Introductory Econometrics

Terence Tai-Leung Chong

April 1, 2010

Contents

Chapter 1

Probability

1.1 Revision of the Summation Operator

The **summation operator** $\sum$ has the following properties:

1. If k is a constant, then $\sum\limits_{t=1}^{T} k = Tk$;
2. If k is a constant, then $\sum\limits_{t=1}^{T} kx_t = k\sum\limits_{t=1}^{T} x_t$;
3. $\sum\limits_{t=1}^{T} (x_t + y_t) = \sum\limits_{t=1}^{T} x_t + \sum\limits_{t=1}^{T} y_t$;
4. $\sum\limits_{t=1}^{T} (x_t - \overline{x}) = 0$;
5. $\sum\limits_{t=1}^{T} (x_t - \overline{x})(y_t - \overline{y}) = \sum\limits_{t=1}^{T} (x_t - \overline{x}) y_t = \sum\limits_{t=1}^{T} (y_t - \overline{y}) x_t$;
6. $\left(\sum\limits_{i=1}^{I} x_i\right)\left(\sum\limits_{j=1}^{J} y_j\right) = \sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J} x_i y_j$
$= x_1y_1 + x_1y_2 + ... + x_1y_J + x_2y_1 + ... + x_2y_J + ... + x_Iy_1 + ... + x_Iy_J$;
7. $\left(\sum\limits_{t=1}^{T} x_t\right)^2 = \sum\limits_{t=1}^{T} x_t^2 + 2\sum\limits_{i=1}^{T-1}\sum\limits_{j>i}^{T} x_i x_j.$

Exercise 1: Compute

(i) $\sum_{i=1}^{3} (i+4)$.

(ii) $\sum_{i=1}^{3} 3^i$.

(iii) $\sum_{i=1}^{3} \sum_{j=1}^{2} ij$.

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.

For example, tossing a coin and throwing a dice are random experiments.

Definition 2: The **sample space** S is defined to be the set of all possible outcomes of the random experiment. The elements of S are called *elementary events.*

For example, when tossing a coin, $S = \{H, T\}$, elementary events are H=head and T=tail.

When throwing a dice, $S = \{1, 2, 3, 4, 5, 6\}$, the elementary events are 1, 2, 3, 4, 5 and 6.

Definition 3: An **event** is a subset of the sample space. Every subset is an event. It may be empty, 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.

For example, when tossing a coin, the subsets of S are ϕ, $\{H\}$, $\{T\}$ and $\{H, T\}$, where ϕ is an empty set. The event "H and T appear at the same time" belongs to ϕ.

Consider the sum of points in throwing two dices, the sample space is

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

The event that the sum is an even number will be

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

The event that the sum is bigger than 13 will be ϕ, or a null event.

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.

Axiom 1: Kolmogorov Axioms of Probability

Let A be an event, then

(i) $0 \leq \Pr(A) \leq 1$;

(ii) $\Pr(S) = 1$;

(iii) $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $A \cap B = \phi$, where " $\cup$ " is the union of sets, meaning "or". " $\cap$ " stands for intersection of sets, meaning "and".

Example 1: For what values of k can

$$\Pr(X = i) = (1 - k) k^i$$

serve as the values of the probability distribution of a random variable with the countably infinite range $i = 0, 1, 2, ...$?

Solution: Since

(i) $0 \leq \Pr(X = i) \leq 1$. Thus, $0 \leq (1 - k) k^i \leq 1$, which implies $0 \leq k \leq 1$.

(ii) $\Pr(X = 0 \text{ or } 1 \text{ or } 2 \text{ or } 3 \text{ or} ....) = 1$;

(iii) Since the event "$X = i$ and $X = j$" $= \phi$ for all $i \neq j$, we have

$$\Pr(X = 0 \text{ or } 1 \text{ or } 2 \text{ or } 3 \text{ or....}) = \Pr(X = 0) + \Pr(X = 1) + ...$$

Further, by using property (ii), we have

$$\sum_{i=0}^{\infty} \Pr(X = i) = 1,$$

$$\begin{aligned} \sum_{i=0}^{\infty} (1-k)k^i &= 1, \\ (1-k)\sum_{i=0}^{\infty} k^i &= 1 \end{aligned}$$

Thus, we rule out the cases where $k = 0$ and $k = 1$, since otherwise the equality will not hold. Since k is strictly bigger than zero and strictly smaller than one, we have

$$(1-k).\frac{1}{1-k} = 1$$

$$1 = 1$$

Thus, any value of k with $0 < k < 1$ is a solution. ■

Definition 4: The **conditional probability** of B occurring, given that A has occurred is

$\Pr(B|A) = \dfrac{\Pr(B \cap A)}{\Pr(A)}$ if $P(A) \neq 0$. If $\Pr(A) = 0$, we define $\Pr(B|A) = 0$. The result implies that

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

For example, consider a card game, let A be the event that a “Heart” appears, B be the event that an “Ace” appears.

$$\Pr(\text{Ace}|\text{Heart}) = \frac{\Pr(\text{Ace} \cap \text{Heart})}{\Pr(\text{Heart})} = \frac{1/52}{13/52} = \frac{1}{13}.$$

Definition 5: Two events A and B are **independent** if and only if $\Pr(A \cap B) = \Pr(A)\Pr(B)$, i.e. $\Pr(B|A) = \Pr(B)$.

The statement “if and only if” is different from “if”. When we say “A if and only if B”, we mean “if A then B” and “if B then A” are both true. Thus “if and only if” is a formal definition.

Therefore if two events are independent, we must have $\Pr(A \cap B) = \Pr(A)\Pr(B)$. If we known $\Pr(A \cap B) = \Pr(A)\Pr(B)$, then A and B must be independent.

Exercise 2: Give two independent events and two dependent events.

Definition 6: A **random variable** X is a real-valued function of the elements of a sample space. 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.

Thus the value of a random variable corresponds to the outcome of an random experiment.

For example, tossing a coin is a random experiment, the outcomes are represented by Heads and Tails. However, Heads and Tails are not real-value numbers, thus Heads and Tails are not random variables. If we define $X = 1$ if a Head appears and $X = 2$ if a Tail appears, then X is a random variable.

1.2 Probability Distribution Function and Density Function

Let X, Y be two continuous random variables.

Definition 7: The **probability distribution function** of X is defined as $F_x(u) = \Pr(-\infty < X \leq u)$, with $F_x(\infty) = 1$.

Definition 8: The **density function** is $f(x) = \dfrac{dF(x)}{dx}$, with $f(x) \geq 0$, and $f(-\infty) = f(\infty) = 0$.

Example 2: Let X be a random variable evenly distributed in zero-one interval, then

$\Pr(X < 0) = 0 \qquad u < 0;$

$\Pr(0 \leq X \leq u) = u \qquad 0 \leq u \leq 1;$

$\Pr(X > u) = 0 \qquad u > 1.$

$$\begin{aligned} F_x(u) &= 0, \qquad u < 0 \\ &= u, \qquad 0 \leq u \leq 1 \\ &= 1, \qquad u > 1 \end{aligned}$$

$$\begin{aligned} f(u) &= 0, \qquad u < 0 \\ &= 1, \qquad 0 \leq u \leq 1 \\ &= 0, \qquad u > 1. \end{aligned}$$

Definition 9: 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 $F(x,y), f(x,y), f(x)$ and $f(y)$ is:

$$
\begin{aligned}
F(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)\,dsdt, \\
f(x,y) &= \frac{\partial^2}{\partial x \partial y} F(x,y), \\
f(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy, \\
f(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx.
\end{aligned}
$$

Further, $F(-\infty,-\infty) = \Pr(X \leq -\infty \text{ and } Y \leq -\infty) = 0$, $F(\infty,\infty) = \Pr(X \leq \infty \text{ and } Y \leq \infty) = 1$, and $f(x,y) \geq 0$. X and Y are independent **if and only if** $f(x,y) = f(x)f(y)$.

Exercise 3: Suppose a continuous random variable X has density function

$f(x;\theta) = \theta x + 0.5$ 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) For what value of θ is the variance of X maximized.

Exercise 4: 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) $\Pr(X>2Y)$.

1.3 Mathematical Expectation

Definition 10: The **first moment, mean** or **expected value** of a random variable X, is defined as:

$$E(X)=\sum_i x_i P(x_i) \qquad \text{if } X \text{ is discrete}$$

$$E(X)=\int_{-\infty}^{\infty} x f(x)\,dx \qquad \text{if } X \text{ is continuous}$$

It has the following properties: For any random variables X, Y and any constants a, b.

(i) $E(a)=a$;

(ii) $E(E(X))=E(X)$;

(iii) $E(aX)=aE(X)$;

(iv) $E(aX+bY)=aE(X)+bE(Y)$.

Other measures of central tendency are the median, which is the value that is exceeded by the random variable with probability one-half, and the mode, which is the value of x at which $f(x)$ takes its maximum.

Exercise 5: Let X and Y be two independent random variables, if $E\left(\frac{X}{Y}\right)>1$, then $\frac{E(X)}{E(Y)}>1$. True/False/Uncertain. Explain.

Definition 11: The **second moment around the mean** or **variance** of a random variable is

$Var\left(X\right)=E\left(X-E\left(X\right)\right)^{2}=E\left(X^{2}\right)-E^{2}\left(X\right)=\sum_{i}\left(x_{i}-E\left(X\right)\right)^{2}P\left(x_{i}\right)$ if X is discrete.

$Var\left(X\right)=\int_{-\infty}^{\infty}\left(x-E\left(X\right)\right)^{2}f\left(x\right)dx$ if X is continuous.

It has the following properties: for any random variables X, Y and any constant a,

(i) $Var\left(a\right)=0$;

(ii) $Var\left(aX\right)=a^{2}Var\left(X\right)$;

(iii) $Var\left(X\pm Y\right)=Var\left(X\right)+Var\left(Y\right)\pm2Cov\left(X,Y\right)$ if X and Y are not independent;

(vi) $Var\left(X\pm Y\right)=Var\left(X\right)+Var\left(Y\right)$ if X and Y are independent.

Note: $Var\left(X-Y\right)\neq Var\left(X\right)-Var\left(Y\right)$!

Definition 12: The **covariance** of two random variables X and Y, is defined to be:

$$Cov\left(X,Y\right)=E\left(X-E\left(X\right)\right)\left(Y-E\left(Y\right)\right)=E\left(XY\right)-E\left(X\right)E\left(Y\right)$$

where

$E\left(XY\right)=\sum_{i}x_{i}y_{i}\Pr\left(x_{i},y_{i}\right)$ if X and Y are discrete.

$E\left(XY\right)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf\left(x,y\right)dxdy$ if X and Y are continuous.

$E\left(XY\right)=E\left(X\right)E\left(Y\right)$ 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.

Example 3: Let X, Y, and Z be three random variables, if $Cov\left(X,Z\right)\neq0$ and $Cov\left(Y,Z\right)\neq0$, then $Cov\left(X,Y\right)\neq0$. True/False/Uncertain. Explain.

Solution: The statement is false. Consider the following counter example:

Define $Z = X + Y$ where X and Y are defined to be independent and $Var(X)$ and $Var(Y) \neq 0$.

$$\begin{aligned} Cov(Z, X) &= Cov(X + Y, X) \\ &= Cov(X, X) + Cov(Y, X) \\ &= Var(X) \neq 0 \\ Cov(Z, Y) &= Var(Y) \neq 0 \text{ similarly.} \\ Cov(X, Y) &= 0 \text{ (given)} \end{aligned}$$

(Note that independence of X and Y implies $Cov(X, Y) = 0$.) ■

Definition 13: The **correlation coefficient** between X and Y is defined as:

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

Example 4: Prove that for any two random variables X and Y, $-1 \leq \rho_{xy} \leq 1$.

Solution: For any random variables X and Y, and any real-valued constant t, we have

$$\begin{aligned} Var(tX+Y) &\geq 0 \\ Var\,(tX) + 2Cov(tX,Y) + Var\,(Y) &\geq 0 \\ Var\,(X)\,t^2 + 2Cov(X,Y)t + Var\,(Y) &\geq 0. \end{aligned}$$

since the variance for any random variable is positive.

Consider the solution of a quadratic equation in t,

$$at^2 + bt + c = 0.$$

The solution is

$$t^* = \frac{-b \pm \sqrt{b^2 - 4ac.}}{2a}$$

There will be two solutions if $b^2 - 4ac > 0$, 1 solutions if $b^2 - 4ac = 0$, and no solution if $b^2 - 4ac < 0$.

In our case, $a = Var\,(X) \geq 0$, $b = 2Cov(X,Y)$, $c = Var\,(Y)$.

Since for any value of t the function $at^2 + bt + c \geq 0$, it means $at^2 + bt + c$ never cross the X-axis, so there is at most 1 solution of t such that $at^2 + bt + c = 0$. When $at^2 + bt + c > 0$, there is no solution.

Hence we have $b^2 - 4ac = 0$ or $b^2 - 4ac < 0$.

It implies that $b^2 - 4ac \leq 0$, or

$(2Cov(X,Y))^2 - 4Var\,(X)\,Var\,(Y) \leq 0$

$\iff (Cov(X,Y))^2 \leq Var\,(X)\,Var\,(Y)$

$\iff \dfrac{(Cov(X,Y))^2}{Var\,(X)\,Var\,(Y)} \leq 1$

$\iff -1 \leq \dfrac{Cov(X,Y)}{\sqrt{Var\,(X)\,Var\,(Y)}} \leq 1.$ ■

Exercise 6: Let X, Y, W, and Z be random variables, and a, b, c, d be constants. Show that:

(a) $Var\left(aX+c\right)=Var\left(-aX-d\right)$.

(b) $Cov\left(aX,bY\right)=abCov\left(X,Y\right)$.

(c) $Cov\left(X,X\right)=Var\left(X\right)$.

(d)$Cov\left(aX+bY,cW+dZ\right)=acCov\left(X,W\right)+adCov\left(X,Z\right)+bcCov\left(Y,W\right)+bdCov\left(Y,Z\right)$.

Suppose $W=3+5X$, and $Z=4-8Y$.

(e) Is $\rho_{yz}=1$? Prove or disprove.

(f) Is $\rho_{wz}=\rho_{xy}$? Prove or disprove.

Exercise 7: Let X and Y be two random variables, if $Cov\left(X^2,Y^2\right)=0$, then $Cov\left(X,Y\right)=0$. True/False/Uncertain. Explain.

Exercise 8: Let X and Y be two random variables, if X and Y are independent, then $Cov\left(X^2,Y^2\right)>Cov\left(X,Y\right)$. True/False/Uncertain. Explain.

Exercise 9: A Poisson random variable X has the following distribution

$$\Pr\left(X=j\right)=\frac{e^{-\lambda}\lambda^j}{j!}\qquad j=0,,1,2,.....$$

where $j!=j\left(j-1\right)\left(j-2\right)...1$.

(a) Graph the distribution for $j=0,1,2,3,4$.

(b) Find the mean of X.

(c) Find the variance of X.

Chapter 2

Special Probability Distributions

2.1 Uniform Distribution

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

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

The distribution function is then

$$\begin{aligned} F\,(x) &= 0 \qquad \text{for } x \leq 0; \\ F\,(x) &= x \qquad \text{for } x \in (0,1)\,; \\ F\,(x) &= 1 \qquad \text{for } x \geq 1. \end{aligned}$$

The mean is obviously equal to $\dfrac{1}{2}$. To calculate the variance, note that

$$\begin{aligned} Var(X) &= E\left(X^2\right) - E^2(X) = E\left(X^2\right) - \left(\frac{1}{2}\right)^2 = \int_0^1 x^2 f(x)\,dx - \frac{1}{4} = \int_0^1 x^2 dx - \frac{1}{4} \\ &= \left[\frac{x^3}{3}\right]_0^1 - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \end{aligned}$$

Exercise 1: If $X \sim U(0,1)$, find

(i) $\Pr(X < 0)$;

(ii)$\Pr(X \leq 1)$;

(iii)$\Pr(X > 0)$;

(iv) $\Pr(X \leq 0.5)$;

(v) $\Pr(X > 0.7)$;

(vi) $\Pr(0.4 < X \leq 0.8)$;

(vii) $\Pr(X = 0.8)$.

Note that the area under the density function has to sum up to 1, so if we have a random variable which is uniformly distributed between 1 and 3, i.e. if $X \sim U(1,3)$, then its density function is

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

The distribution function will be

$$\begin{aligned} F(x) &= 0 \qquad \text{for } x \leq 1; \\ F(x) &= \frac{x-1}{2} \qquad \text{for } x \in (1,3); \\ F(x) &= 1 \qquad \text{for } x \geq 3. \end{aligned}$$

Exercise 2: If $X \sim U(1,2)$, find (i) $f(x)$; (ii) $F(x)$; (iii) $E(X)$; (iv) $Var(X)$.

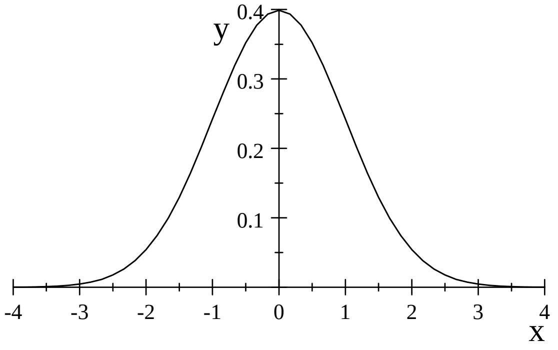
Exercise 3: If $X \sim U(a,b)$, where $a < b$, find (i) $f(x)$; (ii) $F(x)$; (iii) $E(X)$; (iv) $Var(X)$.

2.2 Normal Distribution

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

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

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



N(0,1)

Exercise 4: If $X \sim N(1,4)$, find

(i) $\Pr(X < 0)$;

(ii)$\Pr(X \leq 1)$;

(iii)$\Pr(X > 0)$;

(iv) $\Pr(X \leq -1)$;

(v) $\Pr(X > 2)$;

(vi) $\Pr(1 < X \leq 3)$;

(vii) $\Pr(X = 1)$.

2.3 Standardized Normal Distribution

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

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

Example 1: If $X \sim N(3,4)$, then $Z = \frac{X-3}{2}$ follows $N(0,1)$.

$$\begin{aligned} \Pr(1 \leq X \leq 5) &= \Pr\left(\frac{1-3}{2} \leq \frac{X-3}{2} \leq \frac{5-3}{2}\right) \\ &= \Pr(-1 \leq Z \leq 1) \simeq 0.67. \end{aligned}$$

Exercise 5: If $X \sim N(0,1)$, find

(i) $\Pr(X < 0)$;

(ii)$\Pr(X \leq 1)$;

(iii)$\Pr(X > 0)$;

(iv) $\Pr(X \leq -1)$;

(v) $\Pr(X > 2)$;

(vi) $\Pr(1 < X \leq 3)$;

(vii) $\Pr(X = 1)$.

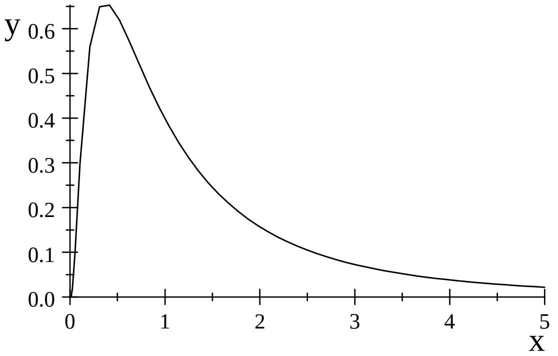
2.4 The Lognormal Distribution

When we study the relationship between a person's IQ score and his income, we find that they are positively correlated. A person with a higher IQ score usually makes more money than a person with a lower IQ score. IQ scores are approximately normally distributed, while the distribution of income is skews to the right and has a long right tail. Thus, it appears that IQ score and income do not have a linear relationship. We use the lognormal distribution to approximate the distribution of income. The lognormal distribution is defined as follows:

If $X \sim N(\mu, \sigma^2)$, and $X = \ln Y$, or equivalently $Y = \exp(X)$, then Y follows a lognormal distribution.

Its density function is:

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



Distribution of Y when lnY is N(0,1).

Thus if X is the IQ score, Y is the income of an individual, then we can treat X as a normally distributed random variable and Y as a lognormally distributed random variable.

Exercise 6: If $X \sim N(0,1)$, $X = \ln Y$, find

(i) $\Pr(Y < 0)$;

(ii)$\Pr(Y \leq 1)$;

(iii)$\Pr(Y > 0)$;

(iv) $\Pr(Y \leq -1)$;

(v) $\Pr(Y > 2)$;

(vi) $\Pr(1 < Y \leq 3)$;

(vii) $\Pr(Y = 1)$.

2.5 Chi-square Distribution

Chi-squared distribution

If $Z \sim N(0,1)$, then Z^2 follows Chi-squared distribution with degree of freedom equals 1.

Example 2: If $Z \sim N(0,1)$, then $U = Z^2$ follows χ_1^2.

$\Pr(0 \leq U \leq 1) = \Pr(-1 \leq Z \leq 1) \simeq 0.67$,

$\Pr(0 \leq U \leq 4) = \Pr(-2 \leq Z \leq 2) \simeq 0.95$,

$\Pr(0 \leq U \leq 9) = \Pr(-3 \leq Z \leq 3) \simeq 0.99$.

Thus a chi-squared 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-squared distribution with k degrees of freedom, and we write it as χ_k^2.

The mean of a chi-squared distribution equals its degrees of freedom. This is because

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

and thus

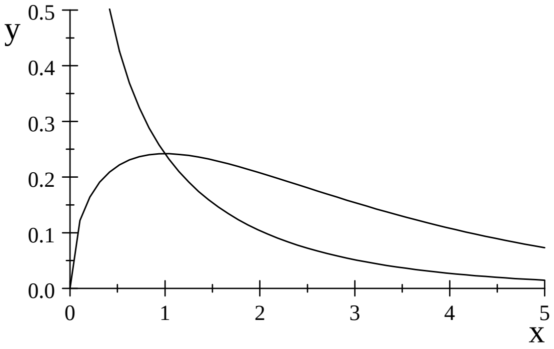
$$E(U) = E\left(Z_1^2 + Z_2^2 + ... + Z_k^2\right) = k.$$

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 $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

A chi-square random variable must take non-negative values, and the distribution has a long right tail.



Chi-square distributions with d.f.=1, 3.

Exercise 7: If $Z \sim N(0,1)$, $U = Z^2$, find

(i) $\Pr(U < 0)$;

(ii)$\Pr(U \leq 1)$;

(iii)$\Pr(U > 0)$;

(iv) $\Pr(U \leq -1)$;

(v) $\Pr(U > 2)$;

(vi) $\Pr(1 < U \leq 3)$;

(vii) $\Pr(U = 1)$.

2.6 Exponential Distribution

For $\theta > 0$, if the random variable X has an exponential distribution with mean θ, then X has the following density function.

$$\begin{aligned} f(x) &= \frac{1}{\theta}e^{-x/\theta}, \qquad 0 < x < \infty \\ f(x) &= 0 \qquad \text{elsewhere} \end{aligned}$$

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

Exercise 8: If X is an exponential distribution with mean 2, find

(i) $\Pr(X < 0)$;

(ii)$\Pr(X \leq 1)$;

(iii)$\Pr(X > 0)$;

(iv) $\Pr(X \leq -1)$;

(v) $\Pr(X > 2)$;

(vi) $\Pr(1 < X \leq 3)$;

(vii) $\Pr(X = 1)$.

2.7 Student's t-Distribution

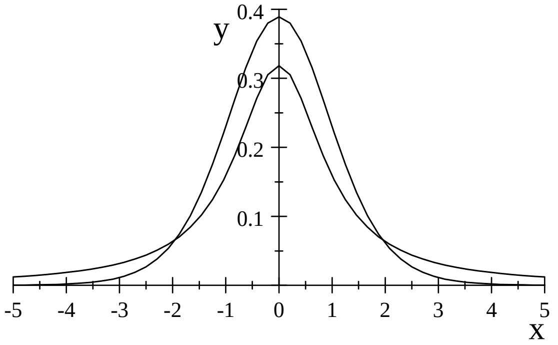
If $Z \sim N(0,1)$, $U \sim \chi^2_k$, and Z and U are independent, then:

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

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". The density function of the t-distribution with degrees of freedom k is given by

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}} \qquad -\infty < t < \infty.$$



t-distributions with d.f.=1,10.

The t-distribution has a thicker tail than the normal distribution. When the degree of freedom goes to infinity, that is when $k \to \infty$, the t-distribution becomes a standardized normal distribution.

This is because as $k \to \infty$, the random variable

$$\frac{U}{k} = \frac{Z_1^2 + Z_2^2 + ... + Z_k^2}{k}$$

which is the sample average of Z_i^2, $(i = 1, 2, ...k)$ will converge to the true mean of Z_i^2, i.e. $E\left(Z_i^2\right)$. Since $E\left(Z_i^2\right) = Var\left(Z_i\right) + E^2\left(Z_i\right) = 1 + 0 = 1$, we have

$$\frac{U}{k} = \frac{Z_1^2 + Z_2^2 + ... + Z_k^2}{k} \to 1.$$

Thus,

$$t = \frac{Z}{\sqrt{U/k}} \to \frac{Z}{\sqrt{1}} = Z \sim N(0,1).$$

Hence a t-distribution with degrees of freedom infinity is a standardize normal distribution. You may check the t-table to see if those critical values for large degrees of freedom are close to the critical values from a $N(0,1)$ table.

Exercise 9: If the random variable t has a t-distribution with degree of freedom 5, find

(i) $\Pr(t \leq 0)$;

(ii)$\Pr(t > 0.267)$;

(iii)$\Pr(t > 0.727)$;

(iv) $\Pr(t \leq 1.476)$;

(v) $\Pr(t > 2.015)$;

(vi) $\Pr(2.571 < t \leq 3.365)$;

(vii) $\Pr(t = 1)$.

2.8 Cauchy Distribution

Let Z_1 and Z_2 be independent and follow $N(0,1)$, then the ratio

$$R = \frac{Z_1}{Z_2}$$

will have a Cauchy distribution. A Cauchy distribution is a t-distribution with 1 degree of freedom.

Its density has the form:

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

For most distributions, the mean and variance are finite. However, the mean and variance of a Cauchy distribution do not exist. In other words, when we draw a sample of size n from a Cauchy distribution, the sample average will not converge to a constant no matter how large the sample size is.

Exercise 10: If the random variable R has a Cauchy distribution, find
(i) $\Pr(R \leq 0)$;
(ii)$\Pr(R > 0.325)$;
(iii)$\Pr(R > 1)$;
(iv) $\Pr(R \leq 3.078)$;
(v) $\Pr(R > 6.314)$;
(vi) $\Pr(12.706 < R \leq 31.821)$;
(vii) $\Pr(R = 1)$.

2.9 F-Distribution

If $U \sim \chi_m^2$ and $V \sim \chi_n^2$, and if U and V are independent of each other, then

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

has an F-distribution with m and n degrees of freedom.

Note that:

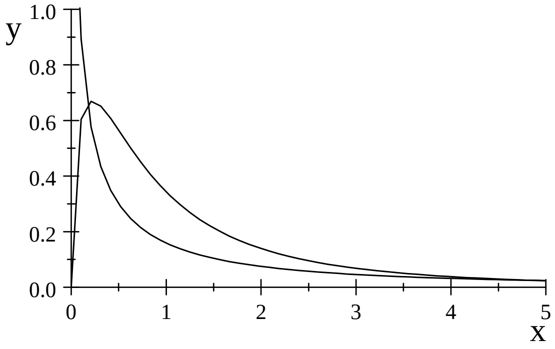
$$F(1,k) = t_k^2.$$

The density function of the F-distribution with degrees of freedom (m,n) is given by

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\left(\frac{m}{2}-1\right)}\left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}} \qquad \text{for } 0 \leq x < \infty,$$

and

$$f(x) = 0 \qquad \text{for } x < 0.$$



F-distributions with d.f.=(1,1) and (3,4).

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

Example 3: 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 $\dfrac{Z_{k+1}}{\sqrt{U/k}}$ and $\dfrac{Z_{k+1}^2}{U/k}$?

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.

Solution:

(a) $U \sim \chi_k^2$.

$$\begin{aligned} E(U) &= E(Z_1^2 + Z_2^2 + ... + Z_k^2) \\ &= E(Z_1^2) + E(Z_2^2) + ... + E(Z_k^2) \\ &= 1 + 1 + ... + 1 \qquad \text{since } E(Z_i^2) = Var(Z_i) + [E(Z_i)]^2 \text{ for } i = 1, 2, ..., k. \\ &= k. \end{aligned}$$

■

(b) Since Z_{k+1} and U are independent, $\dfrac{Z_{k+1}}{\sqrt{U/k}} \sim t_k$ and $\dfrac{Z_{k+1}^2}{U/k} \sim F(1,k)$. ■

(c) This statement is false. It is possible that $Z_{k+1}^2 > U$ and hence $V < 0$. Since, as we know, chi-square distribution should be positive, V does not have a chi-square distribution. ■

Exercise 11: If the random variable F has a F-distribution with degrees of freedom $(1,5)$, find

(i) $\Pr(F \leq 0)$;

(ii)$\Pr(F > 0.071289)$;

(iii)$\Pr(F > 0.528529)$;

(iv) $\Pr(F \leq 2.178576)$;

(v) $\Pr(F > 4.060225)$;

(vi) $\Pr(6.610041 < F \leq 11.323225)$;

(vii) $\Pr(F = 1)$.

Exercise 12: For $k > 4$, let Z_1,..., Z_k be independent $N(0,1)$ random variables, and let

$$U = Z_1^2 + Z_2^2 + Z_3^2,$$

$$V = Z_4^2 + Z_5^2 + Z_6^2 + ... + Z_{k-1}^2 + Z_k^2.$$

a) What are the distributions of U and V? Find $E\left(U\right)$ and $E\left(V\right)$.

b) What is the distribution of $\dfrac{U/3}{V/\left(k-3\right)}$? Find $E\left(\dfrac{U/3}{V/\left(k-3\right)}\right)$ and $E\left(UV\right)$.

Chapter 3

Estimation and Hypothesis Testing

3.1 Point Estimation

Population and sample are very different concepts. We would like to know the mean (μ) and the variance (σ^2) of a population, but we will never know these as we do not have the resources to do a detailed calculation of the population. Even in the case of throwing a dice, we do not know whether the dice is leaded or not. What we can do, however, is to draw a sample from the population. A sample is a subset of a population. hopefully, we can retrieve information about a population from a sample when the sample size is large enough. We usually construct estimators to estimate the population mean and variance.

Definition 1: 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 usually constructed by using 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 2: An **estimate** is the numerical value taken by an estimator, it usually depends on what sample is drawn.

Example 1:

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, then 3.4 is an estimate of the population mean. Thus the estimate differs from sample to sample.

Example 2:

The statistic

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

is also an estimator of the population mean. Note that we usually put a symbol on top of X to indicate that it is an estimator. Conventionally, $\overline{X}$ denotes the sample mean, we may use $\widetilde{X}$, $\widehat{X}$, X^*, etc. to denote other estimators.

Example 3:

An weighted average

$$\widetilde{X} = w_1X_1 + w_2X_2 + ... + w_TX_T \qquad \text{where } \sum_{i=1}^{T} w_i = 1$$

is also estimator of the population mean.

Example 4:

A single observation X_1 is also an estimator of the population mean.

Example 5:

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

can also be an estimator of the population mean.

Example 6:

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 in this case. 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. The problem is which one is the best, and what criteria should be used to define a good estimator.

In choosing the best estimator, we usually use criterion such as linearity, unbiasedness and efficiency.

The first criterion in choosing estimator is linearity, an linear estimator is by construction simpler than a nonlinear estimator.

Definition 2: An estimator $\widehat{X}$ is **linear** if it is a linear combination of the sample observations. i.e.

$$\widehat{X} = a_1X_1 + a_2X_2 + ... + a_TX_T$$

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 examples 1, 2, 3 and 4 are linear, while estimators in example 5 and 6 are not linear.

We reduce the set of all possible estimators to the set linear estimators. Still, there are plenty of linear estimators, so how should they be compared? We introduce the concept of unbiasedness.

Definition 3: 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 note that any single observation from the sample is unbiased. i.e.

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

This is because when an observation is drawn from a population, what is the best and most reasonable guess?

The best and most reasonable guess is to expect the observation to be 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 7: If X_t $(t = 1, 2, ..., T)$ are random variables with $E(X_t) = \mu$ and $Var(X_t) = \sigma^2$. Show that:

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

(b) Find $E(X_t^2)$ and $E\left(\left(\overline{X}\right)^2\right)$ in terms of μ and σ^2.

(c) Show that $\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2 = \sum_{t=1}^{T} X_t^2 - T\left(\overline{X}\right)^2$.

(d) Use (a) and (c), show that $\widehat{\sigma}^2 = \frac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}{T-1}$ is an unbiased estimator for σ^2.

Solution:(a)

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

■

(b)

$$\begin{aligned}
Var\left(X_t\right) &= \sigma^2 = E\left(X_t^2\right) - E^2\left(X_t\right) \\
&= E\left(X_t^2\right) - \mu^2 \\
\Rightarrow E\left(X_t^2\right) &= \sigma^2 + \mu^2
\end{aligned}$$

■

$$
\begin{aligned}
Var(\overline{X}) &= Var\left(\frac{1}{T}\sum_{t=1}^{T} X_t\right) \\
&= \frac{1}{T^2} Var\left(\sum_{t=1}^{T} X_t\right) \\
&= \frac{1}{T^2}\sum_{t=1}^{T} Var\left(X_t\right) \text{ since } X_t \text{ is } i.i.d. \\
&= \frac{T\sigma^2}{T^2} = \frac{\sigma^2}{T}.
\end{aligned}
$$

Also,

$$
\begin{aligned}
Var\left(\overline{X}\right) &= E\left(\overline{X}^2\right) - E^2\left(\overline{X}\right) \\
&= E\left(\overline{X}^2\right) - \mu^2 \\
\Rightarrow E\left(\overline{X}^2\right) &= \frac{\sigma^2}{T} + \mu^2.
\end{aligned}
$$

■

(c)

$$
\begin{aligned}
\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2 &= \sum_{t=1}^{T}\left(X_t^2 - 2X_t\overline{X} + \overline{X}^2\right) \\
&= \sum_{t=1}^{T} X_t^2 - 2\overline{X}\sum_{t=1}^{T} X_t + T\overline{X}^2 \\
&= \sum_{t=1}^{T} X_t^2 - 2T\overline{X}^2 + T\overline{X}^2 \\
&= \sum_{t=1}^{T} X_t^2 - T\overline{X}^2.
\end{aligned}
$$

■

(d)

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

■

Exercise 1: Show that the estimators in examples 1, 3 and 4 are unbiased, and that the estimators in examples 2, 5 and 6 are biased.

We further reduce the set of all possible estimators to the set of linear and unbiased estimators. However, if there are plenty of linear and unbiased estimators, how should we compare them? For this, we introduce the concept of efficiency.

Definition 4: An estimator $\widehat{X}$ is more **efficient** than another estimator X^* if $Var\left(\widehat{X}\right) < Var\left(X^*\right).$

Example 8: If we look at the efficiency criteria, the estimator in example 6 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.

Exercise 2: 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.

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

$$\widehat{X}_b = \frac{X_1 + X_2 + X_3}{3}.$$

Exercise 3: Rank the efficiency of the estimators in examples 1 to 6.

Definition 5: 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. This is because if the estimator truly reveals the value of the population mean μ, the variation of this estimator should become smaller and smaller when the sample is getting larger and larger. 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 an consistent estimator. If the estimator is biased, it may also be consistent, 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 as it involves the concept of infinity. It is very important for an estimator to be consistent, as what we want is to retrieve information about the population mean from the estimator. If an estimator is inconsistent, it tells us nothing about the population no matter how large the sample is.

One of the consistent estimators is the sample mean

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

Why it is consistent? Note first 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

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

Note 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 $T \to \infty$,

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

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

An estimator can also be unbiased but inconsistent. Consider the estimator in example 4, a single observation as an estimator for the population mean. It is unbiased. However, it is inconsistent as we only use one observation from a sample of size T, no matter how large T is. Thus, increasing the number of other observation is of no use in improving the precision of this estimator.

Exercise 4: Construct an estimator which is biased, consistent and less efficient than the simple average $\overline{X}$.

Exercise 5: 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.

3.2 The Law of Large Numbers and the Central Limit Theorem

Definition 6: A sequence of random variables X_t, $(t = 1, 2, ...T)$ follow an **Independent and Identical Distribution (i.i.d.)** if all the X_t have the same distribution and X_i does not depend on X_j for any $i \neq j$.

The **Law of Large Numbers** states that, if X_t is an i.i.d. with finite mean μ and finite variance σ^2, the sample average $\overline{X}$ converges to the true mean μ as the sample size n goes to infinity.

Exercise 6: To illustrate 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, recording the value of $\overline{X}$ and plot a diagram which indicates its behavior as T increases from 1 to 30. Does $\overline{X}$ converge to 3.5?

The **Central Limit Theorem** states that, if X_t is an 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 $\frac{\sigma^2}{T}$, as the sample size T goes to infinity.

It is a powerful theorem because X_t can come from **any** distribution.

Example 9: 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

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

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

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

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.

Exercise 7: To show the Central limit Theorem, let us consider the random experiment of throwing a dice T times in the previous exercise .

(a) Conduct the experiment yourself, with $T = 30$. 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) Repeat part (b) 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 $\overline{X}$.

Exercise 8: Use computers or calculators to generate 36 random numbers from the uniform distribution $U(0,1)$; calculate the sample mean, and repeat this procedure 100 times. 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.

3.3 Testing a Statistical Hypothesis

In the real world, when we observe a phenomenon, we would liket to explain it a hypothesis. We usually post a null hypothesis, and an alternative hypothesis, which is the set of complement of events described in the null hypothesis.

