\sum_{t=1}^{T}\left(\Pi_{j\neq t}f\left(y_j;\theta\right)\frac{\partial}{\partial\theta'}f\left(y_t;\theta\right)\right)dy_1\cdots dy_T &= I\\
\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\sum_{t=1}^{T}\left(\Pi_{j\neq t}f\left(y_j;\theta\right)f\left(y_t;\theta\right)\frac{1}{f\left(y_t;\theta\right)}\frac{\partial}{\partial\theta'}f\left(y_t;\theta\right)\right)dy_1\cdots dy_T &= I\\
\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\sum_{t=1}^{T}\left(\Pi_{j=1}^{T}f\left(y_j;\theta\right)\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right)dy_1\cdots dy_T &= I\\
\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}\Pi_{j=1}^{T}f\left(y_j;\theta\right)\left(\sum_{t=1}^{T}\frac{\partial}{\partial\theta'}\ln f\left(y_t;\theta\right)\right)dy_1\cdots dy_T &= I\\
\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\widehat{\theta}S'\Pi_{j=1}^{T}f\left(y_j;\theta\right)dy_1\cdots dy_T &= I
\end{aligned}$$

Thus

$$E\left(\widehat{\theta}S'\right)=I$$

Using the fact that

$$E\left(\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)'\right)E\left(SS'\right)-E\left(\widehat{\theta}S'\right)$$

is a positive semi-definite matrix, we have

$$E\left(\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)'\right)I_{\theta\theta}-I$$

is a positive semi-definite matrix, or

$$VarCov\left(\widehat{\theta}\right)-I_{\theta\theta}^{-1}$$

is a positive semi-definite matrix. ■

We call $I_{\theta\theta}^{-1}$ the **Cramér-Rao lower bound** of an unbiased estimator $\widehat{\theta}$.

Theorem 193 $I_{\theta\theta}=-E\left(\dfrac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right)$

Proof. Since

$E\left(S\right)=E\left(\dfrac{\partial}{\partial\theta}\ln L\left(y;\theta\right)\right)$

$=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln L\left(y;\theta\right)\right)f\left(y_1,y_2,...,y_T;\theta\right)dy_1\cdots dy_T=0$

Differentiating both sides w.r.t. θ',

$$\frac{\partial}{\partial\theta'}E\left(S\right)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta'}\left(SL\left(y;\theta\right)\right)dy_1\cdots dy_T=0$$

This implies

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(S\frac{\partial}{\partial\theta'}L\left(y;\theta\right)+L\left(y;\theta\right)\frac{\partial}{\partial\theta'}S\right)dy_1\cdots dy_T=0$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}S\frac{\partial}{\partial\theta'}L\left(y;\theta\right)dy_1\cdots dy_T=-\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right)L\left(y;\theta\right)dy_1\cdots dy_T$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}S\frac{1}{L\left(y;\theta\right)}\left(\frac{\partial}{\partial\theta'}L\left(y;\theta\right)\right)L\left(y;\theta\right)dy_1\cdots dy_T=-E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right)$$

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}S\left(\frac{\partial}{\partial\theta'}\ln L\left(y;\theta\right)\right)L\left(y;\theta\right)dy_1\cdots dy_T=-E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right)$$

$$E\left[SS'\right]=-E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right)$$

$$I_{\theta\theta}=-E\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ln L\left(y;\theta\right)\right)$$

■

Theorem 194 *Under the assumptions (A1) to (C1), let* $Q_T\left(y;\theta\right)=\ln L\left(y;\theta\right)$. *The maximum likelihood estimator* $\widehat{\theta}_T$ *satisfies*

$$\sqrt{T}\left(\widehat{\theta}_T-\theta_0\right)\xrightarrow{d}N\left(0,-\left[\lim_{T\to\infty}TE\left(\frac{\partial^2 L\left(y;\theta\right)}{\partial\theta\partial\theta'}\right)_{\theta_0}\right]^{-1}\right)$$

Proof. (exercise).

Nonlinear Least Squares Estimator

A nonlinear regression model is a model of the form

$$y_t=f_t\left(\beta_0\right)+u_t\qquad t=1,2,...,T.$$

The nonlinear least squares estimator is defined as

$$\widehat{\beta}_T=Arg\min_{\beta}S_T$$

where $S_T=\sum_{t=1}^{T}\left[y_t-f_t\left(\beta\right)\right]^2$

Assumptions:

(A3) $f_t\left(\beta\right)$ is continuous in an open neighborhood N of β_0.

(B3) $\dfrac{\partial f_t\left(\beta\right)}{\partial\beta}$ exists and is continuous in N.

(C3) $\dfrac{1}{T}\sum_{t=1}^{T}f_t\left(\beta_1\right)f_t\left(\beta_2\right)$ converges to a non-stochastic function in probability uniformly in $\beta_1,\beta_2\in N$.

(D3) $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left[f_t(\beta)-f_t(\beta_0)\right]^2\neq 0$ if $\beta\neq\beta_0$.

(E3) $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial f_t(\beta)}{\partial\beta}\right)_{\beta_0}\left(\frac{\partial f_t(\beta)}{\partial\beta'}\right)_{\beta_0}=C$, where C is a finite non-singular matrix.

(F3) $\frac{1}{T}\sum_{t=1}^{T}\frac{\partial f_t(\beta)}{\partial\beta}\frac{\partial f_t(\beta)}{\partial\beta'}$ converges to a finite matrix uniformly for all β in N.

(G3) $\frac{\partial f_t^2(\beta)}{\partial\beta_i\partial\beta_j}$ is continuous in β in N uniformly in t.

(H3) $\lim_{T\to\infty}\frac{1}{T^2}\sum_{t=1}^{T}\left[\frac{\partial f_t^2(\beta)}{\partial\beta_i\partial\beta_j}\right]^2=0$ for all β in N.

(I3) $\frac{1}{T}\sum_{t=1}^{T}f_t(\beta_1)\left(\frac{\partial^2 f_t(\beta)}{\partial\beta\partial\beta'}\right)_{\beta_2}$ converges to a finite matrix uniformly for all β_1 and β_2 in N.

Theorem 195 *Under assumptions (A3) to (D3), we have*

$$\widehat{\beta}_T\xrightarrow{p}\beta_0$$

Proof. Let $h(\beta)=\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left[f_t(\beta_0)-f_t(\beta)\right]^2$

by assumption (C3) and (D3), $h(\beta)$ is a function of β that has a local minimum at β_0 uniformly in β.

$$\begin{aligned}
&\sup_{\beta\in N}\left|\frac{1}{T}S_T-\sigma^2-h(\beta)\right|\\
&\leq\sup_{\beta\in N}\left|\frac{1}{T}\sum_{t=1}^{T}\left[f_t(\beta_0)-f_t(\beta)\right]^2-h(\beta)\right|+\sup_{\beta\in N}\left|\frac{1}{T}\sum_{t=1}^{T}u_t^2-\sigma^2\right|\\
&+\sup_{\beta\in N}\left|\frac{1}{T}\sum_{t=1}^{T}\left[f_t(\beta_0)-f_t(\beta)\right]u_t\right|\\
&\leq\sup_{\beta\in N}\left|\frac{1}{T}\sum_{t=1}^{T}\left[f_t(\beta_0)-f_t(\beta)\right]^2-h(\beta)\right|+\left|\frac{1}{T}\sum_{t=1}^{T}u_t^2-\sigma^2\right|\\
&+\left|\frac{1}{T}\sum_{t=1}^{T}f_t(\beta_0)u_t\right|+\sup_{\beta\in N}\left|\frac{1}{T}\sum_{t=1}^{T}f_t(\beta)u_t\right|\\
&=A_1+A_2+A_3+A_4\xrightarrow{p}0
\end{aligned}$$

Obviously A_1 to A_3 are $o_p(1)$. To show that A_4 also tends to 0 in probability, we partition N into n non-overlapping regions $N_1,...,N_n$. By assumption

(B3), for any $\epsilon > 0$, we can find a sufficiently large n such that for each $i = 1, 2, ..., n$

$$|f_t(\beta_1) - f_t(\beta_2)| < \frac{\epsilon}{2\sqrt{\sigma^2 + 1}}$$

for $\beta_1, \beta_2 \in N_i$ and for all t.

