STA5130 High-dimensional Data Analysis, Fall 2009

Terence Tai-Leung Chong

November 20, 2014

Contents

Chapter 1

Probability and Distribution Theory

1.1 Revision of the Summation Operator

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

1. If k is a constant, then $\sum\limits_{i=1}^{n} k = nk$;
2. If k is a constant, then $\sum\limits_{i=1}^{n} kx_i = k\sum\limits_{i=1}^{n} x_i$;
3. $\sum\limits_{i=1}^{n} (x_i + y_i) = \sum\limits_{i=1}^{n} x_i + \sum\limits_{i=1}^{n} y_i$;
4. $\sum\limits_{i=1}^{n} (x_i - \overline{x}) = 0$;
5. $\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum\limits_{i=1}^{n} (x_i - \overline{x}) y_i = \sum\limits_{i=1}^{n} (y_i - \overline{y}) x_i$;
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. $\left(\sum\limits_{i=1}^{n} x_i\right)^2 = \sum\limits_{i=1}^{n} x_i^2 + 2\sum\limits_{i=1}^{n-1}\sum\limits_{j>i}^{n} x_i x_j$.

Exercise 1.1:

(a) Compute

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

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

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

(iv) $\sum\limits_{i=1}^{n} (x_i - \overline{x})\,\overline{x}$

(b) True/False.

(i). $\sum\limits_{i=1}^{n} (x_i - \overline{x}) = 1$.

(ii). $\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} \overline{x}$.

(c) The daily return of a stock is defined as $r_t = \ln P_t - \ln P_{t-1}$, where P_t is the closing price of a stock on day t. Extract the daily closing price of HUIYUAN JUICE [01886] from yahoo finance for the period 31/8/2013 to 31/8/2014. Let 2/9/2014 be day 1. Find the sample mean $\overline{r} = \frac{1}{n} \sum\limits_{t=1}^{n} r_t$ and sample standard deviation $s = \sqrt{\frac{1}{n-1} \sum\limits_{t=1}^{n} (r_t - \overline{r})^2}$, where n is the sample size.

Definition 1.1: 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 1.2: 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 1.3: 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 1.1: 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 1.1: 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 1.4: 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 1.5: 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 1.2: Give two independent events and two dependent events.

Exercise 1.3: The Mark Six lottery is a lottery game conducted by HKJC Lotteries Limited using the facilities of The Hong Kong Jockey Club. Since its inception in 1975, the Mark Six has contributed over HK$24 billion to the Hong Kong SAR Government Treasury and the Lotteries Fund, being a fund that supports charitable causes in Hong Kong. To win the first prize of the Mark Six, one needs to get 6 numbers correct out of a pool of 49 numbers indexed from 1 to 49. Suppose each number has the same chance of being drawn,

(a) Find the probability of winning the first prize of the Mark Six.

(b) Suppose you have to bet 5 dollars for the first prize of 50,000,000 dollars. If there is only one first prize winner, find the expected gain (or loss) of your game.

(c) Suppose Chinese people have preference over the "lucky" numbers 8, 18, 28, 38, and a large proportion of people like to put these numbers on their Mark-Six tickets. Suppose the amount of money for the first the prize is fixed, and has to be shared among winners. Should we avoid these "lucky" numbers when buying Mark Six? Explain.

Definition 1.6: 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 a 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 the outcome is Head and $X = 2$ if the outcome is Tail, then X is a random variable.

1.2 Probability Distribution Function and Density Function

Let X, Y be two continuous random variables.

Definition 1.7: The **probability distribution function** of X is defined as $F_x(u) = \Pr(-\infty < X \leq u)$, with $F_x(\infty) = 1$.

Definition 1.8: The **density function** is $f(x) = \dfrac{dF(x)}{dx}$, with $f(x) \geq 0$, and $f(-\infty) = f(\infty) = 0$.

Example 1.2: Let X be a random variable evenly distributed in zero-one interval, then

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

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

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

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

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

Definition 1.9: The **joint distribution function** of X and Y is defined as $F(x,y) = \Pr(X \le x \text{ and } Y \le y)$. 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, $F(-\infty,-\infty) = \Pr(X \le -\infty \text{ and } Y \le -\infty) = 0$, $F(\infty,\infty) = \Pr(X \le \infty \text{ and } Y \le \infty) = 1$, and $f(x,y) \ge 0$. X and Y are independent **if and only if** $f(x,y) = f(x)f(y)$.

Exercise 1.4: 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 1.5: Suppose the joint density of X and Y is given by:

$f(x, y) = 2$ for $x > 0$, $y > 0$, $x + y < 1$

$f(x, y) = 0$ otherwise

Find

(i) $\Pr\left(X \leq \frac{1}{2} \text{ and } Y \leq \frac{1}{2}\right)$.

(ii) $\Pr\left(X + Y > \frac{2}{3}\right)$.

(iii) $\Pr(X > 2Y)$.

Exercise 1.6: Let X be a discrete random variable with the probability distribution as follows:

$X = -1$ with probability $\dfrac{1}{2}$.

$X = 1$ with probability $\dfrac{1}{2}$.

Suppose we draw two observations, X_1 and X_2 independently from this distribution. For the following Z variables, what are the possible values that Z will take and what is the associate probability of each value?

(a) $Z = X^2$.

(b) $Z = \dfrac{X_1}{X_2}$.

(c) $Z = \overline{X}$.

(d) $Z = \min\{X_1, X_2\}$.

Exercise 1.7: Let X be a discrete random variable with the probability distribution as follows:

$X = -2$ with probability $\frac{1}{3}$;
$X = 0$ with probability $\frac{1}{3}$;
$X = 2$ with probability $\frac{1}{3}$.

Suppose we draw two observations, X_1 and X_2 independently from this distribution.

For the following Z variables,

(a) $Z = \dfrac{X_1 + X_2}{2}$;

(b) $Z = X_1^2 + X_2^2$;

What are the possible values that Z? What is the probability for each possible value? (e.g., write it in the form $\Pr(Z = 0) = 0.5$ and so on).

Exercise 1.8: Let X, Y be two independent identical discrete random variable with the probability distribution as follows:
$X = -1$ with probability $\frac{1}{2}$.
$X = 1$ with probability $\frac{1}{2}$.
$Y = -1$ with probability $\frac{1}{2}$.
$Y = 1$ with probability $\frac{1}{2}$.
Find the distribution of Z if:
a) $Z = X - Y$.
b) $Z = \dfrac{X}{Y}$.
c) $Z = \max\{X, Y\}$.

Exercise 1.9: If X and Y are two continuous random variables, then $X + Y$ must be continuous too. True or false? Explain.

Exercise 1.10: Let X be a random variable with a symmetrical distribution about zero and a finite variance. Give a random variable Y such that X and Y are uncorrelated but not independent.

1.3 Mathematical Expectation

Definition 1.10: 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 1.11: True/False/Uncertain. Explain.

(a). $\dfrac{1}{E(X)} = E\left(\dfrac{1}{X}\right)$.

(b) Let X and Y be two independent random variables, if $E\left(\dfrac{X}{Y}\right) > 1$, then $\dfrac{E(X)}{E(Y)} > 1$.

Definition 1.11: The **second moment around the mean** or **variance** of a random variable is

$Var\left(X\right) = E\left(X - E\left(X\right)\right)^2 = E\left(X^2\right) - E^2\left(X\right) = \sum\limits_i \left(x_i - E\left(X\right)\right)^2 P\left(x_i\right)$ 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) $Var\left(X \pm Y\right) = Var\left(X\right) + Var\left(Y\right) \pm 2Cov\left(X, Y\right)$ 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 1.12: The **covariance** of two random variables X and Y, is defined as $Cov\left(X, Y\right) = E\left(X - E\left(X\right)\right)\left(Y - E\left(Y\right)\right) = E\left(XY\right) - E\left(X\right)E\left(Y\right)$, where

$E\left(XY\right) = \sum\limits_i x_i y_i \Pr\left(x_i, y_i\right)$ 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 1.3: Let X, Y, and Z be three random variables, if $Cov\left(X, Z\right) \neq 0$ and $Cov\left(Y, Z\right) \neq 0$, then $Cov\left(X, Y\right) \neq 0$. 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 1.13: The **correlation coefficient** between X and Y is defined as:

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

Example 1.4: 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

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

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

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

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

Exercise 1.12: The daily return of a stock is defined as $r_t = \ln P_t - \ln P_{t-1}$, where P_t is the closing price of a stock on day t. Extract the daily closing price of [5] Hong Kong Bank and [11] Hang Seng Bank from yahoo finance for the period 31/8/2013 to 31/8/2014. Let 2/9/2013 be day 1. Let $r_{HSBC,t}$ $r_{HSB,t}$ be the daily return of Hong Kong Bank and Hang Seng Bank from 2/9/2013-31/8/2014 respectively.

(a) Plot $(r_{HSBC,t}, r_{HSB,t})$ on the X-Y plane.

(b) Calculate the sample variance of $r_{HSBC,t}$ and $r_{HSB,t}$,

(c) Calculate the sample covariance of $r_{HSBC,t}$ and $r_{HSB,t}$ $\left(= \frac{1}{n}\sum_{t=1}^{n}\left(r_{HSBC,t} - \overline{r_{HSBC}}\right)\left(r_{HSB,t} - \overline{r_{HSB}}\right)\right.$ and the sample correlation coefficient.

Exercise 1.13: Let X, Y, W, and Z be random variables, and a, b, c, d be constants. Show that:

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

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

(c) $Cov(X,X) = Var(X)$.

(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 1.14: True/False/Uncertain. Explain. Let X be a random variable, then

(a). $Cov\left(Var(X), X\right) = 0.$

(b). $E\left(Var(X)\right) = Var\left(X\right).$

(c). $E\left(Var(X)\right) = Var\left(E\left(X\right)\right).$

(d). $Var\left(E(X)\right) = 0.$

(e). $Var\left(\frac{1}{X}\right) = \frac{1}{Var(X)}.$

Exercise 1.15: True/False/Uncertain. Explain. Let X and Y be two random variables.

(a) If $Cov\left(X^2, Y^2\right) = 0$, then $Cov\left(X,Y\right) = 0$.

(b) If X and Y are independent, then $Cov\left(X^2, Y^2\right) > Cov\left(X,Y\right)$. True/False/Uncertain. Explain.

(c). If X is symmetrical about zero, $W = X$, and $Z = \frac{1}{X}$, then $Cov\left(W,Z\right) = 1$.

(d). If X and Y are dependent, let $W = XY$, then $Cov\left(W,X\right) = E(Y)Var(X)$

Exercise 1.16: 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 of X for $j = 0,1,2,3,4$.

(b) Find the mean of X.

(c) Find the variance of X.