For example, when we observe that the death toll in winter is usually higher than the death toll in the other three 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 positively related to temperature.

A hypothesis is not a theorem, which is always true when certain assumptions are held. A hypothesis is just a guess, which may be wrong. Thus, we have to test how likely our hypothesis is going to be correct. In testing a hypothesis, we cannot be sure that it is a correct hypothesis, as otherwise it would become a theorem. As such, we may commit errors when concluding a hypothesis. There are two possible types of errors as described below.

Definition 7: Rejection of the null hypothesis when it is true is called the **Type I Error**; the probability of committing the type I error is denoted by α.

Definition 8: Acceptance of the null hypothesis when it is false is called the **Type II Error**; the probability of committing the type II error is denoted by β.

We want to reduce both Type I and Type II errors as much as possible. However, as there is no free lunch, there is no way to eliminate both errors. Reducing the chance of committing the Type I Error will increase the chance of committing the Type II Error. Reducing the chance of committing the Type II Error will increase the chance of committing the Type I Error.

Exercise 9: In a judicial trial, suppose the null hypothesis is that "the defendant is not guilty".

(a) 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?

3.4 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 $\overline{X} = \frac{1}{T}(X_1 + X_2 + ... + X_T) \sim N\left(\mu_0, \dfrac{\sigma^2}{T}\right)$.

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 (i) the observations X_t come from the normal distribution and (ii) the variance is known. A more commonly used test is the t-test, which is used when the population variance is unknown and the sample size is small.

3.5 The t-Test

Consider a random sample X_1, X_2,...X_T drawn from a **normal** distribution with unknown mean μ and a **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

$$\widehat{\sigma} = \sqrt{\frac{\sum\limits_{t=1}^{T} \left(X_t - \overline{X}\right)^2}{T-1}}.$$

Note that

$$t_{obs} = \frac{\overline{X} - \mu_0}{\widehat{\sigma}/\sqrt{T}} = \frac{\frac{\overline{X}-\mu_0}{\sigma/\sqrt{T}}}{\sqrt{\frac{1}{T-1}\sum\limits_{t=1}^{T}\left(\frac{X_t-\overline{X}}{\sigma}\right)^2}}.$$

Under $H_0 : \mu = \mu_0$, $X_t \sim N\left(\mu_0, \sigma^2\right)$, thus $\overline{X} = \frac{1}{T}\left(X_1 + X_2 + ... + X_T\right) \sim N\left(\mu_0, \dfrac{\sigma^2}{T}\right)$, and

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{T}} \sim N\left(0, 1\right).$$

Further, it can be shown that (very difficult)

$$\sum_{t=1}^{T} \left(\frac{X_t - \overline{X}}{\sigma}\right)^2$$

has a Chi-squared 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,

$$t_{obs}=\frac{\overline{X}-\mu_0}{\widehat{\sigma}/\sqrt{T}}=\frac{\frac{\overline{X}-\mu_0}{\sigma/\sqrt{T}}}{\sqrt{\frac{1}{T-1}\sum\limits_{t=1}^{T}\left(\frac{X_t-\overline{X}}{\sigma}\right)^2}}=\frac{N(0,1)}{\sqrt{\chi^2_{T-1}/(T-1)}}$$

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 the 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 the 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 10: 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, 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.

Exercise 10: 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 11: 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 June 1999.

(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 12: 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 1998 to June 1999.

(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 13: 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 1998 to June 1999.

(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 $\alpha = 5\%$.

3.6 What if $\mathbf{X}_t$ are not Normally Distributed?

Thus far we have assumed 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.

Exercise 14: 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.)

Chapter 4

Simple Regression Models

4.1 Introduction

Regression analysis is a statistical technique used to describe relationships among variables. The simplest case to examine is the one in which a variable Y, referred to as the dependent variable, may be related to another variable X, called an independent or explanatory variable. If the relationship between Y and X is linear, then the equation expressing this relationship will be the equation for a line:

$$Y = \beta_0 + \beta_1 X$$

where β_0 and β_1 are constants.

This is an **exact** or **deterministic** linear relationship. Exact linear relationship may be encountered in various science course. In social sciences as well as in economics, exact linear relationships are the exception rather than the rule. In most cases in economics, X and Y may be linearly related, but is not an exact relationship. There may be some unknown factors that also affect Y, we used u to represent all these unknown factors, thus we write

$$Y = \beta_0 + \beta_1 X + u.$$

For example, when Y is consumption and X is income, then the above model is a consumption function.

β_1 represents when X increases 1 unit, Y will increase β_1 unit(s).

β_0 is the value of Y when $X = 0$.

We would like to estimate the **unknown** parameter β_0 and β_1 based on our observations $\{X_t, Y_t\}_{t=1}^{T}$.

When we plot the observations and try to find a line which fits these observations the best, how should we do it? What criteria should we use?

Of course not all the data lie on our line, so we have to minimize the "distance" between the observations and the line. What distance? Statistically speaking, we may use vertical distance, horizontal distance or distance that are perpendicular to our line. In Economics, we use vertical distance. However, we are not just minimizing the sum of errors, as it is possible that some big positive errors may be cancelled out by some big negative errors, ending up with a small value of the sum of errors. Thus, we have to make the errors positive. We may take absolute values, but we would like to penalize observations which are further away from the line. Thus, we minimize the sum of squared errors. This is the **Ordinary Least Squares** (**OLS**) Estimation method, proposed in the 19th century by the French mathematician Adrien Legendre.

Let $\widehat{\beta}_0$, $\widehat{\beta}_1$ be the OLS estimators for β_0 and β_1 respectively. To ensure that the estimators have the desirable properties such as unbiasedness, efficiency and consistency, we have to make the following assumptions:

4.2 Assumptions

1: The true model (population) is a linear model, i.e.,

$$Y_t = \beta_0 + \beta_1 X_t + u_t.$$

Linearity means **linear in β's**, not necessarily linear in Y and X.

e.g. $Y_t = \beta_0 + \beta_1 X_t^2 + u_t$ is a linear model, but $Y_t = \beta_0 + \beta_1^2 X_t + u_t$ is not a linear model.

This assumption allows us to derive the OLS estimator $\widehat{\beta}_0$, $\widehat{\beta}_1$ via simple calculus. If the model is nonlinear in $\beta' s$, the problem becomes very complicated when taking differentiation on $\beta' s$.

2: $E(u_t) = 0$ for all t.

This assumption is to ensure that the OLS estimators are unbiased, i.e. $E\left(\widehat{\beta}_0\right) = \beta_0$ and $E\left(\widehat{\beta}_1\right) = \beta_1$ if this assumption is made.

3: X_t cannot be all the same.

This assumption is to ensure that we will not obtain a vertical line with infinite slope. If the slope is infinity, the model becomes meaningless.

4: X_t is given and is non-random, in the sense that you can choose the values of X_t. (This assumption can be relaxed later)

This assumption simplifies our analysis when we discuss the unbiasedness of the estimators, since X can be treated as a constant and taken out of the expectation operator. For example, $E(X_t u_t) = X_t E(u_t) = 0$ by assumption 2. This also implies $Cov(X_t, u_t) = 0$.

5: Homoscedasticity, i.e., $Var(u_t) = \sigma^2 \quad$ for all t.

6: Serial Independence, i.e., $Cov(u_t, u_s) = 0 \quad$ for all $t \neq s$.

Assumptions 5 and 6 simplify our calculation of $Var\left(\widehat{\beta}_0\right)$ and $Var\left(\widehat{\beta}_1\right)$, see example 2 below. They also ensure that OLS estimators are the most efficient estimators among all the linear and unbiased estimators.

As far as the estimation of $\beta's$ is concerned, assumptions 1 to 6 ensure the OLS estimators are the best linear unbiased estimators (**BLUE**) .

4.3 Least Squares Estimation

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

$$u_t = Y_t - \beta_0 - \beta_1 X_t.$$

The problem is

$$\min_{\beta_0, \beta_1} \sum_{t=1}^{T} (Y_t - \beta_0 - \beta_1 X_t)^2 .$$

The first order conditions are:

$$\left.\frac{\partial\sum\limits_{t=1}^{T}\left(Y_t-\beta_0-\beta_1X_t\right)^2}{\partial\beta_0}\right|_{\widehat{\beta}_0,\widehat{\beta}_1}=-2\sum_{t=1}^{T}\left(Y_t-\widehat{\beta}_0-\widehat{\beta}_1X_t\right)=0, \qquad (*)$$

$$\left.\frac{\partial\sum\limits_{t=1}^{T}\left(Y_t-\beta_0-\beta_1X_t\right)^2}{\partial\beta_1}\right|_{\widehat{\beta}_0,\widehat{\beta}_1}=-2\sum_{t=1}^{T}\left(Y_t-\widehat{\beta}_0-\widehat{\beta}_1X_t\right)X_t=0. \qquad (**)$$

Solving these two **normal equations** gives the **Ordinary Least Squares Estimators**:

$$\widehat{\beta}_1=\frac{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)\left(Y_t-\overline{Y}\right)}{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)^2},$$

$$\widehat{\beta}_0=\overline{Y}-\widehat{\beta}_1\overline{X}.$$

Note: If X are also assumed to be random, then when sample size increases, $\widehat{\beta}_1$ will converge to $\dfrac{Cov(X,Y)}{Var(X)}$.

Example 1: Show that

$$\widehat{\beta}_1=\beta_1+\frac{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)u_t}{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)^2}.$$

Solution:

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)\left(Y_t-\overline{Y}\right)}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2} \\
&= \frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)Y_t}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2} \\
&= \frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)\left(\beta_0+\beta_1X_t+u_t\right)}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2} \\
&= \beta_0\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}+\beta_1\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)X_t}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}+\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)u_t}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2} \\
&= \beta_0\frac{0}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}+\beta_1\left(1\right)+\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)u_t}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2} \\
&= \beta_1+\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)u_t}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}.
\end{aligned}
$$

Exercise 1: Solve (*) and (**) for $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

4.4 Properties of Estimators

Under the above assumptions 1-6, the Least Squares Estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have the following properties:

(1) They are linear estimators, i.e. they are linear combinations of Y_t.

Proof.

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)\left(Y_t-\overline{Y}\right)}{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)^2} \\
&= \frac{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)Y_t}{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)^2} \\
&= \frac{X_1-\overline{X}}{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)^2}Y_1+\frac{X_2-\overline{X}}{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)^2}Y_2+...+\frac{X_T-\overline{X}}{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)^2}Y_T \\
&= \sum_{i=1}^{T}a_iY_i,
\end{aligned}$$

where

$$a_i=\frac{X_i-\overline{X}}{\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)^2}.$$

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{Y}-\widehat{\beta}_1\overline{X} \\
&= \frac{1}{T}\sum_{i=1}^{T}Y_i-\left(\sum_{i=1}^{T}a_iY_i\right)\overline{X} \\
&= \sum_{i=1}^{T}\frac{1}{T}Y_i-\sum_{i=1}^{T}\overline{X}a_iY_i \\
&= \sum_{i=1}^{T}\left(\frac{1}{T}-\overline{X}a_i\right)Y_i \\
&= \sum_{i=1}^{T}b_iY_i,
\end{aligned}$$

where

$$b_i = \frac{1}{T} - \overline{X} a_i = \frac{1}{T} - \overline{X} \left(\frac{X_i - \overline{X}}{\sum\limits_{t=1}^{T} \left(X_t - \overline{X}\right)^2}. \right).$$

(2) They are unbiased, i.e. $E\left(\widehat{\beta}_0\right) = \beta_0$ and $E\left(\widehat{\beta}_1\right) = \beta_1$.

Proof. From example 1,

$$\widehat{\beta}_1 \quad = \quad \beta_1 + \frac{\sum_{t=1}^{T} \left(X_t - \overline{X}\right) u_t}{\sum_{t=1}^{T} \left(X_t - \overline{X}\right)^2}.$$

Thus

$$\begin{aligned} E\left(\widehat{\beta}_1\right) &= E\left(\beta_1 + \frac{\sum_{t=1}^{T} \left(X_t - \overline{X}\right) u_t}{\sum_{t=1}^{T} \left(X_t - \overline{X}\right)^2}\right) \\ &= \beta_1 + \frac{\sum_{t=1}^{T} \left(X_t - \overline{X}\right) E\left(u_t\right)}{\sum_{t=1}^{T} \left(X_t - \overline{X}\right)^2} \\ &= \beta_1 + \frac{\sum_{t=1}^{T} \left(X_t - \overline{X}\right) \times 0}{\sum_{t=1}^{T} \left(X_t - \overline{X}\right)^2} \\ &= \beta_1. \end{aligned}$$

■

$$\begin{aligned}
E\left(\widehat{\beta}_0\right) &= E\left(\overline{Y}-\overline{X}\widehat{\beta}_1\right)\\
&= E\left(\frac{\sum_{t=1}^{T}Y_t}{T}\right)-\overline{X}E\left(\widehat{\beta}_1\right)\\
&= E\left(\frac{\sum_{t=1}^{T}\left(\beta_0+\beta_1X_t+u_t\right)}{T}\right)-\overline{X}\beta_1\\
& E\left(\beta_0\frac{\sum_{t=1}^{T}1}{T}+\beta_1\frac{\sum_{t=1}^{T}X_t}{T}+\frac{\sum_{t=1}^{T}u_t}{T}\right)-\overline{X}\beta_1\\
&= \beta_0+\overline{X}\beta_1+E\left(\frac{\sum_{t=1}^{T}u_t}{T}\right)-\overline{X}\beta_1\\
&= \beta_0+E\left(\frac{\sum_{t=1}^{T}u_t}{T}\right)\\
&= \beta_0+\frac{1}{T}\sum_{t=1}^{T}E\left(u_t\right)\\
&= \beta_0,\ \text{since } E(u_t)=0 \qquad \blacksquare
\end{aligned}$$

(3) They are consistent, i.e. $\widehat{\beta}_0 \rightarrow \beta_0$ and $\widehat{\beta}_1 \rightarrow \beta_1$ as the sample size goes to infinity.
Proof. Skip.

(4) They are efficient among all the linear unbiased estimators by the Gauss-Markov Theorem.

Gauss−Markov Theorem: Under assumptions 1-6, the Ordinary Least Squares(OLS) estimators are the Best Linear Unbiased Estimators ($BLUE$):
Proof. Skip.

(5) The estimated regression line must pass through the point $(\overline{X},\overline{Y})$.

Proof. Note that the estimated regression line is

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

By the definition of $\widehat{\beta}_0 = \overline{Y} - \overline{X}\widehat{\beta}_1$,

$$\begin{aligned} y &= \overline{Y} - \overline{X}\widehat{\beta}_1 + \widehat{\beta}_1 x \\ y - \overline{Y} &= \widehat{\beta}_1 \left(x - \overline{X}\right) \end{aligned}$$

The question is where the line passes through the point $(\overline{X}, \overline{Y})$, if it does, then the equality should hold when we put $x = \overline{X}$ and $y = \overline{Y}$. This is obvious since

$$\begin{aligned} \overline{Y} - \overline{Y} &= \widehat{\beta}_1 \left(\overline{X} - \overline{X}\right) \\ 0 &= 0 \blacksquare \end{aligned}$$

Although OLS has so many nice properties, it also has shortcomings. If there are observations whose values are extremely large, those observations will dominate other observations in the determination of the OLS estimates. In other words, the OLS estimator is not robust to outliers.

Example 2: Derive $Var\left(\widehat{\beta}_1\right)$.

Solution:

$$
\begin{aligned}
Var\left(\widehat{\beta}_1\right) &= Var\left(\beta_1 + \frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)X_t}\right) \\
&= Var\left(\frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)X_t}\right) \qquad \text{since } \beta_1 \text{ is a constant} \\
&= \frac{1}{\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)X_t\right)^2} Var\left(\sum_{t=1}^{T}\left(X_t - \overline{X}\right)u_t\right) \\
&= \frac{\sum\limits_{t=1}^{T} Var\left(\left(X_t - \overline{X}\right)u_t\right)}{\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)X_t\right)^2} \qquad \text{since Cov}\left(u_i, u_j\right)\text{=0 for all i} \neq \text{j} \\
&= \frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2 Var\left(u_t\right)}{\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2\right)^2} = \frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2 \sigma^2}{\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2\right)^2} = \frac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}.
\end{aligned}
$$

Exercise 2: True/False/Uncertain, explain.

a) OLS estimators are most efficient among all estimators.

b) The R^2 increases with the number of observations.

c) If $E\left(u_t\right) = 2$, $\widehat{\beta}_0$ will be biased.

d) If $E\left(u_t\right) = 2$, $\widehat{\beta}_1$ will be biased.

Exercise 3: Show that $Cov\left(\overline{u}, \widehat{\beta}_1\left(X_t - \overline{X}\right)\right) = 0$.

Exercise 4: Derive $Var\left(\widehat{\beta}_0\right)$ and $Cov\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$.

4.5 Goodness of Fit

To see whether the regression line fits the data, we first define the variation of Y about its mean as the total sum of squares (TSS), where

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

Let

$$\widehat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 X_t$$

be the predicted value of Y_t given X_t. Consider the following identity:

$$Y_t - \overline{Y} \equiv \left(\widehat{Y}_t - \overline{Y}\right) + \left(Y_t - \widehat{Y}_t\right) .$$

Squaring both sides gives

$$\left(Y_t - \overline{Y}\right)^2 = \left(\widehat{Y}_t - \overline{Y}\right)^2 + \left(Y_t - \widehat{Y}_t\right)^2 + 2\left(\widehat{Y}_t - \overline{Y}\right)\left(Y_t - \widehat{Y}_t\right) .$$

Summing up from $t = 1$ to T, we have

$$\sum_{t=1}^{T} \left(Y_t - \overline{Y}\right)^2 = \sum_{t=1}^{T} \left(\widehat{Y}_t - \overline{Y}\right)^2 + \sum_{t=1}^{T} \left(Y_t - \widehat{Y}_t\right)^2 + 2\sum_{t=1}^{T} \left(\widehat{Y}_t - \overline{Y}\right)\left(Y_t - \widehat{Y}_t\right) .$$

We want to show the last item in the R.H.S. is zero. Note that

$$\begin{aligned}
& \sum_{t=1}^{T}\left(\widehat{Y}_t-\overline{Y}\right)\left(Y_t-\widehat{Y}_t\right) \\
= & \sum_{t=1}^{T}\left(\widehat{Y}_t-\overline{Y}\right)\left(Y_t-\widehat{\beta}_0-\widehat{\beta}_1X_t\right) \\
= & \sum_{t=1}^{T}\widehat{Y}_t\left(Y_t-\widehat{\beta}_0-\widehat{\beta}_1X_t\right)-\overline{Y}\sum_{t=1}^{T}\left(Y_t-\widehat{\beta}_0-\widehat{\beta}_1X_t\right) \\
= & \sum_{t=1}^{T}\widehat{Y}_t\left(Y_t-\widehat{\beta}_0-\widehat{\beta}_1X_t\right)-\overline{Y}\times 0 \qquad \text{by the normal equations (*)} \\
= & \sum_{t=1}^{T}\left(\widehat{\beta}_0+\widehat{\beta}_1X_t\right)\left(Y_t-\widehat{\beta}_0-\widehat{\beta}_1X_t\right) \\
= & \widehat{\beta}_0\sum_{t=1}^{T}\left(Y_t-\widehat{\beta}_0-\widehat{\beta}_1X_t\right)+\widehat{\beta}_1\sum_{t=1}^{T}\left(Y_t-\widehat{\beta}_0-\widehat{\beta}_1X_t\right)X_t \\
= & 0 \qquad \text{by the normal equations (*) and (**).}
\end{aligned}$$

Thus we have:

$$\underbrace{\sum_{t=1}^{T}\left(Y_t-\overline{Y}\right)^2}_{TSS}=\underbrace{\sum_{t=1}^{T}\left(\widehat{Y}_t-\overline{Y}\right)^2}_{RSS}+\underbrace{\sum_{t=1}^{T}\left(Y_t-\widehat{Y}_t\right)^2}_{ESS}$$

where

TSS stands for the total sum of squares,

RSS stands for the regression sum of squares, and

ESS stands for the error sum of squares.

Thus the difference between Y_t and $\overline{Y}$ can be decomposed into two parts. The first part is

$$\left(\widehat{Y}_t-\overline{Y}\right)=\left(\widehat{\beta}_0+\widehat{\beta}_1X_t\right)-\left(\widehat{\beta}_0+\widehat{\beta}_1\overline{X}\right)=\widehat{\beta}_1\left(X_t-\overline{X}\right).$$

This part shows that Y_t differs from its average because X_t differs from its average.

The second part $\left(Y_t - \widehat{Y}_t\right)$ is the unknown reason why Y_t varies. It is the residual that remains unexplained by the regressor X_t

We define

$$R^2 = 1 - \frac{ESS}{TSS}.$$

Since ESS and TSS are positive, and $TSS \geq ESS$, the range for R^2 is

$$0 \leq R^2 \leq 1.$$

We use R^2 to measure the goodness of fit of a regression line. If R^2 is close to 0, X and Y do not have linear relationship. If R^2 is close to 1, X and Y are highly linearly correlated.

If X cannot explain Y at all, then $RSS = 0$, $TSS = ESS$, and $R^2 = 0$ in this case, and the regression line does not fit the data.

If there is nothing that remains unexplained, then $ESS = 0$, that means the variation of Y can be totally explained by the variation of X, and $R^2 = 1$ in this case, and all the data must lie on the regression line.

Remark: These abbreviations TSS, ESS and RSS are drawn from Ramanathan text, some other texts and computer programs(e.g. MFIT386) use RSS to represent the residual sum of squares and ESS to denote the explained sum of squares, which are the opposites of Ramanathan's definitions. Therefore be careful when you use these abbreviations.

4.6 Properties of $\mathbf{R}^2$

(1) In the simple regression model (i.e., only one regressor X), R^2 can be written as

$$R^2 = \frac{\left(\sum_{t=1}^{T}\left(X_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)\right)^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2 \sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2}.$$

Proof.

$$\begin{aligned}
R^2 &= 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS} \\
&= \frac{\sum_{t=1}^{T}\left(\widehat{Y}_t - \overline{Y}\right)^2}{\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2} = \frac{\sum_{t=1}^{T}\left(\widehat{\beta}_1\left(X_t - \overline{X}\right)\right)^2}{\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2} \\
&= \widehat{\beta}_1^2 \frac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}{\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2} = \widehat{\beta}_1^2 \frac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}{\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2} \\
&= \left(\frac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}\right)^2 \frac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}{\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2} \\
&= \frac{\left(\sum_{t=1}^{T}\left(X_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)\right)^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2 \sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2}. \blacksquare
\end{aligned}$$

(2) Given the data (X_t, Y_t), $t = 1, 2, ..T$, We run a regression of Y_t on X_t and obtain the following results

$$\widehat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 X_t, \qquad R^2 = a.$$

Now suppose we use the same data and run a regression of X_t on Y_t, and obtain the following regression.

$$\widehat{X}_t = \widehat{\alpha}_0 + \widehat{\alpha}_1 Y_t, \qquad R^2 = b.$$

Then

$$a = b = \widehat{\beta}_1 \widehat{\alpha}_1.$$

Proof.

$$a = \frac{\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)\right)^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2 \sum\limits_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2}$$

$$b = \frac{\left(\sum\limits_{t=1}^{T}\left(Y_t - \overline{Y}\right)\left(X_t - \overline{X}\right)\right)^2}{\sum\limits_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2 \sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2} = a$$

$$\begin{aligned}
\widehat{\beta}_1 \widehat{\alpha}_1 &= \frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2} \times \frac{\left(\sum\limits_{t=1}^{T}\left(Y_t - \overline{Y}\right)\left(X_t - \overline{X}\right)\right)^2}{\sum\limits_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2} \\
&= \frac{\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)\right)^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2 \sum\limits_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2} = a. \quad \blacksquare
\end{aligned}$$

(3) It is possible that R^2 may turn out to be negative or bigger than one if we run a regression without an intercept. See example 4 below.

Example 3: Given the data (X_t, Y_t), $t = 1, 2, ..T$, suppose we know $\overline{X} = 30$. We run a regression of Y_t on X_t and obtain the following results

$$\widehat{Y}_t = 0.8 + 0.9X_t, \qquad R^2 = 0.9.$$

Now suppose we use the same data and run a regression of X_t on Y_t, and obtain the following regression.

$$\widehat{X}_t = a + bY_t, \qquad R^2 = c.$$

Find the values of $\overline{Y}$, a, b, and c.

Solution: Given that $\widehat{Y}_t = 0.8 + 0.9X_t$, $R^2 = 0.9$ and $\overline{X} = 30$.

$$\overline{Y} = 0.8 + 0.9\overline{X} = 0.8 + 0.9\,(30) = 27.8.$$

Regression of Y_t on X_t yields ■

$$R^2 = \frac{\left(\sum_{t=1}^{T}\left(X_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)\right)^2}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2 \sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2} = 0.9.$$

Regression of X_t on Y_t yields

$$c = \frac{\left(\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)\left(X_t - \overline{X}\right)\right)^2}{\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2 \sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}.$$

Thus,

$$c = 0.9.$$

■

Also,

$$\begin{aligned}
R^2 &= \frac{\left(\sum_{t=1}^{T}\left(X_t-\overline{X}\right)\left(Y_t-\overline{Y}\right)\right)^2}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2\sum_{t=1}^{T}\left(Y_t-\overline{Y}\right)^2} \\
&= \frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)\left(Y_t-\overline{Y}\right)}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}\times\frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)\left(Y_t-\overline{Y}\right)}{\sum_{t=1}^{T}\left(Y_t-\overline{Y}\right)^2} \\
0.9 &= (0.9)\,b \\
\Rightarrow b &= 1.
\end{aligned}$$

■

Since $\overline{X}=a+b\overline{Y}$,

$$\begin{aligned}
30 &= a+27.8 \\
\Rightarrow a &= 2.2 \blacksquare
\end{aligned}$$

Example 4: Consider the model: $Y_t=\beta_1 X_t+u_t, \qquad t=1,2,...,T.$

a. Show that the OLS estimator for β_1 is given by $\widehat{\beta}_1=\dfrac{\sum\limits_{t=1}^{T}X_tY_t}{\sum\limits_{t=1}^{T}X_t^2}$;

b. If we have three observations of (X_t,Y_t), $t=1,2,3$.

X_t	0	1	2
Y_t	2	1	0

Calculate the numerical values of:

i) $\widehat{\beta}_1$;

ii) $\widehat{Y}_t=\widehat{\beta}_1X_t$ for $t=1,2,3$;

iii) $ESS=\sum\limits_{t=1}^{3}\left(Y_t-\widehat{Y}_t\right)^2$;

iv) $TSS=\sum\limits_{t=1}^{3}\left(Y_t-\overline{Y}\right)^2$;

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

Solution:

(a) The problem is

$$\min_{\beta_1} \sum_{t=1}^{T} u_t^2 = \min_{\beta_1} \sum_{t=1}^{T} \left(Y_t - X_t\beta_1\right)^2 .$$

The first order condition is

$$\begin{aligned} \frac{\partial \sum_{t=1}^{T} \left(Y_t - X_t\beta_1\right)^2}{\partial \beta_1} &= -2\sum_{t=1}^{T} \left(Y_t - X_t\beta_1\right) X_t = 0 \\ &\Rightarrow \widehat{\beta}_1 = \frac{\sum_{t=1}^{T} X_t Y_t}{\sum_{t=1}^{T} X_t^2}. \end{aligned}$$

■

(b)

t	1	2	3
X_t	0	1	2
Y_t	2	1	0

(i)

$$\widehat{\beta}_1 = \frac{(0)(2) + (1)(1) + (2)(0)}{(0)^2 + (1)^2 + (2)^2} = \frac{1}{5}.$$

■

(ii)

$$\begin{aligned} \widehat{Y}_1 &= \frac{1}{5}(0) = 0, \\ \widehat{Y}_2 &= \frac{1}{5}(1) = \frac{1}{5}, \\ \widehat{Y}_3 &= \frac{1}{5}(2) = \frac{2}{5}. \end{aligned}$$

■

(iii)

$$\begin{aligned} ESS &= \sum_{t=1}^{3}\left(Y_t-\widehat{Y}_t\right)^2 \\ &= (2-0)^2+\left(1-\frac{1}{5}\right)^2+\left(0-\frac{2}{5}\right)^2 \\ &= 4.8 \blacksquare \end{aligned}$$

(iv)

$$\begin{aligned} TSS &= \sum_{t=1}^{3}\left(Y_t-\overline{Y}\right)^2 \\ &= (2-1)^2+(1-1)^2+(0-1)^2 \\ &= 2 \blacksquare \end{aligned}$$

(v)

$$R^2 = 1-\frac{ESS}{TSS} = 1-\frac{4.8}{2} = -1.4 \blacksquare$$

Note that R^2 is negative because the regression line excludes the intercept term and $\sum_{t=1}^{3}\widehat{u}_t \neq 0$.

Exercise 5: Given the data (X_t, Y_t), $t = 1, 2, ..., T$, and $\overline{X} = 10$. Suppose we run a regression of Y_t on X_t with an intercept, and get the following results:

$$\widehat{Y}_t = X_t, \qquad R^2 = 1.$$

Now suppose we use the same data and run a regression of X_t on Y_t with an intercept, and get the following regression:

$$\widehat{X}_t = a + bY_t \qquad R^2 = c.$$

Find the values of $\overline{Y}$, a, b, and c.

Exercise 6: Consider the model: $Y_t = \beta_0 + \beta_1 X_t + u_t, \qquad t = 1, 2, ..., T.$ If we have three observations of (X_t, Y_t), $t = 1, 2, 3$.

X_t	0	1	2
Y_t	2	1	0

Calculate the numerical values of:

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

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

iii) $ESS = \sum_{t=1}^{3} \left(Y_t - \widehat{Y}_t\right)^2$;

iv) $TSS = \sum_{t=1}^{3} \left(Y_t - \overline{Y}\right)^2$;

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

vi) $\overline{R}^2 = 1 - (1 - R^2)\dfrac{T-1}{T-k-1}$.

Exercise 7: 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 8: 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?

4.7 Hypothesis Testing on βs

We run a linear regression for the model

$$Y_t = \beta_0 + \beta_1 X_t + u_t.$$

because we want to examine whether Y is linearly related to X, i.e., we want to test whether β_1 equals zero.

After the estimation, we may perform hypothesis testing. Suppose we find that $\widehat{\beta}_1 = 0.34$ from the sample. We may test whether the true parameter β_1 equals zero or not. That is, we test $H_0 : \beta_1 = 0$. We must perform this test because if we cannot reject H_0, that implies that X cannot explain Y and the regression model will be useless. When we test this hypothesis, we have to form a test statistic and find its distribution. We may use a test statistic which follows a t-distribution. As mentioned in the previous chapter, when using the t-distribution, we have to assume that the observations come from a normal distribution. In the context of regression models, the random elements are u_t.

Note that we have not specified the distribution of u_t. Thus far we have only assumed that u_t are uncorrelated and identically distributed with mean zero and variance σ^2. Therefore we have to make the following assumption when we carry out hypothesis testing:

Assumption 7: Normality of errors: $u_t \sim N\left(0, \sigma^2\right)$.

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

$$W = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{Var\left(\widehat{\beta}_1\right)}}.$$

where

$\widehat{\beta}_1$ is the OLS estimator for the unknown parameter β_1 and

$$Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}$$

from example 1.

Note that since $\widehat{\beta}_1$ is unbiased,

$$E(W) = \frac{E\left(\widehat{\beta}_1 - \beta_1\right)}{\sqrt{\frac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}} = 0,$$

and

$$Var(W) = Var\left(\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}}\right) = \frac{Var\left(\widehat{\beta}_1\right)}{\frac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}} = 1.$$

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 from example 1 that

$$\begin{aligned} W &= \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}} \\ &= \frac{\frac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}{\sqrt{\frac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}} \\ &= \frac{X_1 - \overline{X}}{\sqrt{\sigma^2\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}u_1 + \frac{X_1 - \overline{X}}{\sqrt{\sigma^2\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}u_1 + ... + \frac{X_T - \overline{X}}{\sqrt{\sigma^2\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}u_T, \end{aligned}$$

which is a linear combination of u_t. Since u_t has a normal distribution by assumption 7, if σ^2 is known, then by the property that normal plus normal is still normal, the test statistic W 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:

$$Y_t = \beta_0 + \beta_1 X_t + u_t.$$

Now after we obtain 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

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

We use $\widehat{\sigma}^2$ to estimate σ^2.

You should have two questions here, why use $\sum\limits_{t=1}^{T} \widehat{u}_t^2$ but not $\sum\limits_{t=1}^{T} \left(\widehat{u}_t - \overline{\widehat{u}}\right)^2$? And why must we use $(T-2)$, but not T?

The answer to the first question is $\sum\limits_{t=1}^{T} \widehat{u}_t = \sum\limits_{t=1}^{T} \left(Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_t\right) = 0$ using the first normal equation (*). Thus $\overline{\widehat{u}} = \frac{1}{T} \sum\limits_{t=1}^{T} \widehat{u}_t = 0$.

The reason why we have to use $(T-2)$ is because we want $\widehat{\sigma}^2$ to be an unbiased estimator of σ^2 (see example 5). 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. This is because we want to obtain an unbiased estimator of σ^2.

Now,

$$\begin{aligned}
W &= \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\widehat{\sigma}^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}} \\
&= \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}}\sqrt{\frac{\sigma^2}{\widehat{\sigma}^2}} \\
&= N\left(0,1\right) \times \sqrt{\frac{\sigma^2}{\widehat{\sigma}^2}} \\
&= \frac{N\left(0,1\right)}{\sqrt{\dfrac{\widehat{\sigma}^2}{\sigma^2}}} \\
&= \frac{N\left(0,1\right)}{\sqrt{\dfrac{\frac{1}{T-2}\sum\limits_{t=1}^{T}\widehat{u}_t^2}{\sigma^2}}}. \\
&= \frac{N\left(0,1\right)}{\sqrt{\dfrac{\sum\limits_{t=1}^{T}\left(\dfrac{\widehat{u}_t}{\sigma}\right)^2}{T-2}}}
\end{aligned}$$

It can be shown that (very difficult) $\sum\limits_{t=1}^{T}\left(\dfrac{\widehat{u}_t}{\sigma}\right)^2$ has a chi-squared distribution with degree of freedom $(T-2)$, and that $\sum\limits_{t=1}^{T}\left(\dfrac{\widehat{u}_t}{\sigma}\right)^2$ is independent of $\dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}}}$, thus the test statistic

$$W = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{T-2}}{T-2}}}$$

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.

Example 5: Show that $\widehat{\sigma}^2 = \frac{\sum\limits_{t=1}^{T}\widehat{u}_t^2}{T-2}$ is an unbiased estimator for σ^2, i.e.,

$$E\left(\widehat{\sigma}^2\right) = E\left(\frac{\sum\limits_{t=1}^{T}\widehat{u}_t^2}{T-2}\right) = \sigma^2.$$

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

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

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

$$= E\left(\sum_{t=1}^{T}\left(\beta_0 + \beta_1 X_t + u_t - \left(\beta_0 + \beta_1\overline{X} + \overline{u}\right) - \widehat{\beta}_1\left(X_t - \overline{X}\right)\right)^2\right)$$

$$= E\left(\sum_{t=1}^{T}\left(u_t - \overline{u} - \left(\widehat{\beta}_1 - \beta_1\right)\left(X_t - \overline{X}\right)\right)^2\right)$$

$$= E\left[\sum_{t=1}^{T}\left(u_t - \overline{u}\right)^2 + \left(\widehat{\beta}_1 - \beta_1\right)^2\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2 - 2\left(\widehat{\beta}_1 - \beta_1\right)\sum_{t=1}^{T}\left(X_t - \overline{X}\right)\left(u_t - \overline{u}\right)\right]$$

$$= E\left[\sum_{t=1}^{T}\left(u_t - \overline{u}\right)^2 + \left(\frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}\right)^2\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2 - 2\frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)X_t}\sum_{t=1}^{T}\left(X_t - \overline{X}\right)u_t\right]$$

$$= \sum_{t=1}^{T}E\left(u_t - \overline{u}\right)^2 - \frac{E\left[\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t\right]^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}$$

$$= \sum_{t=1}^{T}E\left(u_t^2\right) - TE\left(\overline{u}^2\right) - \frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2 E\left(u_t^2\right) + 2\sum\limits_{i=1}^{T-1}\sum\limits_{j>i}^{T}\left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)E(u_i u_j)}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}$$

$$= \sum_{t=1}^{T}\sigma^2 - T\left(\frac{\sigma^2}{T}\right) - \frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2\sigma^2+2\sum_{i=1}^{T-1}\sum_{j>i}^{T}\left(X_i-\overline{X}\right)\left(X_j-\overline{X}\right)(0)}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}$$

$$= (T-1)\,\sigma^2 - \sigma^2 = (T-2)\,\sigma^2 \ \blacksquare$$

Exercise 9: Go to the following webpage:

http://osc.universityofcalifornia.edu/journals/journals_a.html.

Let

Y=List Price;

X=ISI impact factor.

Skip the journals without data. For all the journals (A-Z) with data

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 journal list price affected by the impact factor?

vi) Repeat part i) to iii) using the UC online uses as X.

Exercise 10: Below are 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.

4.8 Prediction and Forecasting

If we are just interested in the relationship between X and Y, we can simply use $Cov(X, Y)$ or $Corr(X, Y)$. An important purpose of running a regression is to predict the value of Y at a given value of X. The idea is that the regression line can be extended indefinitely in the XY plane. Thus, for any given value of X, you can find a corresponding value of Y.