Thus

$$\sup_{\beta \in N_i} \left| \frac{1}{T} \sum_{t=1}^{T} f_t(\beta) u_t \right|$$

$$= \sup_{\beta \in N_i} \frac{1}{T} \left| \sum_{t=1}^{T} \left(f_t(\beta_i) u_t + f_t(\beta) - f_t(\beta_i) \right) u_t \right|$$

$$\leq \frac{1}{T} \left| \sum_{t=1}^{T} f_t(\beta_i) u_t \right| + \sup_{\beta \in N_i} \frac{1}{T} \left| \sum_{t=1}^{T} \left(f_t(\beta) - f_t(\beta_i) \right) u_t \right|$$

by Triangle inequality

$$\leq \frac{1}{T} \left| \sum_{t=1}^{T} f_t(\beta_i) u_t \right| + \sqrt{\frac{1}{T} \sum_{t=1}^{T} u_t^2} \sup_{\beta \in N_i} \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left[f_t(\beta) - f_t(\beta_i) \right]^2}$$

by Cauchy-Schwartz inequality

$$\leq \frac{1}{T} \left| \sum_{t=1}^{T} f_t(\beta_i) u_t \right| + \sqrt{\frac{1}{T} \sum_{t=1}^{T} u_t^2} \sup_{\beta \in N_i} \frac{1}{T} \sum_{t=1}^{T} |f_t(\beta) - f_t(\beta_i)|$$

$$\leq \frac{1}{T} \left| \sum_{t=1}^{T} f_t(\beta_i) u_t \right| + \sqrt{\frac{1}{T} \sum_{t=1}^{T} u_t^2} \frac{\epsilon}{2\sqrt{\sigma^2 + 1}}$$

Thus

$$\Pr(A_4 > \epsilon) \leq \sum_{i=1}^{N} \Pr\left(\sup_{\beta \in N_i} \left| \frac{1}{T} \sum_{t=1}^{T} f_t(\beta) u_t \right| > \epsilon \right)$$

$$\leq \sum_{i=1}^{N} \Pr\left(\frac{1}{T} \left| \sum_{t=1}^{T} f_t(\beta_i) u_t \right| + \sqrt{\frac{1}{T} \sum_{t=1}^{T} u_t^2} \frac{\epsilon}{2\sqrt{\sigma^2 + 1}} > \epsilon \right)$$

$$\leq \sum_{i=1}^{N} \Pr\left(\frac{1}{T} \left| \sum_{t=1}^{T} f_t(\beta_i) u_t \right| + \sqrt{\frac{1}{T} \sum_{t=1}^{T} u_t^2} \frac{\epsilon}{2\sqrt{\sigma^2 + 1}} > \epsilon \cap \sqrt{\frac{1}{T} \sum_{t=1}^{T} u_t^2} \frac{\epsilon}{2\sqrt{\sigma^2 + 1}} \leq \tfrac{\epsilon}{2} \right)$$

$$+ \sum_{i=1}^{N} \Pr\left(\frac{1}{T} \left| \sum_{t=1}^{T} f_t(\beta_i) u_t \right| + \sqrt{\frac{1}{T} \sum_{t=1}^{T} u_t^2} \frac{\epsilon}{2\sqrt{\sigma^2 + 1}} > \epsilon \cap \sqrt{\frac{1}{T} \sum_{t=1}^{T} u_t^2} \frac{\epsilon}{2\sqrt{\sigma^2 + 1}} > \tfrac{\epsilon}{2} \right)$$

$$\leq \sum_{i=1}^{N} \Pr\left(\frac{1}{T} \left| \sum_{t=1}^{T} f_t(\beta_i) u_t \right| > \tfrac{\epsilon}{2} \right) + \sum_{i=1}^{N} \Pr\left(\sqrt{\frac{1}{T} \sum_{t=1}^{T} u_t^2} \frac{\epsilon}{2\sqrt{\sigma^2 + 1}} > \tfrac{\epsilon}{2} \right)$$

$$\leq \sum_{i=1}^{N} \Pr\left(\frac{1}{T} \left| \sum_{t=1}^{T} f_t(\beta_i) u_t \right| > \tfrac{\epsilon}{2} \right) + n \Pr\left(\frac{1}{T} \sum_{t=1}^{T} u_t^2 > \sigma^2 + 1 \right)$$

$\to 0$ ■

Theorem 196 *Under assumptions (A3) to (I3), we have*

$$\sqrt{T}\left(\widehat{\beta}_T - \beta_0\right) \xrightarrow{d} N\left(0, \sigma^2 C^{-1}\right)$$

Proof. (exercise).

Questions:

1-6. Greene, Chapter 4, Exercises 7-9, 11, 16-17.

7. Consider the following model.

$$y_t = \alpha x_t^{\beta} + u_t$$

$t = 1, 2, ..., T.$, $x_t \sim U(0,1)$, $u_t \sim N(0,1)$, $\alpha = 2$, $\beta = 2$, x and u are independent.

a) Show that the nonlinear least squares estimators of α and β are consistent.

b) Show that the nonlinear least squares estimators of α and β are asymptotically normal.

c) Use GAUSS to simulate the sampling distribution of the nonlinear least squares estimators for T=50, 100, 1000, using 20000 replications.

Proof.

(i) Since $X_n \xrightarrow{d} X$, we have $X_n + c \xrightarrow{d} X + c$ by Theorem 25. Further, $X_n + Y_n - (X_n + c) = Y_n - c \xrightarrow{p} 0$. From Theorem 27, $X_n + Y_n$ and $X_n + c$ have the same limiting distribution which is $X + c$.

(ii) Since $X_n \xrightarrow{d} X$, we have $cX_n \xrightarrow{d} cX$ by Theorem 25. Further, and $X_n Y_n - cX_n = X_n (Y_n - c) \xrightarrow{p} 0$ by Theorem 28. From Theorem 27, $X_n Y_n$ and cX_n have the same limiting distribution which is cX.

(iii) Since $X_n \xrightarrow{d} X$, we have $\dfrac{X_n}{c} \xrightarrow{d} \dfrac{X}{c}$ by Theorem 25. Further, and $\dfrac{X_n}{Y_n} - \dfrac{X_n}{c} = X_n \left(\dfrac{1}{Y_n} - \dfrac{1}{c} \right) \xrightarrow{p} 0$ by Theorem 20 and Theorem 28. From Theorem 27, $\dfrac{X_n}{Y_n}$ and $\dfrac{X_n}{c}$ have the same limiting distribution which is $\dfrac{X}{c}$.

■

Applying theorem 9 with $k = 2$ yields the expansion

#####

Theorem 197 *If* $E\left(|X|^k\right) < \infty$, *then*

$$\left| \Phi_X(t) - \sum_{j=0}^{k} \frac{(it)^j}{j!} E\left(X^j\right) \right| \leq \min \left\{ E\left(\frac{2|tX|^k}{k!} \right), E\left(\frac{|tX|^{k+1}}{(k+1)!} \right) \right\}.$$

Proof. A function f that is differentiable k times has the expansion

$$f(t) = \sum_{j=0}^{k-1} \frac{(t)^j}{j!} f^{(j)}(0) + f^{(k)}(\alpha t) \frac{(t)^k}{k!}$$

where $f^{(j)}$ is the j'th derivative of f and $0 \leq \alpha \leq 1$.

The expansion of $f(t) = e^{itx}$ gives

$$e^{itx} = \sum_{j=0}^{k-1} \frac{(itx)^j}{j!} + \frac{(itx)^k}{k!} e^{i\alpha tx} = \sum_{j=0}^{k} \frac{(itx)^j}{j!} + \frac{(itx)^k}{k!} \left(e^{i\alpha tx} - 1 \right)$$

or alternatively
$e^{itx} = \sum_{j=0}^{k} \frac{(itx)^j}{j!} + \frac{(itx)^{k+1}}{(k+1)!} e^{i\alpha' tx}$
where $0 \leq \alpha \leq 1$, $0 \leq \alpha' \leq 1$,
$|e^{i\alpha tx} - 1| = \sqrt{(\cos \alpha tx + i \sin \alpha tx - 1)(\cos \alpha tx - i \sin \alpha tx - 1)}$
$= \sqrt{2 - 2\cos \alpha tx} \leq 2$, and $\left|e^{i\alpha' tx}\right| = 1$.
Thus, we can write

$$\left| e^{itx} - \sum_{j=0}^{k} \frac{(itx)^j}{j!} \right| = \left| \frac{(itx)^k}{k!} \left(e^{i\alpha tx} - 1 \right) \right| \leq 2 \frac{|tx|^k}{k!}$$

and

$$\left| e^{itx} - \sum_{j=0}^{k} \frac{(itx)^j}{j!} \right| = \left| \frac{(itx)^{k+1}}{(k+1)!} e^{i\alpha' tx} \right| = \frac{|tx|^{k+1}}{(k+1)!}.$$

Replacing x by the random variable X, taking expectation and using the modulus inequality that $|E(Z)| \leq E|Z|$ for a complex random variable Z, we have
$\left| \Phi_X(t) - \sum_{j=0}^{k} \frac{(it)^j}{j!} E(X^j) \right| = \left| E\left(e^{itX} - \sum_{j=0}^{k} \frac{(itX)^j}{j!} \right) \right|$
$\leq E \left| e^{itX} - \sum_{j=0}^{k} \frac{(itx)^j}{j!} \right| \leq E\left(\frac{2|tX|^k}{k!} \right)$
and
$\left| \Phi_X(t) - \sum_{j=0}^{k} \frac{(it)^j}{j!} E(X^j) \right| \leq E \left| e^{itX} - \sum_{j=0}^{k} \frac{(itx)^j}{j!} \right| = E\left(\frac{|tX|^{k+1}}{(k+1)!} \right)$
Thus
$\left| \Phi_X(t) - \sum_{j=0}^{k} \frac{(it)^j}{j!} E(X^j) \right| \leq \min \left\{ E\left(\frac{2|tX|^k}{k!} \right), E\left(\frac{|tX|^{k+1}}{(k+1)!} \right) \right\}$. ■
Note that there is no need for $E|X|^{k+1}$ to exist for this theorem to hold.
#####
*****

Theorem 198 *If* $E(X^2) < \infty$, *then*

$$\left| \Phi_X(t) - \sum_{j=0}^{2} \frac{(it)^j}{j!} E\left(X^j\right) \right| \leq \min \left\{ E\left(\frac{2|tX|^2}{2!} \right), E\left(\frac{|tX|^3}{3!} \right) \right\}.$$

Proof. A function f that is twice differentiable has the expansion

$$f(t) = \sum_{j=0}^{1} \frac{(t)^j}{j!} f^{(j)}(0) + f^{(2)}(\alpha t) \frac{(t)^2}{2!}$$

where $f^{(j)}$ is the j'th derivative of f and $0 \leq \alpha \leq 1$.