1.4 Special Probability Distributions

1.4.1 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 $\frac{1}{2}$. 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 1.17: 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 1.18:

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

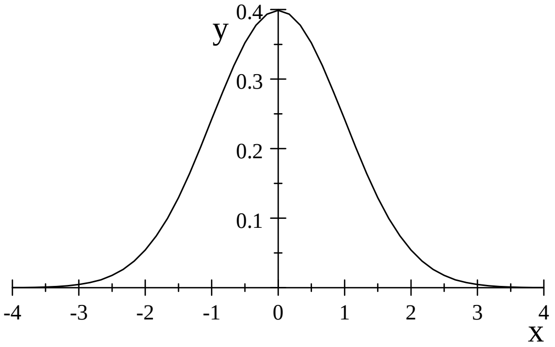
(b) If $X \sim U(a,b)$, where $a < b$, find (i) $f(x)$; (ii) $F(x)$; (iii) $E(X)$; (iv) $Var(X)$.

1.4.2 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 1.19:

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

(b) If two normally distributed random variables are uncorrelated, then they are independent. True/False/Uncertain. Explain.

(c)Let $r_{HSBC,t}$ $r_{HSB,t}$ and $r_{HW,t}$ be the daily return of [5] Hong Kong Bank, [11] Hang Seng Bank and [13] Hutchison from 2/9/2013-31/8/2014 respectively.

(i)With the help of computer, plot the histograms of $r_{HSBC,t}$ $r_{HSB,t}$ and $r_{HW,t}$.

(ii) From visual inspection, are they normally distributed?

1.4.3 Standardized Normal Distribution

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

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

Example 1.5: If $X\sim N(3,4)$, then $Z=\dfrac{X-3}{2}$ 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 1.20: 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) $\Pr(X=1)$.

Exercise 1.21: Let Z_1, Z_2 be independent $N(0,1)$ random variables, let

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

(a) What is the distribution of U?

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

Exercise 1.22: Let Z be a $N(0,1)$ random variable

(a) Write down the distribution of Z^2.
(b) Given that $Var(Z^2) = 2$, find $E(Z^4)$.
(c) Are Z and Z^2 uncorrelated? Explain.

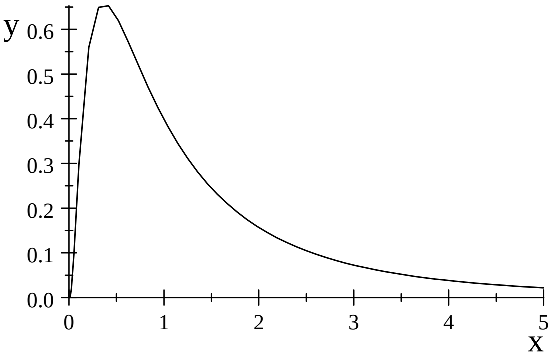
1.4.4 The Lognormal Distribution

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 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 1.23: 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) $\Pr(Y = 1)$.

1.4.5 Chi-square Distribution

Chi-squared distribution

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

Example 1.6: 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

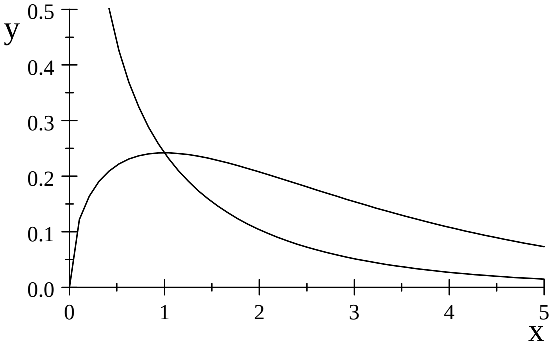
$$E(U) = E\left(Z_1^2 + Z_2^2 + ... + Z_k^2\right) = k.$$

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 1.24: 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) $\Pr(U = 1)$.

1.4.6 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 1.25: 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) $\Pr(X = 1)$.

1.4.7 Student's t-Distribution

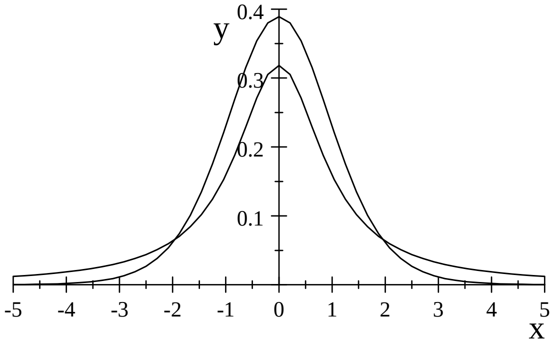
If $Z \sim N(0,1)$, $U \sim \chi^2_k$, and Z and U are independent, then:

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

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,

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

Hence, a t-distribution with degrees of freedom infinity is a standardized 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 $N\left(0, 1\right)$ table.

Exercise 1.26: If the random variable t has a t-distribution with degree of freedom 5, find (i) $\Pr\left(t \leq 0\right)$;(ii)$\Pr\left(t > 0.267\right)$;(iii)$\Pr\left(t > 0.727\right)$;(iv) $\Pr\left(t \leq 1.476\right)$;(v) $\Pr\left(t > 2.015\right)$;(vi) $\Pr\left(2.571 < t \leq 3.365\right)$;(vii) $\Pr\left(t = 1\right)$.

1.4.8 Cauchy Distribution

Let Z_1 and Z_2 be independent and follow $N\left(0, 1\right)$, 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:

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

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 1.27: 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) $\Pr(R = 1)$.

1.4.9 F-Distribution

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$, 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:

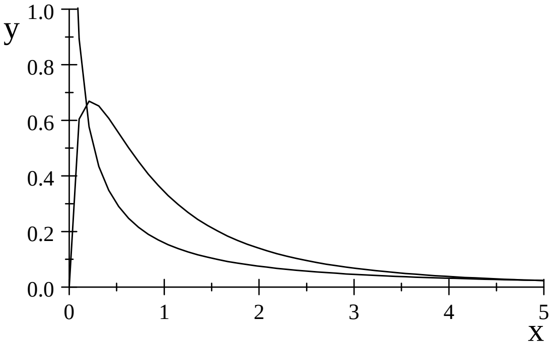
$$F(1,k) = t_k^2.$$

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 1.7: 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 1.28: 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 1.29: Let Z_1, Z_2 be independent $N(0,1)$ random variables, and let

$$\begin{aligned} U &= \frac{Z_1}{Z_2}, \\ V &= Z_1 Z_2. \end{aligned}$$

(a) Write down the distribution of U.

(b) Is the distribution of V a χ_2^2? Why?

Exercise 1.30: 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 $E(U)$ and $E(V)$.

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

Exercise 1.31: True/False.

(a). A Cauchy distribution is a t-distribution with 1 degree of freedom.

(b). A Cauchy distribution is special case of uniform distribution.

(c) An F distribution is a t-distribution with 1 degree of freedom.

Exercise 1.32: Let X be a discrete random variable with the probability distribution as follows:

$$X = 2^n \text{ with probability } \frac{1}{2^n} \text{ for } n = 1, 2, 3, ...$$

(a) Find $E(X)$.

(b) Find $E\left(\frac{1}{X}\right)$ and $\text{Var}\left(\frac{1}{X}\right)$.

Exercise 1.33: True/False/Uncertain. Explain.

(a). $\left(\sum\limits_{i=1}^{n} (x_i - \overline{x})\, y_i\right)^2 \leq \sum\limits_{i=1}^{n} (x_i - \overline{x})^2 \sum\limits_{i=1}^{n} (y_i - \overline{y})^2$.

(b). Let X, Y and Z be three random variables, then

$$Cov\,(XZ, YZ)) = ZCov\,(X, Y)).$$

(c). If two random variables X and Y are independent, then $Cov\,(X^2, Y^2) = 0$.

(d). The Central Limit Theorem states that the sample average has a uniform distribution when sample size is large.

Exercise 1.34: Suppose you are invited to play a game of coin flipping. The possible outcomes are {H, T}. If H appears in the n^{th} trial $(n = 1, 2, ...)$, your payoff is HK$ 2^n and the game stops. Let X be your payoff. It is a discrete random variable with the probability distribution as follows:

$$X = 2^n \text{ with probability } \frac{1}{2^n} \text{ for } n = 1, 2, 3, ...$$

(a) What is the expected payoff $E(X)$ of this game?

(b) Suppose you need to pay an amount of money M in order to play this game. Suppose you will play the game as long as the $E(X) > M$, **should** you play the game if (i) M =HK\$ 2 and (ii) M =HK\$ 2 million?

(c) In reality, **will** you play this game, assuming that there is no budget constraint problem.

(d) Suppose your utility (or happiness) of having X dollar is $U(X) = \log X$, i.e., your have a diminishing utility in money. Suppose you do not have any money to begin with. Show that your expected utility $E(U(X))$ of this game is $E(U(X)) = \sum_{n=1}^{\infty} \frac{1}{2^n} \log 2^n$.

(e) Show that $E(U(X)) = \log 4 < \infty$.

(f) Suppose you will play the game as long as the $E(U(X)) > \log M$, will you play the game if (i) M =HK\$ 2 and (ii) M =HK\$ 2 million? Explain.

Exercise 1.35: Suppose $X \sim N(0, 1)$. We define a new random variable Y, where

$$Y = 1 - X \qquad \text{if } X > 0$$

and

$$Y = -X \qquad \text{if } X \leq 0.$$

(a) Find $Cov(X, Y)$.

(b) Does Y takes continuous or discrete values?

(c) Let $Z = X + Y$, what values will Z take? What is the associated probability for each value? Is Z a discrete or a continuous random variable?

(d) Find $Cov(X, Z)$.

1.5 More Demanding Material

Theorem 1.1: (Chebyshev's Inequality) If X is **any** random variable with finite variance σ^2 and k is a finite positive constant, then

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

Proof. (for continuous random variable)

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

this implies

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

Theorem 1.2: (Jensen's Inequality) Let $g : R \to R$ be a convex function on an interval $B \subset R$ and let Z be a random variable such that $P(Z \in B) = 1$. Then $g(E(Z)) \leq E(g(Z))$.

Proof. (exercise)

Example 1.6: Let $g(z) = |z|$. It follows from Jensen's inequality that $|E(Z)| \leq E|Z|$.

Example 1.7: Let $g(z) = z^2$. It follows from Jensen's inequality that $E^2(Z) \leq E(Z^2)$.

Theorem 1.3: For random sample of size n from an infinite population which has the value $f(x)$ at x, the probability density of the r^{th} **order statistic** Y_r is given by

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

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

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

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

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

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

Chapter 2

Matrix

2.1 Vectors

Definition 2.1: Letting x_i denote the i^{th} observation where i goes from 1 to n, the $n \times 1$ vector $\mathbf{x}$ is represented as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Definition 2.2: The transpose of $\mathbf{x}$ is defined as $\mathbf{x}' = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$.