Make sure that you distinguish the differences between

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

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

and

$$\widehat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 X_t.$$

The first equation is the true model, the second is the estimated model. The actual observed values of Y_t do not necessary lie on the line, so residuals are added to both equations. The last equation represents a regression line, every $\widehat{Y}_t$ is a point in the regression line, no error is needed.

We use the regression line $\widehat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 X_t$ to make predictions, e.g. If $\widehat{\beta}_0 = 1$, $\widehat{\beta}_1 = 1$, the predicted value $\widehat{Y}_t$ at $X_t = 10$ will be 11.

Exercise 11: The following table is adopted from Hong Kong Annual Digest of Statistics 1996 Edition, page 301, table 17.2.

$YEAR$	X (HK$million)	Y (HK$million)
1986	450411	253618
1989	a	312682
1991	612016	359019
1992	650347	386519
1993	690223	b
1994	726709	442025
1995	760728	445302

where

Y=private consumption expenditure at constant (1990) market price;

X=Expenditure-based GDP at constant (1990) market price.

i) Fill in the values of a and b.

For parts ii) to vi), if the **second last** number of your student ID is 1 (e.g. 04567712) , then delete the observation in 1991, if it is 6, then delete the observation in 1986, and so on. If the second last number of your student ID is (7,8,0), then you have to use all the seven observations.

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 Hong Kong Annual Digest of Statistics 1997 Edition, find the (X, Y) for 1996.

vi) Using the estimated model to predict the value of Y in 1996 using X in 1996. Is the predicted value $\widehat{Y}$ different from the actual Y in 1996?

Exercise 12: Consider Table 11.20 in Hong Kong Annual Digest of Statistics 1996 Edition, page 223, the Statistics of Results of Hong Kong Certificate of Education Examination 1995.

Let

Y= % of student getting A.

X=Number sat.

i) If a student wants to get 10 straight As in HKCEE, which 10 subjects would you recommend for him/her to take?

ii) If a student wants to fail 10 subjects in HKCEE, which 10 subjects would you recommend for him/her to take?

For parts iii) and vi), if your last name starts with A (e.g. Au) , then delete the subjects which start with A (Accommodation and Catering Services, Additional Mathematics, Art), and so on. If you don't have to delete any subject, then use all the observations. **Anyone who does not follow this rule will earn no credit for this question.**

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 13: Use 9/00 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 2000, 2 for October 2000, 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 $\mathrm{e}^{\widehat{\ln(HSI)}}$ and use it as the predicted value for HSI.

Finally, obtain 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 14: Let X and Y be random variables, $W = 1 - X$, and $Z = 1 - Y$,

(a) Show that $Cov\left(W, Z\right) = Cov\left(X, Y\right)$.

(b) Suppose we draw a sample size T from the above distributions of X and Y. We run the following two regression models:

$$Y_t = \beta_{0a} + \beta_{1a} X_t + u_t,$$

$$Z_t = \beta_{0b} + \beta_{1b} W_t + u_t,$$

Then the two estimates of β_1 are identical in the two regression models. True or False? Explain.

Exercise 15: Let A, B, C, D be four random variables with zero mean and unit variance.

(a) Is $Cov\left(A, B\right) - Cov\left(C, D\right) = Cov\left(A - B, C - D\right)$?

(b) Suppose we draw a sample size T from the above distributions of A, B, C and D, and run the following three regression models:

$$B_t = \beta_{0a} + \beta_{1a} A_t + u_t,$$

$$D_t = \beta_{0b} + \beta_{1b} C_t + u_t,$$

$$C_t - D_t = \beta_{0c} + \beta_{1c}\left(A_t - B_t\right) + u_t,$$

Is $\widehat{\beta}_{1c} = \widehat{\beta}_{1a} - \widehat{\beta}_{1b}$?

Exercise 16: True/False. Explain.

(a). The $\overline{R}^2$ can be equal to 1.

(b). In a linear regression model $Y_t = \beta_0 + \beta_1 X_t + u_t$, $Var\left(Y_t\right) = Var\left(u_t\right)$.

(c). The OLS estimators are inefficient linear unbiased estimators.

(d). The more the regressors, the lower the R^2 in a regression model.

Exercise 17: Let Z_1, Z_2 be independent $N\left(0, 1\right)$ random variables, let

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

(a) What is the distribution of U?

(b) Find $E(U)$.

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

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

(e) Suppose we draw a sample size T from the above distributions of Z_1 and Z_2. In a linear regression model $Z_{2t}^2 = \beta_0 + \beta_1 Z_{1t}^2 + u_t$, what will $\widehat{\beta}_1$ converge to?

Exercise 18: Let X, Y be two independent identical discrete random variables with the probability distributions 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) $Z = min\{X, Y\}$.

b) $Z = XY$.

c) $Z = X + Y$.

Suppose we draw a sample size T from the above distributions of X, Y and Z, and run the following regressions:

(i) $Y_t = \beta_0 + \beta_1 X_t + u_t$,

(ii) $Z_t = \beta_0 + \beta_1 X_t + u_t$,

(iii) $Z_t = \beta_0 + \beta_1 Y_t + u_t$, When T goes to infinity, what are the values of $\widehat{\beta}_0$, $\widehat{\beta}_1$ in each of the 7 possible cases ?

Exercise 19: True or false? Explain.

If X and Y are two continuous random variables,

(a) then $X+Y$ must be continuous too.

(b) If $X+Y$ is discrete, then the slope estimate of the regression of $(X+Y)$ on the continuous random variable X must converge to zero.

Chapter 5

Multiple Regression

5.1 Introduction

Usually a single explanatory variable is not sufficient to explain the variation of Y, we may have to regress Y on many explanatory variables. A multiple regression is of the following form:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t.$$

The OLS estimated model is:

$$\widehat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1t} + \widehat{\beta}_2 X_{2t} + ... + \widehat{\beta}_k X_{kt}.$$

It should be noted that the number of regressors cannot exceed the number of observations. Here the interpretation of $\widehat{\beta}$'s is a little bit different from the case of simple regression. $\widehat{\beta}_0$ is interpreted as the predicted value of Y if all the X's are zero. Sometimes $\widehat{\beta}_0$ is not interpretable as X cannot be zero physically, or the predicted value of Y is beyond its possible range. $\widehat{\beta}_1$ is interpreted as the increase in the value of $\widehat{Y}$ if X_1 is increased by 1 unit, holding all other X's constant. Similar interpretations hold for $\widehat{\beta}_2$ to $\widehat{\beta}_k$. It is the statement 'holding other X's constant' which sometimes makes the sign of $\widehat{\beta}$ counter-intuitive.

For example, if you regress the price of a house on its size and the number of bedrooms, it may happen that the estimated coefficient associated with the number of bedrooms is negative, although we expect it to be positive. The reason is that we are holding the size of the house constant, but keep adding bedrooms, this may reduce the price of the house.

Again, we use R^2 to measure the goodness of fit of multiple regression models. However, we cannot use R^2 to measure the correlation between Y and X, since we have more than one regressor here. We define $R^2 = 1 - \dfrac{ESS}{TSS}$.

As we increase the number of regressors, the explanatory power of the regression increases, the error sum of squares is reduced. Thus, R^2 is always non-decreasing with the number of X's. In principle, as the number of regressors approach infinity, R^2 should approach 1. Of course we cannot do that due to the limited number of observations. Even if we have a lot of observations, it is not always a good idea to increase the number of regressors.

A good model is a model that is simple and has high explanatory power. Even if we add a garbage variable to the model, we may still increase R^2. Thus, we should not use R^2 to compare models. Instead, we define an adjusted R^2 as follows:

$$\overline{R}^2 = 1 - \frac{T-1}{T-k-1}\left(1 - R^2\right).$$

Note that as k increases, there are two effects. The direct effect is a reduction in $\overline{R}^2$. This is because including an additional regressor reduces the degrees of freedom of the model. The indirect effect is an increase $\overline{R}^2$ via the increase in R^2. Thus, whether $\overline{R}^2$ increases or decreases with k depends critically upon the importance of the additional regressor. If the additional regressor is significantly explaining the variation of Y, then R^2 will increase a lot, and the indirect effect will dominate the direct effect, ending up with an increase in $\overline{R}^2$. However, if the additional variable is a garbage variable, R^2 will only increase by a small amount. Hence, the direct effect dominates the indirect effect, ending up with a decrease in $\overline{R}^2$. In light of this, we normally use $\overline{R}^2$ to compare models.

Example 1: The more the number of explanatory variables, the higher the adjusted R^2. True/False/Uncertain. Explain.

Solution: False

By definition,

$$\overline{R}^2 = 1 - \frac{T-1}{T-k-1}\left(1-R^2\right).$$

Differentiate both sides with respect to k, we have

$$\begin{aligned}
\frac{d\overline{R}^2}{dk} &= -\left(T-1\right)\left[\frac{1}{\left(T-k-1\right)^2}\left(1-R^2\right) - \frac{1}{T-k-1}\frac{dR^2}{dk}\right] \\
&= \frac{T-1}{T-k-1}\left[\frac{dR^2}{dk} - \frac{1-R^2}{T-k-1}\right].
\end{aligned}$$

Thus,

$$\frac{d\overline{R}^2}{dk} \overset{>}{\underset{<}{=}} 0 \Rightarrow \frac{dR^2}{dk} \overset{>}{\underset{<}{=}} \frac{1-R^2}{T-k-1}.$$

■

5.2 Simple Hypothesis Testing

If we are just interesting in one of the coefficients in the multiple regression model, the t-test is performed as usual, the degrees of freedom are $T-k-1$.

For any $i=0,1,2,...,k$, we test:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

We define

$$t_{obs} = \frac{\widehat{\beta}_i}{\widehat{sd}\left(\widehat{\beta}_i\right)}$$

$\widehat{\beta}_i$ $(i=0,1,...,k)$ are obtained by solving the $k+1$ normal equations.

$$\begin{aligned}
\widehat{sd}\left(\widehat{\beta}_i\right) &= \sqrt{\widehat{\sigma}^2 c_{i+1,i+1}} \\
\widehat{\sigma}^2 &= \frac{\sum_{t=1}^{T} \widehat{u}_t^2}{T-k-1} \\
\widehat{u}_t &= Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1t} - \widehat{\beta}_2 X_{2t} - ... - \widehat{\beta}_k X_{kt}
\end{aligned}$$

$c_{i+1,i+1}$ is the $(i+1,i+1)^{th}$ element of the matrix $(X'X)^{-1}$.

$$X = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & & X_{k2} \\ \vdots & & & \ddots & \vdots \\ 1 & X_{1T} & X_{2T} & \cdots & X_{kT} \end{pmatrix}$$

We reject the null at the significance level α if $|t_{obs}| > \left|t_{\frac{\alpha}{2},T-k-1}\right|$.

Example 1: Consider the following data

	$t=1$	$t=2$	$t=3$	$t=4$
X_{1t}	3	1	2	0
X_{2t}	1	2	3	4
Y_t	2	1	4	5

$$X = \begin{pmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ 1 & X_{13} & X_{23} \\ 1 & X_{14} & X_{24} \end{pmatrix} = \begin{pmatrix} 1 & 3 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 0 & 4 \end{pmatrix}$$

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 3 & 1 & 2 & 0 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 3 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 0 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 6 & 10 \\ 6 & 14 & 11 \\ 10 & 11 & 30 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 4 & 6 & 10 \\ 6 & 14 & 11 \\ 10 & 11 & 30 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{299}{36} & -\frac{35}{18} & -\frac{37}{18} \\ -\frac{35}{18} & \frac{5}{9} & \frac{4}{9} \\ -\frac{37}{18} & \frac{4}{9} & \frac{5}{9} \end{pmatrix}$$

:

$$c_{11} = \frac{299}{36}, c_{22} = \frac{5}{9}, c_{33} = \frac{5}{9}.$$

5.3 Joint Hypothesis Testing

Sometimes, we are interested in testing the significance of a set of coefficients. For example,

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0,$$

i.e., we would like to test whether all the X_2, X_3, and X_4 are garbage regressors.

Be careful when you write down the alternative hypothesis H_1. Most students make mistakes here. Remember $H_0 \cup H_1 = S$, where S is the sample space. Thus, H_1 must be the complement of the statement H_0. Some of you may write down $H_1 : \beta_2 = \beta_3 = \beta_4 \neq 0$ or $H_1 : \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$, which are inappropriate, as those statements are not the complements of H_0. The correct statement should be

H_1: at least one of the $\beta_2, \beta_3, \beta_4$ is not equal to zero.

Sometimes, we are interested in the linear relationship among $\beta's$ rather than testing if the $\beta's$ equal some prespecified values. For instance, we may want to test

$$\begin{aligned} H_0 &: \beta_2 = \beta_3 = \beta_4 \\ H_1 &: \beta_2, \beta_3, \text{and } \beta_4 \text{ are not all the same.} \end{aligned}$$

or

$$\begin{aligned} H_0 &: \beta_2 = 2\beta_3 \\ H_1 &: \beta_2 \neq 2\beta_3. \end{aligned}$$

In all the aforementioned situations, the t-test is no longer appropriate, as the hypothesis involves more than one β. We use the F-test in these cases.

The idea behind the F-test is as follows:

We run two regressions, one is the unrestricted model:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t.$$

We obtain the unrestricted error sum of squares from this model, called ESS_U.

Another type is the restricted model, where we impose the restriction of H_0 on the model. For example, if $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$, then our restricted model is:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_5 X_{4t} + ... + \beta_k X_{kt} + u_t.$$

We obtain the restricted error sum of squares from this model, and call it ESS_R. (Note that $ESS_R \geq ESS_U$, why?)

If H_0 is true, the estimates of β_2, β_3, and β_4 in the unrestricted model will converge to zero, and there will be no difference between the restricted and unrestricted models. Thus, their error sum of squares should be the same when the sample size is very large.

If H_0 is false, then at least one of the $\beta_2, \beta_3, \beta_4$ is not equal to zero, and $ESS_U \neq ESS_R$ as a result. We can therefore construct a test statistic based on the difference between ESS_R and ESS_U. We define

$$F_{obs} = \frac{(ESS_R - ESS_U) \,/\, (df_R - df_U)}{ESS_U / df_U}$$

where df_R and df_U are the degrees of freedom for the restricted and unrestricted model respectively.

If H_0 is true, $ESS_R - ESS_u$ will be very small. This implies F_{obs} will be small if H_0 is true. But how small is small? We have to find a critical value.

Now at a given value of α, find out the critical $F-$value at $df = (df_R - df_U, df_U)$ from the F-table. If the observed F-value is bigger than the critical $F-$value, we reject H_0 at α level of significance.

Example 2: Consider the following demand function for chicken.

$$\ln Y_t = \beta_0 + \beta_1 \ln X_{1t} + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_4 \ln X_{4t} + u_t$$

Suppose we run an OLS and obtain

$$\begin{aligned}\widehat{\ln Y_t} &= \underset{(0.1557)}{2.1898} + \underset{(0.0833)}{0.3425} \ln X_{1t} - \underset{(0.1109)}{0.5046} \ln X_{2t} + \underset{(0.0997)}{0.1485} \ln X_{3t} + \underset{(0.1007)}{0.0997} \ln X_{4t} \\ R^2 &= 0.9823\end{aligned}$$

$t = 1, 2, ..., 30.$

where

Y=per capita consumption of chicken (lbs)

X_1=real disposable per capita income ($)

X_2=real retail price of chicken per lb (cents)

X_3=real retail price of pork per lb (cents)

X_4=real retail price of beef per lb (cents)

and the figures in the parentheses are the estimated standard errors.

(a) Interpret each of the above coefficient estimates. Perform the t-test for $H_0 : \beta_i = 0$ v.s. $H_1 : \beta_i \neq 0$, $i = 0, 1, 2, 3, 4$ at $\alpha = 5\%$.

(b) Suppose we want to test the hypothesis that $H_0 : \beta_3 = \beta_4 = 0$. What is the purpose of testing this hypothesis? Now suppose under H_0, we obtain

$$\begin{aligned}\widehat{\ln Y_t} &= \underset{(0.1162)}{2.0328} + \underset{(0.0247)}{0.4515} \ln X_{1t} - \underset{(0.0635)}{0.3722} \ln X_{2t} \\ R^2 &= 0.9801\end{aligned}$$

Perform an F-test for $H_0 : \beta_3 = \beta_4 = 0$ at $\alpha = 5\%$.

Solution: Given

$$\ln Y_t = \beta_0 + \beta_1 \ln X_{1t} + \beta_2 \ln X_{2t} + \beta_3 \ln X_{3t} + \beta_4 \ln X_{4t} + u_t.$$

(a)

$$\begin{aligned}\beta_i &= \frac{\partial \ln Y_t}{\partial \ln X_{it}} = \frac{\partial \ln Y_t}{\partial Y_t}\frac{\partial Y_t}{\partial X_{it}}\frac{\partial X_{it}}{\partial \ln Xi_t} = \frac{\partial Y_t}{\partial X_{it}}\frac{X_{it}}{Y_t} \\ &= \text{elasticity of } Y \text{ with respect to } X_i \text{ for } i = 1, 2, 3, 4\end{aligned}$$

Thus,

$$\begin{aligned}\widehat{\beta}_1 &= \text{estimated elasticity of per capita consumption w.r.t. disposable} \\ &\quad\ \text{per capita income (income elasticity)} \\ \widehat{\beta}_2 &= \text{estimated elasticity of per capita consumption w.r.t. price of chicken} \\ &\quad\ \text{(price elasticity)} \\ \widehat{\beta}_3 &= \text{estimated elasticity of per capita consumption w.r.t. price of pork} \\ &\quad\ \text{(cross price elasticity)} \\ \widehat{\beta}_4 &= \text{estimated elasticity of per capita consumption w.r.t. price of beef} \\ &\quad\ \text{(cross price elasticity)} \\ \exp\left(\widehat{\beta}_0\right) &= \text{estimated autonomous amount of per capita consumption when} \\ &\quad\ X_{1t},\ X_{2t},\ X_{3t} \text{ and } X_{4t} \text{ equal one.}\end{aligned}$$

To test the hypotheses $H_0 : \beta_i = 0$ for $i = 0, 1, 2, 3, 4$, we find out the critical value of the t-statistic at 5% level of significance with degree of freedom $(30 - 5) = 25$.

$$t = 2.060.$$

The calculated t-statistics are

$$\begin{aligned}
\text{When } i &= 0,\ t_{obs}=\frac{\widehat{\beta}_0}{\widehat{sd}\left(\widehat{\beta}_0\right)}=\frac{2.1898}{0.1557}=14.06.\ H_0 \text{ is rejected.}\\
\text{When } i &= 1,\ t_{obs}=\frac{\widehat{\beta}_1}{\widehat{sd}\left(\widehat{\beta}_1\right)}=\frac{0.3425}{0.0833}=4.11.\ H_0 \text{ is rejected.}\\
\text{When } i &= 2,\ t_{obs}=\frac{\widehat{\beta}_2}{\widehat{sd}\left(\widehat{\beta}_2\right)}=\frac{0.5046}{0.1109}=4.55.\ H_0 \text{ is rejected.}\\
\text{When } i &= 3,\ t_{obs}=\frac{\widehat{\beta}_3}{\widehat{sd}\left(\widehat{\beta}_3\right)}=\frac{0.1485}{0.0997}=1.49.\ H_0 \text{ cannot be rejected.}\\
\text{When } i &= 4,\ t_{obs}=\frac{\widehat{\beta}_4}{\widehat{sd}\left(\widehat{\beta}_4\right)}=\frac{0.0997}{0.1007}=0.99.\ H_0 \text{ cannot be rejected.}
\end{aligned}$$

■

(b) The purpose of testing hypothesis $H_0 : \beta_3 = \beta_4 = 0$ is to test the relevance of the variables X_3 and X_4. If the hypothesis cannot be rejected, this implies that we do not need to introduce the variables X_3 and X_4 into the model.

Note that $R^2 = 1 - \dfrac{ESS}{TSS}$. Then,

$$\begin{aligned}
F_{obs} &= \frac{(ESS_R - ESS_U)\ /\ (df_R - df_U)}{ESS_U\ /\ df_U}\\
&= \frac{[TSS\,(1-R_R^2) - TSS\,(1-R_U^2)]\ /\ (df_R - df_U)}{TSS\,(1-R_U^2)\ /\ df_U}\\
&= \frac{(R_U^2 - R_R^2)\ /\ (df_R - df_U)}{(1-R_U^2)\ /\ df_U}\\
&= \frac{(0.9823-0.9801)}{1-0.9823} \times \frac{25}{27-25}\\
&= 1.5537.
\end{aligned}$$

Thus, $F_{obs} < F_{0.05}\,(2,25) = 3.39$. The hypothesis $H_0 : \beta_3 = \beta_4 = 0$ cannot be rejected at 5% level of significance. ■

Exercise 1: A model of deaths due to heart disease is estimated as follows:

$$\begin{aligned}
\widehat{CHD}_t &= 139.68 + 10.71 CIG_t + 3.38 EDFAT_t + 26.75 SPIRITS_t - 4.13 BEER_t \\
T &= \text{Sample size} = 34 \\
k &= 4 = \text{Number of explanatory variables excluding the constant term} \\
ESS &= \sum_{t=1}^{34} \left(CHD_t - \widehat{CHD}_t \right)^2 = 2122 \\
\overline{R}^2 &= 1 - \frac{ESS/(T-k-1)}{TSS/(T-1)} = 0.672
\end{aligned}$$

where

$CHD =$ Death rate (per million population) due to coronary heart disease in the U.S. during each of the years 1947-1980.

$CIG =$Per capita consumption of cigarettes measured in pounds of tobacco.

$EDFAT =$ Per capita intake of edible fats and oil, measured in pounds.

$SPIRITS =$Per capita consumption of distilled spirits in gallons.

$BEER =$ Per capita consumption of malted liquor in gallons.

a) Find the value of R^2, Total Sum of Squares ($TSS = \sum_{t=1}^{34} \left(CHD_t - \overline{CHD} \right)^2$) and the Regression Sum of Squares (RSS) in the above model.

b) Suppose we want to test the joint hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, and run the restricted model as:

$$CHD_t = \beta_0 + u_t.$$

i) Show that the Ordinary Least Squares estimate for β_0 is $\widehat{\beta}_0 = \overline{CHD}$, where $\overline{CHD} = \frac{\sum\limits_{t=1}^{34} CHD_t}{34}$.

ii) Show that $\widehat{CHD}_t = \overline{CHD}$ for all $t = 1, 2, ..., 34$. What is the value of the restricted error sum of squares $ESS_r = \sum\limits_{t=1}^{34}\left(CHD_t - \widehat{CHD}_t\right)^2$?

iii) Perform an F test on $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ at $\alpha = 5\%$ using the F-statistic defined as $F = \frac{(ESS_r - ESS_u)/(df_r - df_u)}{ESS_u/df_u}$.

5.4 The Trivariate Model

Consider the following model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t.$$

Our objective is to

$$\underset{\beta_0,\beta_1,\beta_2}{\text{Min}} \sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta_1 X_{1t} - \beta_2 X_{2t}\right)^2.$$

The first-order conditions are:

$$\frac{\partial \sum\limits_{t=1}^{T} \left(Y_t - \beta_0 - \beta_1 X_{1t} - \beta_2 X_{2t}\right)^2}{\partial \beta_0} = -2\sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta_1 X_{1t} - \beta_2 X_{2t}\right) = 0.$$

$$\frac{\partial \sum\limits_{t=1}^{T} \left(Y_t - \beta_0 - \beta_1 X_{1t} - \beta_2 X_{2t}\right)^2}{\partial \beta_1} = -2\sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta_1 X_{1t} - \beta_2 X_{2t}\right) X_{1t} = 0.$$

$$\frac{\partial \sum\limits_{t=1}^{T} \left(Y_t - \beta_0 - \beta_1 X_{1t} - \beta_2 X_{2t}\right)^2}{\partial \beta_2} = -2\sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta_1 X_{1t} - \beta_2 X_{2t}\right) X_{2t} = 0.$$

Solving these three normal equations gives the Least Squares Estimators:

$$\widehat{\beta}_1 = \frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22} - S_{12}^2},$$

$$\widehat{\beta}_2 = \frac{S_{y2}S_{11} - S_{y1}S_{12}}{S_{11}S_{22} - S_{12}^2},$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X}_1 - \widehat{\beta}_2\overline{X}_2,$$

where

$$S_{y1} = \sum_{t=1}^{T}\left(X_{1t} - \overline{X}_1\right)\left(Y_t - \overline{Y}\right),$$

$$S_{y2} = \sum_{t=1}^{T}\left(X_{2t} - \overline{X}_2\right)\left(Y_t - \overline{Y}\right),$$

$$S_{11} = \sum_{t=1}^{T}\left(X_{1t} - \overline{X}_1\right)^2,$$

$$S_{22} = \sum_{t=1}^{T}\left(X_{2t} - \overline{X}_2\right)^2,$$

$$S_{12} = \sum_{t=1}^{T}\left(X_{1t} - \overline{X}_1\right)\left(X_{2t} - \overline{X}_2\right).$$

Exercise 2: Suppose we have 4 observations of a trivariate model.

	$t=1$	$t=2$	$t=3$	$t=4$
X_{1t}	3	1	2	0
X_{2t}	1	2	3	4
Y_t	2	1	4	5

a) Find S_{y1}, S_{y2}, S_{11}, S_{22}, S_{12};

b) Find $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$;

c) Find $\widehat{u}_t = Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1t} - \widehat{\beta}_2 X_{2t}$for $t = 1, 2, 3, 4$;

d) Find $\widehat{\sigma}^2 = \dfrac{\sum\limits_{t=1}^{T} \widehat{u}_t^2}{T-k-1}$;

e) Find $\widehat{sd}\left(\widehat{\beta}_i\right)$ for $i = 0, 1, 2$;

f) Test

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

for $i = 0, 1, 2$.

Exercise 3: Show that in a trivariate model, the OLS estimates $\widehat{\beta}_2$, $\widehat{\beta}_1$, and $\widehat{\beta}_0$ are unbiased.

Exercise 4: Consider the model:

$$PRICE_t = \beta_0 + \beta_1 SQFT_t + \beta_2 BEDROOM_t + u_t,$$

$t = 1, 2, ..., 19$.

where

$PRICE_t$ is the price of house t (thousands of dollars)

$SQFT_t$ is the living areas of house t. (square feet)

$BEDROOM_t$ is the number of bedrooms in house t

Suppose we estimate the model and get

$$\begin{aligned}
\widehat{PRICE}_t &= \underset{(1.53)}{142.2} + \underset{(6.73)}{0.313} SQFT_t + \underset{(2.545)}{43.9}\, BEDROOM_t, \\
T &= \text{Sample size} = 19, \\
k &= 2 = \text{Number of explanatory variables excluding the constant term}, \\
ESS &= \sum_{t=1}^{19} \left(PRICE_t - \widehat{PRICE}_t \right)^2 = 1332 = \text{Error Sum of Squares}, \\
\overline{R}^2 &= 1 - \frac{ESS/(T-k-1)}{TSS/(T-1)} = 0.75,
\end{aligned}$$

and the figures in the parentheses are **t-ratios**.

a) Interpret each of the above coefficient estimates.

b) Perform the t-test for $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$, $i = 0, 1, 2$ at $\alpha = 5\%$.

c) Find the value of R^2, Total Sum of Squares ($TSS = \sum_{t=1}^{19} \left(PRICE_t - \overline{PRICE} \right)^2$) and the Regression Sum of Squares ($RSS = TSS - ESS$) in the above model.

d) Suppose we want to test the joint hypothesis $H_0 : \beta_1 = \beta_2 = 0$, and run the restricted model as:

$$PRICE_t = \beta_0 + u_t.$$

i) Show that the Ordinary Least Squares estimate for β_0 is $\widehat{\beta}_0 = \overline{PRICE} = \frac{\sum_{t=1}^{19} PRICE_t}{19}$.

ii) Show that $\widehat{PRICE}_t = \overline{PRICE}$ for all $t = 1, 2, ..., 19$. What is the value of the restricted error sum of squares $ESS_r = \sum_{t=1}^{19} \left(PRICE_t - \widehat{PRICE}_t \right)^2$? (3 points)

iii) Perform an F test on $H_0 : \beta_1 = \beta_2 = 0$ at $\alpha = 5\%$ using the F-statistic defined as $F = \frac{(ESS_r - ESS_u) / (df_r - df_u)}{ESS_u / df_u}$.

5.5 Inclusion of an Irrelevant Variable

Suppose the true model is a bivariate model

$$Y_t = \beta_0 + \beta_1 X_{1t} + u_t$$

But we estimate a trivariate model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t.$$

Are the OLS estimators still unbiased? The answer is yes. To see why, recall how we estimate a trivariate model

$$\widehat{\beta}_1 = \frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22} - S_{12}^2}$$

where

$$\begin{aligned} S_{y1} &= \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) Y_t \\ &= \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) \left(\beta_0 + \beta_1 X_{1t} + u_t\right) \\ &= \beta_1 \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) X_{1t} + \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) u_t \\ &= \beta_1 S_{11} + \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) u_t \end{aligned}$$

(note that we always plug in the true Y_t)

Thus

$$E\left(S_{y1}\right)=\beta_1 S_{11}.$$

Similarly,

$$E\left(S_{y2}\right)=\beta_1 S_{12}.$$

Thus

$$E\left(\widehat{\beta}_1\right)=\frac{E\left(S_{y1}\right)S_{22}-E\left(S_{y2}\right)S_{12}}{S_{11}S_{22}-S_{12}^2}=\beta_1.$$

$$E\left(\widehat{\beta}_2\right)=\frac{E\left(S_{y2}\right)S_{11}-E\left(S_{y1}\right)S_{12}}{S_{11}S_{22}-S_{12}^2}=0.$$

$$\begin{aligned}
E\left(\widehat{\beta}_0\right) &= E\left(\overline{Y}\right)-E\left(\widehat{\beta}_1\right)\overline{X}_1-E\left(\widehat{\beta}_2\right)\overline{X}_2 \\
&= E\left(\beta_0+\beta_1\overline{X}_1+\overline{u}\right)-\beta_1\overline{X}_1-0 \\
&= \beta_0.
\end{aligned}$$

Thus all the estimators are unbiased. The reason why the inclusion of an irrelevant variable does no harm (except we have one less degrees of freedom) is that all of the information in the true model are included in the estimated model.

Exercise 4: If the true model is a bivariate model, but we estimate a trivariate model. If $S_{y2}=0$, then β_1 will be over-estimated. True/False/Uncertain. Explain.

5.6 Exclusion of a Pertinent Variable

Suppose the true model is a trivariate model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t.$$

However, we estimate a bivariate model

$$Y_t = \beta_0 + \beta_1 X_{1t} + u_t.$$

The OLS estimators are biased now. To see why, recall how we estimate a bivariate model,

$$\widehat{\beta}_1 = \frac{S_{y1}}{S_{11}},$$

where

$$\begin{aligned} S_{y1} &= \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) Y_t \\ &= \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) \left(\beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t\right) \\ &= \beta_1 \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) X_{1t} + \beta_2 \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) X_{2t} + \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) u_t \\ &= \beta_1 S_{11} + \beta_2 S_{12} + \sum_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right) u_t. \end{aligned}$$

$$E\left(S_{y1}\right) = \beta_1 S_{11} + \beta_2 S_{12}.$$

$$E\left(\widehat{\beta}_1\right) = \frac{E\left(S_{y1}\right)}{S_{11}} = \beta_1 + \beta_2 \frac{S_{12}}{S_{11}} \neq \beta_1$$

in general.

Therefore, all of the estimators are biased in general. Excluding a relevant variable is a serious problem as far as unbiasedness is concerned. The reason why we cannot obtain unbiased estimator is because we lack some information in the true model.

5.7 Retrieving the Trivariate Estimates from Bivariate Estimates

In the past when computers were not available, estimating a trivariate model was a nightmare for researchers, but estimating a bivariate model is relatively easier. As such, people used to retrieve the trivariate estimates from several bivariate models. Note that:

$$\widehat{\beta}_1 = \frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22} - S_{12}^2} = \frac{\frac{S_{y1}}{S_{11}} - \frac{S_{y2}}{S_{22}}\frac{S_{12}}{S_{11}}}{1 - \frac{S_{12}^2}{S_{11}S_{22}}} = \frac{\widehat{\beta}_{y1} - \widehat{\beta}_{y2}\widehat{\beta}_{21}}{1 - r_{12}^2}$$

$$\widehat{\beta}_2 = \frac{S_{y2}S_{11} - S_{y1}S_{12}}{S_{11}S_{22} - S_{12}^2} = \frac{\frac{S_{y2}}{S_{22}} - \frac{S_{y1}}{S_{11}}\frac{S_{12}}{S_{22}}}{1 - \frac{S_{12}^2}{S_{11}S_{22}}} = \frac{\widehat{\beta}_{y2} - \widehat{\beta}_{y1}\widehat{\beta}_{12}}{1 - r_{12}^2}$$

where

$\widehat{\beta}_{y1}$ is the OLS slope estimate when we regress Y on one and X_1,
$\widehat{\beta}_{y2}$ is the OLS slope estimate when we regress Y on one and X_2,
$\widehat{\beta}_{12}$ is the OLS slope estimate when we regress X_1 on one and X_2,
$\widehat{\beta}_{21}$ is the OLS slope estimate when we regress X_2 on one and X_1,
r_{12}^2 is the R^2 when regressing X_1 on an intercept and X_2.

Example 3: Given the data (X_{1t}, X_{2t}, Y_t), $t = 1, 2, ..., T$, suppose we run a regression of X_{1t} on X_{2t}, and obtain the following results:

$$\widehat{X}_{1t} = 1 + 0.8X_{2t} \qquad R^2 = 0.64$$

and we know $\overline{X_2} = 30$. Now suppose we run a regression of X_{2t} on X_{1t}, and obtain the following results:

$$\widehat{X}_{2t} = a + bX_{1t} \qquad R^2 = c$$

Now suppose we run a regression of Y_t on X_{1t}, and obtain the following results:

$$\widehat{Y}_t = 2 + 0.6X_{1t} \qquad R^2 = 0.8$$

Now suppose we run a regression of Y_t on X_{2t}, and obtain the following results:

$$\widehat{Y}_t = d + X_{2t} \qquad R^2 = 0.7$$

Now suppose we run a regression of Y_t on X_{2t}, and obtain the following results:

$$\widehat{Y}_t = e + fX_{1t} + gX_{2t}$$

Find the values of a, b, c, d, e, f,and g.

Solution: Regression of X_{2t} on X_{1t} yields

$$\widetilde{R}^2 = c = \frac{\left(\sum_{t=1}^{T}\left(X_{1t} - \overline{X}_1\right)\left(X_{2t} - \overline{X}_2\right)\right)^2}{\sum_{t=1}^{T}\left(X_{1t} - \overline{X}_1\right)^2 \sum_{t=1}^{T}\left(X_{2t} - \overline{X}_2\right)^2}$$

Regression of X_{1t} on X_{2t} yields

$$R^2 = \frac{\left(\sum_{t=1}^{T}\left(X_{1t} - \overline{X}_1\right)\left(X_{2t} - \overline{X}_2\right)\right)^2}{\sum_{t=1}^{T}\left(X_{1t} - \overline{X}_1\right)^2 \sum_{t=1}^{T}\left(X_{2t} - \overline{X}_2\right)^2} = 0.64$$

Thus,

$$c = 0.64.$$ ■

Also,

$$\begin{aligned}
R^2 &= \frac{\sum_{t=1}^{T}\left(X_{1t}-\overline{X}_1\right)\left(X_{2t}-\overline{X}_2\right)}{\sum_{t=1}^{T}\left(X_{1t}-\overline{X}_1\right)^2} \times \frac{\sum_{t=1}^{T}\left(X_{1t}-\overline{X}_1\right)\left(X_{2t}-\overline{X}_2\right)}{\sum_{t=1}^{T}\left(X_{2t}-\overline{X}_2\right)^2} \\
0.64 &= (0.8)\,b \\
\Rightarrow b &= 0.8.
\end{aligned}$$

■

Since $\overline{X}_1 = 1 + 0.8\overline{X}_2 = 25$ and $\overline{X}_2 = a + b\overline{X}_1$,

$$\begin{aligned}
30 &= a + (0.8)(25) \\
\Rightarrow a &= 10.
\end{aligned}$$

■

As we know $\overline{Y} = 2 + 0.6\overline{X}_1 = 17$ and $\overline{Y} = d + \overline{X}_2$,

$$\begin{aligned}
17 &= d + 30 \\
\Rightarrow d &= -13.
\end{aligned}$$

■

On the other hand, $\widehat{Y}_t = e + fX_{1t} + gX_{2t}$.

$$\begin{aligned}
f &= \frac{0.6 - (1)(0.8)}{1 - 0.64} \\
&= -\frac{1}{18}.
\end{aligned}$$

■

$$\begin{aligned}
g &= \frac{1 - (0.6)(0.8)}{1 - 0.64} \\
&= \frac{13}{9}.
\end{aligned}$$

■

$$\begin{aligned}
e &= \overline{Y} - f\overline{X}_1 - g\overline{X}_2 \\
&= 17 - \left(-\frac{1}{18}\right)(25) - \left(\frac{13}{9}\right)(30) \\
&= -24.94.
\end{aligned}$$

■

Exercise 5: Go to the following webpage:

http://osc.universityofcalifornia.edu/journals/journals_a.html.

Let

Y=List Price;

X_1=Impact factor.

X_2=Online uses.

Skip the journals without data. For all the journals (A-Z) with data

i) Run the following regression model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_1 X_{2t} + u_t,$$

Find the values of $\widehat{\beta}_0$, $\widehat{\beta}_1, \widehat{\beta}_2$ What is the meaning of $\widehat{\beta}_0$ in this case? Interpret $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

ii) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$. Is the journal list price affected by the impact factor and/or the online uses?

iii) Compare your results with those from the simple regressions in the previous chapter. What are the differences. Can the results in this section be applied to extract the trivariate estimates? Why or why not? If not, fix the problem and show that the results apply.