The expansion of $f(t) = e^{itx}$ gives

$e^{itx} = \sum_{j=0}^{1} \frac{(itx)^j}{j!} + \frac{(itx)^2}{2!} e^{i\alpha tx} = \sum_{j=0}^{2} \frac{(itx)^j}{j!} + \frac{(itx)^2}{2!} (e^{i\alpha tx} - 1)$

or alternatively

$e^{itx} = \sum_{j=0}^{2} \frac{(itx)^j}{j!} + \frac{(itx)^3}{3!} e^{i\alpha' tx}$

where $0 \leq \alpha \leq 1$, $0 \leq \alpha' \leq 1$,

$|e^{i\alpha tx} - 1| = \sqrt{(\cos \alpha tx + i \sin \alpha tx - 1)(\cos \alpha tx - i \sin \alpha tx - 1)}$

$= \sqrt{2 - 2\cos \alpha tx} \leq 2$, and $\left|e^{i\alpha' tx}\right| = 1$.

Thus, we can write

$$\left| e^{itx} - \sum_{j=0}^{2} \frac{(itx)^j}{j!} \right| = \left| \frac{(itx)^2}{2!} \left(e^{i\alpha tx} - 1\right) \right| \leq 2 \frac{|tx|^2}{2!}$$

and

$$\left| e^{itx} - \sum_{j=0}^{2} \frac{(itx)^j}{j!} \right| = \left| \frac{(itx)^{2+1}}{(2+1)!} e^{i\alpha' tx} \right| = \frac{|tx|^{2+1}}{(2+1)!}.$$

Replacing x by the random variable X, taking expectation and using the modulus inequality that $|E(Z)| \leq E|Z|$ for a complex random variable Z, we have

$\left| \Phi_X(t) - \sum_{j=0}^{2} \frac{(it)^j}{j!} E(X^j) \right| = \left| E\left(e^{itX} - \sum_{j=0}^{2} \frac{(itX)^j}{j!} \right) \right|$

$\leq E \left| e^{itX} - \sum_{j=0}^{2} \frac{(itx)^j}{j!} \right| \leq E\left(\frac{2|tX|^2}{2!} \right)$

and

$\left| \Phi_X(t) - \sum_{j=0}^{2} \frac{(it)^j}{j!} E(X^j) \right| \leq E \left| e^{itX} - \sum_{j=0}^{2} \frac{(itx)^j}{j!} \right| = E\left(\frac{|tX|^{2+1}}{(2+1)!} \right)$

Thus

$\left| \Phi_X(t) - \sum_{j=0}^{2} \frac{(it)^j}{j!} E(X^j) \right| \leq \min\left\{ E\left(\frac{2|tX|^2}{2!} \right), E\left(\frac{|tX|^3}{3!} \right) \right\}$. ■

Note that there is no need for $E|X|^{2+1}$ to exist for this theorem to hold.

*****

$$\left|\Phi_X\left(\lambda\sigma^{-1}n^{-1/2}\right)-\sum_{j=0}^{2}\frac{(i\lambda)^j}{j!}E\left(\left[n^{-1/2}\left(\left(X_t-\mu\right)/\sigma\right)\right]^j\right)\right|$$

$$\leq \min\left\{E\left(\frac{2\left|\lambda\left(X_t-\mu\right)/\sigma\right|^2}{2!}\right),E\left(\frac{\left|\lambda\left(X_t-\mu\right)/\sigma\right|^3}{3!}\right)\right\}$$

In other words,

$$\left|\Phi_X\left(\lambda\sigma^{-1}n^{-1/2}\right)-1+\frac{\lambda^2}{2n}\right|\leq \min\left\{\frac{\lambda^2}{n},E\frac{\left|\lambda\left(X_t-\mu\right)\right|^3}{6\sigma^3n^{3/2}}\right\}$$

which makes it possible to write, for fixed λ,

$$\Phi_X\left(\lambda\sigma^{-1}n^{-1/2}\right)=1-\frac{\lambda^2}{2n}+O\left(\frac{1}{n^{3/2}}\right)$$

****

Example 199 $\Pr\left(X_n=1\right)=\frac{1}{2}+\frac{1}{n+1}$, $\Pr\left(X_n=2\right)=\frac{1}{2}-\frac{1}{n+1}$. *As* n *goes to infinity, the two probabilities converge to* $\frac{1}{2}$, *since* $\lim_{n\to\infty}\left|\Pr\left(X_n=1\right)-\frac{1}{2}\right|=\lim_{n\to\infty}\frac{1}{n+1}=0$ *and* $\lim_{n\to\infty}\left|\Pr\left(X_n=2\right)-\frac{1}{2}\right|=-\lim_{n\to\infty}\frac{1}{n+1}=0$. *However,* X_n *does not converge to a constant.*

***

Definition 200 *A test of* $H_0:\theta\in\Theta_0$ *against* $H_1:\theta\in\Theta_1$ *as defined by some rejection region* C_1 *is said to be* ***uniformly most powerful*** *(UMP) test of size* α *if*

$(i)\max_{\theta\in\Theta_0}P\left(\theta\right)=\alpha$

$(ii)\,P\left(\theta\right)\geq P^*\left(\theta\right)$ for all $\theta\in\Theta_1$;

where $P^*\left(\theta\right)$ is the power function of any other test of size α.

A test is most powerful if it has greater power than any other test of the same size. It is uniformly most powerful if it has greater power than any other test of the same size for all admissible values of the parameter.

***

Assumptions:

(A) $f(x_t;\theta)$ is continuous in an open neighborhood B of β_0.

(B) $\dfrac{\partial f(x_t;\theta)}{\partial\theta}$ exists and is continuous in B.

(C) $\dfrac{1}{T}\sum_{t=1}^{T} f(x_t;\theta_1) f(x_t;\theta_2)$ converges to a non-stochastic function in probability uniformly in $\theta_1, \theta_2 \in B$.

(D) $\lim_{T\to\infty} \dfrac{1}{T}\sum_{t=1}^{T} [f(x_t;\theta) - f(x_t;\theta_0)]^2 \neq 0$ if $\theta \neq \theta_0$.

(E) $\lim_{T\to\infty} \dfrac{1}{T}\sum_{t=1}^{T} \left(\dfrac{\partial f(x_t;\theta)}{\partial\theta}\right)_{\theta_0} \left(\dfrac{\partial f(x_t;\theta)}{\partial\theta'}\right)_{\theta_0} = C$, where C is a finite non-singular matrix.

(F) $\dfrac{1}{T}\sum_{t=1}^{T} \dfrac{\partial f(x_t;\theta)}{\partial\theta}\dfrac{\partial f(x_t;\theta)}{\partial\theta'}$ converges to a finite matrix uniformly for all θ in B.

(G) $\dfrac{\partial^2 f(x_t;\theta)}{\partial\theta_i\partial\theta_j}$ is continuous in θ in B uniformly in t.

(H) $\lim_{T\to\infty} \dfrac{1}{T^2}\sum_{t=1}^{T} \left[\dfrac{\partial^2 f(x_t;\theta)}{\partial\theta_i\partial\theta_j}\right]^2 = 0$ for all β in B.

(I) $\dfrac{1}{T}\sum_{t=1}^{T} f(x_t;\theta_1) \left(\dfrac{\partial^2 f(x_t;\theta)}{\partial\theta\partial\theta'}\right)_{\theta_2}$ converges to a finite matrix uniformly for all θ_1 and θ_2 in B.

Theorem 201 *Under assumptions (A) to (D), we have*

$$\widehat{\theta}_T \xrightarrow{p} \theta_0$$

Proof. Let $h(\beta) = \lim_{T\to\infty} \dfrac{1}{T}\sum_{t=1}^{T} [f(x_t;\theta_0) - f(x_t;\theta)]^2$

by assumption (C) and (D), $h(\beta)$ is a function of β that has a local minimum at β_0 uniformly in β.

$$\begin{aligned}
&\sup_{\beta\in B} |S_T(\beta) - \sigma^2 - h(\beta)| \\
&\leq \sup_{\beta\in B} \left|\frac{1}{T}\sum_{t=1}^{T} [f(x_t;\theta_0) - f(x_t;\theta)]^2 - h(\beta)\right| + \sup_{\beta\in N} \left|\frac{1}{T}\sum_{t=1}^{T} u_t^2 - \sigma^2\right| \\
&+ \sup_{\beta\in B} \left|\frac{1}{T}\sum_{t=1}^{T} [f(x_t;\theta_0) - f(x_t;\theta)]\, u_t\right| \\
&\leq \sup_{\beta\in B} \left|\frac{1}{T}\sum_{t=1}^{T} [f(x_t;\theta_0) - f(x_t;\theta)]^2 - h(\beta)\right| + \left|\frac{1}{T}\sum_{t=1}^{T} u_t^2 - \sigma^2\right|
\end{aligned}$$

$$+\left|\frac{1}{T}\sum_{t=1}^{T}f\left(x_t;\theta_0\right)u_t\right|+\sup_{\beta\in N}\left|\frac{1}{T}\sum_{t=1}^{T}f\left(x_t;\theta\right)u_t\right|$$
$$=A_1+A_2+A_3+A_4\xrightarrow{p}0$$

Obviously A_1 to A_3 are $o_p\left(1\right)$. To show that A_4 also tends to 0 in probability, we partition B into n non-overlapping regions $B_1,...,B_n$. By assumption (B), for any $\epsilon>0$, we can find a sufficiently large n such that for each $i=1,2,...,n$

$$\left|f\left(x_t;\theta_1\right)-f\left(x_t;\theta_2\right)\right|<\frac{\epsilon}{2\sqrt{\sigma^2+1}}$$

for $\beta_1,\beta_2\in B_i$ and for all t.