The vector $\mathbf{x}$ with n elements represents, geometrically, a point in the n-dimensional Euclidean space. Note that, if the x-y axis rotates, the corresponding value of a vector may change. For example, consider a vector $\mathbf{x} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$. If the x-y axis rotate anti-clockwise such that the original point fall into the new x-axis, then the new vector will be read as $\begin{pmatrix} 5 \\ 0 \end{pmatrix}$.

Definition 2.3: The inner product of two k by 1 vectors $\mathbf{x}$ and $\mathbf{y}$ is

$$\mathbf{x}'\mathbf{y} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = x_1y_1 + x_2y_2 + ... + x_ny_n = \sum_{i=1}^{n} x_iy_i.$$

Definition 2.4: Two k by 1 vectors $\mathbf{x}$ and $\mathbf{y}$ and perpendicular (or called orthogonal) if $\mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x} = 0$.

Definition 2.5: The length of a vector $\mathbf{x}$ is defined as $L_x = (\mathbf{x}'\mathbf{x})^{\frac{1}{2}} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$.

The sum of two $n \times 1$ vectors can be defined as

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}.$$

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent if for some non-zero constants a and b,

$$a\mathbf{x} + b\mathbf{y} = \begin{pmatrix} ax_1 + by_1 \\ ax_2 + by_2 \\ \vdots \\ ax_n + by_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}.$$

Exercise 2.1: Plot the following x and y vectors. Are x and y orthogonal? Are x and y linearly dependent?

(a) $\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$; (b) $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$;

(c) $\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$; (d) $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$;

(e) $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

Exercise 2.2: Let $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} 3 \\ 6 \\ 12 \end{pmatrix}$. Find

(a) $\mathbf{x}'$ and $\mathbf{y}'$.

(b) L_x and L_y.

(c) $\mathbf{x} + \mathbf{y}$.

(d) $\mathbf{x}'\mathbf{y}$ and $\mathbf{y}'\mathbf{x}$. Are $\mathbf{x}$ and $\mathbf{y}$ orthogonal?

(e) Are $\mathbf{x}$ and $\mathbf{y}$ linearly independent.

Exercise 2.3: Consider the P/E and dividend of the following stocks as of 14/9/2011.

	[4]	[19]	[267]
	Wharf Holding	Swire Pacific A	Citic Pacific
P/E	3.32	3.79	5.70
$Dividend(\%)$	2.32	3.63	3.24

(a) Treat the data as three 2 × 1 vectors, plot the three vectors (using P/E as the x-axis and Dividend as the y-axis).

(b) Now treat the data as two 3 × 1 vectors called **PE** and **Dividend**. Let

$$\begin{aligned} \mathbf{h} &= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \\ \mathbf{x} &= \mathbf{PE} - \frac{1}{3}\mathbf{h}'(\mathbf{PE})\,\mathbf{h}, \\ \mathbf{y} &= \mathbf{Dividend} - \frac{1}{3}\mathbf{h}'(\mathbf{Dividend})\,\mathbf{h}. \end{aligned}$$

(i) Find $\mathbf{h}'\mathbf{PE}$. Are $\mathbf{h}$ and $\mathbf{PE}$ orthogonal to each other?

(ii) Find $\mathbf{h}'\mathbf{Dividend}$. Are $\mathbf{h}$ and $\mathbf{Dividend}$ orthogonal to each other?

(iii) Find $\mathbf{h}'\mathbf{x}$. Are $\mathbf{h}$ and $\mathbf{x}$ orthogonal to each other?

(iv) Find $\mathbf{h}'\mathbf{y}$. Are $\mathbf{h}$ and $\mathbf{y}$ orthogonal to each other?

2.2 Matrix

Definition 2.6: A $n \times k$ matrix X is defined as

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}.$$

If $n = k$, the matrix is called a square matrix.

Definition 2.7: The transpose of X is defined as

$$X' = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}' = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{pmatrix}.$$

Definition 2.8: The determinant of a 2 by 2 matrix $X = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ is written as $|X|$ and is equal to $ad - bc$.

The determinant of a k by k matrix is more complicated. One may calculate it with the help of a computer. This note will use 2 by 2 matrices as examples for simplicity.

For two k by k matrices X and Y, the determinant has the following properties:

(a) $|X| = |X'|$.

(b) If there is a zero row or zero column in X, then $|X| = 0$.

(c) If any two rows (columns) are linearly dependent, then $|X| = 0$.

(d) The determinant of XY equals the product of their determinants, i.e., $|XY| = |X|\,|Y|$.

Definition 2.9: The trace of a square matrix, written as tr(X), is the sum of the diagonal elements. In the above 2 by 2 matrix, tr$(X) = a + d$.

For k by k matrices X and Y, the trace has the following properties:

(a) tr(X) =tr(X') .

(b) tr$(X \pm Y)$ =tr(X) $\pm$tr(Y) .

(c) tr(XY) =tr(YX) .

(d) tr$(Y^{-1}XY)$ =tr(X) .

(e) For any constant c, tr$(cX) = c$tr(X).

Exercise 2.4: True/ False. Explain. For a k by k matrices X and Y,

(i)$|X + Y| = |X| + |Y|$.

(ii) tr(XY)=tr(X) $\times$tr(Y) .

Hint: If the statement is true, prove it mathematically. If the statement is false, give a counter example.

Example 2.1: Let $X = \begin{pmatrix} 1 & 2 & 2 \\ 3 & 4 & 5 \end{pmatrix}$, $Y = \begin{pmatrix} 2 & 3 & 1 \\ 6 & 0 & 1 \end{pmatrix}$.

Find

(a) X' and Y'.

(b) $X + Y$.

(c) $X'Y$, $Y'X$. XY' and YX'.

(d) Are **x** and **y** linearly independent.

Solution:

(a) $X' = \begin{pmatrix} 1 & 3 \\ 2 & 4 \\ 2 & 5 \end{pmatrix}$ and $Y' = \begin{pmatrix} 2 & 6 \\ 3 & 0 \\ 1 & 1 \end{pmatrix}$;

(b) $X + Y = \begin{pmatrix} 3 & 5 & 3 \\ 9 & 4 & 6 \end{pmatrix}$;

(c) $X'Y = \begin{pmatrix} 1 & 3 \\ 2 & 4 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 6 & 0 & 1 \end{pmatrix}$

$$= \begin{pmatrix} 1\times 2+3\times 6 & 1\times 3+3\times 0 & 1\times 1+3\times 1 \\ 2\times 2+4\times 6 & 2\times 3+4\times 0 & 2\times 1+4\times 1 \\ 2\times 2+5\times 6 & 2\times 3+5\times 0 & 2\times 1+5\times 1 \end{pmatrix} = \begin{pmatrix} 20 & 3 & 4 \\ 28 & 6 & 6 \\ 34 & 6 & 7 \end{pmatrix};$$

$$Y'X = \begin{pmatrix} 2 & 6 \\ 3 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 \\ 3 & 4 & 5 \end{pmatrix} = \begin{pmatrix} 20 & 28 & 34 \\ 3 & 6 & 6 \\ 4 & 6 & 7 \end{pmatrix} = (X'Y)';$$

$$XY' = \begin{pmatrix} 1 & 2 & 2 \\ 3 & 4 & 5 \end{pmatrix} \begin{pmatrix} 2 & 6 \\ 3 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 10 & 8 \\ 23 & 23 \end{pmatrix};$$

$$YX' = \begin{pmatrix} 2 & 3 & 1 \\ 6 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 4 \\ 2 & 5 \end{pmatrix} = \begin{pmatrix} 10 & 23 \\ 8 & 23 \end{pmatrix} = (XY')'.$$

Exercise 2.5: Let $X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, Y = \begin{pmatrix} 2 & 3 \\ 6 & 0 \end{pmatrix}$. Find

(a) X' and Y'.

(b) $X + Y$.

(c) $X'Y$, $Y'X$. XY' and YX'.

Definition 2.10: The row (column) rank of a matrix is the maximum number of linearly independent rows (columns).

Example 2.2: Both the row rank and column of $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ is 1.

Definition 2.11: A 2 by 2 symmetric matrix is of the form $X = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$. It has the property that $X' = X$.

Definition 2.12: A 2 by 2 diagonal matrix is of the form $X = \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix}$.

Definition 2.13: A 2 by 2 identity matrix is defined as $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Exercise 2.6: $X'X = I$ if and only if $X = I$. True or False? Explain.

Definition 2.14: The inverse of a square matrix X is denoted as X^{-1}, it has the property that $X^{-1}X = XX^{-1} = I$.

How to find the inverse of an matrix? Consider $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and

$$Y = X^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

$$XY = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} aA + bC & aB + bD \\ cA + dC & cB + dD \end{pmatrix} = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We have four equations four unknowns.

$$aA + bC = 1,$$

$$cA + dC = 0,$$

$$aB + bD = 0,$$

$$cB + dD = 1.$$

Multiply the four equations by c, a, d , b respectively, we have

$$acA + bcC = c,$$

$$acA + adC = 0,$$

$$adB + bdD = 0,$$

$$bcB + bdD = b.$$

The first equation minus the second, and the third minus the fourth, we have

$$(bc - ad)\,C = c,$$

$$(ad - bc)\,B = -b.$$

Then we solve

$$B = \frac{-b}{ad - bc},$$

$$C = \frac{-c}{ad - bc}.$$

Using equations 2 and 3, we also have

$$A = \frac{-d}{c}C = \frac{d}{ad - bc},$$

$$D = -\frac{a}{b}B = \frac{a}{ad - bc}.$$

Thus, the inverse of X is equal to

$$X^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Note:

(a) $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \neq \begin{pmatrix} a^{-1} & b^{-1} \\ c^{-1} & d^{-1} \end{pmatrix}.$

(b) A matrix whose determinant equals zero does not have an inverse.

Example 2.3: Consider a regression model

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

We have the following data

	$i=1$	$i=2$	$i=3$	$i=4$
x_{1i}	3	1	2	0
x_{2i}	1	2	3	4
y_i	2	1	4	5

Define

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & x_{13} & x_{23} \\ 1 & x_{14} & x_{24} \end{pmatrix} = \begin{pmatrix} 1 & 3 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 0 & 4 \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \\ 5 \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_1 \end{pmatrix}, \qquad U = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix},$$

and

$$Y = X\beta + U.$$

The least squares estimator for β is obtained by to minimizing $\sum u_i^2 = \min_\beta U'U = \min_\beta (Y - X\beta)' (Y - X\beta)$. The first-order condition is

$$X' (Y - X\beta) = 0$$

and we have

$$\widehat{\beta} = (X'X)^{-1} X'Y.$$

Thus, we need to find the inverse of $X'X$. Note that

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

For a 3 by 3 matrix, the inverse can be calculated by a computer program, we have

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

$$\begin{aligned} \widehat{\beta} &= \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_1 \end{pmatrix} = (X'X)^{-1} X'Y \\ &= \begin{pmatrix} \frac{299}{36} & -\frac{35}{18} & -\frac{37}{18} \\ -\frac{35}{18} & \frac{5}{9} & \frac{4}{9} \\ -\frac{37}{18} & \frac{4}{9} & \frac{5}{9} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 3 & 1 & 2 & 0 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 4 \\ 5 \end{pmatrix} = \begin{pmatrix} -\frac{7}{2} \\ 1 \\ 2 \end{pmatrix}. \end{aligned}$$

Note: The inverse of a 3 by 3 matrix A is complicated. If A is symmetric of the form $A = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}$, then

$$A^{-1} = \frac{1}{fb^2 - 2ebc + dc^2 + ae^2 - adf} \begin{pmatrix} e^2 - df & bf - ce & cd - be \\ bf - ce & c^2 - af & ae - bc \\ cd - be & ae - bc & b^2 - ad \end{pmatrix}$$

In particular, if $b = c = e = 0$, then A is a diagonal matrix of the form $A = \begin{pmatrix} a & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & f \end{pmatrix}$ and $A^{-1} = \begin{pmatrix} a^{-1} & 0 & 0 \\ 0 & d^{-1} & 0 \\ 0 & 0 & f^{-1} \end{pmatrix}$.