5.8 Multicollinearity

Multicollinearity, introduced by Ragnar Frisch in his book "Statistical Confluence Analysis by Means of Complete Regression Systems," published in

1934, nowadays refers to situations where two or more regressors are linearly related, so that it is difficult to disentangle their separate effects on the dependent variable.

As we have mentioned before, in a trivariate model, if the two regressors are orthogonal to each other, in the sense that $S_{12} = 0$, then the OLS estimate $\widehat{\beta}_1$ will be the same in both the bivariate and trivariate models. Thus an additional regressor will be of no impact on the original slope estimates as long as it is orthogonal to all the existing regressors. However, if we add a new regressor which is not totally orthogonal to all the existing regressors, then some distortions on the estimates are unavoidable. In extreme cases, when the new regressor is perfectly linearly related to one or more of the existing regressors, the new model is not estimable. We call this problem the **Perfect Collinearity**.

To show the problem more explicitly, consider the following model:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t.$$

If $X_2 = 2X_1$, the model is reduced to

$$Y_t = \beta_0 + (\beta_1 + 2\beta_2) X_{1t} + u_t.$$

Thus it is a simple regression model, and we can obtain the OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta_1 + 2\beta_2}$. However, we cannot obtain estimates for β_1 and β_2, which means the original trivariate model is not estimable.

Let $r_{12}^2 = \dfrac{S_{12}}{S_{11}S_{22}}$. As long as $r_{12}^2 = 1$, the trivariate model is not estimable, since

$$\widehat{\beta}_1 = \frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22} - S_{12}^2} = \frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}$$

$$\widehat{\beta}_2 = \frac{S_{y2}S_{11} - S_{y1}S_{12}}{S_{11}S_{22} - S_{12}^2} = \frac{S_{y2}S_{11} - S_{y1}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}$$

are undefined.

In general, if our model is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t.$$

The model is not estimable if there are constants $\lambda_0, \lambda_1, \lambda_2, ..., \lambda_k$ (at least some of them are non-zero) such that for all t,

$$\lambda_0 + \lambda_1 X_{1t} + \lambda_2 X_{2t} + ... + \lambda_k X_{kt} = 0$$

5.9 Consequences of near or high Multicollinearity

Recall that if the assumptions of the classical model are satisfied, the OLS estimators of the regression estimators are BLUE. The existence of multicollinearity does not violate any one of the classical assumptions, so if the model is still estimable, the OLS estimator will still be consistent, efficient, linear, and unbiased. So why do we care about multicollinearity? Although **multicollinearity does not affect the estimation, it will affect the hypothesis testing**.

1: Large Variances of OLS Estimators

Consider the trivariate model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

$$Var\left(\widehat{\beta}_1\right) = Var\left(\frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22} - S_{12}^2}\right) = Var\left(\frac{S_{u1}S_{22} - S_{u2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2} Var\left(S_{u1}S_{22} - S_{u_2}S_{12}\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2}\left(Var\left(S_{u1}S_{22}\right) + Var\left(S_{u_2}S_{12}\right) - 2Cov\left(S_{u1}S_{22}, S_{u2}S_{12}\right)\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2}\left(S_{22}^2S_{11}\sigma^2 + S_{12}^2S_{22}\sigma^2 - 2S_{12}S_{22}Cov\left(S_{u1}, S_{u2}\right)\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2}\left(S_{22}^2S_{11}\sigma^2 + S_{12}^2S_{22}\sigma^2 - 2S_{12}^2S_{22}\sigma^2\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2} S_{11}S_{22}^2\left(1 - r_{12}^2\right)\sigma^2 = \frac{\sigma^2}{S_{11}\left(1 - r_{12}^2\right)}.$$

Similarly, it can be shown that

$$Var\left(\widehat{\beta}_2\right) = \frac{\sigma^2}{S_{22}\left(1 - r_{12}^2\right)}.$$

Thus, the variances of the estimators increase as the relationship between regressors increase. In extreme cases, they explode when there is perfect multicollinearity.

2: Wider Confidence Intervals

Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger. Therefore, in cases of high multicollinearity, the chance of accepting the null hypothesis increases, hence Type II error (Accept H_0 when H_0 is false) increases. Even if the explanatory variable does individually explain the dependent variable well, we may still tend to conclude that each of them is not significant if there is multicollinearity.

3: Insignificant t Ratio

Recall that the t statistic for the hypothesis $H_0 : \beta_i = 0$ $(i = 0, 1, 2, ..., k)$ is

$$t = \frac{\widehat{\beta}_i}{\sqrt{\widehat{Var}\left(\widehat{\beta}_2\right)}}$$

In cases of high collinearity, the estimated standard errors increase dramatically, thereby making the t values smaller for any given values of $\widehat{\beta}_i$. Therefore, one will over-accept the null that $\beta_i = 0$.

5.10 Detection of Multicollinearity

Multicollinearity is a question of degree, not of kind. Therefore, we do not test for the presence of multicollinearity, but instead we measure its degree in any particular sample.

Since multicollinearity refers to the condition of the explanatory variables that are assumed to be nonstochastic, it is essentially a sample phenomenon, arising out of the largely nonexperimental data collected in most social sciences, we do not have one unique method of detecting it or measuring its strength.

Our rule of thumb is that, if we run a regression and find **a High R^2 but few significant t Ratios,** then this is a sign of multicollinearity. If R^2 is high, the F test in most cases will reject the hypothesis that the slope coefficients are zero simultaneously. However, very few or even none of the individual t tests will be significant.

Other symptoms of multicollinearity include: (1) Small changes in the data can produce wide swings in the parameter estimates, and (2) Coefficients will have the wrong sign or an implausible magnitude.

5.11 Remedial Measures

What can be done if multicollinearity is serious? The following methods can be tried.

1. **A priori information**

Suppose we consider the model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

Suppose a priori we believe or economic theory suggests that $\beta_1 = 2\beta_2$, then we can run the following regression,

$$Y_t = \beta_0 + 2\beta_2 X_{1t} + \beta_2 X_{2t} + u_t$$

$$Y_t = \beta_0 + \beta_2 X_t + u_t$$

where $X_t = 2X_{1t} + X_{2t}$. Once we obtain $\widehat{\beta}_2$, we can define $\widehat{\beta}_1 = 2\widehat{\beta}_2$.

2. **Using first differences or ratios**

Suppose we have

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

where X_{1t} and X_{2t} are highly collinear. To reduce the degree of collinearity, we can still estimate β_1 and β_2 by the "first difference" model, i.e. we estimate

$$Y_t - Y_{t-1} = \beta_1 \left(X_{1t} - X_{1(t-1)}\right) + \beta_2 \left(X_{2t} - X_{2(t-1)}\right) + \left(u_t - u_{t-1}\right)$$

Although the first difference model may reduce the severity of multicollinearity, it creates some additional problems. In the transformed model, the new error terms $(u_t - u_{t-1})$ is not serially independent as $Cov\left(u_t - u_{t-1}, u_{t-1} - u_{t-2}\right) = -Var\left(u_{t-1}\right) = -\sigma^2 \neq 0$. We will discuss the problem of serial correlation later. Here, we alleviate multicollinearity at the expense of violating one of the classical assumptions "serial independence", this implies that the Gauss-Markov theorem will not hold anymore, and the OLS estimators are not BLUE in the "first difference" model. Further, since the new observations become $\{y_t - y_{t-1}\}_{t=2}^{T}$, there is a loss of one observation due to the differencing procedure, and therefore the degrees of freedom are reduced by one.

The problem is similar if we use ratios and estimate an equation of the form

$$\frac{Y_t}{X_{2t}} = \beta_2 + \beta_0 \frac{1}{X_{2t}} + \beta_1 \frac{X_{1t}}{X_{2t}} + \frac{u_t}{X_{2t}}$$

Now the new residuals will be heteroskedastic.

3. **Dropping a variable(s)**

When faced with severe multicollinearity, the simplest thing to do is to drop one of the collinear variables. However, we may commit a specification error if a variable is dropped from the model. While multicollinearity may prevent precise estimation of the parameters of the model, omitting a variable may make the estimators inconsistent.

4. **Increasing the sample size**

Since multicollinearity is a sample feature, it is possible that in another sample the problem may not be as serious as in the first sample. Sometimes simply increasing the sample size may attenuate the problem, for example, in the trivariate model, we have

$Var\left(\widehat{\beta}_1\right) = \dfrac{\sigma^2}{S_{11}\left(1-r_{12}^2\right)}$ and $Var\left(\widehat{\beta}_2\right) = \dfrac{\sigma^2}{S_{22}\left(1-r_{12}^2\right)}$ since S_{11} and S_{22} increase as the sample size increases, hence $Var\left(\widehat{\beta}_1\right)$ and $Var\left(\widehat{\beta}_2\right)$ will decline as a result.

5. **Benign Neglect**

If we are less interested in interpreting individual coefficients but more interested in forecasting, multicollinearity is not a serious problem. We can simply ignore it.

Chapter 6

Dummy Variables

6.1 Introduction

If a variable is useful in a regression but is not quantifiable, how do we make it a feasible regressor? For example, variables such as gender, race, religion, political background, season and so on are not quantifiable. Consider a simple example. Suppose the wage of a person depends on his/her educational level and gender, we write

$$WAGE_t = \beta_0 + \beta_1 EDUC_t + \beta_2 SEX_t + u_t.$$

Now suppose we define

$$SEX_t = 1 \qquad \text{if person t is a man,}$$

and

$$SEX_t = 0 \qquad \text{if person t is a woman.}$$

Then, what is the meaning of β_2? You may consider β_2 as the amount of wage increases if SEX_t is increased by 1 unit, holding $EDUC_t$ constant. In other words, β_2 is by how much more a man is paid over a woman.

Thus, testing $H_0 : \beta_2 = 0$ is testing whether there is sexual discrimination on workers' compensation.

Theoretically, there is no reason to index male by 1, and female by 0, one can do it the other way around. We do not even need to use 1 and 0, we may pick -1 and 4, or .989 and -108.677675, any two numbers you want. The reason why we use the zero-one combination is totally based on practical consideration, as β_2 can be easily interpreted in this setting.

Suppose we split people into two groups by their gender, then the wage model for men is:

$$WAGE_t = (\beta_0 + \beta_2) + \beta_1 EDUC_t + u_t$$

and the wage model for women is:

$$WAGE_t = \beta_0 + \beta_1 EDUC_t + u_t.$$

After plotting the two regression lines, you will see that one line is parallel to the other, which line is higher depends on the sign of β_2. If $\beta_2 > 0$, the wage of men will generally be higher than that of women of the same education level.

Without using dummy variables, we have to run two regressions. By using them, we only need to run one regression, and we can still distinguish the different features between subgroups.

Example 1: Consider a wage model:

Model A:

$$WAGE_t = \beta_0 + \beta_1 EDU_t + \beta_2 SEX_t + u_t,$$

$t = 1, 2, ..., 40.$

where

$WAGE_t$ is the wage of individual t (dollars).

EDU_t is the years of education of individual t.

SEX_t is the gender of individual t, which defined to be 1 if the individual is a male, and 0 otherwise.

(a) What is the purpose of including SEX_t in the model?

(b) Suppose there 20 men and 20 women in the sample, and the average education for all people in the sample is 10 years. Suppose we run OLS on model A and obtain

$$\widehat{WAGE}_t = 5 + 1.5EDU_t + 10SEX_t.$$

Now suppose we run the model on all the observations without SEX_t, and obtain

$$\widehat{WAGE}_t = 5 + \widehat{\alpha}_1 EDU_t.$$

Find the value of $\widehat{\alpha}_1$.

Solution:

(a) To differentiate the effect of gender on wage income.

(b) Since $\overline{EDU} = 10$ and $\overline{SEX} = \frac{20}{40} = \frac{1}{2}$,

$$\begin{aligned} \overline{WAGE} &= 5 + (1.5)(10) + (10)\left(\frac{1}{2}\right) \\ &= 25. \end{aligned}$$

$$\begin{aligned}
\overline{WAGE} &= 5+\widehat{\alpha}_1(10) \\
\widehat{\alpha}_1 &= \frac{25-5}{10} \\
&= 2.
\end{aligned}$$

6.2 Slope Dummy

Thus far, we have only considered the intercept dummy, i.e. allowing the intercept of the regression lines to be different for different categories, but their slopes are the same. Suppose the value of an additional year of education differs between men and women, how do we reformulate the model to capture this feature? We add an interaction term $EDUCSEX_t$ into the model, where

$$EDUCSEX_t = EDUC_t \times SEX_t.$$

Now we have

$$WAGE_t = \beta_0 + \beta_1 EDUC_t + \beta_2 SEX_t + \beta_3 EDUCSEX_t + u_t.$$

Then the model for male will be

$$WAGE_t = (\beta_0+\beta_2) + (\beta_1+\beta_3)\,EDUC_t + u_t$$

and the model for female is

$$WAGE_t = \beta_0 + \beta_1 EDUC_t + u_t.$$

Now testing $H_0 : \beta_3 = 0$ is testing the gender equality of marginal effect of education on wage.

Exercise 1: Consider a wage model:

Model A:

$$WAGE_t = \beta_0 + \beta_1 EDU_t + \beta_2 SEX_t + \beta_3 EDUSEX_t + u_t$$

$t = 1, 2, ..., 40.$

where

$WAGE_t$ is the wage of individual t (dollars).

EDU_t is the years of education of individual t.

SEX_t is the gender of individual t, which defined to be 1 if the individual is a male, and 0 otherwise.

$EDUSEX_t = EDU_t \times SEX_t$.

(a) What is the purpose of including SEX_t and $EDUSEX_t$ in the model?

(b) Suppose there 20 men and 20 women with the average education of 10 years in the sample, while the average education for all men is 8 years. Suppose we run OLS on model A and obtain

$$\widehat{WAGE}_{tt} = 5 + 1.5EDU_t + 10SEX_t + 2EDUSEX_t.$$

Now suppose we run the model on all the observations without SEX_t and $EDUSEX_t$, and obtain

$$\widehat{WAGE}_t = 5 + \widehat{\alpha}_1 EDU_t.$$

Find the value of $\widehat{\alpha}_1$.

6.3 Seasonal Dummy

Some products like ice-cream, swimming suits, air-conditioners, and clothes are highly seasonally dependent. However, we do not need to run four regressions to tell the difference on sales between seasons. We can create three dummy variables:

$SPRING_t = 1$ if the season is spring, and $= 0$ otherwise;
$SUMMER_t = 1$ if the season is summer, and $= 0$ otherwise;
$FALL_t = 1$ if the season is fall, and $= 0$ otherwise.

The model is:

$$SALES_t = \beta_0 + \beta_1 PRICE_t + \beta_2 SPRING_t + \beta_3 SUMMER_t + \beta_4 FALL_t + u_t.$$

Now, to interpret β_2, β_3, and β_4 and to obtain some additional insight, let us look at the individual models.

The model for Spring is

$$SALES_t = (\beta_0 + \beta_2) + \beta_1 PRICE_t + u_t.$$

The model for Summer is

$$SALES_t = (\beta_0 + \beta_3) + \beta_1 PRICE_t + u_t.$$

The model for Fall is

$$SALES_t = (\beta_0 + \beta_4) + \beta_1 PRICE_t + u_t.$$

The model for Winter is

$$SALES_t = \beta_0 + \beta_1 PRICE_t + u_t.$$

Now, it is clear that β_2 is by how much the sales in the spring are higher than those in the winter, if the prices are the same. Similar interpretations hold for β_3 and β_4. All the first three models are compared to the winter model (the control group).

One may wonder why we do not run the following model:

$$SALES_t = \beta_0 + \beta_1 PRICE_t + \beta_2 SEASON_t + u_t$$

where

$SEASON_t = 1$ if Spring;

$SEASON_t = 2$ if Summer;

$SEASON_t = 3$ if Fall;

$SEASON_t = 4$ if Winter.

The reason is that if you do so, the model will imply:

For Spring,

$$SALES_t = (\beta_0 + \beta_2) + \beta_1 PRICE_t + u_t.$$

For Summer,

$$SALES_t = (\beta_0 + 2\beta_2) + \beta_1 PRICE_t + u_t.$$

For Fall,

$$SALES_t = (\beta_0 + 3\beta_2) + \beta_1 PRICE_t + u_t.$$

For Winter,

$$SALES_t = (\beta_0 + 4\beta_2) + \beta_1 PRICE_t + u_t.$$

Which means for any given price, either

$$SALES_{tspring} > SALES_{tsummer} > SALES_{tfall} > SALES_{twinter} \qquad \text{if } \beta_2 < 0$$

or

$$SALES_{tspring} < SALES_{tsummer} < SALES_{tfall} < SALES_{twinter} \qquad \text{if } \beta_2 > 0$$

or

$$SALES_{tspring} = SALES_{tsummer} = SALES_{tfall} = SALES_{twinter} \qquad \text{if } \beta_2 = 0.$$

This is not very realistic, as there is no reason to presume that $SALES_t$ are either increasing or declining for the four consecutive seasons. Not only are we presuming their $SALES_t$ are in order, but we are also restricting the increment of jump in $SALES_t$ (β_2) to be the same between each consecutive season.

6.4 Dummy Variable Trap

One may ask why we do not create two dummy variables in the gender case and four in the season case. The problem is perfect collinearity.

If we define $MALE_t = 1$ if the person is a male and zero otherwise, and define $FEMALE_t = 1$ if the person is a female, and zero otherwise, and run

$$WAGE_t = \beta_0 + \beta_1 EDUC_t + \beta_2 MALE_t + \beta_3 FEMALE_t + u_t$$

then by definition $MALE_t + FEMALE_t = 1$, and there is perfect collinearity between regressors, and the model is not estimable since

$$WAGE_t = \beta_0 + \beta_1 EDUC_t + \beta_2 MALE_t + \beta_3 (1 - MALE_t) + u_t$$

$$WAGE_t = (\beta_0 + \beta_3) + \beta_1 EDUC_t + (\beta_2 - \beta_3) MALE_t + u_t$$

which means we cannot solve all the $\beta' s$ individually.

Thus, we usually use $N - 1$ dummy variables, where N is the number of categories. In the gender case, $N = 2$, so we use one dummy. In the season's case, $N = 4$, so we use three dummy variables.

If one is not happy with using $N - 1$ dummy and want to use N dummy, he/she may avoid the dummy variable trap by dropping the intercept term. i.e. we run

$$WAGE_t = \beta_1 EDUC_t + \beta_2 MALE_t + \beta_3 FEMALE_t + u_t$$

$$WAGE_t = \beta_1 EDUC_t + \beta_2 MALE_t + \beta_3 (1 - MALE_t) + u_t$$

$$WAGE_t = \beta_3 + \beta_1 EDUC_t + (\beta_2 - \beta_3) MALE_t + u_t.$$

Thus, running a regression of wage on $EDUC_t$, $MALE_t$, and $FEMALE_t$, without an intercept is equivalent to regressing $WAGE_t$ on an intercept, $EDUC_t$, and $MALE_t$.

Therefore, we can obtain the estimates $\widehat{\beta}_3$, $\widehat{\beta}_1$, and $\widehat{\beta_2 - \beta_3}$.

If we define $\widehat{\beta}_2 = \widehat{\beta_2 - \beta_3} + \widehat{\beta}_3$, then all the three $\beta' s$ can be retrieved.

Exercise 2: Suppose the model is

$$MODEL\ A : Y_t = \beta_a Z_{1t} + \beta_b Z_{2t} + u_t$$

where $0 < k < T$, $u_t \sim i.i.d.(0, \sigma^2)$, $t = 1, 2, ..., T$.

$Z_{1t} = 1$ for $t \leq k$;

$Z_{1t} = 0$ for $t > k$;

$Z_{2t} = 0$ for $t \leq k$;

$Z_{2t} = 1$ for $t > k$.

(a) Is the model estimable? Why or why not?

(b) For any given k, with $0 < k < T$, derive the OLS estimators $\widehat{\beta}_a(k)$ and $\widehat{\beta}_b(k)$, and show that they are unbiased estimators for β_a and β_b.

(c) Which of the following models are estimable? Explain.

$$MODEL\ B : Y_t = \beta_a + \beta_b Z_{1t} + u_t$$

$$MODEL\ C : Y_t = \beta_a + \beta_b Z_{1t} + \beta_c Z_{2t} + u_t$$

$$MODEL\ D : Y_t = \beta_a Z_{1t} + \beta_b Z_{2t} + \beta_c Z_{1t} Z_{2t} + u_t$$

$$MODEL\ E : Y_t = \beta_a Z_{1t} + \beta_b Z_{2t} + \beta_c Z_{1t}^2 + u_t$$

(d) Suppose the true model is

$$MODEL\ F : Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

where $0 < k_1 < k_2 < T$, $u_t \sim i.i.d.(0, \sigma^2)$.

$X_{1t} = 1$ for $0 < t \leq k_1$ and $= 0$ otherwise;

$X_{2t} = 1$ for $k_1 < t \leq k_2$ and $= 0$ otherwise;

$X_{3t} = 1$ for $t > k_2$ and $= 0$ otherwise.

However, we misspecify the model and estimate $MODEL\ A$. Find $E\left(\widehat{\beta}_a(k)\right)$ and $E\left(\widehat{\beta}_b(k)\right)$ for the following cases:

i) $0 < k \leq k_1$;

ii) $k_1 < k \leq k_2$;

iii) $k_2 < k \leq T$.

(e) Suppose we know the values of k. Suppose we estimate Model A and obtain

$$\widehat{Y}_t = 10Z_{1t} + 5Z_{2t}$$

Now instead of using a 0-1 dummy, we use a 1-2 dummy defined as follows:

$Z^*_{1t} = 1$ for $t \leq k$

$Z^*_{1t} = 2$ for $t > k$

$Z^*_{2t} = 2$ for $t \leq k$

$Z^*_{2t} = 1$ for $t > k$

and obtain

$$\widehat{Y}_t = \widehat{\beta}_a + \widehat{\beta}_b Z^*_{2t}$$

Find $\widehat{\beta}_a$ and $\widehat{\beta}_b$.

Exercise 3: True/False.

(a) When we regress a variable Y_t on an intercept only, the $\overline{R}^2$ will be negative.

(b) When dummy variables are used, OLS estimators are biased only in large sample.

Exercise 4: Consider the following model

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

where both X_t and Y_t are zero-one dummy variable, how will the followings affect the values of $\widehat{\beta}_0$, $\widehat{\beta}_1$, t-ratio of $\widehat{\beta}_0$, t-ratio of $\widehat{\beta}_1$, and R^2 of the model:

a) X_t is redefined from zero-one to zero-two.

b) X_t is redefined from zero-one to five-ten.

c) Y_t is redefined from zero-one to two-zero.

d) Y_t is redefined from zero-one to two-zero and X_t is redefined from zero-one to zero-two.

e) the sample size T increases.

Exercise 5: Go to the following webpage:

http://osc.universityofcalifornia.edu/journals/journals_a.html.

Let

Y=List Price;

ΔY=average annual increase in price (for the unobserved value, let the price increase be zero)

X_1=Impact factor.

X_2=Online uses.

Define D_i =dummy variable for publisher i (i=Elsevier,Kluwer,...), except Blackwell, which serves as a control dummy.

i) Construct an excel data file, delete the journals without Y, X_1 or X_2 data, as well as the extreme data of ΔY that seem to be unreasonably large.

ii) For all the journals (A-Z) with data, run the following regression models

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \sum \alpha_i D_i + u_t,$$

$$\Delta Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 Y_t + \sum \alpha_i D_i + u_t,$$

and

$$X_{1t} = \beta_0 + \beta_1 Y_t + \beta_2 X_{2t} + \beta_3 Y_t + \sum \alpha_i D_i + u_t,$$

iii) For each of the above models, delete the insignificant variables one at a time and report the model with the highest adjusted R square.

Chapter 7

Heteroskedasticity

7.1 Introduction

Recall that in the OLS estimation, we have an assumption that $Var\left(u_t\right) = \sigma^2$ for all t, which means the errors have the same variance. This is the homoskedasticity assumption. Why do we need this assumption? What is the problem of relaxing it? In fact, this assumption may be quite unrealistic. Consider the consumption model: $C_t = \beta_0 + \beta_1 Y_t + u_t$, where C_t is the consumption of individual t, Y_t is the income of individual t, β_0 can be defined as the autonomous spending, and β_1 can be treated as the marginal propensity to consume. It is quite possible that the fluctuation of consumption may be higher for higher-income group, i.e,. $Var\left(u_t\right)$ may be an increasing function of Y_t, say, $Var(u_t) = \sigma_t = \sigma^2 Y_t$, or $= \sigma^2 Y_t^2$, etc. Also, $Var\left(u_t\right)$ may not depend on Y_t but depend on another variable Z_t. This problem is called heteroskedasticity, meaning that the variance of errors is not a constant. We will study the consequences of heteroskedasticity, the remedies for it and the test for its existence.

7.2 The Consequences of Ignoring Heteroskedasticity

A major consequence heteroskedasticity is that the OLS estimators for β_0 and β_1 will be inefficient. Also, the estimated variances of regression coefficients will be biased and inconsistent, and hence tests of hypothesis are invalid. Fortunately, the estimates will still be unbiased.

The Unbiasedness of OLS Estimator under Heteroskedasticity

To see the unbiasedness of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, it should be noted that

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2}.$$

Thus $E\left(\widehat{\beta}_1\right) = \beta_1$ as long as $E\left(u_t\right) = 0$, so does $E\left(\widehat{\beta}_0\right) = \beta_0$. Therefore, the unbiasedness of the OLS estimators for β does not depend on the variance of u_t.

Inefficiency of OLS Estimator under Heteroskedasticity

However, the OLS estimators will be inefficient since there exists another linear unbiased estimator which has a smaller variance. To see this, consider the model

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

with $Var\left(u_t\right) = \sigma_t^2$.

Recall that the OLS $\widehat{\beta}_1 = \frac{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)Y_t}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2} = \sum_{t=1}^{T} w_t Y_t$, where $w_t = \frac{X_t-\overline{X}}{\sum_{t=1}^{T}\left(X_t-\overline{X}\right)^2}$, we will show that this w_t does not minimize the variance of $\widehat{\beta}_1$, and will find another weight a_t such that the new estimator $\widetilde{\beta}_1 = \sum_{t=1}^{T} a_t Y_t$ is the best linear unbiased estimator for β_1. First, for $\widetilde{\beta}_1$ to be unbiased, we need

$$E\left(\widetilde{\beta}_1\right) = E\left(\sum_{t=1}^{T} a_t Y_t\right) = \sum_{t=1}^{T} a_t\left(\beta_0 + \beta_1 X_t\right) = \beta_0 \sum_{t=1}^{T} a_t + \beta_1 \sum_{t=1}^{T} a_t X_t = \beta_1.$$

In other words,

$$\begin{aligned} \sum_{t=1}^{T} a_t &= 0, \\ \sum_{t=1}^{T} a_t X_t &= 1. \end{aligned}$$

The variance of $\widetilde{\beta}_1$ is given by

$$Var\left(\widetilde{\beta}_1\right) = Var\left(\sum_{t=1}^{T} a_t Y_t\right) = Var\left(\sum_{t=1}^{T} a_t u_t\right) = \sum_{t=1}^{T} a_t^2 \sigma_t^2.$$

Our problem is to choose a series of weight a_t to minimize $Var\left(\widetilde{\beta}_1\right)$ subject to $\sum_{t=1}^{T} a_t = 0$ and $\sum_{t=1}^{T} a_t X_t = 1$. We apply the Lagrangian multiplier method. Let

$$L = \sum_{t=1}^{T} a_t^2 \sigma_t^2 - \lambda_1\left(\sum_{t=1}^{T} a_t\right) - \lambda_2\left(\sum_{t=1}^{T} a_t X_t - 1\right).$$

The first-order conditions are

$$\frac{\partial L}{\partial a_t} = 2a_t\sigma_t^2 - \lambda_1 - \lambda_2 X_t = 0$$

for $t = 1, 2, ..., T$,

and

$$\frac{\partial L}{\partial \lambda_1} = -\sum_{t=1}^{T} a_t = 0,$$

$$\frac{\partial L}{\partial \lambda_2} = -\sum_{t=1}^{T} a_t X_t + 1 = 0.$$

The first T equations give

$$a_t = \frac{1}{2\sigma_t^2}\left(\lambda_1 + \lambda_2 X_t\right)$$

for $t = 1, 2, ..., T$

Adding up all the a_t gives

$$\sum_{t=1}^{T} a_t = \sum_{t=1}^{T} \frac{1}{2\sigma_t^2}\left(\lambda_1 + \lambda_2 X_t\right) = 0.$$

Adding up all the $a_t X_t$ gives

$$\sum_{t=1}^{T} a_t X_t = \sum_{t=1}^{T} \frac{1}{2\sigma_t^2}\left(\lambda_1 X_t + \lambda_2 X_t^2\right) = 1.$$

Solving the two equations above gives

$$\lambda_1 = \frac{-2\sum_{t=1}^{T}\sigma_t^{-2}X_t}{\sum_{t=1}^{T}\sigma_t^{-2}\sum_{t=1}^{T}\sigma_t^{-2}X_t^2 - \left(\sum_{t=1}^{T}\sigma_t^{-2}X_t\right)^2},$$

$$\lambda_2 = \frac{2\sum_{t=1}^{T}\sigma_t^{-2}}{\sum_{t=1}^{T}\sigma_t^{-2}\sum_{t=1}^{T}\sigma_t^{-2}X_t^2 - \left(\sum_{t=1}^{T}\sigma_t^{-2}X_t\right)^2}.$$

Plugging the solutions for λ_1 and λ_2 back into the equations for a_i, $i = 1, 2, ..., T$, we obtain

$$a_i = \frac{-\sigma_i^{-2}\sum_{t=1}^{T}\sigma_t^{-2}X_t + \sigma_i^{-2}X_i\sum_{t=1}^{T}\sigma_t^{-2}}{\sum_{t=1}^{T}\sigma_t^{-2}\sum_{t=1}^{T}\sigma_t^{-2}X_t^2 - \left(\sum_{t=1}^{T}\sigma_t^{-2}X_t\right)^2},$$

which is generally different from the OLS weight w_i unless $\sigma_i^2 = \sigma^2$ for all i. Of course we also have to verify the second order condition to make sure that a_t is actually minimizing $Var\left(\widetilde{\beta}_1\right)$.

Note that if $\sigma_i^2 = \sigma^2$ for all t, the a_i will be reduced to

$$a_i = \frac{-\sum_{t=1}^{T}X_t + X_iT}{T\sum_{t=1}^{T}X_t^2 - \left(\sum_{t=1}^{T}X_t\right)^2} = \frac{X_i - \overline{X}}{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2} = w_i.$$

Now, we know that OLS estimator will be inefficient if there is heteroskedasticity, but which estimator is the most efficient one? Of course the most efficient estimator will be obtained if we rewrite the model such that the error terms become homoskedastic. How should we do it? Suppose that our model is

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

with $Var\left(u_t\right) = \sigma^2 Z_t^2$, where Z is a variable independent of u_t, it may or may not be a function of X. If we divide the whole equation by Z_t, which gives

$$\frac{Y_t}{Z_t} = \beta_0\frac{1}{Z_t} + \beta_1\frac{X_t}{Z_t} + \frac{u_t}{Z_t},$$

$$\frac{Y_t}{Z_t} = \beta_0\frac{1}{Z_t} + \beta_1\frac{X_t}{Z_t} + v_t,$$

where

$$v_t = \frac{u_t}{Z_t}.$$

We claim that v_t is homoskedastic, to see this, note that

$$Var\left(v_t\right) = Var\left(\frac{u_t}{Z_t}\right) = \frac{1}{Z_t^2} Var\left(u_t\right) = \frac{1}{Z_t^2}\sigma^2 Z_t^2 = \sigma^2.$$

Thus, the new error term is homoskedastic, and if we apply OLS on the transformed model, the estimators will be BLUE by the Gauss Markov Theorem. We call this method the Generalized Least Squares method (GLS) or the Weighted Least Squares method (WLS).

Example 1: Suppose the model is

$$y_t = \beta_0 + \beta_1 x_t + u_t \qquad t = 1, 2, ..., T.$$

Suppose we have three observations, i.e. $T = 3$

x_t	8	4	0
y_t	9	2	0

then the OLS estimator of β_1

$$\widehat{\beta}_{1,OLS} = \frac{\sum\limits_{t=1}^{3}\left(x_t - \overline{x}\right) y_t}{\sum\limits_{t=1}^{3}\left(x_t - \overline{x}\right)^2} = \frac{(8-4)\,9 + (4-4)\,2 + (0-4)\,0}{(8-4)^2 + (4-4)^2 + (0-4)^2} = 1.125$$

Suppose there is heteroskedasticity of the form $Var\left(u_t\right) = \sigma^2 z_t$ where z_t is another variable.

If we transform the previous model by dividing all the observations by $\sqrt{z_t}$, the new model becomes

$$\frac{y_t}{\sqrt{z_t}} = \beta_0 \frac{1}{\sqrt{z_t}} + \beta_1 \frac{x_t}{\sqrt{z_t}} + \frac{u_t}{\sqrt{z_t}},$$

and the variance of the new error term will be a constant since

$$Var\left(\frac{u_t}{\sqrt{z_t}}\right) = \frac{1}{z_t} Var\left(u_t\right) = \frac{1}{z_t}\sigma^2 z_t = \sigma^2.$$

Assume that we observe the values of z_t and construct the following table:

x_t	y_t	z_t	$\sqrt{z_t}$	$\frac{x_t}{\sqrt{z_t}}$	$\frac{1}{\sqrt{z_t}}$	$\frac{y_t}{\sqrt{z_t}}$
8	9	1	1	8	1	9
4	2	16	4	1	$\frac{1}{4}$	$\frac{1}{2}$
0	0	9	3	0	$\frac{1}{3}$	0

$$\begin{aligned}
\widehat{\beta}_{1,GLS} &= \frac{\sum_{t=1}^{3} \frac{y_t}{\sqrt{z_t}} \frac{x_t}{\sqrt{z_t}} \sum_{t=1}^{3} \left(\frac{1}{\sqrt{z_t}}\right)^2 - \sum_{t=1}^{3} \frac{y_t}{\sqrt{z_t}} \frac{1}{\sqrt{z_t}} \sum_{t=1}^{3} \frac{1}{\sqrt{z_t}} \frac{x_t}{\sqrt{z_t}}}{\sum_{t=1}^{3} \left(\frac{x_t}{\sqrt{z_t}}\right)^2 \sum_{t=1}^{3} \left(\frac{1}{\sqrt{z_t}}\right)^2 - \left(\sum_{t=1}^{3} \frac{1}{\sqrt{z_t}} \frac{x_t}{\sqrt{z_t}}\right)^2} \\
&= \frac{\sum_{t=1}^{3} \frac{x_t y_t}{z_t} \sum_{t=1}^{3} \frac{1}{z_t} - \sum_{t=1}^{3} \frac{y_t}{z_t} \sum_{t=1}^{3} \frac{x_t}{z_t}}{\sum_{t=1}^{3} \frac{x_t^2}{z_t} \sum_{t=1}^{3} \frac{1}{z_t} - \left(\sum_{t=1}^{3} \frac{x_t}{z_t}\right)^2} \\
&\neq \widehat{\beta}_{1,OLS}.
\end{aligned}$$

Example 2: Suppose the model is

$$y_t = \beta_0 + \beta_1 x_t + u_t \qquad t = 1, 2, ..., T$$

Suppose we have three observations, i.e. $T = 3$,

x_t	1	2	3
y_t	2	3	4

$$\widehat{\beta}_{1,OLS} = \frac{\sum\limits_{t=1}^{3}(x_t - \overline{x})\, y_t}{\sum\limits_{t=1}^{3}(x_t - \overline{x})^2} = \frac{(1-2)\,2 + (2-2)\,3 + (3-2)\,4}{(1-2)^2 + (2-2)^2 + (3-2)^2} = 1,$$

$$\widehat{\beta}_{0,OLS} = \overline{y} - \widehat{\beta}_{0,OLS}\overline{x} = 3 - 1\,(2) = 1.$$

If there is heteroskedasticity of the form $Var\,(u_t) = \sigma^2 x_t^2$, the new model will become

$$\frac{y_t}{x_t} = \beta_1 + \beta_0 \frac{1}{x_t} + \frac{u_t}{x_t}$$

and the variance of the new error term will be a constant since

$$Var\left(\frac{u_t}{x_t}\right) = \frac{1}{x_t^2} Var\,(u_t) = \frac{1}{x_t^2}\sigma^2 x_t^2 = \sigma^2.$$

We can construct the following table:

x_t	y_t	$\frac{1}{x_t}$	$\frac{y_t}{x_t}$
1	2	1	2
2	3	$\frac{1}{2}$	$\frac{3}{2}$
3	4	$\frac{1}{3}$	$\frac{4}{3}$

$$\overline{(1/x)} = (1 + 1/2 + 1/3)\,/3 = \frac{11}{18},$$

$$\widehat{\beta}_{0,GLS} = \frac{\sum\limits_{t=1}^{3}\left(\frac{1}{x_t} - \overline{\left(\frac{1}{x}\right)}\right)\frac{y_t}{x_t}}{\sum\limits_{t=1}^{3}\left(\frac{1}{x_t} - \overline{\left(\frac{1}{x}\right)}\right)^2} = \frac{\left(1 - \frac{11}{18}\right)2 + \left(\frac{1}{2} - \frac{11}{18}\right)\frac{3}{2} + \left(\frac{1}{3} - \frac{11}{18}\right)\frac{4}{3}}{\left(1 - \frac{11}{18}\right)^2 + \left(\frac{1}{2} - \frac{11}{18}\right)^2 + \left(\frac{1}{3} - \frac{11}{18}\right)^2} = 1,$$

$$\widehat{\beta}_{1,GLS} = \overline{\left(\frac{y}{x}\right)} - \widehat{\beta}_{0,GLS}\overline{\left(\frac{1}{x}\right)} = \frac{29}{18} - 1\left(\frac{11}{18}\right) = 1.$$

Thus, the OLS and GLS estimates are identical in this case.