Thus

$$\sup_{\beta\in B_i}\left|\frac{1}{T}\sum_{t=1}^{T}f\left(x_t;\theta\right)u_t\right|$$
$$=\sup_{\beta\in B_i}\frac{1}{T}\left|\sum_{t=1}^{T}\left(f\left(x_t;\theta_i\right)u_t+f\left(x_t;\theta\right)-f\left(x_t;\theta_i\right)\right)u_t\right|$$
$$\leq\frac{1}{T}\left|\sum_{t=1}^{T}f\left(x_t;\theta_i\right)u_t\right|+\sup_{\beta\in B_i}\frac{1}{T}\left|\sum_{t=1}^{T}\left(f\left(x_t;\theta\right)-f\left(x_t;\theta_i\right)\right)u_t\right|$$

by Triangle inequality

$$\leq\frac{1}{T}\left|\sum_{t=1}^{T}f\left(x_t;\theta_i\right)u_t\right|+\sqrt{\frac{1}{T}\sum_{t=1}^{T}u_t^2}\sup_{\beta\in B_i}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[f\left(x_t;\theta\right)-f\left(x_t;\theta_i\right)\right]^2}$$

by Cauchy-Schwartz inequality

$$\leq\frac{1}{T}\left|\sum_{t=1}^{T}f\left(x_t;\theta_i\right)u_t\right|+\sqrt{\frac{1}{T}\sum_{t=1}^{T}u_t^2}\sup_{\beta\in N_i}\frac{1}{T}\sum_{t=1}^{T}\left|f\left(x_t;\theta\right)-f\left(x_t;\theta_i\right)\right|$$
$$\leq\frac{1}{T}\left|\sum_{t=1}^{T}f\left(x_t;\theta\right)u_t\right|+\sqrt{\frac{1}{T}\sum_{t=1}^{T}u_t^2}\frac{\epsilon}{2\sqrt{\sigma^2+1}}$$

Thus

$$\Pr\left(A_4>\epsilon\right)\leq\sum_{i=1}^{n}\Pr\left(\sup_{\beta\in N_i}\left|\frac{1}{T}\sum_{t=1}^{T}f\left(x_t;\theta\right)u_t\right|>\epsilon\right)$$
$$\leq\sum_{i=1}^{n}\Pr\left(\frac{1}{T}\left|\sum_{t=1}^{T}f\left(x_t;\theta_i\right)u_t\right|+\sqrt{\frac{1}{T}\sum_{t=1}^{T}u_t^2}\frac{\epsilon}{2\sqrt{\sigma^2+1}}>\epsilon\right)$$
$$\leq\sum_{i=1}^{n}\Pr\left(\frac{1}{T}\left|\sum_{t=1}^{T}f\left(x_t;\theta_i\right)u_t\right|+\sqrt{\frac{1}{T}\sum_{t=1}^{T}u_t^2}\frac{\epsilon}{2\sqrt{\sigma^2+1}}>\epsilon\cap\sqrt{\frac{1}{T}\sum_{t=1}^{T}u_t^2}\frac{\epsilon}{2\sqrt{\sigma^2+1}}\leq\frac{\epsilon}{2}\right)$$
$$+\sum_{i=1}^{n}\Pr\left(\frac{1}{T}\left|\sum_{t=1}^{T}f\left(x_t;\theta_i\right)u_t\right|+\sqrt{\frac{1}{T}\sum_{t=1}^{T}u_t^2}\frac{\epsilon}{2\sqrt{\sigma^2+1}}>\epsilon\cap\sqrt{\frac{1}{T}\sum_{t=1}^{T}u_t^2}\frac{\epsilon}{2\sqrt{\sigma^2+1}}>\frac{\epsilon}{2}\right)$$

$$\leq \sum_{i=1}^{n} \Pr\left(\frac{1}{T}\left|\sum_{t=1}^{T} f(x_t;\theta_i)\, u_t\right| > \tfrac{\epsilon}{2}\right) + \sum_{i=1}^{N} \Pr\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T} u_t^2}\,\frac{\epsilon}{2\sqrt{\sigma^2+1}} > \tfrac{\epsilon}{2}\right)$$
$$\leq \sum_{i=1}^{n} \Pr\left(\frac{1}{T}\left|\sum_{t=1}^{T} f(x_t;\theta_i)\, u_t\right| > \tfrac{\epsilon}{2}\right) + n \Pr\left(\frac{1}{T}\sum_{t=1}^{T} u_t^2 > \sigma^2+1\right)$$
$\to 0$ ■

***

ECO5120

Econometric Theory and Application

1st Term 1999/2000, Final Exam

Answer all questions.

1. Explain why there can be no random variable X for which the moment generating function $M_X(t) = \dfrac{t}{1-t}$.

2. Find the limits of the following sequences as $n \to \infty$:

(a)

$$c_n = \frac{n^2}{2^n};$$

(b)

$$c_1 = \sqrt{1}, c_2 = \sqrt{1+\sqrt{1}}, c_3 = \sqrt{1+\sqrt{1+\sqrt{1}}}, c_4 = ...$$

(16 points)

3. Suppose you know that, conditional upon Z, X is distributed as $N(Z,1)$. If Z is a $U(0,1)$ random variable, find $E(X)$ and $E(X^2)$.

(15 points)

4. Define $X_t = u_t - u_{t-1}$, $\overline{X} = \dfrac{\sum_{t=1}^{T} X_t}{T}$. Find $E\left(\overline{X}\right), Var\left(\overline{X}\right)$ and examine whether the central limit theorem applies to $\overline{X}$ in the following cases:

a) $u_t = u_{t-1} + \varepsilon_t$, where $\{\varepsilon_t\}_{t=0}^{T} \sim i.i.d.\,(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon^2 < \infty$.

b) $\{u_t\}_{t=0}^{T} \sim i.i.d.\,(0, \sigma_u^2)$, $\sigma_u^2 < \infty$.

5. Consider the following density function of a random variable X.

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta} \quad \text{for } 0 < x < \theta; \\ &= 0 \quad \text{elsewhere.} \end{aligned}$$

i) Find the moment generating function of x.

ii) Sketch the graph of $f(x;1)$, $f(x;2)$ and $f(x;3)$.

Let $X_1, X_2, ..., X_T$ constitute a random sample of size T from the above population.

iii) Find the joint density of $X_1, X_2, ..., X_T$.

iv) Find the likelihood function $L(x;\theta)$ and the log-likelihood function $\ln L(x;\theta)$.

v) Find the score $S = \dfrac{\partial}{\partial\theta} \ln L(x;\theta)$, does the score have zero expectation?

vi) Find the ML estimator $\widehat{\theta}$.

vii) Find the Fisher's information matrix using $I(\theta)$ using

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2} \ln L(x;\theta)\right).$$

viii) Suppose we would like to test $H_0 : \theta = 1$ versus $H_1 : \theta > 1$.

Define a Wald test

$$W\left(\widehat{\theta}\right) = \left(\widehat{\theta} - 1\right)^2 I\left(\widehat{\theta}\right),$$

where $I\left(\widehat{\theta}\right)$ is the Fisher's Information Matrix evaluated at $\widehat{\theta}$.

Since we have only one restriction, $W\left(\widehat{\theta}\right)$ has an asymptotic chi-square distribution with 1 degree of freedom. Thus at $\alpha = 5\%$, we reject H_0 when $W\left(\widehat{\theta}\right) > 3.84146$;

Now consider the case where the sample size $T = 1$;

a) show that $\widehat{\theta} = X_1$.

b) if $\widehat{\theta} = \frac{1}{3}$, intuitively, should we reject or not reject H_0? Now compute $W\left(\widehat{\theta}\right)$ at $\widehat{\theta} = \dfrac{1}{3}$. Is H_0 rejected at $\alpha = 5\%$?

c) if H_0 is true, can $\widehat{\theta} = 2$? Intuitively, should we reject or not reject H_0 if $\widehat{\theta} = 2$? Now compute $W\left(\widehat{\theta}\right)$ at $\widehat{\theta} = 2$. Is H_0 rejected at $\alpha = 5\%$?

d) plot $W\left(\widehat{\theta}\right)$ at $\widehat{\theta} = 1, 2, 3, 4, \infty$.

ix) For $\theta \geq 1$, plot the power functions of this test for $T = 1, 2, 3, 4, \infty$.

x) Explain why the test is not properly behaved. Design a test for above hypothesis.