Definition 2.15: A square matrix is orthogonal if $X^{-1} = X'$.

An orthogonal matrix has the following properties:

(a) $X'X = I$.

(b) The columns are vectors with length equal one and are mutually perpendicular.

Let $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be an orthogonal matrix, then

$$X'X = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

A 2 by 2 orthogonal matrix must satisfy the followings.

$$a^2 + c^2 = 1.$$

$$ab + cd = 0.$$

$$b^2 + d^2 = 1.$$

The are many solutions. For example, $a = \frac{4}{5}$, $b = c = \frac{3}{5}$, $d = -\frac{4}{5}$ satisfy the above conditions. Therefore, $X = \begin{pmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{3}{5} & -\frac{4}{5} \end{pmatrix}$ is an orthogonal matrix since $X^{-1} = \begin{pmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{3}{5} & -\frac{4}{5} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{3}{5} & -\frac{4}{5} \end{pmatrix} = X'$.

Exercise 2.7:

(a) Verify that $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ is an orthogonal matrix.

(b) Is $X = \frac{1}{2}\begin{pmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix}$ an orthogonal matrix?

(c) If $X = \begin{pmatrix} \frac{4}{5} & \frac{3}{5} \\ \frac{3}{5} & -\frac{4}{5} \end{pmatrix}$, find $X^2 = XX$ and X^{100}.

Definition 2.16: Let $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be a 2 by 2 matrix, its eigenvalues can be found by setting the determinant of $(X - \lambda I)$ to zero. i.e.,

$$\left| \begin{pmatrix} a & b \\ c & d \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right| = 0,$$

$$\left| \begin{pmatrix} a - \lambda & b \\ c & d - \lambda \end{pmatrix} \right| = 0,$$

$$(a - \lambda)(d - \lambda) - bc = 0,$$

$$\lambda^2 - (a + d)\lambda + ad - bc = 0.$$

The solutions are:

$$\lambda_1 = \frac{1}{2}\left(a + d + \sqrt{(a + d)^2 - 4(ad - bc)}\right),$$

$$\lambda_2 = \frac{1}{2}\left(a + d - \sqrt{(a + d)^2 - 4(ad - bc)}\right).$$

The roots can be simplified to

$$\lambda_1 = \frac{1}{2}\left(a + d + \sqrt{(a - d)^2 + 4bc}\right),$$

Note that the eigenvalues may not be real numbers. The eigenvalues of a matrix has many nice properties.

(1) The determinant of a 2 by 2 matrix is $\lambda_1\lambda_2$.

(2) The trace of a 2 by 2 matrix is equals to $\lambda_1 + \lambda_2$.

In our case

$$\begin{aligned}\lambda_1\lambda_2 &= \frac{1}{2}\left(a+d+\sqrt{(a-d)^2+4bc}\right)\frac{1}{2}\left(a+d-\sqrt{(a-d)^2+4bc}\right)\\ &= \frac{1}{4}\left((a+d)^2-(a-d)^2+4bc\right)\\ &= ad-bc.\end{aligned}$$

$$\lambda_1+\lambda_2=a+d.$$

In general, for a k by k matrix

(1) The determinant is $\lambda_1\lambda_2...\lambda_k$.

(2) The trace of a 2 by 2 matrix is equal to $\lambda_1+\lambda_2+...+\lambda_k$.

Example 2.4: Find the eigenvalues of $A=\begin{pmatrix}6&2\\2&3\end{pmatrix}$

Solution:

$$\left|\begin{pmatrix}6&2\\2&3\end{pmatrix}-\begin{pmatrix}\lambda&0\\0&\lambda\end{pmatrix}\right|=0,$$

$$\left|\begin{pmatrix}6-\lambda&2\\2&3-\lambda\end{pmatrix}\right|=0,$$

$$(6-\lambda)(3-\lambda)-2(2)=0,$$

$$\lambda^2-9\lambda+14=0.$$

$$\lambda_1=\frac{1}{2}\left(9+\sqrt{(-9)^2-4(1)(14)}\right)=7.$$

$$\lambda_2=\frac{1}{2}\left(9-\sqrt{(-9)^2-4(1)(14)}\right)=2.$$

(1) The determinant of A is $\lambda_1\lambda_2=14$.

(2) The trace of Ais equal to $\lambda_1+\lambda_2=9$.

Definition 2.17: Let A be a k by k matrix and λ be its eigenvalue. If $\mathbf{x}$ is a nonzero vector such that $A\mathbf{x} = \lambda\mathbf{x}$, then $\mathbf{x}$ is said to be an eigenvector of A.

Example 2.5:

$$A = \begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix},$$

$$A\mathbf{x} = \lambda\mathbf{x},$$

$$\begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \end{pmatrix},$$

$$6x_1 + 2x_2 = \lambda x_1,$$

$$2x_1 + 3x_2 = \lambda x_2.$$

When $\lambda = 7$,

$$6x_1 + 2x_2 = 7x_1,$$

$$2x_1 + 3x_2 = 7x_2.$$

Thus, we have

$$x_1 = 2x_2,$$

which gives $x_1 = 2x_2$ and there are infinite number of solutions. To normalize the solutions, we impose the condition that $\sqrt{x_1^2 + x_2^2} = 1$. i.e., we require the eigenvectors to have the unit length. Under this condition and $x_1 = 2x_2$, we have $\sqrt{(2x_2)^2 + x_2^2} = 1, x_2 = \frac{1}{\sqrt{5}}$ and $x_1 = \frac{2}{\sqrt{5}}$. So one of the eigenvector is

$$x = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}.$$

Similarly, when $\lambda = 2$,

$$6x_1 + 2x_2 = 2x_1,$$

$$2x_1 + 3x_2 = 2x_2.$$

We have

$$x_2 = -2x_1.$$

The eigenvector is

$$x = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{pmatrix}.$$

Note that the two eigenvectors are orthogonal.

Definition 2.17: The spectral decomposition of a k by k symmetric matrix A can be expressed as

$$A = \sum_{i=1}^{k} \lambda_i e_i e_i',$$

where e is the eigenvector.

Note that $\sum_{i=1}^{k} e_i e_i' = I$.

Example 2.6: Find the spectral decomposition of $A = \begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix}$.

Solution:

$$\begin{aligned}\sum_{i=1}^{k}\lambda_i e_i e_i' &= 7\begin{pmatrix}\frac{2}{\sqrt{5}}\\ \frac{1}{\sqrt{5}}\end{pmatrix}\begin{pmatrix}\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}}\end{pmatrix}+2\begin{pmatrix}\frac{1}{\sqrt{5}}\\ \frac{-2}{\sqrt{5}}\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{5}}\end{pmatrix}\\ &= 7\begin{pmatrix}\frac{4}{5} & \frac{2}{5}\\ \frac{2}{5} & \frac{1}{5}\end{pmatrix}+2\begin{pmatrix}\frac{1}{5} & -\frac{2}{5}\\ -\frac{2}{5} & \frac{4}{5}\end{pmatrix}=\begin{pmatrix}6 & 2\\ 2 & 3\end{pmatrix}=A.\end{aligned}$$

Exercise 2.8: Let $A=\begin{pmatrix}6 & 2\\ 2 & 3\end{pmatrix}$,

(a) Show that $A^{-1}=\begin{pmatrix}\frac{3}{14} & -\frac{1}{7}\\ -\frac{1}{7} & \frac{3}{7}\end{pmatrix}$.

(b) Find the spectral decomposition of A^{-1}.

Exercise 2.9: Let $P_{CP,i}$ $P_{SPA,i}$ $(i=1,2,3,4,5)$ be the daily closing price of [267] Citic Pacific and [19] Swire Pacific A from 15/9/2014-19/9/2014 respectively.

(a) Plot $(P_{CP,i},P_{SPA,i})$ on the $X-Y$ plane.

(b) Calculate the sample variance of $P_{CP,i}$ and $P_{SPA,i}$, called them $s_{11}=\frac{1}{4}\sum_{i=1}^{5}\left(P_{CP,t}-\overline{P_{CP}}\right)^2$ and $s_{22}=\frac{1}{4}\sum_{i=1}^{5}\left(P_{SPA,i}-\overline{P_{SPA}}\right)^2$ respectively.

(c) Calculate the sample covariance $s_{12}=s_{21}=\frac{1}{4}\sum_{i=1}^{5}\left(P_{CP,i}-\overline{P_{CP}}\right)\left(P_{SPA,i}-\overline{P_{SPA}}\right)$.

(d) Let $X=\begin{pmatrix}s_{11} & s_{12}\\ s_{21} & s_{22}\end{pmatrix}$. Find X^{-1}.

(e) Find the spectral decomposition of X.

Exercise 2.10: Let $\mathbf{x}=\begin{pmatrix}6\\ -14\\ 8\end{pmatrix}$, $\mathbf{y}=\begin{pmatrix}-9\\ 6\\ 3\end{pmatrix}$. Find

(a) $\mathbf{x}'$ and $\mathbf{y}'$.

(b) L_x and L_y.

(c) $\mathbf{x}+\mathbf{y}$.

(d) $\mathbf{x}'\mathbf{y}$. Are $\mathbf{x}$ and $\mathbf{y}$ orthogonal?

(e) Repeat (a) to (d) if $\mathbf{x} = \begin{pmatrix} -7 \\ 4 \\ 3 \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}.$

Exercise 2.11: Let $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$,

(a) Show that $A^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$.

(b) Find the spectral decomposition of A^{-1}.

Exercise 2.12: Let

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

(a) Find A^2, A^3 and A^n.

(b) Write down A^{-1}. Verify that $A^{-1}A = I$.

(c) Find the spectral decomposition of $A'A$.

Exercise 2.13: Let $A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$,

(a) Find A^{-1}.

(b) Find the spectral decomposition of A^{-1}.