7.3 Testing for Heteroskedasticity

The Goldfeld-Quandt (G-Q) test

The Goldfeld-Quandt (G-Q) test begins with the idea that the variance can be related monotonically to a variable Z. Therefore, if we sort our data by values of Z so that $t = 1$ corresponds to the smallest value of Z while T is the largest, it follows that σ_t^2 should increase monotonically with t. Thus, we need to determine whether σ_t^2 is larger for large t than for small t. The G-Q test is a good test for heteroskedasticity, but it does have a few problems. First, for the test to work well, one must be able to order the σ_t^2. This may be presumptuous since we know so little about heteroskedasticity, though. Second, it relies upon one being able to create a sample in which there is a difference between the first and the second part. For example, if there were 11 observations with $\sigma_t^2 = \sigma^2 + \gamma\,(t-6)^2$, then clearly if we split $\{1, ..., 5\}$ and $\{6, ..., 11\}$ we would obtain the same $\sum \sigma_t^2$ in both samples. Hence, even though there is heteroskedasticity, it does not show up. This is because the heteroskedasticity is not monotonic in t.

The G-Q test is basically an F test, where

$H_0 : \sigma_A^2 = \sigma_B^2$

$H_1 : \sigma_A^2 < \sigma_B^2$

$$F_{obs} = \frac{\widehat{\sigma}_B^2}{\widehat{\sigma}_A^2} = \frac{ESS_B/(T_2 - k - 1)}{ESS_A/(T_1 - k - 1)},$$

where $ESS_A = \sum_{t=1}^{T_1} \widehat{u}_t^2$ and $ESS_B = \sum_{t=T-T_2+1}^{T} \widehat{u}_t^2$, k is the number of regression coefficients excluding the intercept term.

We divide the sample of T observations into first T_1 and the last T_2 observations, and estimate separate regressions for both subsamples. We omit the middle $T_1 + 1$ through $T - T_2$ observations. The number of observations to be omitted is arbitrary and is usually between one-sixth and one-third of total observations. Johnston suggests one-third. Of course, T_1 and T_2 must be greater than the number of coefficients to be estimated.

If $F_{obs} > F^*$,where F^* is the point on the F-distribution such that the area to the right is 5%, then reject the null.

Performing the G-Q Test

Suppose the model is

$$y_t = \beta_0 + \beta_1 x_t + u_t,$$

$t = 1, 2, ..., T$.

Assume that we have ten observations, i.e. $T = 10$.

x_t	1	3	4	2	5	2	1	4	5	3
y_t	0	0	0	0	0	4	2	8	10	6

If there is heteroskedasticity of the form $Var(u_t) = \sigma^2 x_t$, we first arrange the observations according to increasing values of x_t, i.e.,

x_t	1	1	2	2	3	3	4	4	5	5
y_t	0	2	0	4	0	6	0	8	0	10

Now suppose we drop the middle 2 observations and divide the data into two groups, the first group is

x_t	1	1	2	2
y_t	0	2	0	4

$$\begin{aligned}
\widehat{\beta}_{1,OLS} &= \frac{\sum\limits_{t=1}^{4}(x_t-\overline{x})\,y_t}{\sum\limits_{t=1}^{4}(x_t-\overline{x})^2} \\
&= \frac{(1-1.5)\,0+(1-1.5)\,2+(2-1.5)\,0+(2-1.5)\,4}{(1-1.5)^2+(1-1.5)^2+(2-1.5)^2+(2-1.5)^2} \\
&= 1,
\end{aligned}$$

$$\widehat{\beta}_{0,OLS}=\overline{y}-\widehat{\beta}_{0,OLS}\overline{x}=1.5-1\,(1.5)=0,$$

thus

$$\widehat{y}_t=\widehat{\beta}_{0,OLS}+\widehat{\beta}_{1,OLS}x_t=x_t.$$

The error sum of squares for the first group is

$$\begin{aligned}
ESS_A &= \sum_{t=1}^{4}(y_t-\widehat{y}_t)^2=\sum_{t=1}^{4}(y_t-x_t)^2 \\
&= (0-1)^2+(2-1)^2+(0-2)^2+(4-2)^2=10.
\end{aligned}$$

For the second group

x_t	4	4	5	5
y_t	0	8	0	10

$$\widehat{\beta}_{1,OLS}=\frac{\sum\limits_{t=1}^{4}(x_t-\overline{x})\,y_t}{\sum\limits_{t=1}^{4}(x_t-\overline{x})^2}=\frac{(4-4.5)\,0+(4-4.5)\,8+(5-4.5)\,0+(5-4.5)\,10}{(4-4.5)^2+(4-4.5)^2+(5-4.5)^2+(5-4.5)^2}=1,$$

$$\widehat{\beta}_{0,OLS} = \overline{y} - \widehat{\beta}_{0,OLS}\overline{x} = 4.5 - 1\,(4.5) = 0,$$

$$\widehat{y}_t = \widehat{\beta}_{0,OLS} + \widehat{\beta}_{1,OLS}x_t = x_t.$$

The error sum of squares for the second group is

$$ESS_B = \sum_{t=1}^{4}(y_t - \widehat{y}_t)^2 = \sum_{t=1}^{4}(y_t - x_t)^2 = (0-4)^2 + (8-4)^2 + (0-5)^2 + (10-5)^2 = 82.$$

We would like to test

H_0: homoskedasticity

H_1: heteroskedasticity

or

$H_0 : \sigma_B^2 = \sigma_A^2$

$H_1 : \sigma_B^2 > \sigma_A^2$

$$F_{obs} = \frac{\widehat{\sigma}_B^2}{\widehat{\sigma}_A^2} = \frac{ESS_B/\,(T_2 - k - 1)}{ESS_A/\,(T_1 - k - 1)} = \frac{82/\,(4-1-1)}{10/\,(4-1-1)} = 8.2$$

From the F-Table, the critical F-value at 5% level of significance with d.f. (2,2), $F^*_{5\%}(2,2) = 19.00$. Since $F_{obs} < F^*_{5\%}(2,2)$, so we do not reject $H_0 : \sigma_B^2 = \sigma_A^2$ at $\alpha = 5\%$. i.e. We cannot conclude that the variance of u_t is increasing with x_t.

Breusch-Pagan test(B-P test)

Let the model be

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t,$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + ... + \alpha_p Z_{pt},$$

where Z may include some of the variables of X.

We would like to test

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_p = 0.$$

Step 1: Estimate the first model above by OLS and compute

$$\widehat{u}_t = Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1t} - \widehat{\beta}_2 X_{2t} - ... - \widehat{\beta}_k X_{kt},$$

and define

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

Step 2: Run another regression

$$\frac{\widehat{u}_t^2}{\widehat{\sigma}^2} = \alpha_0 + \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + ... + \alpha_p Z_{pt} + v_t.$$

Step 3: Breusch and Pagan show that for large samples, under the null hypothesis

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_p = 0,$$

one-half of the regression sum of squares, $\frac{RSS}{2}$, follows the Chi-square distribution with p degrees of freedom. We reject H_0 if

$$\frac{RSS}{2} > \chi_p^2(\alpha),$$

where α is the level of significance.

White's Test

The B-P test has been shown to be sensitive to any violation of the normality assumption. Also, the previous test assumes prior knowledge of heteroskedasticity. White (1980) has proposed a direct test for heteroskedasticity that is very closely related to the B-P test. The advantage of White's test over the B-P test is that White's test is not sensitive to the normality assumption and we do not need any prior knowledge of the heteroskedasticity. Suppose the model is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t,$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{1t}^2 + \alpha_4 X_{2t}^2 + \alpha_5 X_{1t} X_{2t}.$$

We estimate the first model by OLS, then obtain the estimated residuals $\widehat{u_t} = Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1t} - \widehat{\beta}_2 X_{2t}$, and square it. Regress $\widehat{u}_t^2$ against a constant one, $X_{1t}, X_{2t}, X_{1t}^2, X_{2t}^2$, and $X_{1t} X_{2t}$. This is the auxiliary regression corresponding to the second model above. We now calculate the TR^2 where T is the sample size, and R^2 is the unadjusted R-squared from the auxiliary regression. We reject the null $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ at the significance level α if $TR^2 > \chi_5^2(\alpha)$, $d.f. = 5$. The White's test for models with more than two variables can be extended easily.

Exercise 1: Consider the model $Y_t = \beta_0 + \beta_1 X_t + u_t$, $Var(u_t) = X_t$, X_t is a $0 - 1$ dummy variable. How do we obtain the most efficient estimators for β_0 and β_1 under this kind of heteroskedasticity? Be careful that dividing the whole equation by $\sqrt{X_t}$ may not work as X_t may be zero.

Exercise 2: Consider a simple linear regression model: $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$, $Var(u_t) = \sigma^2 X_{1t}^2$, $t = 1, 2, ..., T$.

i) Are the OLS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ still unbiased in the presence of heteroskedasticity?

ii) Describe how you can obtain the BLUE estimator in the above model.

Exercise 3: Consider the model

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

where

X_t is a single explanatory variable,

β is a scalar parameter,

$E(u_t) = 0$, $Var(u_t) = \sigma_t^2 = Z_t^2\sigma^2$, $Cov(u_t, u_s) = 0$ for all $t \neq s$.

a) Describe how we can transform the model and obtain the Best Linear Unbiased Estimator(BLUE).

b) Under what Z_t is each of the following estimators best linear unbiased? Explain.

(i) $\widehat{\beta} = \dfrac{\sum\limits_{t=1}^{T} X_t Y_t}{\sum\limits_{t=1}^{T} X_t^2}$;

(ii) $\widehat{\beta} = \dfrac{\sum\limits_{t=1}^{T} Y_t}{\sum\limits_{t=1}^{T} X_t}$;

(iii) $\widehat{\beta} = \dfrac{1}{T}\sum\limits_{t=1}^{T}\left(\dfrac{Y_t}{X_t}\right)$.

c) Let $\widehat{\beta}_{GLS}$ be the generalized least squares estimator, and $\widehat{\beta}_{OLS}$ the ordinary least squares estimator. Suppose $\sigma_t^2 = \sigma^2 X_t^2$, show that

(i) $Var\left(\widehat{\beta}_{OLS}\right) = \dfrac{\sigma^2 \sum\limits_{t=1}^{T} X_t^4}{\left(\sum\limits_{t=1}^{T} X_t^2\right)^2}$;

(ii) $Var\left(\widehat{\beta}_{GLS}\right) = \dfrac{\sigma^2}{T}$;

(iii) $Var\left(\widehat{\beta}_{GLS}\right) \leq Var\left(\widehat{\beta}_{OLS}\right)$.

d) Consider the model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t \qquad t = 1, 2, ..., T$$

If $Var(u_t) = (X_{1t} + X_{2t})^2 \sigma^2$, is it possible to obtain BLUE estimators when (i) $\beta_0 \neq 0$, and (ii) $\beta_0 = 0$? Why or why not?

Exercise 4: Consider the model

$$Y_t = \beta + u_t \qquad t = 1, 2, ..., T$$

where

$E(u_t) = 0$, $Var(u_t) = \sigma_t^2 = t\sigma^2$, $0 < \sigma^2 < \infty$, $Cov(u_t, u_s) = 0$ for all $t \neq s$.

Show that

(i) $Var\left(\widehat{\beta}_{OLS}\right) = \dfrac{\sigma^2 (T+1)}{2T}$ and $\lim_{T\to\infty} Var\left(\widehat{\beta}_{OLS}\right) = \dfrac{\sigma^2}{2}$. Is $\widehat{\beta}_{OLS}$ a consistent estimator for β? If yes, why? If not, why not and what does $\widehat{\beta}_{OLS}$ converge to?

(ii) $Var\left(\widehat{\beta}_{GLS}\right) = \dfrac{\sigma^2}{\sum_{t=1}^{T} t^{-1}}$ and $\lim_{T\to\infty} Var\left(\widehat{\beta}_{GLS}\right) = 0$. Is $\widehat{\beta}_{GLS}$ a consistent estimator for β? If yes, why? If not, why not and what does $\widehat{\beta}_{GLS}$ converge to?

Exercise 5: Consider the following model

$$WAGE_t = \beta EDU_t + u_t$$

$t = 1, 2, ..., T$

Suppose we estimate the model by OLS and obtain $\widehat{\beta}_{OLS} = 2$ and $R^2 = 1$. Now suppose there is heteroskedasticity of the form $Var(u_t) = \sigma^2 Z_t^2$ where

Z_t is any variable. If we use GLS to estimate the model, can we say that $\widehat{\beta}_{GLS} = 2$ and R^2 in the transformed model also equals one? If yes, prove it. If not, give a counter example.

Exercise 6: When there is multicollinearity, there is heteroskedasticity.

Exercise 7: When dummy variables are used, OLS estimators are biased only in large sample.

Exercise 8: Consider the model

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

where

X_t is a single explanatory variable; β is a scalar parameter; $E(u_t) = 0$, $Var(u_t) = \sigma_t^2 = \frac{\sigma^2}{X_t^2}$, $Cov(u_t, u_s) = 0$ for all $t \neq s$. Find the Best Linear Unbiased Estimator (BLUE).

Exercise 9: Consider the following model

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

where both X_t and Y_t are zero-one dummy variable, how will the followings affect the values of $\widehat{\beta}_0$, $\widehat{\beta}_1$, t-ratio of $\widehat{\beta}_0$, t-ratio of $\widehat{\beta}_1$, and R^2 of the model:

a) X_t is redefined from zero-one to zero-two.

b) X_t is redefined from zero-one to five-ten.

c) Y_t is redefined from zero-one to two-zero.

d) Y_t is redefined from zero-one to two-zero and X_t is redefined from zero-one to zero-two.

e) the sample size T increases.

Exercise 10: Suppose the model is

$$y_t = \beta_0 + \beta_1 x_t + u_t,$$

$t = 1, 2, ..., T.$

Assume that we have eight observations, i.e. $T = 8$.

x_t	1	-3	4	-2	-5	2	1	4
y_t	0	0	0	0	0	4	2	8

Suppose there is heteroskedasticity of the form $Var\left(u_t\right) = \sigma^2 x_t^2$, perform the Goldfeld-Quandt $(G - Q)$ test without deleting any obseravtion.

Chapter 8

Serial Correlation

8.1 Introduction

In discussing the problem of heteroskedasticity, we have learned that the estimators are still unbiased but will be inefficient. Inefficiency, however, is not the most serious problem. The most problematic issue is inconsistency, which means that the estimator does not converge to the true parameter even if the sample size goes to infinity. One possible cause for inconsistency is the misspecification of the model. As discussed previously, if the true model is a trivariate model, but we estimate a bivariate model, then the OLS estimator is biased, and is inconsistent too. Another possible cause for inconsistency is the violation of the assumption of serial independence.

Consider a simple bivariate model

$$Y_t = \beta_0 + \beta_1 X_t + u_t.$$

The error term u_t is said to be serially dependent if $Cov\left(u_t, u_s\right) \neq 0$ for some $t \neq s$. Consider a simple case where u_t is generated by the process

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

where $|\rho| < 1$ and $\varepsilon_t \sim iid\left(0, \sigma_\varepsilon^2\right)$.

$$\begin{aligned} Cov\left(u_t, u_{t-1}\right) &= E\left(u_t u_{t-1}\right) \\ &= E\left(\left(\rho u_{t-1} + \varepsilon_t\right) u_{t-1}\right) \\ &= \rho E\left(u_{t-1}^2\right) + E\left(\varepsilon_t u_{t-1}\right) \\ &= \rho \sigma^2. \end{aligned}$$

$$Corr\left(u_t, u_{t-1}\right) = \frac{Cov\left(u_t, u_{t-1}\right)}{\sqrt{Var\left(u_t\right) Var\left(u_{t-1}\right)}} = \frac{\rho\sigma^2}{\sigma^2} = \rho.$$

One can easily show that $Corr\left(u_t, u_{t-k}\right) = \rho^{|k|}$ for all k and t.

Example 1: Consider the model:

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

$$u_t = \rho u_{t-2} + \varepsilon_t,$$

$t = 1, 2, ..., T$, $|\rho| < 1$, $\varepsilon_t \sim i.i.d.\left(0, \sigma_\varepsilon^2\right)$.

Find $Corr\left(u_t, u_{t+k}\right)$ in terms of ρ and k, for $k = ... - 2, -1, 0, 1, 2, ...$

Solution: Given that

$$u_t = \rho u_{t-2} + \varepsilon_t.$$

Lead the expression by k periods and we have

$$\begin{aligned}
u_{t+k} &= \rho u_{t+k-2}+\varepsilon_{t+k} \\
u_{t+k}u_t &= \rho u_{t+k-2}u_t+\varepsilon_{t+k}u_t \\
E\left(u_{t+k}u_t\right) &= \rho E\left(u_{t+k-2}u_t\right)+E\left(\varepsilon_{t+k}u_t\right) \\
&= \rho E\left(u_{t+k-2}u_t\right) \text{ since } \varepsilon_t \sim iid.
\end{aligned}$$

When k is even and $k \geq 0$,

$$\begin{aligned}
E\left(u_{t+k}u_t\right) &= \rho^2 E\left(u_{t+k-4}u_t\right) \\
&= \rho^3 E\left(u_{t+k-6}u_t\right) \\
&= \rho^{k/2} E\left(u_t^2\right) \\
&= \rho^{k/2}\sigma^2.
\end{aligned}$$

When k is even and $k < 0$,

$$\begin{aligned}
E\left(u_{t+k}u_t\right) &= E\left(u_t u_{t+k}\right) \text{ since } u_t \text{ is a stationary process.} \\
&= \rho^{-k/2}\sigma^2.
\end{aligned}$$

Hence,

$$E\left(u_{t+k}u_t\right)=\rho^{|k/2|}\sigma^2.$$

Then,

$$\begin{aligned}
Corr\left(u_{t+k}u_t\right) &= \frac{E\left(u_{t+k}u_t\right)}{Var\left(u_t\right)} \\
&= \frac{\rho^{|k/2|}\sigma^2}{\sigma^2} \\
&= \rho^{|k/2|}.
\end{aligned}$$

Similarly, when k is odd and $k > 0$,

$$E\left(u_{t+k}u_t\right) = \rho^{(k-1)/2} E\left(u_{t+1}u_t\right).$$

Now, $E\left(u_{t+1}u_t\right) = \rho E\left(u_{t-1}u_t\right)$. Since u_t is a stationary process, $E\left(u_{t+1}u_t\right) = E\left(u_{t-1}u_t\right)$. We have

$$E\left(u_{t+1}u_t\right)\left(1-\rho\right) = 0.$$

Since $|\rho| < 1$, $E\left(u_{t+1}u_t\right) = 0$. This result can also be applied to the case $k < 0$. Thus, $E\left(u_{t+k}u_t\right) = 0$ and $Corr\left(u_{t+k}u_t\right) = 0$ when k is odd. ■

We will now examine the properties of estimators in the above model. Recall that

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}.$$

Thus, $E\left(\widehat{\beta}_1\right) = \beta_1$ as long as $Cov\left(X_t, u_t\right) = 0$ and $E\left(u_t\right) = 0$.

However, the variance of the estimator is not easy to figure out now.

$$
\begin{aligned}
Var\left(\widehat{\beta}_1\right) &= Var\left(\beta_1 + \frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}\right) \\
&= \frac{Var\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t\right)}{\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2\right)^2} \\
&= \frac{\sum\limits_{t=1}^{T} Var\left(\left(X_t - \overline{X}\right)u_t\right) + \sum\limits_{i=1}^{T}\sum\limits_{j\neq i} Cov\left(\left(X_i - \overline{X}\right)u_i, \left(X_j - \overline{X}\right)u_j\right)}{\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2\right)^2} \\
&= \frac{\sigma^2}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2} + \frac{\sum\limits_{i=1}^{T}\sum\limits_{j\neq i}\left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)Cov\left(u_i, u_j\right)}{\left(\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2\right)^2}.
\end{aligned}
$$

Note *very carefully* that by saying the OLS estimator is inefficient, we are not saying the OLS estimator has a larger variance in the case of serial correlation. We may even obtain a smaller variance as the covariance terms above may end up with a negative value. The variance of the estimator in the presence of serial correlation may be bigger or smaller than in the case of serial independence. The main issue is that even though the OLS estimator has a smaller variance in the case of serial correlation than the OLS estimator in the absence of serial correlation, it does not achieve the global minimum. This result is obvious as the objective Lagrangian function is different. Referring to the chapter on heteroskedasticity, for a linear estimator $\sum\limits_{t=1}^{T} a_t Y_t$ to be the best unbiased estimator in the presence of serial correlation, we apply the Lagrangian multiplier method to minimize:

$$L = \sigma^2 \sum_{t=1}^{T} a_t^2 + \sum_{i \neq j} a_i a_j Cov\left(u_i, u_j\right) - \lambda_1 \left(\sum_{t=1}^{T} a_t \right) - \lambda_2 \left(\sum_{t=1}^{T} a_t X_t - 1 \right).$$

Thus, the solution for a_t will be different from the OLS weight

$$w_t = \frac{X_t - \overline{X}}{\sum\limits_{i=1}^{T} \left(X_i - \overline{X}\right)^2}$$

which minimizes

$$L = \sigma^2 \sum_{t=1}^{T} w_t^2 - \lambda_1 \left(\sum_{t=1}^{T} w_t \right) - \lambda_2 \left(\sum_{t=1}^{T} w X_t - 1 \right).$$

8.2 Cases where $\widehat{\beta}$ is Inconsistent

The problem becomes more serious when the regressors include the lag of Y_t. Consider the following model:

$$Y_t = \beta Y_{t-1} + u_t,$$

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

where $y_0 = 0,\ -1 < \rho < 1$, and $\varepsilon_t \sim iid\left(0, \sigma_\varepsilon^2\right)$.

Now since

$$\widehat{\beta} = \beta + \frac{\sum\limits_{t=1}^{T} Y_{t-1} u_t}{\sum\limits_{t=1}^{T} Y_{t-1}^2}.$$

The estimator is inconsistent as the term $\frac{\sum\limits_{t=1}^{T} Y_{t-1}u_t}{\sum\limits_{t=1}^{T} Y_{t-1}^2}$ does not converge to zero, i.e. $\widehat{\beta}$ does not converge to β even if T goes to infinity. To see this,

$$Y_t = \beta Y_{t-1} + u_t = \beta \left(\beta Y_{t-2} + u_{t-1}\right) + u_t = \sum_{i=0}^{t} \beta^i u_{t-i}.$$

Thus as $T \to \infty$

$$\frac{\sum\limits_{t=1}^{T} Y_{t-1}u_t}{\sum\limits_{t=1}^{T} Y_{t-1}^2} = \frac{\frac{1}{T}\sum\limits_{t=1}^{T} Y_{t-1}u_t}{\frac{1}{T}\sum\limits_{t=1}^{T} Y_{t-1}^2} \xrightarrow{p} \frac{Cov\left(Y_{t-1}, u_t\right)}{Var\left(Y_t\right)}.$$

$$\begin{aligned} Cov\left(Y_{t-1}, u_t\right) &= Cov\left(\sum_{i=0}^{t-1} \beta^i u_{t-1-i}, u_t\right) \\ &= E\left(\sum_{i=0}^{t-1} \beta^i u_{t-1-i} u_t\right) \\ &= \sum_{i=0}^{t-1} \beta^i E\left(u_{t-1-i} u_t\right) \\ &= \sigma^2 \sum_{i=0}^{t-1} \beta^i \rho^{1+i} \\ &\neq 0 \end{aligned}$$

and $Var\left(Y_t\right) > 0$ as Y_t is not a constant.

Exercise 1: True/False/Uncertain. Explain.

a. If there is serial correlation, the OLS estimators will be biased.

b. If there is serial correlation, the OLS estimators will be inconsistent.

Exercise 2: Show that in the model

$$\begin{aligned} Y_t &= \beta Y_{t-1} + u_t, \\ u_t &= \rho u_{t-1} + \varepsilon_t, \\ |\rho| &< 1, \\ u_0 &= 0, \end{aligned}$$

$$\widehat{\beta} \xrightarrow{p} \beta + \frac{\rho\left(1-\beta^2\right)}{1+\beta\rho} \qquad \text{as } T \to \infty.$$

8.3 Estimation under Serial Correlation

Cochrane-Orcutt Iterative Procedure (COIP)

Recall that in the chapter on heteroskedasticity, the way to get rid of heteroskedasticity is to transform the model so that the new error term becomes homoskedastic. In the case of serial correlation, we transform the model until the new error term does not have serial correlation. Consider the following model:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t,$$

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad -1 < \rho < 1.$$

If we use the lag of the first model and multiple it by ρ, we get

$$\rho Y_{t-1} = \beta_0 \rho + \beta_1 \rho X_{1(t-1)} + \beta_2 \rho X_{2(t-1)} + \ldots + \beta_k \rho X_{k(t-1)} + \rho u_{t-1}.$$

Subtract this model from the first model, we get

$$\begin{aligned} Y_t - \rho Y_{t-1} = \beta_0 (1-\rho) + \beta_1 \left(X_{1t} - \rho X_{1(t-1)} \right) + \beta_2 \left(X_{2t} - \rho X_{2(t-1)} \right) + ... \\ + \beta_k \left(X_{kt} - \rho X_{k(t-1)} \right) + u_t - \rho u_{t-1}. \end{aligned}$$

$$Y_t^* = \beta_0^* + \beta_1 X_{1t}^* + \beta_2 X_{2t}^* + ... + \beta_k X_{kt}^* + \varepsilon_t.$$

where

$$\begin{aligned} \beta_0^* &= \beta_0 (1-\rho), \\ Y_t^* &= Y_t - \rho Y_{t-1}, \\ X_{it}^* &= X_{it} - \rho X_{i(t-1)}, \qquad i = 1, 2, ..., k. \end{aligned}$$

The new error term ε_t is now serially independent as we have already assumed it to be i.i.d..

If ρ is known, then we can perform the OLS on the **quasi-differencing** model above, and obtain the best linear unbiased estimators. Of course, ρ is rarely known and has to be estimated. We estimate the original model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t$$

using OLS first, obtain the OLS estimators $\widehat{\beta}' s$ and define

$$\widehat{u}_t = Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1t} - \widehat{\beta}_2 X_{2t} - ... - \widehat{\beta}_k X_{kt}.$$

Then we run a regression using OLS on

$$\widehat{u}_t = \rho \widehat{u}_{t-1} + \varepsilon_t$$

and obtain the OLS estimator $\widehat{\rho}$.

Then we can replace the unknown parameter ρ by $\widehat{\rho}$ in the quasi-differencing model, i.e. we run

$$Y_t^* = \beta_0^* + \beta_1 X_{1t}^* + \beta_2 X_{2t}^* + ... + \beta_k X_{kt}^* + \varepsilon_t,$$

where

$$\begin{aligned} \beta_0^* &= \beta_0 (1 - \widehat{\rho}), \\ Y_t^* &= Y_t - \widehat{\rho} Y_{t-1}, \\ X_{it}^* &= X_{it} - \widehat{\rho} X_{i(t-1)} \qquad i = 1, 2, ..., k. \end{aligned}$$

Now we can obtain the new estimators $\widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_k$ and define $\widehat{\beta}_0 = \dfrac{\widehat{\beta}_0^*}{1 - \widehat{\rho}}$.

The procedure, however, does not end here. Since we have better estimates for $\beta's$ we can now use these new estimates to obtain a better estimate for u_t, and hence ρ, by repeating the above procedure. Getting a better estimate for ρ enables us to obtain an even better estimate for $\beta's$. The procedure is repeated until the estimate of ρ from two successive iterations differ by no more than some prespecified value such as 0.000001.

The Cochrane-Orcutt Iterative procedure is a fast way to obtain efficient estimates. However, it has a deficiency. Like most iterative procedures, the COIP only brings us to the local maximum/minimum. If there is more than one local extremum, we may miss the global maximum/minimum. To correct this deficiency, another estimation method in the presence of serial correlation is proposed.

Hildreth-Lu Search Procedure

The basic idea behind the Hildreth-Lu search procedure is to grid search a value of ρ between -1 an 1 such that the error sum of squares in the regression is minimized. First, we choose a value of ρ, say ρ_1, and use this value to run the Quasi-differencing model. We then record the value of the error sum of squares, $ESS(\rho_1)$. We next choose a different ρ_2 and find out $ESS(\rho_2)$. For example, we may systemically define $\rho_k = \rho_{k-1} + 0.01$. We then find which ρ_k minimizes the error sum of squares. i.e. we calculate $\underset{-1<\rho<1}{Argmin}\ ESS(\rho)$.

8.4 Tests for Serial Correlation

Suppose our model is

$$Y_t = \beta_0 + \beta_1 X_t + u_t,$$

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad -1 < \rho < 1.$$

A test for the first order serial correlation is to test the hypothesis that $H_0 : \rho = 0$. But how to test it? If we can observe the value of $\{u_t\}_{t=1}^T$, then we can run a regression of u_t on u_{t-1}, and perform the t-test. However, u_t is not observable. The only thing observable is $\{X_t, Y_t\}_{t=1}^T$, so we have to extract the information of $\{u_t\}_{t=1}^T$ from the $\{X_t, Y_t\}_{t=1}^T$, which means we have to estimate β_0 and β_1 first.

The first step to test a hypothesis is to identify the null hypothesis, i.e., what are you interested in? The second step is to find out the estimator for the parameter of interest in the null hypothesis. The third step is to construct a test-statistic by transforming or standardizing the estimator. The last step is to find out the theoretical distribution for the test-statistic. It is not an easy task to derive the asymptotic distribution of the estimator and the test-statistic. Even if we know the theoretical result, we may not know the shape of the distribution, and have to rely on high-powered computers to simulate the distribution. After we obtain the distribution for the test-statistic, we will be able to perform the test.

The Durbin-Watson (D-W) test

The most commonly used test for serial correlation is the Durbin-Watson test. The D-W test statistic is defined as

$$d = \frac{\sum\limits_{t=2}^{T} \left(\widehat{u}_t - \widehat{u}_{t-1}\right)^2}{\sum\limits_{t=1}^{T} \widehat{u}_t^2}.$$

Let's investigate why this test-statistic can be used to test serial correlation. Recall that the null hypothesis is that there is no first order serial correlation, i.e. $H_0 : \rho = 0$.

$$d = \frac{\sum\limits_{t=2}^{T} \left(\widehat{u}_t - \widehat{u}_{t-1}\right)^2}{\sum\limits_{t=1}^{T} \widehat{u}_t^2} = \frac{\sum\limits_{t=2}^{T} \widehat{u}_t^2 + \sum\limits_{t=2}^{T} \widehat{u}_{t-1}^2 - 2\sum\limits_{t=2}^{T} \widehat{u}_t \widehat{u}_{t-1}}{\sum\limits_{t=1}^{T} \widehat{u}_t^2} \simeq 2 - 2\frac{\sum\limits_{t=2}^{T} \widehat{u}_t \widehat{u}_{t-1}}{\sum\limits_{t=1}^{T} \widehat{u}_t^2}.$$

Suppose the assumption $Cov\left(X_t, u_t\right) = 0$ still hold, then the $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are consistent estimators for β_0 and β_1 respectively. Thus, $\frac{1}{T}\sum\limits_{t=2}^{T} \widehat{u}_t \widehat{u}_{t-1}$ will converge to $E\left(u_t u_{t-1}\right)$ and $\frac{1}{T}\sum\limits_{t=2}^{T} \widehat{u}_t^2$ will converge to $Var\left(u_t\right)$.

Therefore, as $T \to \infty$, the D-W test statistic

$$d \xrightarrow{p} 2 - 2\frac{E\left(u_t u_{t-1}\right)}{Var\left(u_t\right)} = 2 - 2\frac{\rho\sigma^2}{\sigma^2} = 2\left(1 - \rho\right).$$

Thus, under $H_0 : \rho = 0$, d will converge to 2.

If $\rho > 0$, d will converge to a number less than 2.

If $\rho < 0$, d will converge to a number greater than 2

Thus, we can tell the direction of serial correlation by observing the value of the D-W statistic d. The problem again, is still "how close is close?". The D-W statistic is tabulated in most Econometrics texts. However, people always have difficulties in reading the table. Once you have the number of observations and number of explanatory variables, the D-W table will give you a 5%(and 1%) critical upper and lower bound values d_U and d_L.

To test $H_0 : \rho = 0$ **against** $H_1 : \rho > 0$**.**

$$\textbf{If } d \leq d_L, \textbf{ we reject } H_0.$$

$$\textbf{If } d \geq d_U, \textbf{ we cannot reject } H_0.$$

$$\textbf{If } d_L < d < d_U, \textbf{ the test is inconclusive}.$$

To test $H_0 : \rho = 0$ **against** $H_1 : \rho < 0$**,**

$$\textbf{If } 4 - d \leq d_L, \textbf{ we reject } H_0.$$

$$\textbf{If } 4 - d \geq d_U, \textbf{ we cannot reject } H_0.$$

$$\textbf{If } d_L < 4 - d < d_U, \textbf{ the test is inconclusive}.$$

The major shortcoming of the D-W test is that there is an inconclusive region.

Sometimes the autocorrelation may not be of first order. If a variable is seasonally dependent, and if we are using quarterly data, then we may specify the data generating process of u_t as

$$u_t = \rho_4 u_{t-4} + \varepsilon_t.$$

This extension of the D-W test was given by Wallis (1972). He also provided tables similar to the D-W tables for the test-statistic

$$d_4 = \frac{\sum\limits_{t=5}^{T} \left(\widehat{u}_t - \widehat{u}_{t-4}\right)^2}{\sum\limits_{t=1}^{T} \widehat{u}_t^2}.$$

Example 2: If $T = 40$ (sample size), $k = 4$ (number of explanatory variables excluding the constant term), then $d_L = 1.285$, $d_U = 1.721$.

Exercise 3: A least squares regression based on 24 observations produces the following results:

$$\widehat{Y}_t = \underset{(0.1)}{.3} + \underset{(0.2)}{1.21} X_t, R^2 = 0.982, DW = 1.31.$$

Test the hypothesis that the disturbances are not autocorrelated.

The Lagrange Multiplier (LM) Test

The LM test for the null $H_0 : \rho = 0$ is performed as follows:

Suppose our model is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t,$$

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad -1 < \rho < 1.$$

This model is equivalent to

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + \rho u_{t-1} + \varepsilon_t.$$

Thus, the test for $\rho = 0$ can be treated as a LM test for the additional variable u_{t-1}.

We estimate the original model by OLS and obtain the estimated residuals $\widehat{u}_t$.

We then regress $\widehat{u}_t$ on a constant, all the X's and $\widehat{u}_{t-1}$.

Compute $(T-1)\,R^2$ from this auxiliary regression.

We **reject the null at the significance level** α **if**

$$(T-1)\,R^2 > \chi_1^2\,(\alpha)\,.$$

The LM test does not have the inconclusiveness of the D-W test. However, it is a large-sample test and would need at least 30 degrees of freedom for the test to be meaningful.

Example 3: A model of demand for ice cream is estimated below:

$$\begin{aligned}
\widehat{DEMAND}_t &= \underset{(0.5)}{0.157} - \underset{(-1.1)}{0.892} PRICE_t + \underset{(2.07)}{0.0032} INCOME_t + \underset{(6.42)}{0.00356} TEMP_t, \\
T &= \text{Sample size} = 29, \\
k &= 3 = \text{Number of explanatory variables excluding the constant term}, \\
ESS &= \sum_{t=1}^{29} \left(DEMAND_t - \widehat{DEMAND}_t \right)^2 = 124, \\
\overline{R}^2 &= 1 - \frac{ESS/(T-k-1)}{TSS/(T-1)} = 0.72, \\
DW &= \text{Durbin-Watson Statistic} = 1.55,
\end{aligned}$$

where

$DEMAND$ = per capita consumption of ice cream in pints,

$PRICE$ = price per pint in dollars,

$INCOME$ = weekly family income in dollars,

$TEMP =$ mean temperature in Fahrenheit,

and the figures in the parentheses are the *t-ratio*s

a) Interpret each of the above coefficient estimates. Perform the t-test for $H_0 : \beta_i = 0$ v.s. $H_1 : \beta_i \neq 0$, $i = 0, 1, 2, 3$ at $\alpha = 5\%$.

b) Find the value of R^2, Total Sum of Squares= $\sum_{t=1}^{29} \left(DEMAND_t - \overline{DEMAND}\right)^2$ and the Regression Sum of Squares (RSS) in the above model.

c) Suppose we want to test the joint hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, and run the restricted model as:

$$DEMAND_t = \beta_0 + u_t.$$

i) Show that the Ordinary Least Squares estimate for β_0 is $\widehat{\beta}_0 = \overline{DEMAND}$, where $\overline{DEMAND} = \frac{\sum_{t=1}^{29} DEMAND_t}{29}$.

ii) Show that $\widehat{DEMAND}_t = \overline{DEMAND}$ for all $t = 1, 2, ..., 29$. What is the value of the restricted error sum of squares= $\sum_{t=1}^{29} \left(DEMAND_t - \widehat{DEMAND}_t\right)^2$?

iii) Perform an F-test on $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at $\alpha = 5\%$ using the F-statistic defined as $F = \frac{(ESS_r - ESS_u) / (df_r - df_u)}{ESS_u / df_u}$.

d) We suspect that the error term u_t has a first order serial correlation , i.e. $u_t = \rho u_{t-1} + \varepsilon_t$, where ε_t are i.i.d. random variables. Perform the Durbin-Watson (DW) Test on $H_0 : \rho = 0$ v.s. $H_1 : \rho > 0$ at $\alpha = 5\%$.

e) If we use the residual $\widehat{u}_t$ from the above unrestricted model, and estimate an auxiliary regression:

$$\widehat{\widehat{u}}_t = 0.3 - 1.62PRICE_t + 0.01INCOME_t + 0.078TEMP_t + 0.26\widehat{u}_{t-1},$$

with $R^2 = 0.348$. Perform the Lagrange Multiplier (LM) Test on $H_0 : \rho = 0$ v.s. $H_1 : \rho \neq 0$ at $\alpha = 5\%$.