6. Write a GAUSS program to generate 1000 independent

a) N(0,1) random variables. (3 points)

b) N(1,2) random variables. (4 points)

c) U(0,1) random variables. (3 points)

d) U(-2,2) random variables. (4 points)

7. Consider the model $Y_t = \beta_0 + \beta_1 X_t + u, t = 1, 2, ..., T$. If the dependent variable is upper-truncated at c and lower-censored at 0, for any constants $0 < c < \infty$.

If the error term has a logistic distribution with density and distribution function

$$
\begin{aligned}
f(u) &= \frac{\exp(u)}{(1+\exp(u))^2}, \\
F(u) &= \frac{\exp(u)}{1+\exp(u)}.
\end{aligned}
$$

Show that the log-likelihood function is given by

$$
\ln L = \sum_{Y_t>0} \ln \frac{\exp(Y_t - c)\left(1+\exp(c-\beta_0-\beta_1 X_t)\right)}{\left(1+\exp(Y_t-\beta_0-\beta_1 X_t)\right)^2} + \sum_{Y_t=0} \ln \frac{\exp(-c)\left(1+\exp(c-\beta_0-\beta_1 X_t)\right)}{1+\exp(-\beta_0-\beta_1 X_t)}
$$

Threshold Model

Consider the following model:

$$y_t = \beta_1' x_t + \left(\beta_2' - \beta_1'\right) x_t \Psi_t\left(\gamma^0\right) + \varepsilon_t.$$

where β_1 and β_2 are the pre-shift and post-shift regression slope parameters respectively, with $\beta_i = \left(\beta_{1i}\ \beta_{2i}\ ...\ \beta_{Ki}\right)'$ being a K by 1 matrix of true parameters, $i = 1, 2$;

y_t is a T by 1 vector of dependent variable;

x_t is a T by K matrix of covariates;

Z_t is a T by m vector of threshold variables, where $0 < m < \infty$;

$\left(\varepsilon_1\ \varepsilon_2\ ...\ \varepsilon_T'\right)$ is a T by 1 vector of error term ε_t with $E\left|\varepsilon_t\right|^{4r} < \infty$ for some $r > 1$. The errors are assumed to be independent of the regressors and the threshold variables for simplicity purpose. The observed sample $\left\{y_t, x_t, Z_t\right\}_{t=1}^{T}$ are real-valued;

$\gamma = \left(\gamma_{1,...,}\gamma_m\right)'$ is a vector of m threshold parameters to be estimated;

$\Psi_t\left(\gamma^0\right)$ is the threshold condition, which equals one when the threshold variables satisfy some required conditions, and equals zero otherwise. For example, if the parameters change when all the threshold variables exceed some critical values, then we have:

$$\Psi_t\left(\gamma^0\right)^{and} = I\left(z_{1t} > \gamma_1^0, ..., z_{mt} > \gamma_m^0\right),$$

where z_{jt} is the j^{th} threshold variable. In the example of financial crisis, the crisis will not be triggered until all the threshold variables exceed the critical thresholds. We call this case the "and" case.

If the condition is that at least one threshold variable exceeds the critical value, then

$$\Psi_t\left(\gamma^0\right)^{or} = 1 - I\left(z_{1t} \leq \gamma_1^0, ..., z_{mt} \leq \gamma_m^0\right)$$

We call this the "or" case. It turns out that the "or" case can be rewritten in the form of the "and" case. Let $w_{jt} = -z_{jt}$, then

$$\Psi_t\left(\gamma^0\right)^{or} = 1 - I\left(w_{1t} > -\gamma_1^0, ..., w_{mt} > -\gamma_m^0\right) = \Psi_t\left(W_t, -\gamma^0\right)^{and}.$$

An important application of a threshold model with multiple threshold variables is the prediction of financial crises. Studies in the financial crises literature indicate that the occurrence of financial crises depends critically on the values of several threshold variables. It has long been observed that some economic variables, such as the foreign debt level and interest rate, did cross a certain threshold value before a currency crisis occur.

We focus on the case where $m = 2$. The methods extend in a straightforward manner to models with more than two threshold variables. Define

$$F(a, b) = \Pr(z_1 \leq a, z_2 \leq b),$$

$$\overline{F}(a, b) = \Pr(z_1 > a, z_2 > b).$$

For ease of illustration, we let $x = 1$. We estimate the model via Ordinary Least Squares method, the residual sum of squares is

$$\sum_{t=1}^{T} \left(y_t - \beta_1 \left(1 - \Psi_t(\gamma)\right) - x_t \Psi_t(\gamma)\right)^2.$$

We study the case where

$$\Psi_t\left(\gamma^0\right) = I\left(z_{1t} > \gamma_1^0, z_{2t} > \gamma_2^0\right).$$

Given γ_1 and γ_2, the OLS estimator for β are

$$\widehat{\beta}_1(\gamma) = \frac{\sum_{t=1}^{T} y_t \left(1 - \Psi_t(\gamma)\right)}{\sum_{t=1}^{T} \left(1 - \Psi_t(\gamma)\right)}$$

and

$$\widehat{\beta}_2(\gamma) = \frac{\sum_{t=1}^{T} y_t \Psi_t(\gamma)}{\sum_{t=1}^{T} \Psi_t(\gamma)}.$$

The residual sum of squares is

$$S_T(\gamma) = \sum_{t=1}^{T} \left(y_t - \widehat{\beta}_1 (1 - \Psi_t(\gamma)) - \widehat{\beta}_2 \Psi_t(\gamma) \right)^2,$$

$$\widehat{\gamma} = \left(\widehat{\gamma}_1, \widehat{\gamma}_2 \right) = \arg\min S_T \left(\gamma_1, \gamma_2 \right).$$

The final structural estimators are then defined as

$$\widehat{\beta}_1(\widehat{\gamma}) = \frac{\sum_{t=1}^{T} y_t (1 - \Psi_t(\widehat{\gamma}))}{\sum_{t=1}^{T} (1 - \Psi_t(\widehat{\gamma}))}$$

and

$$\widehat{\beta}_2(\widehat{\gamma}) = \frac{\sum_{t=1}^{T} y_t \Psi_t(\widehat{\gamma})}{\sum_{t=1}^{T} \Psi_t(\widehat{\gamma})}.$$

Note that

$$\begin{aligned} \widehat{\beta}_1 \left(\gamma_1, \gamma_2 \right) &= \beta_1 + \delta \times \\ &\quad \frac{\sum_{t=1}^{T} \left[I(z_{1t} > \gamma_1^0, z_{2t} > \gamma_2^0) - I(z_{1t} > \max\{\gamma_1^0, \gamma_1\}, z_{2t} > \max\{\gamma_2^0, \gamma_2\}) \right]}{\sum_{t=1}^{T} (1 - I(z_{1t} > \gamma_1, z_{2t} > \gamma_2))} \\ &\quad + o_p(1) \end{aligned}$$

Similarly, we have

$$\begin{aligned} \widehat{\beta}_2 \left(\gamma_1, \gamma_2 \right) &= \beta_2 - \delta \times \\ &\quad \frac{\sum_{t=1}^{T} \left[I(z_{1t} > \gamma_1, z_{2t} > \gamma_2) - I(z_{1t} > \max\{\gamma_1^0, \gamma_1\}, z_{2t} > \max\{\gamma_2^0, \gamma_2\}) \right]}{\sum_{t=1}^{T} I(z_{1t} > \gamma_1, z_{2t} > \gamma_2)} \\ &\quad + o_p(1), \end{aligned}$$

where

$$\delta = \beta_2 - \beta_1.$$

It can be shown that

$$\sup_{(\gamma_1, \gamma_2) \in R^2} \left| \frac{1}{T} S_T \left(\gamma_1, \gamma_2 \right) - g(\gamma_1, \gamma_2) \right| = o_p(1).$$

Let

$$b\left(\gamma_1,\gamma_2\right)=T^{2\alpha}\left(g\left(\gamma_1,\gamma_2\right)-\sigma^2\right).$$

We discuss four cases:

Case 1: $\gamma_1\leq\gamma_1^0,\gamma_2\leq\gamma_2^0$;

$$\widehat{\beta}_1\left(\gamma_1,\gamma_2\right)\xrightarrow{p}\beta_1,$$

$$\widehat{\beta}_2\left(\gamma_1,\gamma_2\right)\xrightarrow{p}\beta_2-\delta\left(1-\frac{\overline{F}\left(\gamma_1^0,\gamma_2^0\right)}{\overline{F}\left(\gamma_1,\gamma_2\right)}\right),$$

$$b_1\left(\gamma_1,\gamma_2\right)=c^2\overline{F}\left(\gamma_1^0,\gamma_2^0\right)\left(1-\frac{\overline{F}\left(\gamma_1^0,\gamma_2^0\right)}{\overline{F}\left(\gamma_1,\gamma_2\right)}\right)\geq b_1\left(\gamma_1^0,\gamma_2^0\right)=0.$$