Exercise 2.14: Let $A = \begin{pmatrix} 2^{-1} & 0 \\ 0 & 2 \end{pmatrix}$,

(a) Find A^n.

(b) Find A^{-n}.

(c) Find the Eigenvalues of A^n and A^{-n}.

(d) Find the spectral decomposition of A^n and A^{-n}.

Exercise 2.15: True/False.

(i). If the eigenvalue of a square matrix equals zero, then the matrix is of full rank.

(ii) Let X be a k by k matrix, then $X'X = I$ if and only if $X = I$.

(iii) Let X and Y be two square matrices, and $|X|$ and $|Y|$ be their determinants respectively, then $|X + Y| = |X| + |Y|$.

(iv) Let X and Y be two square matrices, and then $trace\,(XY) = trace\,(X) \times trace\,(Y)$.

Chapter 3

Inference about a Mean Vector

3.1 Point Estimation

Population and sample are two different concepts. We would like to estimate the unknown mean (μ) and the unknown variance (σ^2) of a population. Given limited resources, what we can do is to draw a sample from the population. A sample is a subset of a population. We hope that the sample will be representive enough for us to retrieve the information of a population. One can construct estimators to estimate the population mean and variance.

Definition 3.1: An **estimator** is a rule or formula to estimate an unknown population quantity, such as the population mean and population variance.

An estimator is usually constructed based on the sample information. It is a random variable since it takes different values under different samples. As a random variable, an estimator itself has a mean, a variance and a distribution.

Definition 3.2: An **estimate** is the numerical value taken by an estimator, it usually depends on the sample drawn.

Example 3.1: Suppose we have a sample of size n, the sample mean

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

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 3.2: The statistic

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

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

Example 3.3: A weighted average

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

is also estimator of the population mean.

Example 3.4: A single observation X_1 is also an estimator of the population mean.

Example 3.5: 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.

Example 3.6:

$$X^* = \frac{X_1^2 + X_2^2 + ... + X_n^2}{n}$$

can also be estimator of the population mean.

Thus, there are a lot of estimators for the population mean. The problem is how to select the best one, and what criteria should be used to evaluate an estimator. In choosing the best estimator, we usually use criterion such as linearity, unbiasedness and efficiency. The first criterion in choosing estimator is linearity, a linear estimator is by construction simpler than a nonlinear estimator. The mean and variance of a linear estimator are easier to compute compared to those of a nonlinear estimator.

Definition 3.2: An estimator $\widehat{X}$ is **linear** if it is a linear combination of the sample observations. i.e.,

$$\widehat{X} = a_1 X_1 + a_2 X_2 + ... + a_n X_n,$$

where a_i $(i = 1, 2, ..., n)$ takes a value between zero and one. In some cases, they can be negative or larger than 1, and some of them can be zero. If all a_i are zero, then $\widehat{X}$ is no longer an estimator. Thus, estimators in examples 3.1-3.4 are linear, while estimators in example 3.5 and 3.6 are not linear. The reason why the linear estimator is a desirable estimator because its mean and variance are easy to calculate. For example, the estimator in example 3.6 is nonlinear, and its mean and variance are difficult to obtain. We reduce the set of all possible estimators to the set of linear estimators. Still, there are plenty of linear estimators, so how should they be compared? We introduce the concept of unbiasedness.

Definition 3.3: An linear 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.,

$$E(X_i) = \mu, \qquad i = 1, 2, ..., n.$$

This is because when an observation is drawn from a population, we expect it 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 3.7: If X_i $(i = 1, 2, ..., n)$ are random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Show that:

(a) $\overline{X} = \frac{\sum\limits_{i=1}^{n} X_i}{n}$ is an unbiased estimator for μ.

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

(c) Show that $\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum\limits_{i=1}^{n} X_i^2 - n\left(\overline{X}\right)^2$.

(d) Use (a) and (c), show that $s^2 = \frac{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$ is an unbiased estimator for σ^2.

Solution:(a)

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n\mu}{n} = \mu. \qquad \blacksquare$$

(b)

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

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

Also,

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

■

(c)

$$\begin{aligned} \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 &= \sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right) \\ &= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\overline{X}^2. \end{aligned}$$

■

(d)

$$\begin{aligned} E\left(s^2\right) &= E\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}\right) \\ &= E\left(\frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}\right) \\ &= \frac{\sum_{i=1}^{n} E\left(X_i^2\right) - nE\left(\overline{X}^2\right)}{n-1} \\ &= \frac{n\left(\sigma^2 + \mu^2\right) - n\left(\sigma^2 \ / \ n + \mu^2\right)}{n-1} \\ &= \frac{n-1}{n-1}\sigma^2 \\ &= \sigma^2. \end{aligned}$$

■

Exercise 3.1: Show that the estimators in examples 3.1, 3.3 and 3.4 are unbiased, and that the estimators in examples 3.2, 3.5 and 3.6 are biased.

Still, there are many linear and unbiased estimators, how should we compare them? Here, we introduce the concept of efficiency.

Definition 3.4: An estimator $\widehat{X}$ is more **efficient** than another estimator X^* if $Var\left(\widehat{X}\right) < Var\left(X^*\right).$

Example 3.8: If we look at the efficiency criteria, the estimator in example 3.5 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 gives us no information about the population mean. Thus, despite the fact that it is efficient, it is not a good estimator.

Exercise 3.2: 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 3.3: Rank the efficiency of the estimators in examples 3.1 to 3.5.

Definition 3.5: An estimator $\widehat{X}$ is a **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 $Var\left(\widehat{X}\right) \to 0$ as the sample size goes to infinity. If the estimator truly reveals the value of the population mean μ, the variation of this estimator should 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 a 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 understanding of asymptotics. It is very important for an estimator to be consistent since we would like 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

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}.$$

Note that it is unbiased as

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

Second, suppose the variance of X_i, $Var\left(X_i\right) = \sigma^2 < \infty$ for $i = 1, 2, ...n$, then

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

Note that consistency and unbiasedness do not imply each other. An estimator can be biased but consistent. Consider the estimator in example 3.2,

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

For any given value of sample size n,

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

The bias is

$$\frac{1}{n}\mu$$

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

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

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

An estimator can also be unbiased but inconsistent. Consider the estimator in example 3.4, 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 n, no matter how large n is. Thus, increasing the number of other observations cannot improve the precision of this estimator.

In general, consistency is a concept for both linear and nonlinear estimators, while unbiasedness is a concept for linear estimators only. This is because it is hard to evaluate the expected value of a nonlinear estimator.

Exercise 3.4: Construct an estimator which is biased, consistent and less efficient than the simple average $\overline{X}$.

Exercise 3.5: 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 n observations $X_1, X_2, ..., X_n$ on people's life span, construct a consistent estimator for c and explain why it is consistent.

3.2 The Law of Large Numbers and the Central Limit Theorem

Definition 3.6: A sequence of random variables X_i, $(i = 1, 2, ...n)$ follow an **Independent and Identical Distribution (i.i.d.)** if all the X_i 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_i 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 3.6: To illustrate the Law of Large Numbers, consider the random experiment of throwing a dice n times. Let X_i be the outcome at the i trial, $i = 1, 2, .., n$. Let $\overline{X}$ be the sample average of these X_i.

(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 $n = 1$? $n = 2$? $n = 3$?

(c) Conduct the experiment, record the value of $\overline{X}$ and plot a diagram which indicates its behavior as n increases from 1 to 30. Does $\overline{X}$ converge to 3.5?

Theorem 3.1: The **Central Limit Theorem** states that, if X_i 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 $\dfrac{\sigma^2}{n}$, as the sample size n goes to infinity.

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

Example 3.9: 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

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

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

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

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 3.7: To illustrate the Central Limit Theorem, let us consider the random experiment of throwing a dice n times in the previous exercise.

(a) Conduct the experiment yourself with $n = 30$. 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 obtain 20 values of $\overline{X}$.

(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 four other classmates and pool the result of 100 values of $\overline{X}$.

Exercise 3.8: Use a computer or a calculator 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 $Y_i = \sqrt{36}\left(\overline{X}_i - 0.5\right)$, $i = 1, 2, ..., 100$. Now make two frequency tables of Y_i with the length of each interval 0.01 and 0.1 respectively. Plot the two histograms.

3.3 Testing a Statistical Hypothesis

When we observe a phenomenon, we would like to explain it by a hypothesis. We usually post a null hypothesis, and an alternative hypothesis. The two hypotheses should be complementary. For example, when we observe that the death toll in winter is usually higher than the death toll in the other seasons, we may conjecture that the death toll is negatively related to temperature. The alternative hypothesis would be that the death toll has nothing to do with or is positively related to temperature. A hypothesis is not a theorem. A theorem is always true under certain assumptions. A hypothesis is just a conjecture, we have to test how likely a hypothesis is going to be correct. In testing a hypothesis, we may commit errors when making conclusion. There are two possible types of errors:

Definition 3.7: The 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 3.8: The 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 would like to reduce both Type I and Type II errors as much as we can. However, as there is no free lunch, there is no way to reduce both errors at the same time. Reducing the chance of committing Type I Error will increase the chance of committing Type II Error, and vice versa.

Exercise 3.9: 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 the Type I Error in this case? How will this affect the chance of committing the Type II Error?

(e) How can you fully eliminate the Type II Error in this case? How will this affect the chance of committing the 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 Test for mean when σ^2 is known

Consider a random sample X_1, X_2,...X_n 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

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Under $H_0 : \mu = \mu_0$, $X_i \sim N(\mu_0, \sigma^2)$. Since the sum of normal random variable is also normal, as a result, $\overline{X}$ is also normally distributed for all sample size n, no matter n is small or large. Thus, $\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$. 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

$$\left(\overline{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right).$$

If μ_0 does not fall into this interval, we reject H_0 at the significance level α.

This test is of limited use since we have two very strong assumptions: (i) the observations X_i 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 Test for mean when σ^2 is unknown

Consider a random sample X_1, X_2,...X_n drawn from a **normal** distribution with unknown mean μ and **unknown variance** σ^2. We would like to test whether μ equals a particular value μ_0.

$$H_0 : \mu = \mu_0.$$

We construct a test statistic, defined as

$$t_{obs} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}},$$

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

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}.$$

Note that

$$t_{obs} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2}}.$$

Under $H_0 : \mu = \mu_0$, $X_i \sim N\left(\mu_0, \sigma^2\right)$. As a result,

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

and

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

Further, it can be shown that (difficult)

$$\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

has a Chi-squared distribution with degrees of freedom $(n-1)$, and that (also difficult)

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

and

$$\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

are independent. Recall the definition of t-distribution that,

$$t_{obs} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2}} = \frac{N\left(0,1\right)}{\sqrt{\chi^2_{n-1}/\left(n-1\right)}}$$

will have a t-distribution with degrees of freedom $(n-1)$.