Solution:

(a)

$$\begin{aligned}
\beta_1 &= \text{ Marginal Effect of change in price on the demand for ice-cream} \\
\beta_2 &= \text{ Marginal Effect of change in income on the demand for ice-cream} \\
\beta_3 &= \text{ Marginal Effect of change in temperature on the demand for ice-cream} \\
\beta_0 &= \text{ Effect on the demand for ice-cream when the other variables are zero}
\end{aligned}$$

To test the hypotheses $H_0 : \beta_i = 0$ for $i = 0, 1, 2, 3$, we find out the critical value of the t-statistic at 5% level of significance with degree of freedom $(29 - 4) = 25$.

$$t = 2.060.$$

The calculated t-statistics are

$$\begin{aligned}
\text{When } i &= 0,\ t_{obs} = 0.5.\ H_0 \text{ cannot be rejected.} \\
\text{When } i &= 1,\ t_{obs} = -1.1.\ H_0 \text{ cannot be rejected.} \\
\text{When } i &= 2,\ t_{obs} = 2.07.\ H_0 \text{ is rejected.} \\
\text{When } i &= 3,\ t_{obs} = 6.42.\ H_0 \text{ is rejected.}
\end{aligned}$$

(b) Since $\overline{R}^2 = 1 - \dfrac{T-1}{T-k-1}(1 - R^2)$,

$$\begin{aligned} R^2 &= 1 - \frac{T-k-1}{T-1}\left(1-\overline{R}^2\right) \\ &= 1 - \frac{29-3-1}{29-1}(1-0.72) \\ &= 0.75. \end{aligned}$$

$$\begin{aligned} R^2 &= 1 - \frac{ESS}{TSS} \\ \Rightarrow TSS &= \frac{ESS}{1-R^2} \\ &= \frac{124}{1-0.75} \\ &= 496. \end{aligned}$$

$$\begin{aligned} R^2 &= \frac{RSS}{TSS} \\ \Rightarrow RSS &= TSS \times R^2 \\ &= 496 \times 0.75 \\ &= 372, \end{aligned}$$

or

$$\begin{aligned} RSS &= TSS - ESS \\ &= 496 - 124 \\ &= 372. \end{aligned}$$

(c)(i) The OLS estimate of β_0 is given by

$$\begin{aligned}
\widehat{\beta}_0 &= \frac{\sum_{t=1}^{29} 1 \times DEMAND_t}{\sum_{t=1}^{29} 1^2} \\
&= \frac{1}{29}\sum_{t=1}^{29} DEMAND_t \\
&= \overline{DEMAND}.
\end{aligned}$$

(c)(ii)

$$\widehat{DEMAND}_t = \widehat{\beta}_0 = \overline{DEMAND},$$

by the result in (i).

$$\begin{aligned}
ESS_r &= \sum_{t=1}^{29}\left(DEMAND_t - \widehat{DEMAND}_t\right)^2 \\
&= \sum_{t=1}^{29}\left(DEMAND_t - \overline{DEMAND}\right)^2 \\
&= TSS \\
&= 496.
\end{aligned}$$

(c)(iii)

$$F_{obs} = \frac{(496-124)/3}{124/(29-3-1)} = 25 > F_{3,25} = 2.99$$

Then, we can reject the null hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ at 5% level of significance.

(d) $d = 1.55, d_L = 1.198, d_U = 1.650.$ Since $d_L < d < d_U$, the test is inconclusive.

(e)

$$(T-1)\,R^2 = (28)\,(0.348) = 9.774 > \chi_1^2 = 3.841.$$

Thus, the null hypothesis of no serial correlation is rejected at 5% level of significance. ■

Exercise 4: A model of deaths due to heart disease is estimated below:

$$\begin{aligned}
\widehat{CHD}_t &= 139.68 + 10.71 CIG_t + 3.38 EDFAT_t + 26.75 SPIRITS_t - 4.13 BEER_t, \\
T &= \text{sample size} = 34, \\
k &= 4 = \text{number of explanatory variables excluding the constant term}, \\
ESS &= \sum_{t=1}^{34} \left(CHD_t - \widehat{CHD}_t \right)^2 = 2122, \\
\overline{R}^2 &= 1 - \frac{ESS/(T-k-1)}{TSS/(T-1)} = 0.672, \\
DW &= \text{Durbin-Watson Statistic} = 1.485,
\end{aligned}$$

where

$CHD =$ death rate (per million population) due to coronary heart disease in the U.S. during each of the years 1947-1980,

$CIG =$per capita consumption of cigarettes measured in pounds of tobacco,

$EDFAT =$ per capita intake of edible fats and oil, measured in pounds,

$SPIRITS =$per capita consumption of distilled spirits in gallons,

$BEER =$ per capita consumption of malted liquor in gallons.

a) We suspect that the error term u_t has a first order serial correlation , i.e. $u_t = \rho u_{t-1} + \varepsilon_t$, where ε_t are i.i.d. random variables. Perform the Durbin-Watson (DW) Test on $H_0 : \rho = 0$ v.s. $H_1 : \rho < 0$ at $\alpha = 5\%$.

b) If we use the residual $\widehat{u}_t$ from the above unrestricted model, and estimate an auxiliary regression:

$$\begin{aligned}
\widehat{\widehat{u}}_t &= 113.63 - 4.68CIG_t - 1.58EDFAT_t + 0.36SPIRITS_t + 0.21BEER_t + 0.26\widehat{u}_{t-1}, \\
R^2 &= 0.137.
\end{aligned}$$

Perform the Lagrange Multiplier(LM) Test on $H_0 : \rho = 0$ v.s. $H_1 : \rho \neq 0$ at $\alpha = 5\%$.

Exercise 5: A model of annual demand for ice-cream during the period 1967-1996 is estimated below:

$$\widehat{DEMAND}_t = \underset{(0.5)}{0.157} - \underset{(-1.1)}{0.892}PRICE_t + \underset{(2.07)}{0.0032}INCOME_t + \underset{(6.42)}{0.00356}TEMP_t - \underset{(0.2)}{0.5}D_t$$

$$\begin{aligned}
T &= \text{Sample size} = 30 \\
k &= 4 = \text{Number of explanatory variables excluding the constant term} \\
ESS &= \sum_{t=1}^{30}\left(DEMAND_t - \widehat{DEMAND}_t\right)^2 = 125 \\
\overline{R}^2 &= 1 - \frac{ESS/(T-k-1)}{TSS/(T-1)} = 0.5 \\
DW &= \text{Durbin-Watson Statistic} = 2.51
\end{aligned}$$

where

$DEMAND_t$ = Consumption of ice-cream in pints in year t.

$PRICE_t$ = Price of ice-cream per pint in year t. (dollars)

$INCOME$ = GDP per capita in year t. (dollars)

$TEMP_t$ = Mean temperature in Fahrenheit in year t.

$D_t = 1$ if the year is after 1981, $D_t = 0$ if the year is in or before 1981.

The figures in the parentheses are the **t-ratio**.

a) Interpret each of the above coefficient estimates. Perform the t-test for $H_0 : \beta_i = 0$ v.s. $H_1 : \beta_i \neq 0$, $i = 0, 1, 2, 3, 4$, at $\alpha = 5\%$.

b) Find the value of R^2, Total Sum of Squares= $\sum\limits_{t=1}^{30} \left(DEMAND_t - \overline{DEMAND}\right)^2$ and the Regression Sum of Squares (RSS) in the above model.

c) Suppose we want to test the joint hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, and get the restricted model as:

$$\widehat{DEMAND}_t = 6 - 2D_t$$

i) Show that $\widehat{DEMAND}_t = \overline{DEMAND}$ -1 after 1981 and $\widehat{DEMAND}_t = \overline{DEMAND}$ $+1$ in or before 1981, where $\overline{DEMAND} = \dfrac{\sum\limits_{t=1}^{30} DEMAND_t}{30}$.

ii) Let the average demand for ice-cream after 1981 be $\overline{DEMAND}_A$. Show that the restricted error sum of squares can be written as

$$ESS_r = TSS + 60\overline{DEMAND}_A - 270$$

iii) If $\overline{DEMAND}_A = 3$. Perform an F test on $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at $\alpha = 5\%$ using the F-statistic defined as $F = \dfrac{(ESS_r - ESS_u) / (df_r - df_u)}{ESS_u / df_u}$.

d) If we use the residual $\widehat{u}_t$ from the above unrestricted model, and estimate an auxiliary regression:

$$\widehat{\widehat{u}}_t = 0.3 - 1.62PRICE_t + 0.01INCOME_t + 0.078TEMP_t - 0.03D_t + 0.26\widehat{u}_{t-1}$$

with $R^2 = 0.236$. Perform the Lagrange Multiplier (LM) Test on $H_0 : \rho = 0$ v.s. $H_1 : \rho \neq 0$ at $\alpha = 5\%$.

e) We suspect that the error term u_t has a first order serial correlation , i.e. $u_t = \rho u_{t-1} + \varepsilon_t$, where ε_t are i.i.d. random variables. Perform the Durbin-Watson (DW) Test on $H_0 : \rho = 0$ v.s. $H_1 : \rho > 0$ at $\alpha = 5\%$.

Exercise 6: True/False

(a). When there is serial correlation, the OLS estimators will be BLUE.

(b). The Durbin-Watson Test is a test for Heteroskedasticity.

Chapter 9

Discrete and Limited Dependent Variable Models

9.1 Introduction

Thus far, we have assumed that the dependent variable in a model takes continuous values. However, this is not always the case. For example, assume that we are just interested in whether people participate in the labor force; whether people are married or not; whether people own a car or not, etc. All of these yes-no decisions are not easily quantifiable. In chapter 6 we have studied situations where the independent variables are qualitative. Thus, we can also use a similar technique here. For example, if a person is married, we assign a value of 1 to him/her, and assign 0 otherwise.

9.2 Linear Probability Model

Suppose Y is a $0-1$ 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 obtain 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 sports car while the other purchase family car. We define $Y_t = 1$ if a family car is purchased and $Y_t = 0$ if a sports 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, what will be the increase in probability of buying a family car? An advantage of using the linear probability model is that it is very convenient to carry out. By running a regression we can obtain 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 $Var\left(u_t\right)$ to be a constant in this framework. To see why, note that

$$\begin{aligned}
Var\left(u_t\right) &= E\left(u_t^2\right) - E^2\left(u_t\right) = E\left(u_t^2\right) \\
&= \left(1-\beta_0-\beta_1X_t\right)^2 \Pr\left(Y_t=1\right) + \left(-\beta_0-\beta_1X_t\right)^2 \Pr\left(Y_t=0\right) \\
&= \left(1-\beta_0-\beta_1X_t\right)^2 \Pr\left(Y_t=1\right) + \left(\beta_0+\beta_1X_t\right)^2 \Pr\left(Y_t=0\right) \\
&= \left(1-\Pr\left(Y_t=1\right)\right)^2 \Pr\left(Y_t=1\right) + \Pr\left(Y_t=1\right)^2 \Pr\left(Y_t=0\right) \\
&= \Pr\left(Y_t=0\right)^2 \Pr\left(Y_t=1\right) + \Pr\left(Y_t=1\right)^2 \Pr\left(Y_t=0\right) \\
&= \Pr\left(Y_t=0\right) \Pr\left(Y_t=1\right) \left[\Pr\left(Y_t=0\right) + \Pr\left(Y_t=1\right)\right] \\
&= \Pr\left(Y_t=0\right) \Pr\left(Y_t=1\right) \\
&= \left(1-\beta_0-\beta_1X_t\right)\left(\beta_0+\beta_1X_t\right),
\end{aligned}$$

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 GLS to obtain efficient estimates. If $0 < \widehat{Y}_t < 1$ for all t, we can obtain GLS estimators by dividing all the observations by $\sqrt{\left(1-\widehat{\beta}_0-\widehat{\beta}_1X_t\right)\left(\widehat{\beta}_0+\widehat{\beta}_1X_t\right)} = \sqrt{\left(1-\widehat{Y}_t\right)\widehat{Y}_t}$.

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 generally. Therefore, in large samples the statistical inference of the LPM will follow the usual OLS procedure under the normality assumption.

Questionable value of R^2 as a measure of goodness of fit

The conventionally computed R^2 is of limited value 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 LPM is expected to fit such a scatter well. As a result, the conventionally computed R^2 is likely to be much lower than 1 for such models. In most practical applications the R^2 ranges from 0.2 to 0.6.

Nonfulfillment of $\mathbf{0} <\mathrm{Pr}\,\widehat{(Y_t = 1)}< \mathbf{1}$.

The other problem is on prediction and forecasting. Since

$$\widehat{Y}_t = \widehat{\beta}_0 + \widehat{\beta}_1 X_t = \mathrm{Pr}\,\widehat{(Y_t = 1)}$$

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. The obvious solution to 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 perfect, as it suggests that we might predict an occurrence with a probability of 1 when it is entirely possible that it may not occur, or we might predict an occurrence with probability 0 when it 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. However, the predicted value is the value in a regression curve, so in order to fulfil this restriction, we must find a function $\widehat{Y_t} = g\left(X_t, \beta\right)$ such that $0 \leq g\left(X_t, \beta\right) \leq 1$ for all β and X_t. Clearly $g\left(X_t, \beta\right)$ cannot be linear in either β or X, i.e. $g\left(X_t, \beta\right) = \beta_0 + \beta_1 X_t$ 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 functions 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

$$g\left(X_t, \beta\right) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 X_t\right)\right]}.$$

Note that as $\beta_1 X_t \to -\infty$, $g\left(X_t, \beta\right) \to 0$, and as $\beta_1 X_t \to \infty$, $g\left(X_t, \beta\right) \to 1$. Since $g\left(X_t, \beta\right)$ 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.

Example 1: Consider the following linear probability model:

$$Y_t = \beta_0 + \beta_1 INCOME_t + \beta_2 MARRIED_t + u_t,$$

where

$Y_t = 1$ if individual t purchased a car in the year of the survey and $Y_t = 0$ if not.

$INCOME_t$ =monthly income of individual t (in dollars).

$MARRIED_t = 1$ if individual t is married and $MARRIED_t = 0$ if not.

a) Show that $E(Y_t) = \Pr(Y_t = 1)$.

b) Show that $E(u_t) = 0$ implies

$$\Pr(Y_t = 1) = \beta_0 + \beta_1 INCOME_t + \beta_2 MARRIED_t.$$

c) Show that $\text{Var}(u_t) = \Pr(Y_t = 1)\Pr(Y_t = 0)$.

d) Suppose we estimate the model by OLS and obtain:

$$\widehat{Y}_t = -.1 + 0.0001 INCOME_t + 0.3 MARRIED_t.$$

Interpret each of the above coefficient estimates.

e) Referring to the estimated model in part d), what is the chance of purchasing a car for:

i) an individual who is married and has a monthly income of 5000 dollars.

ii) an individual who is married and has a monthly income of 10000 dollars.

iii) an individual who is not married and has a monthly income of 1000 dollars.

f) State the advantages and shortcomings of the linear probability model.

Solution:

(a)

$$E(Y_t) = 0 \times \Pr(Y_t = 0) + 1 \times \Pr(Y_t = 1) = \Pr(Y_t = 1).$$

(b)

$$\begin{aligned} E(u_t) &= 0 \\ \Rightarrow E(Y_t) &= \beta_0 + \beta_1 INCOME_t + \beta_2 MARRIED_t. \end{aligned}$$

By using the result of part (a), i.e. $E(Y_t) = \Pr(Y_t = 1)$, we have

$$\Pr(Y_t = 1) = \beta_0 + \beta_1 INCOME_t + \beta_2 MARRIED_t.$$

(c)

When $Y_t = 1$,

$$\begin{aligned} u_t &= 1 - \beta_0 - \beta_1 INCOME_t - \beta_2 MARRIED_t \\ &= 1 - \Pr(Y_t = 1) \\ &= \Pr(Y_t = 0). \end{aligned}$$

When $Y_t = 0$,

$$\begin{aligned} u_t &= 0 - \beta_0 - \beta_1 INCOME_t - \beta_2 MARRIED_t \\ &= -\Pr(Y_t = 1). \end{aligned}$$

Now,

$$\begin{aligned} Var(u_t) &= E\left(u_t^2\right) \text{ since } E(u_t) = 0 \\ &= \Pr(Y_t = 0)^2 \times \Pr(Y_t = 1) + (-\Pr(Y_t = 1))^2 \times \Pr(Y_t = 0) \\ &= \Pr(Y_t = 1)\Pr(Y_t = 0)\left[\Pr(Y_t = 0) + \Pr(Y_t = 1)\right] \\ &= \Pr(Y_t = 1)\Pr(Y_t = 0). \end{aligned}$$

(d)

$\beta_1 =$ Marginal Effect of change in monthly income on the probability of $Y_t = 1$.

$\beta_2 =$ Marginal Effect of change in marriage on the probability of $Y_t = 1$.

$\beta_0 =$ Effect on the probability of $Y_t = 1$ when the other variables are zero.

(e)

(i)

$$\begin{aligned} \widehat{Y} &= -0.1 + (0.0001)(5000) + (0.3)(1) \\ &= 0.7. \end{aligned}$$

■

(ii)

$$\begin{aligned} \widehat{Y} &= -0.1 + (0.0001)(10000) + (0.3)(1) \\ &= 1.2. \end{aligned}$$

(iii)

$$\begin{aligned} \widehat{Y} &= -0.1 + (0.0001)(1000) + (0.3)(0) \\ &= 0. \end{aligned}$$

(f) Advantage : It is convenient to carry out. Disadvantage : $0 < \widehat{Y}_t < 1$ may not be satisfied. ■

9.3 Random Utility Model

Suppose you have to make a decision on two alternatives. For example, whether to buy a sports car or a family car. Given the characteristics X_t of individual t , for example, his/her family size, income, etc. Let

$$\begin{aligned} U_{1t} &= \alpha_0 + \alpha_1 X_t + \varepsilon_{1t}, \\ U_{2t} &= \gamma_0 + \gamma_1 X_t + \varepsilon_{2t}. \end{aligned}$$

where U_{1t} is the utility derived from a family car, and U_{2t} is the utility derived from a sports car. The individual will buy a family car if $U_{1t} > U_{2t}$, or $U_{1t} - U_{2t} > 0$. Subtracting the second equation from the first equation gives

$$U_{1t} - U_{2t} = \alpha_0 - \gamma_0 + (\alpha_1 - \gamma_1) X_t + \varepsilon_{1t} - \varepsilon_{2t}.$$

Suppose we define $Y_t^* = U_{1t} - U_{2t}$, $\beta_0 = \alpha_0 - \gamma_0$, $\beta_1 = \alpha_1 - \gamma_1$, $u_t = \varepsilon_{1t} - \varepsilon_{2t}$. We can rewrite the model as

$$Y_t^* = \beta_0 + \beta_1 X_t + u_t.$$

However, we cannot observe the exact value of Y_t^*, what we observe is whether the individual buy a family car or not. That is, we only observe whether $Y_t^* > 0$ or $Y_t^* < 0$. If $Y_t^* > 0$, the individual will buy a family car, we assign a value $Y_t = 1$ for this observation, and assign $Y_t = 0$ otherwise. In other words, we have $Y_t = 1$ if $Y_t^* > 0$ and $Y_t = 0$ if $Y_t^* < 0$. Denote the density function and distribution function of u_t by $f(\cdot)$ and $F(\cdot)$ respectively, and suppose it is symmetric about zero, i.e. $f(u_t) = f(-u_t)$, and $F(u_t) = 1 - F(-u_t)$. We then have:

$$\begin{aligned}
\Pr(Y_t = 1) &= \Pr(Y_t^* > 0) \\
&= \Pr(\beta_0 + \beta_1 X_t + u_t > 0) \\
&= \Pr(u_t > -\beta_0 - \beta_1 X_t) \\
&= \Pr(-u_t < \beta_0 + \beta_1 X_t) \\
&= \Pr(u_t < \beta_0 + \beta_1 X_t) \qquad \text{since } u_t \text{ is symmetrically distributed about zero,} \\
&= F(\beta_0 + \beta_1 X_t),
\end{aligned}$$

and

$$\Pr(Y_t = 0) = 1 - \Pr(Y_t = 1) = 1 - F(\beta_0 + \beta_1 X_t).$$

9.4 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

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta_0} &= \sum_{t=1}^{T} Y_t \frac{f(\beta_0+\beta_1 X_t)}{F(\beta_0+\beta_1 X_t)} - \sum_{t=1}^{T} (1-Y_t) \frac{f(\beta_0+\beta_1 X_t)}{1-F(\beta_0+\beta_1 X_t)} = 0, \\ \frac{\partial \ln L}{\partial \beta_1} &= \sum_{t=1}^{T} Y_t X_t \frac{f(\beta_0+\beta_1 X_t)}{F(\beta_0+\beta_1 X_t)} - \sum_{t=1}^{T} (1-Y_t) X_t \frac{f(\beta_0+\beta_1 X_t)}{1-F(\beta_0+\beta_1 X_t)} = 0.\end{aligned}$$

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 obtain the estimator of $\beta' s$ by simple substitution. We may use grid-search method and a computer algorithm to solve them. The MLE procedure has a number of desirable properties. When sample size is large, all parameter estimators are consistent and also efficient if there is no misspecification in 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 $N(0,\sigma^2)$, i.e.,

$$\begin{aligned}f(\beta_0+\beta_1 X_t) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\beta_0+\beta_1 X_t)^2}{2\sigma^2}\right), \\ F(\beta_0+\beta_1 X_t) &= \int_{-\infty}^{\beta_0+\beta_1 X_t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt.\end{aligned}$$

then we have the **Probit Model**.

The first order condition can be simplified to

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta_0} &= \sum_{Y_t=1} \frac{\exp\left(-\frac{(\beta_0+\beta_1 X_t)^2}{2\sigma^2}\right)}{\int_{-\infty}^{\beta_0+\beta_1 X_t} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt} - \sum_{Y_t=0} \frac{\exp\left(-\frac{(\beta_0+\beta_1 X_t)^2}{2\sigma^2}\right)}{\int_{\beta_0+\beta_1 X_t}^{\infty} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt} = 0. \\ \frac{\partial \ln L}{\partial \beta_1} &= \sum_{Y_t=1} \frac{X_t \exp\left(-\frac{(\beta_0+\beta_1 X_t)^2}{2\sigma^2}\right)}{\int_{-\infty}^{\beta_0+\beta_1 X_t} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt} - \sum_{Y_t=0} \frac{X_t \exp\left(-\frac{(\beta_0+\beta_1 X_t)^2}{2\sigma^2}\right)}{\int_{\beta_0+\beta_1 X_t}^{\infty} \exp\left(-\frac{t^2}{2\sigma^2}\right) dt} = 0.\end{aligned}$$

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

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta_0} &= \sum_{Y_t=1} \frac{1}{1 + \exp(\beta_0 + \beta_1 X_t)} - \sum_{Y_t=0} \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_t)} = 0, \\ \frac{\partial \ln L}{\partial \beta_1} &= \sum_{Y_t=1} \frac{X_t}{1 + \exp(\beta_0 + \beta_1 X_t)} - \sum_{Y_t=0} \frac{X_t}{1 + \exp(-\beta_0 - \beta_1 X_t)} = 0. \end{aligned}$$

We only discuss a simple model with two $\beta' s$ for simplicity purpose. Of course, one can easily extend this to multiple-parameter models.

9.5 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 may only observe the sign of the dependent variable. If we only observe a subpopulation such as individuals with income above a certain level, then we say the data is being 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 $\displaystyle\int_{-\infty}^{\infty} f(Y)\,dY = 1$. 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 $Y > a$.

$$\begin{aligned}
F(Y < c|Y > a) &= \Pr(Y < c|Y > a) = \frac{Pr(Y < c \cap Y > a)}{P(Y > a)} = \frac{\int_a^c f(Y)\,dY}{\int_a^\infty f(Y)\,dY}, \\
f(Y = c|Y > a) &= \frac{dF(Y < c|Y > a)}{dc} = \frac{f(c)}{\int_a^\infty f(Y)\,dY}.
\end{aligned}$$

Example 2: 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

$$\Pr\left(Y > \frac{3}{4}\middle| |Y > \frac{1}{2}\right) = \frac{\Pr\left(Y > \frac{3}{4} \cap Y > \frac{1}{2}\right)}{\Pr\left(Y > \frac{1}{2}\right)} = \frac{\Pr\left(Y > \frac{3}{4}\right)}{\Pr\left(Y > \frac{1}{2}\right)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

9.6 Moments of Truncated Distributions

Note that

$$\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} Yf(Y)\,dY \\
&= \int_{-\infty}^{a} Yf(Y)\,dY + \int_{a}^{\infty} Yf(Y)\,dY \\
&= \int_{-\infty}^{a} Y\frac{f(Y)}{\Pr(Y<a)}dY\Pr(Y<a) + \int_{a}^{\infty} Y\frac{f(Y)}{\Pr(Y>a)}dY\Pr(Y>a) \\
&= \int_{-\infty}^{a} Yf(Y|Y<a)\,dY\Pr(Y<a) + \int_{a}^{\infty} Yf(Y|Y>a)\,dY\Pr(Y>a) \\
&= E(Y|Y<a)\Pr(Y<a) + E(Y|Y>a)\Pr(Y>a)\,.
\end{aligned}$$

Thus, $E(Y)$ is a weighted average of $E(Y|Y<a)$ and $E(Y|Y>a)$, 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

$$\begin{aligned}
E(Y|Y\geq a) &= \int_{a}^{\infty} Yf(Y|Y\geq a)\,dY \geq E(Y)\,, \\
E(Y|Y<a) &= \int_{-\infty}^{a} Yf(Y|Y<a)\,dY \leq E(Y)\,.
\end{aligned}$$

Further, as the truncated density function has a narrower dispersion, we have:

$$\begin{aligned}
Var(Y|Y\geq a) &= \int_{a}^{\infty} [Y - E(Y|Y\geq a)]^2 f(Y|Y\geq a)\,dY \leq Var(Y)\,, \\
Var(Y|Y<a) &= \int_{-\infty}^{a} [Y - E(Y|Y<a)]^2 f(Y|Y<a)\,dY \leq Var(Y)\,.
\end{aligned}$$

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. Truncation reduces the variance compared to the variance in the untruncated distribution.

Example 3: Find $E\left(u|u>1\right)$ and $Var\left(u|u>1\right)$ if $f\left(u\right)=\exp\left(-u\right)$, $u>0$, and compare them to their unconditional means and variances.

Solution:

$$\begin{aligned}
E\left(u \mid u>1\right) &= \int_1^\infty uf\left(u \mid u>1\right)du \\
&= \frac{1}{1-F(1)}\int_1^\infty uf\left(u\right)du \\
&= \frac{1}{1-F(1)}\int_1^\infty u\exp\left(-u\right)du \\
&= \frac{1}{1-F(1)}\left\{\left[-u\exp\left(-u\right)\right]_1^\infty+\int_1^\infty \exp\left(-u\right)du\right\} \\
&= \frac{e^{-1}}{1-F(1)}+\frac{1-F\left(1\right)}{1-F\left(1\right)}.
\end{aligned}$$

Now,

$$\begin{aligned}
1-F(1) &= \int_1^\infty \exp\left(-u\right)du \\
&= -\left[\exp\left(-u\right)\right]_1^\infty \\
&= e^{-1}.
\end{aligned}$$

Thus,

$$E\left(u \mid u>1\right)=2>E\left(u\right)=1.$$

$$\begin{aligned}
& Var\left(u \mid u>1\right) \\
= & E\left(u^2 \mid u>1\right)-\left[E\left(u \mid u>1\right)\right]^2 \\
= & \int_1^{\infty} u^2 f\left(u \mid u>1\right) du-4 \\
= & \frac{1}{1-F\left(1\right)} \int_1^{\infty} u^2 f\left(u\right) du-4 \\
= & e \int_1^{\infty} u^2 f\left(u\right) du-4 \\
= & e \int_1^{\infty} u^2 \exp\left(-u\right) du-4 \\
= & e\left[\left[-u^2 \exp\left(-u\right)\right]_1^{\infty}+2 \int_1^{\infty} u \exp\left(-u\right) du\right]-4 \\
= & e\left[e^{-1}+2 \times 2 e^{-1}\right]-4 \\
= & 1=Var\left(u\right). \blacksquare
\end{aligned}$$

9.7 Maximum Likelihood Estimation of the Truncated Model

Consider the simple model

$$Y_t=\beta_0+\beta_1 X_t+u_t>a.$$

$$\begin{aligned}
\Pr\left(Y_t>a\right) & = \Pr\left(\beta_0+\beta_1 X_t+u_t>a\right) \\
& = \Pr\left(u_t>a-\beta_0-\beta_1 X_t\right) \\
& = 1-F\left(a-\beta_0-\beta_1 X_t\right).
\end{aligned}$$

The likelihood function is

$$\begin{aligned}
L &= f\left(Y_1 = y_1, Y_2 = y_2, ..., Y_T = y_T | Y_1 > a, Y_2 > a, ..., Y_T > a\right) \\
&= f\left(y_1 - \beta_0 - \beta_1 X_1 | Y_1 > a\right) f\left(y_2 - \beta_0 - \beta_1 X_2 | Y_2 > a\right) ... f\left(y_T - \beta_0 - \beta_1 X_T | Y_T > a\right)
\end{aligned}$$

$$\begin{aligned}
\ln L &= \ln\left[f\left(y_1 - \beta_0 - \beta_1 X_1 | Y_1 > a\right) f\left(y_2 - \beta_0 - \beta_1 X_2 | Y_2 > a\right) ... f\left(y_T - \beta_0 - \beta_1 X_T | Y_T >\right.\right. \\
&= \sum_{t=1}^{T} \ln f\left(y_t - \beta_0 - \beta_1 X_t | Y_t > a\right) = \sum_{t=1}^{T} \ln \frac{f\left(y_t - \beta_0 - \beta_1 X_t\right)}{\Pr\left(Y_t > a\right)} \\
&= \sum_{t=1}^{T} \ln f\left(y_t - \beta_0 - \beta_1 X_t\right) - \sum_{t=1}^{T} \ln\left[1 - F\left(a - \beta_0 - \beta_1 X_t\right)\right].
\end{aligned}$$

First order conditions:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \beta_0} &= -\sum_{t=1}^{T} \frac{f'\left(y_t - \beta_0 - \beta_1 X_t\right)}{f\left(y_t - \beta_0 - \beta_1 X_t\right)} - \sum_{t=1}^{T} \frac{f\left(a - \beta_0 - \beta_1 X_t\right)}{1 - F\left(a - \beta_0 - \beta_1 X_t\right)} = 0, \\
\frac{\partial \ln L}{\partial \beta_1} &= -\sum_{t=1}^{T} X_t \frac{f'\left(y_t - \beta_0 - \beta_1 X_t\right)}{f\left(y_t - \beta_0 - \beta_1 X_t\right)} - \sum_{t=1}^{T} X_t \frac{f\left(a - \beta_0 - \beta_1 X_t\right)}{1 - F\left(a - \beta_0 - \beta_1 X_t\right)} = 0.
\end{aligned}$$

9.8 Censored Data

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 of these over-demand values are reported as the total number of rooms in this hotel. We may also observe people either work at a certain hour or do not work at all. If people do not work at all, their optimal working hours may be negative. But we will never observe a negative working hour, we will observe zero working hour instead. Suppose the data is lower-censored at zero.

$$\begin{aligned} Y_t^* &= \beta_0 + \beta_1 X_t + u_t, \\ Y_t &= 0 \text{ if } Y_t^* \leq 0, \\ Y_t &= Y_t^* \text{ if } Y_t{}^* > 0. \end{aligned}$$

Y_t^* is not observable, and 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 $f(Y_t - \beta_0 - \beta_1 X_t)$. If the observation is censored, we use the probability of observing a censored value $\Pr(Y_t = 0)$. Note that:

$$\begin{aligned} \Pr(Y_t = 0) &= \Pr(\beta_0 + \beta_1 X_t + u_t \leq 0) \\ &= \Pr(u_t \leq -\beta_0 - \beta_1 X_t) \\ &= 1 - F(\beta_0 + \beta_1 X_t). \end{aligned}$$

The likelihood function is

$$L = \prod_{Y_t > 0} f(Y_t - \beta_0 - \beta_1 X_t) \prod_{Y_t = 0} \Pr(Y_t = 0).$$

$$\begin{aligned} \ln L &= \ln\left[\prod_{Y_t > 0} f(Y_t - \beta_0 - \beta_1 X_t) \prod_{Y_t = 0} \Pr(Y_t = 0)\right] \\ &= \sum_{Y_t > 0} \ln f(Y_t - \beta_0 - \beta_1 X_t) + \sum_{Y_t = 0} \ln\left[1 - F(\beta_0 + \beta_1 X_t)\right]. \end{aligned}$$

First order condition:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \beta_0} &= -\sum_{Y_t>0} \frac{f'\left(Y_t-\beta_0-\beta_1 X_t\right)}{f\left(Y_t-\beta_0-\beta_1 X_t\right)} - \sum_{Y_t=0} \frac{f\left(\beta_0+\beta_1 X_t\right)}{1-F\left(\beta_0+\beta_1 X_t\right)} = 0, \\
\frac{\partial \ln L}{\partial \beta_1} &= -\sum_{Y_t>0} X_t \frac{f'\left(Y_t-\beta_0-\beta_1 X_t\right)}{f\left(Y_t-\beta_0-\beta_1 X_t\right)} - \sum_{Y_t=0} X_t \frac{f\left(\beta_0+\beta_1 X_t\right)}{1-F\left(\beta_0+\beta_1 X_t\right)} = 0.
\end{aligned}$$

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.

$$f\left(Y_t-\beta_0-\beta_1 X_t\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(Y_t-\beta_0-\beta_1 X_t\right)^2}{2\sigma^2}\right) = \frac{1}{\sigma}\phi\left(\frac{Y_t-\beta_0-\beta_1 X_t}{\sigma}\right).$$

$$f'\left(Y_t-\beta_0-\beta_1 X_t\right) = \frac{1}{\sigma^2}\phi'\left(\frac{Y_t-\beta_0-\beta_1 X_t}{\sigma}\right).$$

$$f\left(\beta_0+\beta_1 X_t\right) = \frac{1}{\sigma}\phi\left(\frac{\beta_0+\beta_1 X_t}{\sigma}\right).$$

$$F\left(\beta_0+\beta_1 X_t\right) = \Phi\left(\frac{Y_t-\beta_0-\beta_1 X_t}{\sigma}\right).$$

Then the log-likelihood can be rewritten as

$$\ln L = \sum_{Y_t>0} \ln \frac{1}{\sigma}\phi\left(\frac{Y_t-\beta_0-\beta_1 X_t}{\sigma}\right) + \sum_{Y_t=0} \ln\left[1-\Phi\left(\frac{Y_t-\beta_0-\beta_1 X_t}{\sigma}\right)\right].$$

We have the well-known **Tobit Model.**

Example 4: 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 1: Find $E\left(u|u > 1\right)$ and $Var\left(u|u > 1\right)$ if $u \sim N\left(0, 1\right)$, and compare them to their unconditional means and variances.

Exercise 2: 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 a divorce in the year of the survey, and $DIVORCE_i = 0$ if not.

$INCOME_i =$ 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 had an extramarital affair, and $AFFAIR_i = 0$ if not.

$CHILDREN_i =$ number of children of couple i.

a) Show that $E(DIVORCE_i) = \Pr(DIVORCE_i = 1)$.

b) Interpret each of the above coefficients $\beta_0, ..., \beta_4$.

c) Show that $E(u_i) = 0$ implies

$$\begin{aligned}\Pr(DIVORCE_i = 1) &= \beta_0 + \beta_1 INCOME_i + \beta_2 YEARMARRIED_i + \beta_3 AFFAIR_t \\ &\quad + \beta_4 CHILDREN_i\end{aligned}$$

d) Show that $\mathrm{Var}(u_i) = \Pr(DIVORCE_i = 1)\Pr(DIVORCE_i = 0)$.

e) Suppose the we estimate the model by OLS and obtain:

$$\begin{aligned}\widehat{DIVORCE}_i &= .5 - .0002 INCOME_i - .015 YEARMARRIED_i + .9 AFFAIR_i \\ &\quad - .03 CHILDREN_i.\end{aligned}$$

What is the chance of getting divorce for:

i) a couple married for 6 years, with 2 children, a monthly income of 1000 dollars, and no extramarital affairs.

ii) a couple married for 1 year, with no children, a monthly income of 2000 dollars, where the husband has had an extramarital affairs.

iii) a couple married for 30 years, with 3 children, a monthly income of 4000 dollars, where the wife has had an extramarital affairs.

f) State an advantage and a shortcoming of the linear probability model.

Chapter 10

Simultaneous Equation Models

10.1 Introduction

In the previous chapters, we have only discussed the estimation of a single equation. We will now discuss the method of estimating a system of equations. For example, suppose we would like to estimate a demand function of the form

$$Q_t = \alpha_0 + \alpha_1 P_t + u_t.$$

One should be careful that the data we observe $\{P_t, Q_t\}_{t=1}^{T}$ are actually the equilibrium price and quantity over time. Therefore, we are observing the intersections of the demand and supply curves. Neither the demand nor supply curve can be observed.