Case 2: $\gamma_1>\gamma_1^0,\gamma_2\leq\gamma_2^0$;

$$\widehat{\beta}_1\left(\gamma_1,\gamma_2\right)\xrightarrow{p}\beta_1+\delta\frac{\overline{F}\left(\gamma_1^0,\gamma_2^0\right)-\overline{F}\left(\gamma_1,\gamma_2^0\right)}{1-\overline{F}\left(\gamma_1,\gamma_2\right)},$$

$$\widehat{\beta}_2\left(\gamma_1,\gamma_2\right)\xrightarrow{p}\beta_2-\delta\left(1-\frac{\overline{F}\left(\gamma_1,\gamma_2^0\right)}{\overline{F}\left(\gamma_1,\gamma_2\right)}\right),$$

$$b_2\left(\gamma_1,\gamma_2\right)=c^2\left[\overline{F}\left(\gamma_1^0,\gamma_2^0\right)-\frac{\left(\overline{F}\left(\gamma_1^0,\gamma_2^0\right)-\overline{F}\left(\gamma_1,\gamma_2^0\right)\right)^2}{1-\overline{F}\left(\gamma_1,\gamma_2\right)}-\frac{\left(\overline{F}\left(\gamma_1,\gamma_2^0\right)\right)^2}{\overline{F}\left(\gamma_1,\gamma_2\right)}\right]\geq b_2\left(\gamma_1^0,\gamma_2^0\right)=0.$$

Case 3: $\gamma_1\leq\gamma_1^0,\gamma_2>\gamma_2^0$;

$$\widehat{\beta}_1\left(\gamma_1,\gamma_2\right)\xrightarrow{p}\beta_1+\delta\frac{\overline{F}\left(\gamma_1^0,\gamma_2^0\right)-\overline{F}\left(\gamma_1^0,\gamma_2\right)}{1-\overline{F}\left(\gamma_1,\gamma_2\right)},$$

$$\widehat{\beta}_2\left(\gamma_1,\gamma_2\right)\xrightarrow{p}\beta_2-\delta\left(1-\frac{\overline{F}\left(\gamma_1^0,\gamma_2\right)}{\overline{F}\left(\gamma_1,\gamma_2\right)}\right),$$

$$\begin{aligned}b_3\left(\gamma_1,\gamma_2\right) &= c^2\left[\begin{array}{c}\overline{F}\left(\gamma_1^0,\gamma_2^0\right)-\frac{\left(\overline{F}\left(\gamma_1^0,\gamma_2^0\right)-\overline{F}\left(\gamma_1^0,\gamma_2\right)\right)^2}{1-\overline{F}\left(\gamma_1,\gamma_2\right)}-\frac{\left(\overline{F}\left(\gamma_1,\gamma_2^0\right)\right)^2}{\overline{F}\left(\gamma_1,\gamma_2\right)}\\+\left(\overline{F}\left(\gamma_1^0,\gamma_2^0\right)-\overline{F}\left(\gamma_1^0,\gamma_2\right)\right)\frac{\overline{F}\left(\gamma_1^0,\gamma_2^0\right)}{\overline{F}\left(\gamma_1,\gamma_2\right)}\end{array}\right]\\ &\geq b_3\left(\gamma_1^0,\gamma_2^0\right)=0.\end{aligned}$$

Case 4: $\gamma_1 > \gamma_1^0, \gamma_2 > \gamma_2^0$;

$$\widehat{\beta}_1\left(\gamma_1,\gamma_2\right) \xrightarrow{p} \beta_1 + \delta\frac{\overline{F}\left(\gamma_1^0,\gamma_2^0\right) - \overline{F}\left(\gamma_1,\gamma_2\right)}{1-\overline{F}\left(\gamma_1,\gamma_2\right)},$$

$$\widehat{\beta}_2\left(\gamma_1,\gamma_2\right) \xrightarrow{p} \beta_2,$$

$$b_4\left(\gamma_1,\gamma_2\right) = c^2\left(\overline{F}\left(\gamma_1^0,\gamma_2^0\right) - \overline{F}\left(\gamma_1,\gamma_2\right)\right)\frac{1-\overline{F}\left(\gamma_1^0,\gamma_2^0\right)}{1-\overline{F}\left(\gamma_1,\gamma_2\right)} \geq b_4\left(\gamma_1^0,\gamma_2^0\right) = 0.$$

Combining the results above, we conclude that the function $b\left(\gamma_1,\gamma_2\right)$ has a global minimum at the true threshold values, i.e.,

$$\underset{(\gamma_1,\gamma_2)\in R^2}{Arg\min} b\left(\gamma_1,\gamma_2\right) = \left(\gamma_1^0,\gamma_2^0\right).$$

Since $b\left(\gamma_1,\gamma_2\right) \geq b\left(\gamma_1^0,\gamma_2^0\right)$ in all cases, the estimators converge to (γ_1^0,γ_2^0).

When z_{1t} and z_{2t} are independent, it can be shown that

Case 1: $\gamma_1 \leq \gamma_1^0, \gamma_2 \leq \gamma_2^0$;

$$\frac{\partial b_1\left(\gamma_1,\gamma_2\right)}{\partial\gamma_1} \leq 0,$$

$$\frac{\partial b_1\left(\gamma_1,\gamma_2\right)}{\partial\gamma_2} \leq 0.$$

Case 2: $\gamma_1 > \gamma_1^0, \gamma_2 \leq \gamma_2^0$;

$$\frac{\partial b_2\left(\gamma_1,\gamma_2\right)}{\partial\gamma_1} \geq 0,$$

$$\frac{\partial b_2\left(\gamma_1,\gamma_2\right)}{\partial\gamma_2} \leq 0.$$

Case 3: $\gamma_1 \leq \gamma_1^0, \gamma_2 > \gamma_2^0$;

$$\frac{\partial b_3\left(\gamma_1,\gamma_2\right)}{\partial\gamma_1} \leq 0,$$

$$\frac{\partial b_3\left(\gamma_1,\gamma_2\right)}{\partial\gamma_2} \geq 0.$$

Case 4: $\gamma_1 > \gamma_1^0, \gamma_2 > \gamma_2^0$.

$$\frac{\partial b_4(\gamma_1,\gamma_2)}{\partial \gamma_1} = c^2 \left(\frac{1-\overline{F}_1(\gamma_1^0)\overline{F}_2(\gamma_2^0)}{1-\overline{F}_1(\gamma_1)\overline{F}_2(\gamma_2)} \right)^2 \overline{F}_2(\gamma_2) f_1(\gamma_1) \geq 0,$$

$$\frac{\partial b_4(\gamma_1,\gamma_2)}{\partial \gamma_2} = c^2 \left(\frac{1-\overline{F}_1(\gamma_1^0)\overline{F}_2(\gamma_2^0)}{1-\overline{F}_1(\gamma_1)\overline{F}_2(\gamma_2)} \right)^2 \overline{F}_1(\gamma_1) f_2(\gamma_2) \geq 0,$$

Note that given γ_2, the value of $b(\gamma_1,\gamma_2)$ reduces whenever γ_1 approaches γ_1^0 from both directions. Similarly, given γ_1, the value of $b(\gamma_1,\gamma_2)$ reduces whenever γ_2 approaches γ_2^0. This implies that

$$\underset{\gamma_1 \in R}{Arg\min}\ b(\gamma_1,\gamma_2) = \gamma_1^0 \qquad \forall \gamma_2$$

and

$$\underset{\gamma_2 \in R}{Arg\min}\ b(\gamma_1,\gamma_2) = \gamma_2^0 \qquad \forall \gamma_1$$

In general, the two threshold variable will not be independent. We assume that the joint distribution of z_1 and z_2 are continuous and differentiable with respect to both variables.

Define the moment functionals

$$\overline{M}\gamma = \overline{M}(\gamma_1,\gamma_2) = E\left(x_t x_t' I(z_{1t} > \gamma_1, z_{2t} > \gamma_2)\right),$$

$$\overline{M}_0 = \overline{M}\left(\gamma_1^0,\gamma_2^0\right),$$

$$M = E\left(x_t x_t'\right),$$

$$D(\gamma_1,\gamma_2) = E\left(x_t x_t' | z_{1t} = \gamma_1, z_{2t} = \gamma_2\right),$$

$$D = D\left(\gamma_1^0,\gamma_2^0\right),$$

$$\overline{V}(\gamma_1,\gamma_2) = E\left(x_t x_t' \varepsilon_t^2 | z_{1t} = \gamma_1, z_{2t} = \gamma_2\right),$$

$$\overline{G}(\gamma_1, \gamma_2) = M^{-1}\overline{M}(\gamma_1, \gamma_2).$$

$$\overline{F}_i(\gamma_1, \gamma_2) = \frac{\partial}{\partial \gamma_i}\overline{F}(\gamma_1, \gamma_2) \qquad i = 1, 2.$$

$$\overline{F}_i^0 = \overline{F}_i\left(\gamma_1^0, \gamma_2^0\right) \qquad i = 1, 2.$$

Let

$$x_t(\gamma) = x_t \Psi_t(\gamma)$$

and let X and X_γ be T by m matrixes formed by stacking the vectors x_t' and $x_t(\gamma)'$.