In the two-sided case (i.e., $H_1 : \mu \neq \mu_0$), we reject H_0 at a significance level α if $|t| > t_{\frac{\alpha}{2},n-1}$. For example, $t_{0.025,9} = 2.262$. In the one-sided case (i.e., $H_1 : \mu > (<)\mu_0$), we reject H_0 at a significance level α if $t > t_{\alpha,n-1}$ $(t < -t_{\alpha,n-1})$.

A $100\left(1-\alpha\right)\%$ **confidence interval** for μ is

$$\left(\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right).$$

If μ_0 does not fall into this interval, we reject H_0 at the significance level α.

Example 3.10: Suppose the body height of the population of Hong Kong is normally distributed $N\left(\mu, \sigma^2\right)$. Suppose we would like to test the hypothesis that the mean height of the population of Hong Kong is μ =160cm. We test this based on a sample of 10 individuals, 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 s =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

$$t_{obs} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{165 - 160}{5/\sqrt{10}} = 3.163.$$

t_{obs} will have a t-distribution with degrees of freedom equal $n-1$. In the two-sided case, we reject H_0 at the significance level α if $|t_{obs}| > t_{\frac{\alpha}{2},n-1}$. Now, let $\alpha = 5\%$, then

$$t_{0.025,9} = 2.262.$$

Since $|t_{obs}| > t_{0.025,9}$, we reject H_0 at $\alpha = 5\%$. Thus, we are 95% sure that the population mean is not equal to 160cm.

A 95% **confidence interval** for μ is

$$\overline{X} \mp t_{0.025,9}\left(\frac{s}{\sqrt{10}}\right) = 165 \mp 2.262\left(\frac{5}{\sqrt{10}}\right) = (161.4, 168.6)\,.$$

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 3.10: A random sample of size $n = 12$ from a normal population has the sample mean $\overline{X} = 28$ and sample variance $s^2 = 3$.

(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 3.11: Let $r_i = \ln P_i - \ln P_{i-1}$ be the daily return of [1] Cheung Kong on day i. Assume that $r_i \sim N(\mu, \sigma^2)$. Consider a sample of r_i from 22/9/14 to 26/9/14.

(a) Find $\overline{r}$ and s^2

(b) Use t-test to test the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ at $\alpha = 5\%$.

(c) Construct a 95% confidence interval for the population mean μ.

Exercise 3.12: Let X_i be the monthly total number of deaths in Hong Kong. Assume that $X_i \sim N(\mu, \sigma^2)$. Consider a sample of X_i from September 2013 to August 2014.

(a) Find $\overline{X}$ and s^2.

(b) Use t-test to test the hypothesis $H_0 : \mu = 3000$ against $H_1 : \mu < 3000$ at $\alpha = 5\%$.

Exercise 3.13: Let X_i be the monthly total number of marriages in Hong Kong. Assume that $X_i \sim N(\mu, \sigma^2)$. Consider a sample of X_i from September 2013 to August 2014.

(a) Find $\overline{X}$ and s^2.

(b) Use t-test to test the hypothesis $H_0 : \mu = 3000$ against $H_1 : \mu > 3000$ at $\alpha = 5\%$.

3.6 Bivariate Normal Distribution

Recall that 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

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

When there are two independent random variables which are jointly normally distributed, their joint density can be expressed as

$$\begin{aligned} f(x_1,x_2) &= f(x_1)\,f(x_2) \\ &= \frac{1}{\sqrt{2\pi\sigma_{11}}}\exp\left(-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2\right)\times\frac{1}{\sqrt{2\pi\sigma_{22}}}\exp\left(-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2\right) \\ &= \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}}}\exp\left(-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2\right]\right). \end{aligned}$$

If the two variables are not independent but have a correlation ρ_{12} , we have

$$\begin{aligned} &f(x_1,x_2) \\ = \;& \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}\left(1-\rho_{12}^2\right)}}\times \\ &\exp\left(-\frac{1}{2\left(1-\rho_{12}^2\right)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right). \end{aligned}$$

Let

$$\mathbf{\Omega} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix},$$

$$\begin{aligned} \mathbf{\Omega}^{-1} &= \frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{pmatrix} \\ &= \frac{1}{\sigma_{11}\sigma_{22}\left(1-\rho_{12}^2\right)}\begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{pmatrix}. \end{aligned}$$

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}.$$

Exercise 3.14: Let x_1 and x_2 be jointly normal $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix},\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

(a) Find the joint density $f(x_1,x_2)$.

(b) Use the computer to plot $f(x_1,x_2)$ for $\rho = 0,\ 0.8,\ -0.8,\ 1,\ -1$.

3.7 Multivariate Normal Distribution

In general, for a random vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$, if the variables are jointly normally distributed $N(\boldsymbol{\mu}, \boldsymbol{\Omega})$, we have

$$f(x_1,x_2,...,x_{\mathrm{p}}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Omega}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right),$$

where $|\boldsymbol{\Omega}|$ is the determinant of $\boldsymbol{\Omega}$.

Contours of constant density for the p dimensional normal distribution are ellipsoids defined by $\mathbf{x}$ such that

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$

The solid ellipsoid of $\mathbf{x}$ values satisfying

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \chi_p^2 (\alpha)$$

has probability $1 - \alpha$.

Example 3.11: Contours of constant density for the one dimensional normal distribution are ellipsoids defined by x such that

$$\left(\frac{x - \mu}{\sigma} \right)^2 = c^2.$$

The solid ellipsoid of x values satisfying

$$\left(\frac{x - \mu}{\sigma} \right)^2 \leq \chi_1^2 (\alpha).$$

Suppose $\alpha = 5\%$, $\mu = 2$, $\sigma^2 = 9$, then the solid ellipsoid of x is the values of x such that

$$\left(\frac{x - 2}{3} \right)^2 \leq \chi_1^2 (0.05) = 3.84.$$

For example, $x = 11$ will not be in this solid ellipsoid, while $x = 5$ will be in this ellipsoid.

Example 3.12: Let x_1 and x_2 be jointly normal $N\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right)\right)$. Contours of constant density for this two dimensional normal distribution are ellipsoids defined by $\mathbf{x}$ such that

$$(\mathbf{x} - \mathbf{0})' \, \mathbf{I}^{-1} \, (\mathbf{x} - \mathbf{0}) = c^2,$$

or

$$\mathbf{x}'\mathbf{x} = c^2.$$

This implies

$$x_1^2 + x_2^2 = c^2,$$

which is a circle on the plane of x_2 vs x_1. The solid ellipsoid of $\mathbf{x}$ values satisfying

$$x_1^2 + x_2^2 \leq \chi_2^2 \, (\alpha)$$

has probability $1 - \alpha$.
Suppose $\alpha = 5\%$, then the solid ellipsoid of x is the values of x such that

$$x_1^2 + x_2^2 \leq \chi_2^2 \, (0.05) = 5.99.$$

For example, $\mathbf{x} = \left(\begin{array}{c} 2 \\ 2 \end{array}\right)$ will not be in this solid ellipsoid, while $\mathbf{x} = \left(\begin{array}{c} 1 \\ -1 \end{array}\right)$ will be in this ellipsoid.

Exercise 3.15: Let x_1 and x_2 be jointly normal $N\left(\left(\begin{array}{c} 1 \\ 2 \end{array}\right), \left(\begin{array}{cc} 2 & 2 \\ 2 & 4 \end{array}\right)\right)$.
(a) Find the joint density $f\,(x_1, x_2)$.

(b) Use the computer to plot $f(x_1, x_2)$.

(c) Is the point (10,-10) in the ellipsoid with $\alpha = 5\%$?

Exercise 3.16: For a random vector $\mathbf{x} = (x_1,x_2,...,x_{20})'$, if the variables are jointly normally distributed $N(\boldsymbol{\mu}, \boldsymbol{\Omega})$, then the joint density function is $f(x_1,x_2,...,x_{20}) = \dfrac{1}{(2\pi)^{10} |\boldsymbol{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x}'\boldsymbol{\Omega}^{-1}\mathbf{x}\right)$, where $|\boldsymbol{\Omega}|$ is the determinant of $\boldsymbol{\Omega}$. True/False?

3.8 Hotelling's T^2

Now consider testing the mean vector of a bivariate normal distribution. Our null hypothesis is $H_0 : \mu = \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}_0$ is a 2 by 1 vector. The data matrix is $X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{21} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$ is n by 2. A natural generalization is to use

$$T^2 = n\left(\overline{X} - \boldsymbol{\mu}_0\right)' S^{-1} \left(\overline{X} - \boldsymbol{\mu}_0\right),$$

where

$$\overline{X} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_{i1} \\ \frac{1}{n}\sum_{i=1}^n x_{i2} \end{pmatrix},$$

$$S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix},$$

$$s_{11} = \frac{1}{n-1}\sum_{i=1}^n \left(x_{i1} - \overline{x_1}\right)^2,$$

$$s_{12} = s_{21} = \frac{1}{n-1}\sum_{i=1}^n \left(x_{i1} - \overline{x_1}\right)\left(x_{i2} - \overline{x_2}\right),$$

$$s_{22} = \frac{1}{n-1}\sum_{i=1}^n \left(x_{i2} - \overline{x_2}\right)^2.$$

The statistic T^2 is called Hotelling's T^2. It is distributed as

$$\frac{2\,(n-1)}{n-2}F_{2,n-2}.$$

Example 3.13: Let the data matrix for a random sample of size $n = 3$ from a bivariate normal population be

$$X = \begin{pmatrix} 6 & 9 \\ 10 & 6 \\ 8 & 3 \end{pmatrix}.$$

Evaluate the observed T^2 for $H_0 : \boldsymbol{\mu} = \begin{pmatrix} 9 \\ 5 \end{pmatrix}$. What is the sampling distribution of T^2 in this case? Should we reject H_0 at 5% level?