How can we identify the demand and supply curves? To identify the demand curve, we have to shift the supply curve. Similarly, to identify the supply curve, we have to shift the demand curve. To shift the supply curve, we can add some variables affecting supply, e.g., weather conditions, to the supply equation. For the demand curve to be shifted, we can add a factor, such as income, to the demand equation. Consider the following model:

$$\begin{aligned} Q_d &= \alpha_0 + \alpha_1 P + \alpha_2 Y + u, \\ Q_s &= \beta_0 + \beta_1 P + \beta_2 R + v, \end{aligned}$$

where Y stands for income and R denotes the amount of rainfall. The first equation is the demand equation while the second is the supply equation. The two equations above are called structural equations. The variables P and Q are called endogenous variables as they are determined within the system. Solving both equations gives us the equilibrium price and quantity. The variables Y and R are called exogenous variables which are determined outside the system. We know that at equilibriumm, $Q_d = Q_s = Q$, i.e.,

$$\alpha_0 + \alpha_1 P + \alpha_2 Y + u = \beta_0 + \beta_1 P + \beta_2 R + v,$$

$$P = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} - \frac{\alpha_2}{\alpha_1 - \beta_1} Y + \frac{\beta_2}{\alpha_1 - \beta_1} R + \frac{v - u}{\alpha_1 - \beta_1}.$$

Substituting P back to the demand-supply model, we can solve for Q

$$Q = \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} - \frac{\beta_1 \alpha_2}{\alpha_1 - \beta_1} Y + \frac{\alpha_1 \beta_2}{\alpha_1 - \beta_1} R + \frac{\alpha_1 v - \beta_1 u}{\alpha_1 - \beta_1}.$$

These two equations are known as reduced form equations, which are obtained by solving each of the endogenous variables in terms of the exogenous variables. We can also obtain the reduced form by making use of matrix algebra. Note that the structural model can be rewritten as

$$\begin{aligned} Q - \alpha_1 P &= \alpha_0 + \alpha_2 Y + u, \\ Q - \beta_1 P &= \beta_0 + \beta_2 R + v. \end{aligned}$$

In matrix notation, the above is equal to

$$\begin{pmatrix} 1 & -\alpha_1 \\ 1 & -\beta_1 \end{pmatrix} \begin{pmatrix} Q \\ P \end{pmatrix} = \begin{pmatrix} \alpha_0 & \alpha_2 & 0 \\ \beta_0 & 0 & \beta_2 \end{pmatrix} \begin{pmatrix} 1 \\ Y \\ R \end{pmatrix} + \begin{pmatrix} u \\ v \end{pmatrix}.$$

Therefore, the reduced form is:

$$\begin{pmatrix} Q \\ P \end{pmatrix} = \begin{pmatrix} 1 & -\alpha_1 \\ 1 & -\beta_1 \end{pmatrix}^{-1} \begin{pmatrix} \alpha_0 & \alpha_2 & 0 \\ \beta_0 & 0 & \beta_2 \end{pmatrix} \begin{pmatrix} 1 \\ Y \\ R \end{pmatrix} + \begin{pmatrix} 1 & -\alpha_1 \\ 1 & -\beta_1 \end{pmatrix}^{-1} \begin{pmatrix} u \\ v \end{pmatrix}.$$

Why do we need to rewrite the whole system in reduced form? Why not just estimate the structural equations directly by OLS? The problem is, if we estimate the structural equations directly, the assumption that $Cov(P, u) = 0$ will be violated in the demand equation. While in the supply equation, the assumption that $Cov(P, v) = 0$ will be violated. To see this, replace P by its reduced form, and if we assume $Cov(u, v) = 0$, then

$$\begin{aligned} Cov\,(P, u) &= Cov\left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} - \frac{\alpha_2}{\alpha_1 - \beta_1}Y + \frac{\beta_2}{\alpha_1 - \beta_1}R + \frac{v - u}{\alpha_1 - \beta_1}, u\right) \\ &= -\frac{Cov\,(u, u)}{\alpha_1 - \beta_1} = -\frac{\sigma_u^2}{\alpha_1 - \beta_1} \neq 0. \end{aligned}$$

and

$$\begin{aligned} Cov\,(P, v) &= Cov\left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} - \frac{\alpha_2}{\alpha_1 - \beta_1}Y + \frac{\beta_2}{\alpha_1 - \beta_1}R + \frac{v - u}{\alpha_1 - \beta_1}, v\right) \\ &= \frac{Cov\,(v, v)}{\alpha_1 - \beta_1} = \frac{\sigma_v^2}{\alpha_1 - \beta_1} \neq 0. \end{aligned}$$

This is similar to the violation of the assumption $Cov(X, u)$ in the regression model, which will cause inconsistent estimates. To show this, suppose we estimate the model by OLS

$$Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 Y_t + u_t.$$

The OLS estimator for α_1 is

$$\begin{aligned}
\widehat{\alpha}_1 &= \frac{S_{QP}S_{YY} - S_{QY}S_{PY}}{S_{PP}S_{YY} - S_{PY}^2} = \alpha_1 + \frac{S_{Pu}S_{YY} - S_{Yu}S_{PY}}{S_{PP}S_{YY} - S_{PY}^2} \\
&= \alpha_1 + \frac{\frac{S_{Pu}}{T}\frac{S_{YY}}{T} - \frac{S_{Yu}}{T}\frac{S_{PY}}{T}}{\frac{S_{PP}}{T}\frac{S_{YY}}{T}\left(1 - r_{PY}^2\right)} \xrightarrow{p} \alpha_1 + \frac{p\lim \frac{S_{Pu}}{T}}{p\lim \frac{S_{PP}}{T}\left(1 - r_{PY}^2\right)}.
\end{aligned}$$

where

$$\begin{aligned}
S_{QP} &= \sum_{t=1}^{T}\left(Q_t - \overline{Q}\right)\left(P_t - \overline{P}\right), \\
S_{YY} &= \sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2
\end{aligned}$$

and so on.

Note that

$$\begin{aligned}
p\lim \frac{S_{Pu}}{T} &= p\lim \frac{1}{T}\sum_{t=1}^{T}\left(P_t - \overline{P}\right)\left(u_t - \overline{u}\right) \\
&= Cov\left(P, u\right) \\
&= -\frac{\sigma_u^2}{\alpha_1 - \beta_1} \\
&\neq 0.
\end{aligned}$$

Thus $\widehat{\alpha}_1$ does not converge to α_1. Similarly, all the estimates of α's and β's will be inconsistent. Therefore, the OLS estimator is biased and inconsistent if we ignore simultaneity. Thus, estimating the structural equation directly will give us inconsistent estimates. The problem can be avoided if write the model in reduced form. Let the reduced form be:

$$\begin{aligned} P &= \pi_1 + \pi_2 Y + \pi_3 R + error, \\ Q &= \pi_4 + \pi_5 Y + \pi_6 R + error. \end{aligned}$$

However, the reduced form equations will not give us the parameters of interest $(\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2)$ directly. We have to recover them by letting

$$\begin{aligned} \widehat{\alpha}_1 &= \frac{\widehat{\pi}_6}{\widehat{\pi}_3}, \\ \widehat{\beta}_1 &= \frac{\widehat{\pi}_5}{\widehat{\pi}_2}, \\ \widehat{\beta}_2 &= \widehat{\pi}_3 \left(\widehat{\alpha}_1 - \widehat{\beta}_1 \right), \\ \widehat{\alpha}_2 &= -\widehat{\pi}_2 \left(\widehat{\alpha}_1 - \widehat{\beta}_1 \right), \\ \widehat{\alpha}_0 &= \widehat{\pi}_4 - \widehat{\alpha}_1 \widehat{\pi}_1, \\ \widehat{\beta}_0 &= \widehat{\pi}_4 - \widehat{\beta}_1 \widehat{\pi}_1. \end{aligned}$$

This is called the **indirect least-squares method (ILS)**.

The parameters of interest are all identified in the above case, and we call this the exact identification. The demand curve is identified because by varying R, we are able to shift the supply curve and trace out the demand curve. By analogy, changing the values of Y allows us to shift the demand curve and trace out the supply curve. Sometimes it is not possible to identify all the structural parameters, this problem is known as **under-identification**. While sometimes we may obtain more than one set of solution for the structural parameters, this situation is known as **over-identification**.

10.2 Under-identification

Suppose the demand-supply model is

$$\begin{aligned} Q_d &= \alpha_0 + \alpha_1 P + \alpha_2 Y + u, \\ Q_s &= \beta_0 + \beta_1 P + v. \end{aligned}$$

The reduced form is

$$\begin{aligned} P &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} - \frac{\alpha_2}{\alpha_1 - \beta_1} Y + \frac{v - u}{\alpha_1 - \beta_1}, \\ Q &= \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} - \frac{\beta_1 \alpha_2}{\alpha_1 - \beta_1} Y + \frac{\alpha_1 v - \beta_1 u}{\alpha_1 - \beta_1}, \end{aligned}$$

or we can write it as

$$\begin{aligned} P &= \pi_1 + \pi_2 Y + error, \\ Q &= \pi_3 + \pi_4 Y + error. \end{aligned}$$

The estimable structural parameters are $\widehat{\beta}_1 = \frac{\widehat{\pi}_4}{\widehat{\pi}_2}$, and $\widehat{\beta}_0 = \widehat{\pi}_3 - \widehat{\beta}_1 \widehat{\pi}_1$. However, there is no way to identify $\alpha's$, i.e. we cannot identify the demand equation. This is because there is no factor to shift the supply curve, and therefore the demand curve is not identifiable. Analogously, if our system is

$$\begin{aligned} Q_d &= \alpha_0 + \alpha_1 P + u, \\ Q_s &= \beta_0 + \beta_1 P + \beta_2 R + v, \end{aligned}$$

then the demand curve is identifiable while the supply curve is not.

10.3 Over-identification

Sometimes we may end up with more than one set of solution for the parameters of interest. Consider the following model:

$$\begin{aligned} Q_d &= \alpha_0 + \alpha_1 P + \alpha_2 Y + u, \\ Q_s &= \beta_0 + \beta_1 P + \beta_2 R + \beta_3 F + v, \end{aligned}$$

where F is another exogenous variable. The reduced form is

$$P = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} - \frac{\alpha_2}{\alpha_1 - \beta_1} Y + \frac{\beta_2}{\alpha_1 - \beta_1} R + \frac{\beta_3}{\alpha_1 - \beta_1} F + \frac{v - u}{\alpha_1 - \beta_1},$$

$$Q = \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} - \frac{\beta_1 \alpha_2}{\alpha_1 - \beta_1} Y + \frac{\alpha_1 \beta_2}{\alpha_1 - \beta_1} R + \frac{\alpha_1 \beta_3}{\alpha_1 - \beta_1} F + \frac{\alpha_1 v - \beta_1 u}{\alpha_1 - \beta_1},$$

or

$$\begin{aligned} P &= \pi_1 + \pi_2 Y + \pi_3 R + \pi_4 F + error, \\ Q &= \pi_5 + \pi_6 Y + \pi_7 R + \pi_8 F + error. \end{aligned}$$

Thus, we can estimate α_1 by $\frac{\widehat{\pi}_7}{\widehat{\pi}_3}$ or $\frac{\widehat{\pi}_8}{\widehat{\pi}_4}$, which are two different values in general. For each possible estimate of α_1, we can obtain the estimates for the rest of the structural parameters. Thus, there is more than one way to recover the structural equations, leaving us with the problem of over-identification.

10.4 Over-identification and under-identification at the same time

Suppose our model is

$$\begin{aligned} Q_d &= \alpha_0 + \alpha_1 P + \alpha_2 Y + \alpha_3 R + u, \\ Q_s &= \beta_0 + \beta_1 P + v. \end{aligned}$$

It is not difficult to show that the supply function is over-identified and the demand function is under-identified.

10.5 The Order Condition

The necessary condition for the model to be identified is called the order condition.

Let G be the number of structural equations, and let K be the number of variables excluded from an equation. The order condition for an equation to be identified is $K \geq G - 1$. e.g.,

$$\begin{aligned} Q_d &= \alpha_0 + \alpha_1 P + \alpha_2 Y + \alpha_3 R + u, \\ Q_s &= \beta_0 + \beta_1 P + v. \end{aligned}$$

Then $G = 2$ and $G - 1 = 1$. For the demand equation, the number of variables excluded is $K = 0$, so the order condition is not satisfied and the demand equation is under-identified. For the supply equation, the number of excluded variables is $K = 2$ (Y and R), so the supply equation is identified. The order condition is not sufficient because if this condition is not satisfied, the model is under-identified. Even if the order condition is satisfied, we may still be unable to identify the equation.

10.6 The Rank Condition

The rank condition is a necessary and sufficient condition for identification. To understand the rank condition, one has to be familiar with matrix algebra. The rank of a matrix is the number of linearly independent rows of the matrix.

Example 1: The rank of

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 2 & 7 & 5 \end{pmatrix}$$

is 2, since row two is 2 times row 1. The rank of

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 1 \\ 2 & 7 & 5 \end{pmatrix}$$

is 3.

We will only provide a simple illustration of the rank condition here. Consider the following model with three endogenous variables Y_1, Y_2, and Y_3, and three exogenous variables X_1, X_2, and X_3.

	Y_1	Y_2	Y_3	X_1	X_2	X_3
Equation 1	γ_{11}	0	γ_{13}	γ_{14}	0	γ_{16}
Equation 2	γ_{21}	0	0	γ_{24}	0	γ_{26}
Equation 3	0	γ_{32}	γ_{33}	γ_{34}	γ_{35}	0

To determine whether the rank condition for identifying equation i ($i = 1, 2, 3$) is satisfied, delete row i and pick up the columns corresponding to the elements that have zeros in that row. If we can form a matrix of rank $G - 1$

from these columns, then the equation is identified, otherwise not. We first check the order condition, in the above case, $G = 3$, so $G - 1 = 2$.

For equation 1, $K = 2 = G - 1$, so the order condition is satisfied and this eqaution is exactly identified.

For equation 2, $K = 3 > G - 1$, so the order condition is satisfied and this eqaution is over-identified.

For equation 3, $K = 2 = G - 1$, so the order condition is satisfied and this eqaution is exactly identified.

Thus the order conditions are satisfied for all equations. Now consider the rank condition.

For equation 1, the resulting matrix is

$$\begin{pmatrix} 0 & 0 \\ \gamma_{32} & \gamma_{35} \end{pmatrix}.$$

The rank of this matrix is 1 because the first row has both element zero. Thus, the rank condition is not satisfied and therefore the first equation is not identified.

For equation 2, the resulting matrix is

$$\begin{pmatrix} 0 & \gamma_{13} & 0 \\ \gamma_{32} & \gamma_{33} & \gamma_{35} \end{pmatrix}.$$

The rank of this matrix is 2. Thus, the rank condition is satisfied and equation 2 is identified.

For equation 3, the resulting matrix is

$$\begin{pmatrix} \gamma_{11} & \gamma_{16} \\ \gamma_{21} & \gamma_{26} \end{pmatrix}.$$

The rank of this matrix is 2 provided that $\frac{\gamma_{11}}{\gamma_{21}} \neq \frac{\gamma_{16}}{\gamma_{26}}$. Thus, the rank condition is satisfied and equation 3 is identified.

Note that the failure of the order condition implies the failure of the rank condition, but the converse is not true. If the rank condition is satisfied, then the order condition should be satisfied.

10.7 Two-Stage Least Squares (2SLS)

For the indirect least squares method discussed in the previous section, we have to estimate the reduced form and recover the structural parameters from the reduced-form estimates. We now present the two-stage least squares method, a method which enables us to obtain consistent estimates of the structural parameters. In practice, ILS is not a widely used technique since it is rare for an equation to be exactly identified. 2SLS is perhaps the most important and widely used procedure. It is applicable to equations which are over-identified or exactly identified. Moreover, when the model is exactly identified, the ILS and the 2SLS will give the same estimates.

The 2SLS method is an instrumental variable method. It is used in situations where the explanatory variable is not easily observed or when the assumption $Cov(X, u) = 0$ is violated.

In the demand-supply system, $Cov(P, u) \neq 0$ in the structural equation. The idea of instrumental variable method is to find a variable Z as a proxy for P so that $Cov(Z, P) \neq 0$ and $Cov(Z, u) = 0$. The 2SLS uses $Z = \widehat{P}$, where $\widehat{P}$ is the predicted values of P obtained from the reduced form. We use $\widehat{P}$ as an instrument because $\widehat{P}$ and P are correlated, and $\widehat{P}$ and the errors are orthogonal(uncorrelated).

Suppose the structural equations are

$$\begin{aligned} Q_d &= \alpha_0 + \alpha_1 P + \alpha_2 Y + u, \\ Q_s &= \beta_0 + \beta_1 P + \beta_2 R + v. \end{aligned}$$

We first estimate the reduced form

$$\widehat{P} = \widehat{\pi}_1 + \widehat{\pi}_2 Y + \widehat{\pi}_3 R.$$

We then replace P by $\widehat{P}$ in the structural equations and estimate

$$\begin{aligned} Q_d &= \alpha_0 + \alpha_1 \widehat{P} + \alpha_2 Y + u_*, \\ Q_s &= \beta_0 + \beta_1 \widehat{P} + \beta_2 R + v_*. \end{aligned}$$

The 2SLS method gives us consistent estimates of the structural parameters $\alpha' s$ and $\beta' s$.

Example 2: Consider the following two-equation model in which the y's are endogenous and the x's are exogenous:

$$\begin{aligned} y_{1t} &= \alpha_1 y_{2t} + \alpha_2 x_{1t} + u_t, \\ y_{2t} &= \beta_1 y_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + v_t. \end{aligned}$$

Suppose that $u_t \sim iid\,(0, \sigma_u^2)$, $v_t \sim iid\,(0, \sigma_v^2)$, $\mathrm{Cov}(u_s, v_t) = 0$ for all s, t. u and v are uncorrelated with x_1, x_2 and x_3.

a) Explicitly derive the reduced form equations for y_{1t} and y_{2t}.

b) Find $Cov\,(y_{2t}, u_t)$ and $Cov\,(y_{1t}, v_t)$. What is the problem with estimating the above structural equations directly by OLS?

c) Check if the order condition is satisfied for each of the structural equations.

d) Describe the Indirect Least Squares estimation procedure for *each* of the structural parameters $\alpha's$ and $\beta's$.

e)Briefly describe the Two-Stage Least Squares estimation procedure in this example.

f)Can we apply OLS directly to the structural equations if we know that $\beta_1 = 0$? Why or why not?

Solution:

(a) The reduced form of y_{1t} is given by

$$\begin{aligned} y_{1t} &= \alpha_1 \left(\beta_1 y_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + v_t\right) + \alpha_2 x_{1t} + u_t \\ &= \frac{\alpha_2}{1-\alpha_1\beta_1} x_{1t} + \frac{\alpha_1\beta_2}{1-\alpha_1\beta_1} x_{2t} + \frac{\alpha_1\beta_3}{1-\alpha_1\beta_1} x_{3t} + \frac{\alpha_1 v_t + u_t}{1-\alpha_1\beta_1} \\ &= \Pi_1 x_{1t} + \Pi_2 x_{2t} + \Pi_3 x_{3t} + w_{1t}. \end{aligned}$$

where $\Pi_1 = \dfrac{\alpha_2}{1-\alpha_1\beta_1}$, $\Pi_2 = \dfrac{\alpha_1\beta_2}{1-\alpha_1\beta_1}$, $\Pi_3 = \dfrac{\alpha_1\beta_3}{1-\alpha_1\beta_1}$ and $w_{1t} = \dfrac{\alpha_1 v_t + u_t}{1-\alpha_1\beta_1}$. The reduced form of y_{2t} is given by

$$\begin{aligned} y_{2t} &= \beta_1 \left(\alpha_1 y_{2t} + \alpha_2 x_{1t} + u_t\right) + \beta_2 x_{2t} + \beta_3 x_{3t} + v_t \\ &= \frac{\alpha_2\beta_1}{1-\alpha_1\beta_1} x_{1t} + \frac{\beta_2}{1-\alpha_1\beta_1} x_{2t} + \frac{\beta_3}{1-\alpha_1\beta_1} x_{3t} + \frac{v_t + \beta_1 u_t}{1-\alpha_1\beta_1} \\ &= \Pi_4 x_{1t} + \Pi_5 x_{2t} + \Pi_6 x_{3t} + w_{2t}. \end{aligned}$$

where $\Pi_4 = \dfrac{\alpha_2\beta_1}{1-\alpha_1\beta_1}$, $\Pi_5 = \dfrac{\beta_2}{1-\alpha_1\beta_1}$, $\Pi_6 = \dfrac{\beta_3}{1-\alpha_1\beta_1}$ and $w_{2t} = \dfrac{v_t + \beta_1 u_t}{1-\alpha_1\beta_1}$.

(b) $Cov\,(y_{1t}, v_t) = \dfrac{\alpha_1}{1-\alpha_1\beta_1}\sigma_v^2 \neq 0$ and $Cov\,(y_{2t}, u_t) = \dfrac{\beta_1}{1-\alpha_1\beta_1}\sigma_u^2 \neq 0$. Thus, the OLS estimates are inconsistent

(c) In this model, $G-1=1$. The order conditions are:

For equation 1, $K=2>G-1$, it is over-identified.

For equation 2, $K=1=G-1$, it is exactly identified.

(d)

Step 1 : Regress y_{1t} on x_{1t}, x_{2t} and x_{3t} to obtain estimates $\widehat{\Pi}_1, \widehat{\Pi}_2$ and $\widehat{\Pi}_3$.

Step 2 : Regress y_{2t} on x_{1t}, x_{2t} and x_{3t} to obtain estimates $\widehat{\Pi}_4, \widehat{\Pi}_5$ and $\widehat{\Pi}_6$.

Step 3 : Solve the relationships among the $\widehat{\Pi}$'s, $\widehat{\alpha}$'s and $\widehat{\beta}$'s. Then, we can obtain the estimates $\widehat{\alpha}$'s and $\widehat{\beta}$'s.

(e)

Step 1 : Regress y_{1t} on x_{1t}, x_{2t} and x_{3t} to obtain $\widehat{y}_{1t}$.

Step 2 : Regress y_{2t} on x_{1t}, x_{2t} and x_{3t} to obtain $\widehat{y}_{2t}$.

Step 3 : Regress y_{1t} on $\widehat{y}_{2t}$, x_{1t}, x_{2t} and x_{3t} to obtain estimates $\widehat{\alpha}_0$, $\widehat{\alpha}_1$, $\widehat{\alpha}_2$ and $\widehat{\alpha}_3$.

Step 4 : Regress y_{2t} on $\widehat{y}_{2t}$, x_{1t}, x_{2t} and x_{3t} to obtain estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$.

(f) The first equation can be estimated by OLS. Since $\beta_1 = 0$, we can obtain consistent OLS estimates $\widehat{\alpha}$'s. The second equation can also be estimated by OLS directly because endogenous variable y_{2t} is absent from the second equation.

Exercise 1: The structural form of a two-equation model is as follows (the t-subscript is omitted) :

$$\begin{aligned} P &= \alpha_1 + \alpha_2 N + \alpha_3 S + \alpha_4 A + u, \\ N &= \beta_1 + \beta_2 P + \beta_3 M + v, \end{aligned}$$

where P and N are endogenous and S, A and M are exogenous.

a. For each equation, examine whether it is underidentified, overidentified, or exactly identified.

b. What explanatory variables, if any, are correlated with u ? What explanatory variables, if any, are correlated with v ?

c. What happens if OLS is used to estimate the α's and the β's ?

d. Can the α's be estimated by ILS ? If yes, derive the estimates. Answer the same question about the β's.

e. Explain step by step how the 2SLS method can be applied on the second equation.

Exercise 2: Consider the following simple three-equation model :

$$\begin{array}{ll} A = X - M & \text{Endogenous : } A, M, X \\ M = \alpha_1 + \alpha_2 Y + \alpha_3 P + \alpha_4 U + u & \text{Exogenous : } Y, P, U \\ X = \beta_1 + \beta_2 P + \beta_3 A + v & \text{Error terms : } u, v \end{array}$$

a. Check whether the order condition is satisfied for the second and third equations. What is your conclusion?

b. Derive the reduced form equations.

c. How would you use the TSLS estimation procedure on the third equation?

d. Explain why we can use the OLS method on the second equation. What properties do the estimates have?

e. Suppose we had use OLS to estimate the third equation. What properties will those estimates have?

Exercise 3: The failure of the rank condition implies the failure of the order condition. True or False? Explain.

Exercise 4: The following are models in three equations with three endogenous variables Y and three exogenous variables X:

a)

Equation no.	Y_1	Y_2	Y_3	X_1	X_2	X_3
1	γ_{11}	0	γ_{13}	γ_{14}	0	γ_{16}
2	γ_{21}	γ_{22}	0	γ_{24}	γ_{25}	0
3	γ_{31}	γ_{32}	γ_{33}	γ_{34}	γ_{35}	0

b)

Equation no.	Y_1	Y_2	Y_3	X_1	X_2	X_3
1	γ_{11}	γ_{12}	0	γ_{14}	0	γ_{16}
2	0	γ_{22}	0	γ_{24}	γ_{25}	γ_{26}
3	0	γ_{32}	γ_{33}	γ_{34}	0	γ_{36}

Determine the identifiability of each equation for each model with the aid of the order and rank conditions of identification.

Exercise 5: Consider the following two-equation model in which the y's are endogenous and the x's are exogenous:

$$\begin{aligned} y_{1t} &= \alpha_0 y_{2t} + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \alpha_3 x_{3t} + u_t \\ y_{2t} &= \beta_0 y_{1t} + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + v_t \end{aligned}$$

Suppose that $u_t \sim iid\,(0, \sigma_u^2)$, $v_t \sim iid\,(0, \sigma_v^2)$, $Cov\,(u_s, v_t) = 0$ for all s and t, u and v are uncorrelated with x_1, x_2 and x_3.

a) Find $Cov\,(y_{2t}, u_t)$ and $Cov\,(y_{1t}, v_t)$. What is the problem of estimating the above structural equations directly by OLS?

b) Verify that neither equation is identified.

c) Explicitly derive the reduced form equations for y_{1t} and y_{2t}.

d) Establish whether or not the following restrictions are sufficient to identify (or partially identify) the model:

i) $\alpha_2 = \beta_3 = 0$,
ii) $\beta_1 = \beta_2 = 0$,
iii) $\alpha_0 = 0$,
iv) $\alpha_0 = \beta_0$, and $\beta_3 = 0$,
v) $\alpha_2 + \beta_2 = 1$,
vi) $\alpha_3 = 0$,
vii) $\alpha_2 = \alpha_3 = \beta_1 = 0$,
viii) $\alpha_2 = \alpha_3 = \beta_2 = \beta_3 = 0$,
ix) $\alpha_1 = \alpha_2 = \alpha_3 = \beta_2 = \beta_3 = 0$.

e) Suppose $\alpha_3 = \beta_3 = 0$. The model becomes

$$\begin{aligned} y_{1t} &= \alpha_0 y_{2t} + \alpha_1 x_{1t} + \alpha_2 x_{2t} + u_t \\ y_{2t} &= \beta_0 y_{1t} + \beta_1 x_{1t} + \beta_2 x_{2t} + v_t \end{aligned}$$

Find $\text{cov}(y_{2t}, u_t)$ and $\text{cov}(y_{1t}, v_t)$. Explicitly derive the reduced form equations for y_{1t} and y_{2t}.

f) Establish whether or not the following restrictions are sufficient to identify (or partially identify) the model in part e).

i) $\beta_1 = \beta_2 = 0$,

ii) $\alpha_0 = 0$,

iii) $\alpha_2 + \beta_2 = 1$,

iv) $\alpha_2 = \beta_2 = 0$,

v) $\alpha_1 = \alpha_2 = \beta_2 = 0$.

Chapter 11

Large Sample Theory

11.1 Introduction

Recall that an estimator is constructed by the observations Y and X, and Y has a random component u. That means the estimator is a combination of all the error residuals, once we have assumed a certain distributional properties on u, we may be able to find out the distribution for the estimator. For example, consider the simple regression model

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum\limits_{t=1}^{T} \left(X_t - \overline{X}\right) u_t}{\sum\limits_{t=1}^{T} \left(X_t - \overline{X}\right)^2}$$

If we assume $u_t \sim N(0, \sigma^2)$, then $\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum\limits_{t=1}^{T} \left(X_t - \overline{X}\right)^2}\right)$, or

$\dfrac{\widehat{\beta}_1 - \beta_1}{\sqrt{Var\left(\widehat{\beta}_1\right)}} \sim N(0,1)$. However, different assumptions of u and X affect the asymptotic behavior of the estimators. To understand the asymptotics of the estimators, we first have to understand some basic large sample theory.

Definition 1: Let Ω be a sample space, and E be events in Ω. The collection $\Im$ of subsets of Ω is called $\sigma-$**algebra** if it satisfies the following properties:

(i) $\Omega \in \Im$

(ii) $E \in \Im \Rightarrow E^c \in \Im$

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

(iii) $E_i, E_j \in \Im \Rightarrow E_i \cup E_j \in \Im$ for all i, j.

(iv) $E_j \in \Im$, $j = 1, 2, ... \Rightarrow \cup_{j=1}^{\infty} E_j \in \Im$.

Definition 2: 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 $P(\cup_{j=1} E_j) = \sum_j P(E_j)$.

Definition 3: 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 4: 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, $\{\omega | X(\omega) < x\} \in \Im$.

Limits of Sequences

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

$$\lim_{T \to \infty} c_n = c$$

or equivalently,

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

e.g.

a) If $c_n = \frac{1}{n}$, $\lim\limits_{n \to \infty} c_n = 0$

b) If $c_n = \left(1 + \frac{a}{n}\right)^n$, $\lim\limits_{n \to \infty} c_n = \exp(a)$

c) If $c_n = n^2$, $\lim\limits_{n \to \infty} c_n = \infty$

b) If $c_n = (-1)^n$, Then no limit exists.

A sequence of deterministic matrices $\mathbf{C}_n$ converges to $\mathbf{C}$ if each element of $\mathbf{C}_n$ converges to the corresponding element of $\mathbf{C}$.

Definition 6: The **supremum** of the sequence, denoted by

$$\sup_{n\to\infty} c_n$$

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

Definition 7: The **infimum** of the sequence, denoted by

$$\inf_{n\to\infty} c_n$$

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

Definition 8: The sequence $\{c_n\}_{n=1}^{\infty}$ is said to be a **monotone non-increasing** sequence if

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

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

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

Definition 9: Let $\{c_n\}_{n=1}^{\infty}$ be a sequence of real numbers and put

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

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

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

Thus we write

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

We immediately have

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

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

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

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

e.g. If $c_n = (-1)^n$, then $\limsup c_n = 1$, $\liminf c_n = -1$, so the limit does not exist.

e.g. If $c_n = \left(-\frac{1}{n}\right)^n$, then $\limsup c_n = 0$, $\liminf c_n = 0$, so the limit exists and is equal to zero.

Definition 10: The sequence $\{b_n\}$ is **at most of order $\mathbf{n}^\lambda$**, 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 $n \geq N$, $\left|\frac{b_n}{n^\lambda}\right| < \Delta$.

Definition 11: The sequence $\{b_n\}$ is **of order smaller than $\mathbf{n}^\lambda$**, denoted $o\left(n^\lambda\right)$, if and only if for *every* real number $\delta > 0$, $0 < \delta < \infty$, there exists a finite integer $N(\delta)$ such that for all $n \geq N(\delta)$, $\left|\frac{b_n}{n^\lambda}\right| < \delta$.

In other words, $\{b_n\}$ is $O\left(n^\lambda\right)$ if $\frac{b_n}{n^\lambda}$ is eventually bounded, where as $\{b_n\}$ is $o\left(n^\lambda\right)$ if $\frac{b_n}{n^\lambda} \to 0$.

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 $b_n \to 0$.

e.g.

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

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

(iii)Let $b_n = \exp(-n)$. Then $\{b_n\}$ is $o\left(n^{-\delta}\right)$ for every $\delta > 0$ and also $O\left(n^{-\delta}\right)$.

Proposition 1:

(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 12: The sequence $\{b_n(\omega)\}$ is **at most of order $\mathbf{n}^\lambda$ in probability**, denoted $O_p(n^\lambda)$, if there exists an $O(1)$ nonstochastic sequence a_n such that $\frac{b_n(\omega)}{n^\lambda} - a_n \xrightarrow{p} 0$.

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

e.g. If $u_t \sim iid(0, \sigma^2)$ with $\sigma^2 < \infty$, then u_t is $O_p(1)$.

e.g. Note that

$$\widehat{\beta}_1 - \beta_1 = \frac{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2} = \frac{1}{\sqrt{T}}\frac{\frac{1}{\sqrt{T}}\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)u_t}{\frac{1}{T}\sum\limits_{t=1}^{T}\left(X_t - \overline{X}\right)^2}$$

Since

$$\begin{aligned} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(X_t - \overline{X}\right)u_t &= O_p(1) \\ \frac{1}{T}\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^2 &= O(1) \end{aligned}$$

therefore

$$\widehat{\beta}_1 - \beta_1 = O_p\left(T^{-1/2}\right) = o_p(1)$$

which means $\widehat{\beta}_1 - \beta_1 \xrightarrow{p} 0$ or equivalently $\widehat{\beta}_1 \xrightarrow{p} \beta_1$, i.e. $\widehat{\beta}_1$ is a consistent estimator for β_1.

We can also write $\sqrt{T}\left(\widehat{\beta}_1 - \beta_1\right) = O_p(1)$, as $\sqrt{T}\left(\widehat{\beta}_1 - \beta_1\right)$ converges in distribution to a normal distribution, mathematically, we write

$$\sqrt{T}\left(\widehat{\beta}_1-\beta_1\right)\xrightarrow{d} N\left(0,\frac{\sigma^2}{\lim\limits_{T\to\infty}\frac{1}{T}\sum\limits_{t=1}^{T}\left(X_t-\overline{X}\right)^2}\right)$$

Definition 13: The sequence $\{b_n(\omega)\}$ is **of order smaller than n$^\lambda$ in probability**, denoted $o_p\left(n^\lambda\right)$, if $\frac{b_n(\omega)}{n^\lambda}\xrightarrow{p}0.$

Proposition 2: Let a_n and b_n be random scalars.

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

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

Definition 14: Let $b_n(\omega)$ be a sequence of real-valued random variables. If there exists a real number b such that for every $\epsilon>0, P\left(\left|b_n(\omega)-b\right|>\epsilon\right)\to 0$ as $n\to\infty$, then $b_n(\omega)$ **converges in probability** to b, written $b_n(\omega)\xrightarrow{p}b.$

Definition 15: Let $b_n(\omega)$ be a sequence of real-valued random variables. If there exists a real number b such that $E\left(\left|b_n(\omega)-b\right|^r\right)\to 0$ as $n\to\infty$ for some $r>0$, then $b_n(\omega)$ **converges in the rth mean** to b, written $b_n(\omega)\overset{r.m.}{\to}b.$

The most commonly encountered occurrence is that in which $r=2$, in which case convergence is said to occur in quadratic mean, denoted $b_n(\omega)\overset{q.m.}{\to}$

$b.$

A useful property of convergence in the rth mean is that it implies convergence in the sth mean for $s < r$. To prove this, we first introduce the Jensen's inequality.

Proposition 3: (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 $g(E(Z)) \leq E(g(Z))$. If g is concave on B, then $g(E(Z)) \geq E(g(Z))$.

e.g. 1: Let $g(z) = |z|$. It follows from the Jensen's inequality that $|E(Z)| \leq E|Z|$.

e.g. 2: Let $g(z) = z^2$. It follows from the Jensen's inequality that $E^2(Z) \leq E(Z^2)$.

Theorem 1: If $b_n(\omega) \overset{r.m.}{\to} b$ and $r > s$, then $b_n(\omega) \overset{s.m.}{\to} b$.

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

$E(|b_n(\omega) - b|^s) = E\left(\{|b_n(\omega) - b|^r\}^{s/r}\right) \leq \{E(|b_n(\omega) - b|^r)\}^{s/r}$

Since $E(|b_n(\omega) - b|^r) \to 0$, it follows that $E(|b_n(\omega) - b|^s) \to 0$, $b_n(\omega) \overset{s.m.}{\to} b$.

Q.E.D.

Convergence in the rth 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 2: (Chebyshev's Inequality)

If Z is a r.v. with finite variance σ^2, then

$$\begin{aligned} P\left(|Z-\mu|<k\sigma\right) &\geq 1-\frac{1}{k^2} \\ P\left(|Z-\mu|\geq k\sigma\right) &\leq \frac{1}{k^2} \end{aligned}$$

Proof: (for continuous r.v.)

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty}(z-\mu)^2 f(z)\,dz \\ &\geq \int_{-\infty}^{\mu-k\sigma}(z-\mu)^2 f(z)\,dz+\int_{\mu+k\sigma}^{\infty}(z-\mu)^2 f(z)\,dz \\ &\geq \int_{-\infty}^{\mu-k\sigma}k^2\sigma^2 f(z)\,dz+\int_{\mu+k\sigma}^{\infty}k^2\sigma^2 f(z)\,dz \\ &= k^2\sigma^2 P\left(Z\leq\mu-k\sigma\right)+k^2\sigma^2 P\left(Z\geq\mu+k\sigma\right) \\ &= k^2\sigma^2 P\left(|Z-\mu|\geq k\sigma\right) \end{aligned}$$

this implies

$$P\left(|Z-\mu|\geq k\sigma\right)\leq\frac{1}{k^2}$$

Q.E.D.