Thus, our model can be rewritten as

$$Y = X\beta_1 + X_\gamma \boldsymbol{\delta} + \varepsilon.$$

Given $\gamma_{1,...,}\gamma_m$, the OLS estimator for β are

$$\widehat{\beta}_1'(\gamma) = \left(\sum_{t=1}^{T} x_t x_t'(1 - \Psi_t(\gamma))\right)^{-1} \sum_{t=1}^{T} y_t x_t'(1 - \Psi_t(\gamma))$$

and

$$\widehat{\beta}_2'(\gamma) = \left(\sum_{t=1}^{T} x_t x_t' \Psi_t(\gamma)\right)^{-1} \sum_{t=1}^{T} y_t x_t' \Psi_t(\gamma),$$

Define

$$S_T(\gamma) = \sum_{t=1}^{T} \left(y_t - \widehat{\beta}_1' x_t - \left(\widehat{\beta}_2' - \widehat{\beta}_1'\right) x_t \Psi_t(\gamma)\right)^2,$$

$$\widehat{\gamma} = \left(\widehat{\gamma}_{1,...,}\widehat{\gamma}_m\right) = \arg\min_{\gamma \in \Gamma_T} S_T\left(\gamma_{1,...,}\gamma_m\right),$$

where

$$\Gamma_T = \Pi_{j=1}^{m}\left(\left[\underline{\gamma_j}, \overline{\gamma_j}\right] \cap \{z_{j1}, ..., z_{jT}\}\right).$$

The final structural estimators are then defined as

$$\widehat{\beta}_1\left(\widehat{\gamma}_{1,...,}\widehat{\gamma}_m\right)$$

and

$$\widehat{\beta}_2\left(\widehat{\gamma}_{1,...,}\widehat{\gamma}_m\right).$$

The residual sum of squares can also be written as

$$S_T(\gamma) = Y\left('I - P_\gamma\right)Y$$

where

$$P_\gamma = \widetilde{X}_\gamma\left(\widetilde{X}'_\gamma\widetilde{X}_\gamma\right)^{-1}\widetilde{X}'_\gamma,$$

$$\widetilde{X}_\gamma = \left[\begin{array}{cc} X & X_\gamma \end{array}\right].$$

Let $X_0 = X_{\gamma_0}$, then since $Y - X\beta_1 - X_\gamma\boldsymbol{\delta}$ and X lies in the space spanned by P_γ,

$$S_T(\gamma) - \varepsilon'\varepsilon = -\varepsilon' P_\gamma\varepsilon + 2\boldsymbol{\delta}X'_0\left(I - P_\gamma\right)\varepsilon + \boldsymbol{\delta}X'_0\left(I - P_\gamma\right)X_0\boldsymbol{\delta}$$

$$\frac{1}{T^{1-2\alpha}}\left(S_T(\gamma) - \varepsilon'\varepsilon\right) = \frac{1}{T}c'\left(X'_0\left(I - P_\gamma\right)X_0\right)c + o_p(1)$$

We discuss four cases. In each case, $\frac{1}{T^{1-2\alpha}}\left(S_T(\gamma) - \varepsilon'\varepsilon\right) \xrightarrow{p} b_i(\gamma)$, with $b_i(\gamma_0) = 0$, $i = 1, 2, 3, 4$.

Case 1: $\gamma_1 \leq \gamma_1^0, \gamma_2 \leq \gamma_2^0$;

$b_1(\gamma) = c'\left(\overline{M}_0 - \overline{M}_0\overline{M}\gamma^{-1}\overline{M}_0\right)c \equiv b_1(\gamma) \geq 0.$

Case 2: $\gamma_1 > \gamma_1^0, \gamma_2 \leq \gamma_2^0$;

$b_2(\gamma) =$

$$c'\left(\begin{array}{c} \overline{M}_0 - \left(\overline{M}_0 - \overline{M}\left(\gamma_1, \gamma_2^0\right)\right)\left(M - \overline{M}\gamma\right)^{-1} \\ \left(\overline{M}_0 - \overline{M}\left(\gamma_1, \gamma_2^0\right)\right) - \overline{M}\left(\gamma_1, \gamma_2^0\right)\overline{M}\gamma^{-1}\overline{M}\left(\gamma_1, \gamma_2^0\right) \end{array}\right)c$$

$\geq 0.$

Case 3: $\gamma_1 \leq \gamma_1^0, \gamma_2 > \gamma_2^0$;

$$b_3(\gamma) = c' \begin{pmatrix} \overline{M}_0 - \left(\overline{M}_0 - \overline{M}(\gamma_1^0, \gamma_2)\right)\left(M - \overline{M}\gamma\right)^{-1} \\ \left(\overline{M}_0 - \overline{M}(\gamma_1^0, \gamma_2)\right) - \overline{M}(\gamma_1^0, \gamma_2)\overline{M}\gamma^{-1}\overline{M}(\gamma_1^0, \gamma_2) \end{pmatrix} c \geq 0.$$

Case 4: $\gamma_1 > \gamma_1^0, \gamma_2 > \gamma_2^0$

$$b_4(\gamma) = c'\left(M - \overline{M}_0 - \left(M - \overline{M}_0\right)\left(M - \overline{M}\gamma\right)^{-1}\left(M - \overline{M}_0\right)\right)c.$$

Since all the four functions of b are minimized at the true thresholds, the threshold estimators are therefore consistent.

The threshold estimators are similar to the change point in the structural-change model. As is well known, because of the superconsistency of the change-point estimator, the distribution of the change-point estimator will degenerate to the true change point for any fixed magnitude of change. To generate a meaningful distribution, the usual practice is to let the magnitude of change to go to zero at an appropriate rate. When the change is small enough, there is a variation of the change point estimate even it is superconsistent.

In the threshold model, in order to obtain the distribution of the threshold estimators, we let $\delta = cT^{-\alpha}$, $\alpha < \frac{1}{2}$. It can be shown that

$$T^{1-2\alpha}\frac{(c'Dc)^2}{c'Vc}\left(\left(\widehat{\gamma}_1 - \gamma_1^0\right)\overline{F}_1^0, \left(\widehat{\gamma}_2 - \gamma_2^0\right)\overline{F}_2^0\right) = (\widehat{r}_1, \widehat{r}_2) \xrightarrow{d} \underset{(r_1,r_2)\in R^2}{\arg\max} \sum_{j=1}^{2}\left(-\frac{1}{2}|r_j| + W_j(r_j)\right).$$

To find the joint distribution in the close form, note that the selection of r_1 does not depend on the choice of r_2 and vice versa, it can be shown that

$$F_{(\widehat{r}_1,\widehat{r}_2)}(a_1, a_2) = \Pi_{j=1}^{2}\left(1 + \sqrt{\frac{a_j}{2\pi}}\exp\left(-\frac{a_j}{8}\right) + \frac{3}{2}\exp(a_j)\Phi\left(-\frac{3\sqrt{a_j}}{2}\right) - \frac{a_j+5}{2}\Phi\left(-\frac{\sqrt{a_j}}{2}\right)\right).$$

Thus, the joint density function can be found as

$$f_{(\widehat{r}_1,\widehat{r}_2)}(a_1, a_2) = \Pi_{j=1}^{2}\left(\frac{3}{2}\exp(a_j)\Phi\left(-\frac{3\sqrt{a_j}}{2}\right) - \frac{1}{2}\Phi\left(-\frac{\sqrt{a_j}}{2}\right)\right),$$

where $\Phi(\cdot)$ is the cdf of a standard normal distribution.

In cases when some of the $a_j < 0$, we can replace those items in the above expression by $F_{\widehat{r}_j}(a_j) = 1 - F_{\widehat{r}_j}(-a_j)$ and $f_{\widehat{r}_j}(a_j) = f_{\widehat{r}_j}(-a_j)$.

In general, if we have m threshold variables,

$$-T^{1-2\alpha}\frac{(c'Dc)^2}{c'Vc}(\widehat{\gamma}-\gamma_0)\circ\frac{\partial F(\gamma_1^0,....,\gamma_m^0)}{\partial\gamma}\xrightarrow{d}\underset{(r_1,...,r_m)\in R^m}{\arg\max}\sum_{j=1}^{m}\left(-\frac{1}{2}|r_j|+W_j(r_j)\right).$$

where $\circ$ is the Hadamard product operator that multiplies on an element by element basis, and

$$\begin{aligned}&F_{(\widehat{r}_1,...,\widehat{r}_m)}(a_1,...,a_m)\\ =\;&\Pi_{j=1}^{m}\left(1+\sqrt{\frac{a_j}{2\pi}}\exp\left(\frac{-a_j}{8}\right)+\frac{3\exp(a_j)}{2}\Phi\left(\frac{-3\sqrt{a_j}}{2}\right)-\frac{a_j+5}{2}\Phi\left(\frac{-\sqrt{a_j}}{2}\right)\right),\end{aligned}$$

$$f_{(\widehat{r}_1,...,\widehat{r}_m)}(a_1,...,a_m)=\Pi_{j=1}^{m}\left(\frac{3}{2}\exp(a_j)\,\Phi\left(-\frac{3\sqrt{a_j}}{2}\right)-\frac{1}{2}\Phi\left(-\frac{\sqrt{a_j}}{2}\right)\right).$$

It should be noted that if the threshold variables are dependent, it may be impossible to derive the joint distribution of the threshold estimators, although we may still get the consistency result. For example, if $z_2 = -z_1$, we may not be able to partition the data into four groups according to the values of the two threshold variables, so the above distributional result will not hold.