Solution:

The mean vector is

$$\overline{X} = \begin{pmatrix} \frac{6+10+8}{3} \\ \frac{9+6+3}{3} \end{pmatrix} = \begin{pmatrix} 8 \\ 6 \end{pmatrix},$$

$$\sum_{i=1}^{3} (x_{i1} - \overline{x_1})^2 = (6-8)^2 + (10-8)^2 + (8-8)^2\,,$$

$$\sum_{i=1}^{3} (x_{i1} - \overline{x_1})\,(x_{i2} - \overline{x_2}) = (6-8)\,(9-6) + (10-8)\,(6-6) + (8-8)\,(3-6)\,,$$

$$\sum_{i=1}^{3} (x_{i2} - \overline{x_2})^2 = (9-6)^2 + (6-6)^2 + (3-6)^2\,.$$

$$\begin{aligned} S &= \begin{pmatrix} \frac{1}{2}\sum_{i=1}^{3} (x_{i1} - \overline{x_1})^2 & \frac{1}{2}\sum_{i=1}^{3} (x_{i1} - \overline{x_1})\,(x_{i2} - \overline{x_2}) \\ \frac{1}{2}\sum_{i=1}^{3} (x_{i1} - \overline{x_1})\,(x_{i2} - \overline{x_2}) & \frac{1}{2}\sum_{i=1}^{3} (x_{i2} - \overline{x_2})^2 \end{pmatrix} \\ &= \begin{pmatrix} 4 & -3 \\ -3 & 9 \end{pmatrix}. \end{aligned}$$

$$S^{-1}=\begin{pmatrix}4 & -3\\ -3 & 9\end{pmatrix}^{-1}=\frac{1}{4\times 9-(-3)(-3)}\begin{pmatrix}9 & 3\\ 3 & 4\end{pmatrix}=\begin{pmatrix}\frac{1}{3} & \frac{1}{9}\\ \frac{1}{9} & \frac{4}{27}\end{pmatrix},$$

$$\begin{aligned}T^2 &= 3\left(\begin{pmatrix}8\\6\end{pmatrix}-\begin{pmatrix}9\\5\end{pmatrix}\right)'\begin{pmatrix}\frac{1}{3} & \frac{1}{9}\\ \frac{1}{9} & \frac{4}{27}\end{pmatrix}\left(\begin{pmatrix}8\\6\end{pmatrix}-\begin{pmatrix}9\\5\end{pmatrix}\right)\\ &= 3\begin{pmatrix}-1 & 1\end{pmatrix}\begin{pmatrix}\frac{1}{3} & \frac{1}{9}\\ \frac{1}{9} & \frac{4}{27}\end{pmatrix}\begin{pmatrix}-1\\1\end{pmatrix}\\ &= \frac{7}{9}.\end{aligned}$$

The sampling distribution of T^2 is

$$\frac{2(3-1)}{3-2}F_{2,3-2}=4F_{2,1}.$$

Note that at $\alpha=5\%$, $F_{2,1}=199.5$, and $4F_{2,1}=798$. Since $\frac{7}{9}<798$, we do not reject H_0 at $\alpha=5\%$.

In general, if there are p variables and n observations, the sampling distribution of T^2 is

$$\frac{p(n-1)}{n-p}F_{p,n-p}.$$

Exercise 3.17: Let $\mathbf{X}$ be the data matrix for a random sample of size $n=3$ from a bivariate normal population. Find the sampling distribution of T^2 and evaluate the observed T^2 for $\boldsymbol{\mu}_0$ when

(a) $\mathbf{X}=\begin{pmatrix}0 & -5\\ 9 & 5\\ 18 & 15\end{pmatrix}$, $\boldsymbol{\mu}_0=\begin{pmatrix}9\\5\end{pmatrix}$

(b) $\mathbf{X}=\begin{pmatrix}6 & -9\\ 14 & 6\\ 10 & -3\end{pmatrix}$, $\boldsymbol{\mu}_0=\begin{pmatrix}8\\-1\end{pmatrix}$

(c) $\mathbf{X} = \begin{pmatrix} 6 & -9 \\ -1 & 6 \\ -2 & 3 \end{pmatrix}$, $\boldsymbol{\mu}_0 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

(d) $\mathbf{X} = \begin{pmatrix} -7 & 2 \\ 4 & 1 \\ 3 & 3 \end{pmatrix}$. $\boldsymbol{\mu}_0 = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$.

(e) $\mathbf{X} = \begin{pmatrix} 6 & -9 \\ -14 & 6 \\ 8 & 3 \end{pmatrix}$. $\boldsymbol{\mu}_0 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

3.9 What if $\mathbf{X}_i$ 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 with observations X_1, X_2,...X_n drawn from **any** distribution with unknown finite mean μ and a finite **unknown variance** σ^2. We would like to test whether μ equals a particular value μ_0.

$$H_0 : \mu = \mu_0.$$

If the sample size is small, say if $n < 30$, then the hypothesis cannot be easily tested since we do not know the behavior of the sample mean $\overline{X}$ and sample variance s^2 if X_i is not normally distributed. However, if the sample size is large, say $n > 30$, we can apply the Central Limited Theorem that $\overline{X}$ is normally distributed and the Law of Large Numbers that s^2 will converge to the population variance σ^2. Then, the test statistic

$$Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

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

$$\overline{X} \mp Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

If μ_0 does not fall into this interval, we reject H_0 at the significance level α. Thus, if the observations X_i are not normal, we need a large sample to perform the test.

Exercise 3.18: A random sample of size $n = 100$ from a population has the sample mean $\overline{X} = 28$ and sample variance $s^2 = 3$.

(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 the observations come from a normal distribution.)

Exercise 3.19: True/False.

(a) Rejection of the null hypothesis when it is true is called the Type I Error.

(b) In general, if there are p variables and n observations, the sampling distribution of T^2 is $\frac{p\,(n-1)}{n-p} F_{p,n-p}$.

(c). The Central Limit Theorem states that the sample average has a uniform distribution when sample size is large.

Chapter 4

Regression

4.1 Introduction

Suppose a variable Y, referred to as the dependent variable, is related to another variable X, called independent or explanatory variable. If the relationship between Y and X is linear, then we have:

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

where β_0 and β_1 are constants.

This is an **exact** (or **deterministic**) linear relationship. An exact linear relationship is the exception rather than rule. In most situations, X and Y may not be perfectly linearly related. There may be other unknown factors that also affect Y, we use u to represent all these unknown factors, and estimate the following regression model

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

Regression is a statistical technique that is used to explain the relationship among variables. For example, if Y is consumption and X is income, then the above model is a consumption function. The value of β_1 indicates that if income increases 1 by dollar, consumption will increase by β_1 dollar. β_0 is the consumption when income is zero.

We would like to estimate the **unknown** parameter β_0 and β_1 based on our sample observations $\{X_i, Y_i\}_{i=1}^n$. We plot the observations and draw a line which fits these observations the best. What criteria should we use? In general, we minimize the "distance" between the observations and the line. We may use vertical distance, horizontal distance or a distance perpendicular to the line. In regression analysis, we use the vertical distance, since Y is the variable of interest. However, we are not just minimizing the sum of errors, as it is possible that the positive errors and negative errors may cancel out each other, ending up with a small value of net errors. We may take absolute values, but we cannot find the optimal estimator in that case by using simple calculus. In addition, we would like to penalize observations which are far away from the line. Thus, we minimize the sum of squared errors. This is called the **Ordinary Least Squares** (**OLS**) estimation method, proposed by Adrien Legendre, a French mathematician in the 19th century. 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 make the following assumptions:

4.1.1 Assumptions

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

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

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

e.g., $Y_i = \beta_0 + \beta_1 X_i^2 + u_i$ is a linear model, while $Y_i = \beta_0 + \beta_1^2 X_i + u_i$ is not.

This assumption allows us to derive the OLS estimator $\widehat{\beta}_0$and $\widehat{\beta}_1$ via simple calculus.

2: $E(u_i) = 0 \qquad$ for all i.

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_i cannot be all the same.

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

4: X_i is given and is non-random, in the sense that one can choose the values of X_i.

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, $E\left(X_i u_i\right) = X_i E\left(u_i\right) = 0$ by assumption 2. This also implies $Cov\left(X_i, u_i\right) = 0$.

5: Homoscedasticity, i.e., $Var\left(u_i\right) = \sigma^2 \qquad \text{for all } i$.

6: Serial Independence, i.e., $Cov\left(u_i, u_s\right) = 0 \qquad \text{for all } i \neq s$.

Assumptions 5 and 6 simplify the calculation of $Var\left(\widehat{\beta}_0\right)$ and $Var\left(\widehat{\beta}_1\right)$. They also ensure that the 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**).

4.2 Least Squares Estimation

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

$$u_i = Y_i - \beta_0 - \beta_1 X_i.$$

The problem is

$$\min_{\beta_0,\beta_1}\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2.$$

The first-order conditions are:

$$\left.\frac{\partial\sum\limits_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2}{\partial\beta_0}\right|_{\widehat{\beta}_0,\widehat{\beta}_1}=-2\sum_{i=1}^{n}\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1X_i\right)=0, \qquad (*)$$

$$\left.\frac{\partial\sum\limits_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2}{\partial\beta_1}\right|_{\widehat{\beta}_0,\widehat{\beta}_1}=-2\sum_{i=1}^{n}\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1X_i\right)X_i=0. \qquad (**)$$

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

$$\widehat{\beta}_1=\frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)Y_i}{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2},$$

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

Note: If X is also a random variable, then when sample size increases, $\widehat{\beta}_1$ will converge to $\frac{Cov(X,Y)}{Var(X)}$.

Example 4.1: Show that

$$\widehat{\beta}_1=\beta_1+\frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)u_i}{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}.$$

Solution:

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right) Y_i}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(\beta_0 + \beta_1 X_i + u_i\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\
&= \beta_0 \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} + \beta_1 \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right) X_i}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} + \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right) u_i}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\
&= \beta_0 \frac{0}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} + \beta_1 (1) + \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right) u_i}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\
&= \beta_1 + \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right) u_i}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}.
\end{aligned}$$

Exercise 4.1: Solve (*) and (**) for $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

4.3 Properties of OLS 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_i.

Proof.

$$\widehat{\beta}_1 = \frac{\sum\limits_{i=1}^n \left(X_i - \overline{X}\right) Y_i}{\sum\limits_{i=1}^n \left(X_i - \overline{X}\right)^2} = \frac{X_1 - \overline{X}}{\sum\limits_{i=1}^n \left(X_i - \overline{X}\right)^2} Y_1 + \frac{X_2 - \overline{X}}{\sum\limits_{i=1}^n \left(X_i - \overline{X}\right)^2} Y_2 + ... + \frac{X_n - \overline{X}}{\sum\limits_{i=1}^n \left(X_i - \overline{X}\right)^2} Y_n = \sum_{i=1}^n a_i Y_i,$$

where

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

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

where

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

(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 4.1,

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)u_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}.$$

Thus

$$\begin{aligned}
E\left(\widehat{\beta}_1\right) &= E\left(\beta_1 + \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)u_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right) = \beta_1 + \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)E\left(u_i\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\times 0}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \beta_1.
\end{aligned}$$

■

$$
\begin{aligned}
E\left(\widehat{\beta}_0\right) &= E\left(\overline{Y}-\overline{X}\widehat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^n Y_i}{n}\right) - \overline{X}E\left(\widehat{\beta}_1\right) \\
&= E\left(\frac{\sum_{i=1}^n \left(\beta_0+\beta_1 X_i+u_i\right)}{n}\right) - \overline{X}\beta_1 \\
&= E\left(\beta_0\frac{\sum_{i=1}^n 1}{n}+\beta_1\frac{\sum_{i=1}^n X_i}{n}+\frac{\sum_{i=1}^n u_i}{n}\right) - \overline{X}\beta_1 \\
&= \beta_0+\overline{X}\beta_1+E\left(\frac{\sum_{i=1}^n u_i}{n}\right) - \overline{X}\beta_1 \\
&= \beta_0+E\left(\frac{\sum_{i=1}^n u_i}{n}\right) \\
&= \beta_0+\frac{1}{n}\sum_{i=1}^n E\left(u_i\right) \\
&= \beta_0, \text{ since } E(u_i)=0
\end{aligned}
$$

■

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

Proof. Skip.