Proposition 4: (Generalized Chebyshev Inequality)

Let Z be a random variable such that $E\left|Z\right|^r<\infty$, $r>0$. Then for every $\epsilon>0$,

$$P\left(|Z|\geq\epsilon\right)\leq\frac{E\left|Z\right|^r}{\epsilon^r}$$

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

Theorem 3: If $b_n(\omega) \overset{r.m.}{\rightarrow} b$ and $r > s$, then $b_n(\omega) \xrightarrow{p} b$.

Proof: Let $Z = b_n(\omega) - b$, and apply the Generalized Chebyshev inequality, for every $\epsilon > 0$,

$$P(|b_n(\omega) - b| \geq \epsilon) \leq \frac{E|b_n(\omega) - b|^r}{\epsilon^r}$$

Since $b_n(\omega) \overset{r.m.}{\rightarrow} b$, we have $E|b_n(\omega) - b|^r \rightarrow 0$, and as a result $P(|b_n(\omega) - b| \geq \epsilon) \rightarrow 0$. Thus, we have $b_n(\omega) \xrightarrow{p} b$.

Chapter 12

Stationarity

Definition 1 *A process Y_t is said to be* ***covariance-stationary*** *or* ***weakly stationary*** *if*

$$\begin{aligned} E\left(Y_t\right) &= \mu < \infty \qquad \text{for all } t \\ E\left(Y_t^2\right) &< \infty \qquad \text{for all } t \\ E\left(Y_t - \mu\right)\left(Y_{t-j} - \mu\right) &= \gamma_{|j|} < \infty \qquad \text{for all } t \text{ and any } j = \pm 1, \pm 2, ... \end{aligned}$$

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 2 *A process is said to be* ***strictly stationary*** *if the joint distribution of*

$\left(Y_t,\ Y_{t+j_1,},\ Y_{t+j_2,}...,\ Y_{t+j_n,}\right)$ depends only on $\left(j_1,\ j_2, ...,\ j_n\right)$, for all t, and any j_1, j_2, ..., j_n, n. i.e. The joint density

$$f\left(Y_t,\ Y_{t+j_1,},\ Y_{t+j_2,}...,\ Y_{t+j_n,}\right) = f\left(Y_s,\ Y_{s+j_1,},\ Y_{s+j_2,}...,\ Y_{s+j_n,}\right)$$

for any s and t.

Weakly stationary and strictly stationary 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, so it is not weakly stationary. But as long as its distribution does not change over time, it is strongly stationary. It is also possible to imagine a process that is covariance-stationary but not strictly stationary, e.g., the mean and covariance are not functions of time, but perhaps higher moment such as $E\left(Y_t^4\right)$ and $E\left(Y_t^5\right)$ are.

If a process is strictly stationary *with finite second moments*, then it must be covariance-stationary.

12.1 AR(1) Process

Consider an autoregressive process of order 1

$$\begin{aligned} Y_t &= \beta Y_{t-1} + u_t \\ Y_0 &= 0 \\ u_t &\sim iid\left(0, \sigma^2\right) \end{aligned}$$

We are interested in finding the mean and variance of the process Y_t. Note that by repeating substitution, we can show that

$$\begin{aligned} &Y_t = \beta\left(\beta Y_{t-2} + u_{t-1}\right) + u_t = \beta^2 Y_{t-2} + \beta u_{t-1} + u_t \\ &= \beta^2\left(\beta Y_{t-3} + u_{t-2}\right) + \beta u_{t-1} + u_t \\ &= ..... \\ &= u_t + \beta u_{t-1} + \beta^2 u_{t-2} + \beta^3 u_{t-3} + ... + \beta^{t-1} u_1 \end{aligned}$$

$$= \sum_{k=0}^{t-1} \beta^k u_{t-k}$$

$$E\left(Y_t\right) = E\left(\sum_{k=0}^{t-1} \beta^k u_{t-k}\right) = \sum_{k=0}^{t-1} \beta^k E\left(u_{t-k}\right) = 0$$

$$Var\left(Y_t\right) = Var\left(\sum_{k=0}^{t-1} \beta^k u_{t-k}\right) = \sum_{k=0}^{t-1} \beta^{2k} Var\left(u_{t-k}\right) = \sigma^2 \sum_{k=0}^{t-1} \beta^{2k}$$
$$= \sigma^2 \frac{1-\beta^{2t}}{1-\beta^2} \rightarrow \frac{\sigma^2}{1-\beta^2} \text{ as } t \rightarrow \infty$$

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 is undefined as it explodes to infinity, and Y_t is *nonstationary.* To see this

$$Var\left(Y_t\right) = Var\left(\sum_{k=0}^{t-1} u_{t-k}\right) = \sum_{k=0}^{t-1} Var\left(u_{t-k}\right) = t\sigma^2 \rightarrow \infty \text{ as } t \rightarrow \infty$$

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..

The $I(1)$ process is widely used in Economics, for example, in the stock market, many people believe that the stock price is a random walk process, in the sense that we cannot predict the future price of stock. Given the information set Ω_t in period t, the prediction of tomorrow stock price is today's price, mathematically speaking,

$$\begin{aligned} P_{t+1} &= P_t + u_{t+1} \\ E\left(P_{t+1} | \Omega_t\right) &= E\left(P_t | \Omega_t\right) + E\left(u_{t+1} | \Omega_t\right) = P_t + 0 = P_t \end{aligned}$$

12.2 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 iid and independent of u_{t-1}.

We know that the OLS estimate for β is biased and inconsistent in this case. We want to see what will it converge to. Note that

$$\widehat{\beta} = \beta + \frac{\sum_{t=1}^{T} Y_{t-1} u_t}{\sum_{t=1}^{T} Y_{t-1}^2} = \beta + \frac{\sum_{t=1}^{T} Y_{t-1} u_t / T}{\sum_{t=1}^{T} Y_{t-1}^2 / T} \xrightarrow{p} \beta + \frac{\lim_{t\to\infty} E\left(Y_{t-1} u_t\right)}{\lim_{t\to\infty} E\left(Y_{t-1}^2\right)}$$

$$Y_{t-1} = u_{t-1} + \beta u_{t-2} + \beta^2 u_{t-3} + \beta^3 u_{t-4} + ... + \beta^{t-2} u_1 = \sum_{k=0}^{t-2} \beta^k u_{t-k-1}$$

$$Y_{t-1} u_t = \sum_{k=0}^{t-2} \beta^k u_{t-k-1} u_t$$

$$E\left(Y_{t-1} u_t\right) = E\left(\sum_{k=0}^{t-2} \beta^k u_{t-k-1} u_t\right) = \sum_{k=0}^{t-2} \beta^k E\left(u_{t-k-1} u_t\right)$$
$$= \sum_{k=0}^{t-2} \beta^k \rho^{k+1} \sigma_u^2 = \rho \sigma_u^2 \sum_{k=0}^{t-2} \left(\beta\rho\right)^k = \frac{\rho\sigma_u^2\left(1 - \left(\beta\rho\right)^{t-1}\right)}{1 - \beta\rho}$$

$$\lim_{t\to\infty} E\left(Y_{t-1} u_t\right) = \lim_{t\to\infty} \frac{\rho\sigma_u^2\left(1 - \left(\beta\rho\right)^{t-1}\right)}{1 - \beta\rho} = \frac{\rho\sigma_u^2}{1 - \beta\rho}$$

$$Y_{t-1}^2 = \left(\sum_{k=0}^{t-2} \beta^k u_{t-k-1}\right)^2 = \sum_{k=0}^{t-2} \beta^{2k} u_{t-k-1}^2 + 2 \sum_{j=0}^{t-3} \sum_{i=j+1}^{t-2} \beta^{i+j} u_{t-i-1} u_{t-j-1}$$

$$E\left(Y_{t-1}^2\right)=\sum_{k=0}^{t-2}\beta^{2k}E\left(u_{t-k-1}^2\right)+2\sum_{j=0}^{t-3}\sum_{i=j+1}^{t-2}\beta^{i+j}E\left(u_{t-i-1}u_{t-j-1}\right)$$
$$=\sum_{k=0}^{t-2}\beta^{2k}\sigma_u^2+2\sum_{j=0}^{t-3}\sum_{i=j+1}^{t-2}\beta^{i+j}\rho^{i-j}\sigma_u^2$$
$$=\sigma_u^2\sum_{k=0}^{t-2}\beta^{2k}+2\sigma_u^2\sum_{j=0}^{t-3}\beta^{2j}\sum_{i=j+1}^{t-2}\beta^{i-j}\rho^{i-j}$$

$$\lim_{t\to\infty}E\left(Y_{t-1}^2\right)=\sigma_u^2\sum_{k=0}^{\infty}\beta^{2k}+2\sigma_u^2\left(\sum_{j=0}^{\infty}\beta^{2j}\right)\left(\sum_{i=j+1}^{\infty}\beta^{i-j}\rho^{i-j}\right)$$
$$=\frac{\sigma_u^2}{1-\beta^2}+2\sigma_u^2\left(\frac{1}{1-\beta^2}\right)\left(\frac{\beta\rho}{1-\beta\rho}\right)=\frac{1+\beta\rho}{\left(1-\beta^2\right)\left(1-\beta\rho\right)}\sigma_u^2$$

Thus,

$$\widehat{\beta}\xrightarrow{p}\beta+\frac{\lim\limits_{t\to\infty}E\left(Y_{t-1}u_t\right)}{\lim\limits_{t\to\infty}E\left(Y_{t-1}^2\right)}=\beta+\frac{\dfrac{\rho\sigma_u^2}{1-\beta\rho}}{\dfrac{1+\beta\rho}{\left(1-\beta^2\right)\left(1-\beta\rho\right)}\sigma_u^2}=\beta+\frac{\rho\left(1-\beta^2\right)}{1+\beta\rho}$$

Q.E.D.

Questions

1. Explain why the variance of an I(1) process does not exist.

2. (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) $\left\{b_n : a_n=(-1)^n\,\dfrac{2}{n},\ n\geq 1\right\}$

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) $o(1)$?

3. 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.

****

Thus far I have distributed three handouts to you. Each chapter is not simply a summary of the texts. It is written based on my past teaching experiences on the similar courses. I would like to make the handout as self-contained as possible. When I teach this course, I will make every concept as clear as I can. When you study, do not just memorize the formula of an estimator. You must ask yourself what is the purpose of this estimator, how do you derive it, what is its distributional properties, what assumptions are made, and what are the consequences of relaxing any one of the assumptions. Does the model make economics sense? How could I improve the model? This course used to be one of the most difficult undergraduate course in our department as well as in most US schools. Students didn't study this course well largely because they do not pay attention in the very beginning. If you understand the handout, pay attention in class, and finish the homework on your own, you will be fine. Don't skip classes unless you have special reasons for doing so. If you find any difficulties in studying this course, please approach me and speak out the problems. If you find any typo and mistakes in the handout, please inform me also.

****

12.3 Revision of Optimization

If $f(x)$ is a function of x, its local minimum or local maximum x^* is obtained by solving $\frac{df(x)}{dx} = 0$.

Evaluate at x^* , if $\frac{d^2 f(x)}{dx^2} > 0$, it is a local minimum. If $\frac{d^2 f(x)}{dx^2} < 0$, it is a local maximum.

If $f(x, y)$ is a function of x and y, its local minimum or local maximum (x^* , y^*) is obtained by solving $\frac{\partial f(x,y)}{\partial x} = 0$ and $\frac{\partial f(x,y)}{\partial y} = 0$.

Evaluate at (x^* , y^*), if $\frac{d^2 f(x,y)}{dx^2} > 0$, $\frac{d^2 f(x,y)}{dy^2} > 0$, and $\frac{d^2 f(x,y)}{dx^2}\frac{d^2 f(x,y)}{dy^2} -$

$\left(\frac{d^2 f(x,y)}{dxdy}\right)^2 > 0$, it is a local minimum.

If $\frac{d^2 f(x,y)}{dx^2} < 0$, $\frac{d^2 f(x,y)}{dy^2} < 0$, and $\frac{d^2 f(x,y)}{dx^2}\frac{d^2 f(x,y)}{dy^2} - \left(\frac{d^2 f(x,y)}{dxdy}\right)^2 > 0$, it is a local maximum.

$*************$

Chapter 13

Multicollinearity

13.1 Introduction

Multicollinearity, introduced by Ragnar Frisch in his book "Statistical Confluence Analysis by Means of Complete Regression Systems," published in 1934, nowadays refers to situations where there are two or more regressors being linearly related, so that it is difficult to disentangle their separate effects on the dependent variable.

As we have mentioned before that, in a trivariate model, if the two regressors are orthogonal to each other, in the sense that $S_{12} = 0$, then the OLS estimate $\widehat{\beta}_1$ will be the same in both the bivariate and trivariate models. Thus an additional regressor will be of no impact on the original slope estimates as long as it is orthogonal to all the existing regressors. However, if we add a new regressor which is not totally orthogonal to all the existing regressors, then some distortions on the estimates are unavoidable. In the extreme case, when the new regressor is perfectly linearly related to one or more of the existing regressors, the new model is not estimable. We call this problem the *Perfect Collinearity.*

To show the problem more explicitly, consider the following model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

If $X_2 = 2X_1$

The model is reduced to

$$Y_t = \beta_0 + (\beta_1 + 2\beta_2) X_{1t} + u_t$$

Thus it is a simple regression model, and we can obtain the OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta_1 + 2\beta_2}$. However, we cannot obtain estimates for β_1 and β_2, which means the original trivariate model is not estimable.

Let $r_{12}^2 = \dfrac{S_{12}}{S_{11}S_{22}}$. As long as $r_{12}^2 = 1$, the trivariate model is not estimable, since

$$\widehat{\beta}_1 = \frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22} - S_{12}^2} = \frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}$$

$$\widehat{\beta}_2 = \frac{S_{y2}S_{11} - S_{y1}S_{12}}{S_{11}S_{22} - S_{12}^2} = \frac{S_{y2}S_{11} - S_{y1}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}$$

are undefined. In general, if our model is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t$$

The model is not estimable if there are constants $\lambda_0, \lambda_1, \lambda_2, ..., \lambda_k$ (at least some of them are non-zero) such that for all t,

$$\lambda_0 + \lambda_1 X_{1t} + \lambda_2 X_{2t} + ... + \lambda_k X_{kt} = 0$$

Example 3 *If there is multicollinearity, the OLS estimators will be biased. True/False/Uncertain. Explain.*

Solution: False.

Consider the following model :

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t.$$

Then, OLS estimators are given by

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)} = \beta_1 + \frac{S_{u1}S_{22} - S_{u2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}, \\
\widehat{\beta}_2 &= \frac{S_{y2}S_{11} - S_{y1}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)} = \beta_2 + \frac{S_{u2}S_{11} - S_{u1}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}, \\
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X}_1 - \widehat{\beta}_2 \overline{X}_2.
\end{aligned}$$

Taking expectations on all estimators, we have

$$\begin{aligned}
E\left(\widehat{\beta}_1\right) &= \beta_1 + \frac{E\left(S_{u1}\right) S_{22} - E\left(S_{u2}\right) S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}, \\
E\left(\widehat{\beta}_2\right) &= \beta_2 + \frac{E\left(S_{u2}\right) S_{11} - E\left(S_{u1}\right) S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}, \\
E\left(\widehat{\beta}_0\right) &= E\left(\overline{Y}\right) - E\left(\widehat{\beta}_1\right)\overline{X}_1 - E\left(\widehat{\beta}_2\right)\overline{X}_2.
\end{aligned}$$

Since $E\left(S_{u1}\right) = E\left(\sum_{t=1}^{T}\left(X_{1t} - \overline{X}_1\right) u_t\right) = \sum_{t=1}^{T}\left(X_{1t} - \overline{X}_1\right) E\left(u_t\right) = 0$ and $E\left(S_{u2}\right) = 0$,

$$E\left(\widehat{\beta}_1\right) = \beta_1,\ E\left(\widehat{\beta}_2\right) = \beta_2 \text{ and } E\left(\widehat{\beta}_0\right) = \beta_0.$$

Thus, $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are unbiased estimators even though $r_{12}^2 \neq 0$. ■

Example 4 *If there is multicollinearity, the OLS estimators will be inconsistent. True/False/Uncertain. Explain.*

Solution: False.

Consider the following model :

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t.$$

Then, OLS estimators are given by

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)} = \beta_1 + \frac{S_{u1}S_{22} - S_{u2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}, \\
\widehat{\beta}_2 &= \frac{S_{y2}S_{11} - S_{y1}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)} = \beta_2 + \frac{S_{u2}S_{11} - S_{u1}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}, \\
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X}_1 - \widehat{\beta}_2 \overline{X}_2.
\end{aligned}$$

Taking probability limits on all estimators, we have

$$\begin{aligned}
plim\ \widehat{\beta}_1 &= \beta_1 + plim\ \frac{S_{u1}S_{22} - S_{u2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}, \\
plim\ \widehat{\beta}_2 &= \beta_2 + plim\ \frac{S_{u2}S_{11} - S_{u1}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}, \\
plim\ \widehat{\beta}_0 &= plim\ \left(\overline{Y} - \widehat{\beta}_1 \overline{X}_1 - \widehat{\beta}_2 \overline{X}_2\right).
\end{aligned}$$

In particular, we consider the term $plim\ \frac{S_{u1}S_{22} - S_{u2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}$ only.

$$\begin{aligned}
plim\ \frac{S_{u1}S_{22} - S_{u2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)} &= \frac{plim\left(S_{u1}S_{22} - S_{u2}S_{12}\right)/T^2}{plim\left(S_{11}S_{22}\left(1 - r_{12}^2\right)\right)/T^2} \\
&= \frac{plim\frac{S_{u1}}{T}plim\frac{S_{22}}{T} - plim\frac{S_{u2}}{T}plim\frac{S_{12}}{T}}{plim\frac{S_{11}}{T}plim\frac{S_{22}}{T}\left(1 - r_{12}^2\right)} \\
&= \frac{Cov\left(X_{1t}, u_t\right)Var\left(X_{2t}\right) - Cov\left(X_{2t}, u_t\right)Cov\left(X_{1t}, X_{2t}\right)}{Var\left(X_{1t}\right)Var\left(X_{2t}\right)\left(1 - r_{12}^2\right)}
\end{aligned}$$

By the assumption of OLS, $Cov\left(X_{1t}, u_t\right) = Cov\left(X_{2t}, u_t\right) = 0$. Then,

$$plim\ \widehat{\beta}_1 = \beta_1.$$

Similarly, we can show that

$$plim\ \widehat{\beta}_2 = \beta_2.$$

The probability limit of $\widehat{\beta}_0$ is given by

$$\begin{aligned} plim\ \widehat{\beta}_0 &= plim\ \left(\beta_0 + \left(\beta_1 - \widehat{\beta}_1\right)\overline{X}_1 + \left(\beta_2 - \widehat{\beta}_2\right)\overline{X}_2 + \overline{u}\right) \\ &= \beta_0 + \overline{X}_1 plim\ \left(\beta_1 - \widehat{\beta}_1\right) + \overline{X}_2 plim\ \left(\beta_2 - \widehat{\beta}_2\right) + plim\ \frac{1}{T}\sum_{t=1}^{T} u_t \\ &= \beta_0 \text{ since } plim\ \widehat{\beta}_1 = \beta_1,\ plim\ \widehat{\beta}_2 = \beta_2 \text{ and } plim\ \frac{1}{T}\sum_{t=1}^{T} u_t = E\left(u_t\right) = 0. \end{aligned}$$

Thus, $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are consistent estimators even though $r_{12}^2 \neq 0$. ■

13.2 Consequences of near or high Multicollinearity

Recall that if the assumptions of the classical model are satisfied, the OLS estimators of the regression estimators are BLUE. The existence of multicollinearity does not violate any one of the classical assumptions, so if the model is still estimable, the OLS estimator will still be consistent, efficient, linear, and unbiased. So why do we care about multicollinearity? Although multicollinearity does not affect the estimation, it will affect the hypothesis testing.

1: Large Variances of OLS Estimators

Consider the trivariate model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

$$Var\left(\widehat{\beta}_1\right) = Var\left(\frac{S_{y1}S_{22} - S_{y2}S_{12}}{S_{11}S_{22} - S_{12}^2}\right) = Var\left(\frac{S_{u1}S_{22} - S_{u2}S_{12}}{S_{11}S_{22}\left(1 - r_{12}^2\right)}\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2} Var\left(S_{u1}S_{22} - S_{u_2}S_{12}\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2}\left(Var\left(S_{u1}S_{22}\right) + Var\left(S_{u_2}S_{12}\right) - 2Cov\left(S_{u1}S_{22}, S_{u2}S_{12}\right)\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2}\left(S_{22}^2S_{11}\sigma^2 + S_{12}^2S_{22}\sigma^2 - 2S_{12}S_{22}Cov\left(S_{u1}, S_{u2}\right)\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2}\left(S_{22}^2S_{11}\sigma^2 + S_{12}^2S_{22}\sigma^2 - 2S_{12}^2S_{22}\sigma^2\right)$$
$$= \frac{1}{\left[S_{11}S_{22}\left(1-r_{12}^2\right)\right]^2} S_{11}S_{22}^2\left(1 - r_{12}^2\right)\sigma^2 = \frac{\sigma^2}{S_{11}\left(1 - r_{12}^2\right)}$$

Similarly, it can be shown that

$$Var\left(\widehat{\beta}_2\right) = \frac{\sigma^2}{S_{22}\left(1 - r_{12}^2\right)}$$

Thus, the variances of the estimators increase as the relationship between regressors increase. In the extreme case, they explode when there is perfect multicollinearity.

2: Wider Confidence Intervals

Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger. Therefore, in cases of high multicollinearity, the chance of accepting the null hypothesis increases, hence Type II error (Accept H_0 when H_0 is false) increases. Therefore, even if the explanatory variable does individually explain the dependent variable well, we may still tend to conclude that each of them is not significant if there is multicollinearity.

3: Insignificant t Ratio

Recall that the t statistic for the hypothesis $H_0 : \beta_i = 0$ $(i = 0, 1, 2, ..., k)$ is

$$t = \frac{\widehat{\beta}_i}{\sqrt{\widehat{Var}\left(\widehat{\beta}_2\right)}}$$

In cases of high collinearity, the estimated standard errors increase dramatically, thereby making the t values smaller for any given values of $\widehat{\beta}_i$. Therefore, one will over-accept the null that $\beta_i = 0$.

13.3 Detection of Multicollinearity

Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between its presence or absence, but between its various degrees. Therefore, we do not test for multicollinearity but instead, measure its degree in any particular sample.

Since multicollinearity refers to the condition of the explanatory variables that are assumed to be nonstochastic, it is essentially a sample phenomenon, arising out of the largely nonexperimental data collected in most social sciences, we do not have one unique method of detecting it or measuring its strength.

Our rule of thumb is, if we run a regression and find **a High $\mathbf{R}^2$ but few significant t Ratios,** then this is a symptom of multicollinearity. If R^2 is high, the F test in most cases will reject the hypothesis that the slope coefficients are zero simultaneously. However, very few or even none of the individual t tests will be significant.

Other symptoms of multicollinearity include: (1) Small changes in the data can produce wide swings in the parameter estimates, and (2) Coefficients will have the wrong sign or an implausible magnitude.

13.4 Remedial Measures

What can be done if multicollinearity is serious? The following methods can be tried.

1. **A priori information**

Suppose we consider the model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

Suppose a priori we believe or economic theory suggests that $\beta_1 = 2\beta_2$, then we can run the following regression,

$$Y_t = \beta_0 + 2\beta_2 X_{1t} + \beta_2 X_{2t} + u_t$$

$$Y_t = \beta_0 + \beta_2 X_t + u_t$$

where $X_t = 2X_{1t} + X_{2t}$. Once we obtain $\widehat{\beta}_2$, we can define $\widehat{\beta}_1 = 2\widehat{\beta}_2$.

2. **Using first differences or ratios**

Suppose we have

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

where X_{1t} and X_{2t} are highly collinear. To reduce the degree of collinearity, we can still estimate β_1 and β_2 by the “first difference” model, i.e. we estimate

$$Y_t - Y_{t-1} = \beta_1 \left(X_{1t} - X_{1(t-1)} \right) + \beta_2 \left(X_{2t} - X_{2(t-1)} \right) + (u_t - u_{t-1})$$

Although the first difference model may reduce the severity of multicollinearity, it creates some additional problems. In the transformed model, the new error terms $(u_t - u_{t-1})$ is not serially independent as $Cov\left(u_t - u_{t-1}, u_{t-1} - u_{t-2}\right) = -Var\left(u_{t-1}\right) = -\sigma^2 \neq 0$. We will discuss the problem of serial correlation in the next chapter. But here we alleviate multicollinearity at the expense of violating one of the classical assumptions "serial independence", this implies that the Gauss-Markov theorem will not hold anymore, and the OLS estimators are not BLUE in the "first difference" model. Further, since the new observations become $\{y_t - y_{t-1}\}_{t=2}^{T}$, there is a loss of one observation due to the difference procedure, and therefore the degrees of freedom are reduced by one.

The problem is similar if we use ratios and estimate an equation of the form

$$\frac{Y_t}{X_{2t}} = \beta_2 + \beta_0 \frac{1}{X_{2t}} + \beta_1 \frac{X_{1t}}{X_{2t}} + \frac{u_t}{X_{2t}}$$

Now the new residuals will be heteroskedastic.

3. **Dropping a variable(s)**

When faced with severe multicollinearity, the simplest thing to do is to drop one of the collinear variables. However, we may commit a specification error if a variable is dropped from the model. While multicollinearity may prevent precise estimation of the parameters of the model, omitting a variable may make the estimators inconsistent.

4. **Increasing the sample size**

Since multicollinearity is a sample feature, it is possible that in another sample the problem may not be as serious as in the first sample. Sometimes simply increasing the sample size may attenuate the problem, for example, in the trivariate model, we have $Var\left(\widehat{\beta}_1\right) = \dfrac{\sigma^2}{S_{11}\left(1-r_{12}^2\right)}$ and $Var\left(\widehat{\beta}_2\right) = \dfrac{\sigma^2}{S_{22}\left(1-r_{12}^2\right)}$ since S_{11} and S_{22} increase as the sample size increases, hence $Var\left(\widehat{\beta}_1\right)$ and $Var\left(\widehat{\beta}_2\right)$ will decline as a result.

5. **Benign Neglect**

If we are less interested in interpreting individual coefficients but more interested in forecasting, multicollinearity is not a serious problem. We can simply ignore it.

Questions:

1. Suppose the true model is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

u_t are iid$(0, \sigma^2)$, we have observations $\{Y_t, X_{1t}, X_{2t}\}_{t=1}^{T}$ and run the following models:

$$\widehat{Y}_t = \widehat{\mu}_0 + \widehat{\mu}_1 X_{1t} + \widehat{\mu}_2 X_{2t}$$

$$\widehat{Y}_t = \widehat{\alpha}_0 + \widehat{\alpha}_1 X_{1t}$$

$$\widehat{Y}_t = \widehat{\gamma}_0 + \widehat{\gamma}_2 X_{2t}$$

Let $S_{12} = \sum\limits_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right)\left(X_{2t} - \overline{X}_2\right)$,

a. Compare the values of $\widehat{\mu}_1$ and $\widehat{\alpha}_1$ when $S_{12} = 0$, $S_{12} > 0$, $S_{12} < 0$.

b. Compare the values of $\widehat{\mu}_2$ and $\widehat{\gamma}_2$ when $S_{12} = 0$, $S_{12} > 0$, $S_{12} < 0$.

c. Compare the values of $\widehat{\mu}_0$, $\widehat{\alpha}_0$ and $\widehat{\gamma}_0$ when $S_{12} = 0$, $S_{12} > 0$, $S_{12} < 0$.

d. Redo (a), (b), and (c), by comparing their expectations.

e. Redo (a), (b), and (c), by comparing their variances.

2. Suppose we have observations $\{Y_t, X_{1t}, X_{2t}\}_{t=1}^{T}$ with

$S_{11} = \sum\limits_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right)^2 = 200$

$S_{12} = \sum\limits_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right)\left(X_{2t} - \overline{X}_2\right) = 150$

$S_{22} = \sum\limits_{t=1}^{T} \left(X_{2t} - \overline{X}_2\right)^2 = 113$

$S_{y1} = \sum\limits_{t=1}^{T} \left(X_{1t} - \overline{X}_1\right)\left(Y_t - \overline{Y}\right) = 350$

$S_{y2} = \sum\limits_{t=1}^{T} \left(X_{2t} - \overline{X}_2\right)\left(Y_t - \overline{Y}\right) = 263$

and run the following models:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

a. Calculate $\widehat{\beta}_1$, $\widehat{\beta}_2$, and r_{12}^2. Is the problem of multicollinearity between X_1 and X_2 serious?

b. Suppose we drop an observation and obtain $S_{11} = 199$, $S_{12} = 149$, $S_{22} = 112$, $S_{y1} = 347.5$, $S_{y2} = 261.5$. Are the new estimates close to the previous estimates in part (a) when using the full sample?

Example 3: Let X and Y be two random variables with

$$\begin{aligned} F(x,y) &= \left(1-e^{-x}\right)\left(1-e^{-y}\right) \qquad \text{for } x>0 \text{ and } y>0 \\ &= 0 \qquad \text{elsewhere} \end{aligned}$$

then

$$\begin{aligned} f(x,y) &= \frac{\partial^2}{\partial x \partial y} F(x,y) = e^{-x}e^{-y} \qquad \text{for } x>0 \text{ and } y>0 \\ &= 0 \qquad \text{elsewhere} \end{aligned}$$

$$\begin{aligned} f(x) &= \frac{\partial}{\partial x} F(x,y) = e^{-x}\left(1-e^{-y}\right) \qquad \text{for } x>0 \text{ and } y>0 \\ &= 0 \qquad \text{elsewhere} \end{aligned}$$

$$\begin{aligned} f(y) &= \frac{\partial}{\partial y} F(x,y) = \left(1-e^{-x}\right)e^{-y} \qquad \text{for } x>0 \text{ and } y>0 \\ &= 0 \qquad \text{elsewhere} \end{aligned}$$

Since $f(x,y) \neq f(x)\,f(y)$, X and Y are not independent.
***

13.5 Functions of Random Variable

Suppose we would like to find the density function of a function of a particular variable, say suppose X is a standard normal random variable, we know that X^2 will have a χ_1^2 distribution. How do we derive the density function of a χ_1^2 from a $N(0,1)$?

Consider a random variable X, let $g(\cdot)$ be a continuous, differentiable and monotonic function, suppose the distribution function and density function of X are $F_X(x)$ and $f_X(x)$ respectively, what is the density function of $Y = g(X)$?

Denote the capital X as a random variable, and small letter x be a particular value. We know that

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

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

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

Substitute x by $g^{-1}(y)$

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

Example 5: If $X = \ln(Y) \sim N(\mu, \sigma^2)$, then $Y = \exp(X)$ will follow a lognormal distribution. To find its density function, note that:

$$\left|\frac{dx}{dy}\right| = \left|\frac{1}{y}\right|.$$

$$f_X\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

$$g^{-1}\left(y\right)=\ln y.$$

We have

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

Example 6: What is the density function of $Y=X^2$ if: (a). $X\sim U\left(0,1\right)$; (b). $X\sim N\left(0,1\right)$.

Solution:

(a) Given $Y=X^2$ and $X\sim U(0,1)$, we have

$$X=Y^{1/2}\text{ and }\frac{dx}{dy}=\frac{1}{2}y^{-1/2}.$$

By the transformation of random variable, the density function of Y is given by

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

In particular,

$$f_Y(y)=\frac{1}{2}y^{-1/2}\text{ when }1>y>0.$$

When $y\leq 0$, $f_Y(y)=0$. ■

(b) Given $Y=X^2$ and $f_X(x)=\dfrac{1}{\sqrt{2\pi}}\exp\left[-\dfrac{x^2}{2}\right]$,

$$\begin{aligned} F_Y(y) &= \Pr\left(Y \leq y\right) \\ &= \Pr\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) \\ &= F_X\left(x\right) - F_X\left(-x\right) \quad \text{where } x = \sqrt{y}. \end{aligned}$$

By differentiation, we have

$$f_Y(y) = \left[f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)\right]\left(\frac{dx}{dy}\right).$$

Since y cannot be negative, $f_Y(y) = 0$ when $y < 0$; the value of $f_Y(0)$ is set equal to 0 arbitrarily. As we know $f_X(x) = f_X(-x)$, it follows that

$$\begin{aligned} f_Y(y) &= 2f_X\left(\sqrt{y}\right)\left(\frac{dx}{dy}\right) \\ &= 2\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{y}{2}\right]\left(\frac{1}{2}y^{-1/2}\right) \\ &= \frac{1}{\sqrt{2\pi}}y^{-1/2}\exp\left[-\frac{y}{2}\right] \qquad \text{when } y > 0. \end{aligned}$$

Note that Y has a chi-square distribution. ■

Exercise 4: Let $X \sim U\left(0,1\right)$, suppose $g\left(X\right) \sim N\left(0,1\right)$, what is the functional form of $g\left(X\right)$?

******

Example 2: Let X and Y be two independent standardized normal random variables. Find:

(i) $\text{Cov}(X, \max\left\{X, Y\right\})$; (ii) $\text{Var}(\max\left\{X, Y\right\})$.

Solution:

(i)

$$\begin{aligned}
& Cov(X, \max\{X,Y\}) \\
= \; & E\left[(X-E(X))\,(X-E(X))|\, X>Y\right]\Pr(X>Y) \\
& \qquad\qquad + E\left[(X-E(X))\,(Y-E(Y))|\, X<Y\right]\Pr(X<Y) \\
= \; & E\left[(X-E(X))\,(X-E(X))\right]\Pr(X>Y) \\
& + E\left[(X-E(X))\,(Y-E(Y))\right]\Pr(X<Y) \\
& \qquad\qquad \text{(since } X \text{ and } Y \text{ are independent)} \\
= \; & Var\,(X)\left(\frac{1}{2}\right) = (1)\,\frac{1}{2} = \frac{1}{2}. \qquad \text{(since } X \sim N\,(0,1)\,\text{)} \qquad \blacksquare
\end{aligned}$$

(ii)$Var(\max\{X,Y\})$

$= E\left[(Y-E(Y))^2\big|\, X<Y\right]\Pr(X<Y)+E\left[(X-E(X))^2\big|\, X>Y\right]\Pr(X>Y)$

$= E\left[(Y-E(Y))^2\right]\Pr(X<Y)+E\left[(X-E(X))^2\right]\Pr(X>Y)$

(since X and Y are independent)

$= 1.$ $\blacksquare$

Exercise 2: Let X and Y be two independent standardized normal random variables. Find:

i) Cov$(X, \min\{X,Y\})$;

ii) Cov$(\min\{X,Y\}, \max\{X,Y\})$;

iii) Var$(\min\{X,Y\})$.

****

Dependent but Identical Distribution

If all the x_i have the same distribution, but x_i depends on x_j for some $i \neq j$.

e.g. if x_t is symmetrically distributed around zero for all t and $x_t = -x_{t-1}$.

Independent but Non-Identical Distribution

If x_i does not depend on x_j for any $i \neq j$, but x_i and x_j have different distributions.

e.g. $Var\left(x_t\right) = t$, in this case, the $Var\left(x_i\right) \neq Var\left(x_j\right)$ for all $i \neq j$.

Dependent and Non-Identical Distribution

If x_i depends on x_j for some $i \neq j$, and x_i and x_j have different distributions.

e.g. $x_t = -2x_{t-1}$.

****

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_t. e.g., We may need $E\left(X^4\right) < \infty$, etc.. Note that we only require the existence of the first and second moments in the simplest version above.

*****

Of course, since $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are estimators which are subjected to errors, the predicted value of Y_t also has error. The error depends on the location of X, and it will be shown that the error will be minimal when X_t is at $\overline{X}$.

Now we look at $Var\left(\widehat{Y}_t\right)$,

$$Var\left(\widehat{Y}_t\right) = Var\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_t\right) = Var\left(\overline{Y} - \widehat{\beta}_1 \overline{X} + \widehat{\beta}_1 X_t\right) = Var\left(\overline{Y} + \widehat{\beta}_1\left(X_t - \overline{X}\right)\right).$$

Note that $\overline{Y}$ and $\widehat{\beta}_1$ are random variables since they depend on u_t.

Substituting $\overline{Y} = \beta_0 + \beta_1 \overline{X} + \overline{u}$ gives
$Var\left(\beta_0 + \beta_1 \overline{X} + \overline{u} + \widehat{\beta}_1 \left(X_t - \overline{X}\right)\right)$
$= Var\left(\overline{u} + \widehat{\beta}_1 \left(X_t - \overline{X}\right)\right)$ since $\beta_0 + \beta_1 \overline{X}$ is constant and fixed
$= Var\left(\overline{u}\right) + \left(X_t - \overline{X}\right)^2 Var\left(\widehat{\beta}_1\right)$ since $Cov\left(\overline{u}, \widehat{\beta}_1 \left(X_t - \overline{X}\right)\right) = 0$ why?
$= \dfrac{\sigma^2}{T} + \left(X_t - \overline{X}\right)^2 \dfrac{\sigma^2}{\sum\limits_{i=1}^{T} \left(X_i - \overline{X}\right)^2}.$
Thus

$$Var\left(\widehat{Y}_t\right) = \left(\frac{1}{T} + \frac{\left(X_t - \overline{X}\right)^2}{\sum\limits_{i=1}^{T} \left(X_i - \overline{X}\right)^2}\right) \sigma^2$$

which is minimized when X_t equals $\overline{X}$, and getting larger as X_t getting farther away from $\overline{X}$.

Sometimes X and/or Y are meaningless in some regions. For example, if Y is the quantity of any tangible product, it must be non-negative. But the predicted value of Y may be negative. Another example is when Y is the probability of something happening given the value of X, we may predict Y to be negative or of values bigger than one. In such cases, we should use other estimation methods instead of linear regression to make the forecast sensible, e.g. Maximum Likelihood method or non-linear regressions.