Our model can be extended to incorporate **panel data**. The observed data are from a balanced panel with n individuals over T periods. Following the lead of Hansen (1999), we assume that all individuals have the same threshold value for each threshold variable. Note that the model in the previous section can be a cross-sectional or a time series model. In the panel model here, n is the cross-sectional sample size. The analysis is asymptotic with fixed T and as $n \to \infty$.

We let

$$\Psi_{it}(\gamma) = I(z_{1it} > \gamma_1, ..., z_{mit} > \gamma_m).$$

The observations are divided into two regimes depending on whether the threshold variable vector Z satisfies the threshold conditions. We assume that x_{it} and Z_{it} are not time invariant. The model is

$$y_{it} = \mu_i + \beta_1' x_{it} + \varepsilon_{it}, \qquad \Psi_{it}(\gamma) = 0,$$

$$y_{it} = \mu_i + \beta_2' x_{it} + \varepsilon_{it}, \qquad \Psi_{it}(\gamma) = 1.$$

Let

$$x_{it}(\gamma) = x_{it}\Psi_{it}(\gamma)$$

$$y_{it} = \mu_i + \beta_1' x_{it} + \delta' x_{it}\Psi_{it}(\gamma) + \varepsilon_{it}.$$

Averaging the above panel equation over t, we have

$$\overline{y}_i = \mu_i + \beta_1' \overline{x}_i + \delta' \overline{x}_i(\gamma) + \overline{\varepsilon}_t,$$

where

$$\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it},$$

$$\overline{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it},$$

$$\overline{x}_i(\gamma) = \frac{1}{T}\sum_{t=1}^{T} x_{it}\Psi_{it}(\gamma),$$

$$\overline{\varepsilon}_t = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it}.$$

Taking the difference, we have

$$y_{it}^* = \beta_1' x_{it}^* + \delta' x_{it}^*(\gamma) + \varepsilon_{it}^*,$$

where

$$y_{it}^{*} = y_{it} - \overline{y}_{i},$$

$$x_{it}^{*} = x_{it} - \overline{x}_{i},$$

$$x_{it}^{*}(\gamma) = x_{it}(\gamma) - \overline{x}_{i}(\gamma),$$

$$\varepsilon_{it}^{*} = \varepsilon_{it} - \overline{\varepsilon}_{i}.$$

Let

$$y_{i}^{*} = \begin{bmatrix} y_{i2}^{*} \\ \vdots \\ y_{iT}^{*} \end{bmatrix}, x_{i}^{*} = \begin{bmatrix} x_{i2}^{*} \\ \vdots \\ x_{iT}^{*} \end{bmatrix}, x_{i}^{*}(\gamma) = \begin{bmatrix} x_{i2}^{*}(\gamma) \\ \vdots \\ x_{iT}^{*}(\gamma) \end{bmatrix}, \varepsilon_{i}^{*} = \begin{bmatrix} \varepsilon_{i2}^{*} \\ \vdots \\ \varepsilon_{iT}^{*} \end{bmatrix}$$

denote the stacked data and errors for an individual, with one time period deleted. Let Y^{*}, $X^{*}(\gamma)$ and ε^{*} denote the data stacked over all individual, i.e.,

$$Y^{*} = \begin{bmatrix} y_{1}^{*} \\ \vdots \\ y_{i}^{*} \\ \vdots \\ y_{n}^{*} \end{bmatrix}, X^{*} = \begin{bmatrix} x_{1}^{*} \\ \vdots \\ x_{i}^{*} \\ \vdots \\ x_{n}^{*} \end{bmatrix}, X^{*}(\gamma) = \begin{bmatrix} x_{1}^{*}(\gamma) \\ \vdots \\ x_{i}^{*}(\gamma) \\ \vdots \\ x_{n}^{*}(\gamma) \end{bmatrix}, \varepsilon^{*} = \begin{bmatrix} \varepsilon_{1}^{*} \\ \vdots \\ \varepsilon_{i}^{*} \\ \vdots \\ \varepsilon_{n}^{*} \end{bmatrix}.$$

Thus, our model becomes

$$Y^{*} = X^{*}\beta_{1} + X^{*}(\gamma)\boldsymbol{\delta} + \varepsilon^{*}.$$

Hence, the estimation method and the asymptotic results in the previous part apply in the panel model.

$$S_{nT}(\gamma) = (Y - X^{*}\beta_{1} - X^{*}(\gamma)\boldsymbol{\delta})'(Y - X^{*}\beta_{1} - X^{*}(\gamma)\boldsymbol{\delta})$$

$$\widehat{\gamma} = \left(\widehat{\gamma}_{1,...,}\widehat{\gamma}_m\right) = \arg\min_{\gamma\in\Gamma_n} S_T\left(\gamma_{1,...,}\gamma_m\right).$$

$$\Gamma_n = \Pi_{j=1}^m \left(\left[\underline{\gamma_j}, \overline{\gamma_j}\right] \cap \left(\cup_{i=1}^n \left\{z_{ji1}, ..., z_{jiT}\right\}\right)\right)$$

The final structural estimators are then defined as

$$\widehat{\beta}_1\left(\widehat{\gamma}_{1,...,}\widehat{\gamma}_m\right)$$

and

$$\widehat{\beta}_2\left(\widehat{\gamma}_{1,...,}\widehat{\gamma}_m\right).$$

and the residual variance is

$$\widehat{\sigma}^2 = \frac{1}{n\left(T-1\right)} S_{nT}\left(\widehat{\gamma}\right).$$

Testing for values of the thresholds

$$H_0 : \gamma = \gamma^0.$$

We borrow the Likelihood Ratio test of Hansen (1999, 2000). Under the assumption that ε_t is i.i.d. $N\left(0, \sigma^2\right)$, we have

$$LR_T\left(\gamma_1, \gamma_2\right) = T\frac{S_T\left(\gamma_1, \gamma_2\right) - S_T\left(\widehat{\gamma}_1, \widehat{\gamma}_2\right)}{S_T\left(\widehat{\gamma}_1, \widehat{\gamma}_2\right)}.$$

H_0 is reject for large $LR_T\left(\gamma_1^0, \gamma_2^0\right)$.

If the threshold variables are independent, we can extend the result of Hansen (2000) to show that

$$LR_T\left(\gamma_1^0, \gamma_2^0\right) \xrightarrow{d} \eta^2\xi,$$

where

$$\xi = \xi_1 + \xi_2,$$

$$\xi_1 = \max_{-\infty<r_1<\infty} \left(-|r_1| + 2W_1(r_1)\right),$$

$$\xi_2 = \max_{-\infty<r_2<\infty} \left(-|r_2| + 2W_2(r_2)\right)$$

and

$$\eta^2 = \frac{(c'Vc)}{\sigma^2 c'Dc}.$$

The distribution of ξ_i $(i = 1, 2)$ is

$$\Pr(\xi_i \leq x) = \left(1 - \exp\left(-\frac{x}{2}\right)\right)^2,$$

$$f_{\xi_i}(x) = \left(1 - e^{-\frac{1}{2}x}\right) e^{-\frac{1}{2}x}.$$

Thus,

$$\begin{aligned}
\Pr(\xi \leq x) &= \Pr(\xi_1 + \xi_2 \leq x) \\
&= \int_0^x \Pr(\xi_1 \leq x - y) f_{\xi_2}(y)\, dy \\
&= \int_0^x \left(1 - \exp\left(-\frac{x-y}{2}\right)\right)^2 \left(1 - e^{-\frac{1}{2}y}\right) e^{-\frac{1}{2}y} dy \\
&= 1 - 5e^{-x} - 2xe^{-\frac{1}{2}x} - e^{-x}x + 4e^{-\frac{1}{2}x},
\end{aligned}$$

$$f_\xi(x) = 4e^{-x} - 4e^{-\frac{1}{2}x} + xe^{-\frac{1}{2}x} + e^{-x}x.$$

END

1. Andrews D.W.K.(1993)"Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821-856.

2. Chong, T.T.L. (1995) "Partial Parameter Consistency in a Misspecified Structural Change Model," *Economics Letters*, 49, 351-357.

3. Chong, T.T.L. and C.S. Lui (1999)"Estimating the Fractionally Integrated Process in the Presence of Measurement Errors," *Economics Letters*, 63, 285-294.

4. Chong, T.T.L. (2000) "Estimating the Differencing Parameter via the Partial Autocorrelation Function," *Journal of Econometrics,* 97, 365-381.

5. Chong, T.T.L. (2001a) "Structural Change in AR(1) Models," *Econometric Theory*, 17, 87-155.

6. Chong, T.T.L. (2001b) "Estimating the Locations and Number of Change Points by the Sample-Splitting Method," *Statistical Papers*, 42, 53-79.

7. Chong, T.T.L. (2003) "Generic Consistency of the Break-Point Estimator under Specification Errors," *Econometrics Journal*, 6, 167-192.

8. Hansen, B.E. (1999) " Threshold effect in Non-dynamic Panels: Estimation, Testing, and Inference," *Journal of Econometrics* 93," 345-368.

9. Hansen, B.E. (2000) "Sample splitting and threshold estimation," *Econometrica*, 68, 575-603.

10. Tieslau, M.A., P. Schmidt and R.T. Baillie (1996)"A Minimum Distance Estimator for Long-Memory Processes," *Journal of Econometrics,* 71, 249-64.