(4) They are efficient among all the linear unbiased estimators.

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

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

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

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

Theorem 4.1: **Gauss−Markov Theorem:** Under assumptions 1-6, the Ordinary Least Squares(OLS) estimators are the Best Linear Unbiased Estimators ($BLUE$):
Proof. Skip.

If we are just interested in the relationship between X and Y, we can simply use $Cov(X, Y)$ or $Corr(X, Y)$. A regression line can also be used to predict the value of Y at a given value of X. For any given value of X, you can find a corresponding value of Y. Make sure that you can distinguish the differences between

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{u}_i$$

and

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i.$$

The first equation is the true model, the second is the estimated model. The actual observed values of Y_i do not necessary lie on the line, so there are residuals in both equations. The last equation represents a regression line, every $\widehat{Y}_i$ is a point in the regression line, no error term is needed. We use the regression line $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ to make predictions, e.g., if $\widehat{\beta}_0 = 1$, $\widehat{\beta}_1 = 1$, the predicted value $\widehat{Y}_i$ at $X_i = 10$ will be 11.

Although the OLS method has 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.

Exercise 4.2: True/False/Uncertain. Explain.

(a) The OLS estimators are most efficient among all estimators.

(b) The OLS estimators are the best linear unbiased estimators.

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

(d). In a linear regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, $Var(Y_i) = Var(u_i)$.

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

(f) If $E(u_i) = 2$, $\widehat{\beta}_0$ will be biased.

(g) If $E(u_i) = 2$, $\widehat{\beta}_1$ will be biased.

(h) In a linear regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, we have $\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right) \widehat{Y}_i = 0$.

(i). In a linear regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, we have $\sum_{i=1}^{n} Y_i \widehat{Y}_i = 0$.

4.4 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

$$TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 .$$

Let

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

be the predicted value of Y_i given X_i. Consider the following identity:

$$Y_i - \overline{Y} \equiv \left(\widehat{Y}_i - \overline{Y}\right) + \left(Y_i - \widehat{Y}_i\right).$$

Squaring both sides gives

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

Summing up from $i = 1$ to n, we have

$$\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 + 2\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)\left(Y_i - \widehat{Y}_i\right).$$

The last item in the R.H.S. can be shown to be zero. Thus, we have:

$$\underset{TSS}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = \underset{RSS}{\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2} + \underset{ESS}{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2}$$

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_i and $\overline{Y}$ can be decomposed into two parts:

$$Y_i - \overline{Y} = \left(Y_i - \widehat{Y}_i\right) + \left(\widehat{Y}_i - \overline{Y}\right).$$

The first part is

$$\left(\widehat{Y}_i - \overline{Y}\right) = \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right) - \left(\widehat{\beta}_0 + \widehat{\beta}_1 \overline{X}\right) = \widehat{\beta}_1 \left(X_i - \overline{X}\right).$$

This part shows that the predicted value $\widehat{Y}_i$ differs from $\overline{Y}$ because X_i differs from $\overline{X}$. The second part $\left(Y_i - \widehat{Y}_i\right)$ is the residual that remains unexplained by the regressor X_i. 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 a linear relationship. If R^2 is close to 1, then X and Y are highly linearly correlated. If X cannot explain Y at all, then $RSS = 0$, $TSS = ESS$, and $R^2 = 0$, and the regression line does not fit the data in this case. If there is nothing that remains unexplained, then $ESS = 0$. This implies the variation of Y can be totally explained by the variation of X, and $R^2 = 1$, and all the data must lie on the regression line in this case.

Example 4.2: Given the data (X_i, Y_i), $i = 1, 2, ..n$, suppose we know $\overline{X} = 30$. We run a regression of Y_i on X_i and obtain the following results

$$\widehat{Y_i} = 0.8 + 0.9X_i, \qquad R^2 = 0.9.$$

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

$$\widehat{X_i} = a + bY_i, \qquad R^2 = c.$$

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

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

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

Regression of Y_i on X_i yields ■

$$R^2 = \frac{\left(\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right)^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2} = 0.9.$$

Regression of X_i on Y_i yields

$$c = \frac{\left(\sum_{i=1}^n \left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)\right)^2}{\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2 \sum_{i=1}^n \left(X_i - \overline{X}\right)^2}.$$

Thus,

$$c = 0.9.$$ ■

Moreover,

$$\begin{aligned}
R^2 &= \frac{\left(\sum_{i=1}^n \left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)\right)^2}{\sum_{i=1}^n \left(X_i-\overline{X}\right)^2 \sum_{i=1}^n \left(Y_i-\overline{Y}\right)^2} \\
&= \frac{\sum_{i=1}^n \left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sum_{i=1}^n \left(X_i-\overline{X}\right)^2} \times \frac{\sum_{i=1}^n \left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sum_{i=1}^n \left(Y_i-\overline{Y}\right)^2} \\
0.9 &= (0.9)\, b \\
\Rightarrow b &= 1. \quad \blacksquare
\end{aligned}$$

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

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

Example 4.3: Consider the model: $Y_i = \beta_1 X_i + u_i, \qquad i = 1, 2, ..., n.$

(a) Show that the OLS estimator for β_1 is given by $\widehat{\beta}_1 = \dfrac{\sum\limits_{i=1}^{n} X_i Y_i}{\sum\limits_{i=1}^{n} X_i^2}$;

(b) If we have three observations of (X_i, Y_i), $i = 1, 2, 3$.

X_i	0	1	2
Y_i	2	1	0

Calculate the numerical values of:

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

ii) $\widehat{Y}_i = \widehat{\beta}_1 X_i$ for $i = 1, 2, 3$;

iii) $ESS = \sum\limits_{i=1}^{3} \left(Y_i - \widehat{Y}_i\right)^2$;

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

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

Solution:

(a) The problem is

$$\min_{\beta_1} \sum_{i=1}^{n} u_i^2 = \min_{\beta_1} \sum_{i=1}^{n} \left(Y_i - X_i \beta_1\right)^2 .$$

The first-order condition is

$$\frac{\partial \sum_{i=1}^{n} \left(Y_i - X_i\beta_1\right)^2}{\partial \beta_1} = -2\sum_{i=1}^{n} \left(Y_i - X_i\beta_1\right) X_i = 0 \Rightarrow \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} X_iY_i}{\sum_{i=1}^{n} X_i^2}. \quad \blacksquare$$

(b)

i	1	2	3
X_i	0	1	2
Y_i	2	1	0

(i)

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

(ii)

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

(iii)

$$ESS = \sum_{i=1}^{3} \left(Y_i - \widehat{Y}_i\right)^2 = (2-0)^2 + \left(1 - \frac{1}{5}\right)^2 + \left(0 - \frac{2}{5}\right)^2 = 4.8 \ \blacksquare$$

(iv)

$$TSS = \sum_{i=1}^{3} \left(Y_i - \overline{Y}\right)^2 = (2-1)^2 + (1-1)^2 + (0-1)^2 = 2 \ \blacksquare$$

(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 $\sum_{i=1}^{3} \widehat{u}_i \neq 0$.

Exercise 4.3: Given the data (X_i, Y_i), $i = 1, 2, ..n$, We run a regression of Y_i on X_i and obtain the following results

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

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

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

Show that

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

Exercise 4.4: Suppose we run a regression of Y_i on X_i with an intercept, and get the slope estimate of 0.8. Using the same data, if we run a regression of X_i on Y_i with an intercept, is it possible to get a slope estimate of -0.8?

Exercise 4.5: Given the data (X_i, Y_i), $i = 1, 2, ..., n$, and $\overline{X} = 10$. Suppose we run a regression of Y_i on X_i with an intercept, and obtain the following results:

$$\widehat{Y}_i = X_i, \qquad R^2 = 1.$$

Now, suppose we use the same data and run a regression of X_i on Y_i with an intercept, and obtain the following regression:

$$\widehat{X}_i = a + bY_i \qquad R^2 = c.$$

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

Exercise 4.6: Given the data (X_i, Y_i), $i = 1, 2, ..., n$. Suppose we run a regression of Y_i on X_i with an intercept, and get the following results:

$$\widehat{Y}_i = X_i, \qquad R^2 = 0.5$$

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

$$\widehat{X}_i = 1 + aY_i \qquad R^2 = b.$$

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

Exercise 4.7: Consider the model: $Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, 2, ..., n$.

If we have three observations of (X_i, Y_i), $i = 1, 2, 3$.

X_i	0	1	2
Y_i	2	1	0

Calculate the numerical values of:
i) $\widehat{\beta}_0, \widehat{\beta}_1$;
ii) $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ for $i = 1, 2, 3$;
iii) $ESS = \sum_{i=1}^{3} \left(Y_i - \widehat{Y}_i\right)^2$;
iv) $TSS = \sum_{i=1}^{3} \left(Y_i - \overline{Y}\right)^2$;
v) $R^2 = 1 - \dfrac{ESS}{TSS}$;
vi) $\overline{R}^2 = 1 - (1 - R^2)\dfrac{n-1}{n-k-1}$.

Exercise 4.8: Consider the model: $Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, 2, ..., n$
(a) Suppose we have four observations of (X_i, Y_i), $i = 1, 2, 3, 4$.

X_i	0	1	c	$1-c$
Y_i	0	1	1	0

Find the followings in term of c:

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

ii) $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ for $i = 1, 2, 3, 4$

iii) $ESS = \sum_{i=1}^{4} \left(Y_i - \widehat{Y}_i\right)^2$

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

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

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

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

Exercise 4.9: If we have four observations of (X_i, Y_i), $i = 1, 2, 3, 4$.

	$i=1$	$i=2$	$i=3$	$i=4$
X_i	-1	1	-1	1
Y_i	1	1	-1	-1

(a) Calculate the numerical values of:

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

ii) $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ for $i = 1, 2, 3, 4$.

iii) $ESS = \sum_{i=1}^{4} \left(Y_i - \widehat{Y}_i\right)^2$.

iv) $TSS = \sum_{i=1}^{4} \left(Y_i - \overline{Y}\right)^2$.

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

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

(b) Plot the four observations and draw the estimated regression line.

(c) Suppose there are two additional observation $(X_5, Y_5) = (0, 1)$ and $(X_6, Y_6) = (0, -1)$ How will this affect the regression line in (b)?

Exercise 4.10: Let X and Y be random variables, $W = 1 - X$, and $Z = 1 - Y$,

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

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

$$Y_i = \beta_{0a} + \beta_{1a} X_i + u_i,$$

$$Z_i = \beta_{0b} + \beta_{1b} W_i + u_i,$$

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

Exercise 4.11: Let A, B, C, D be four random variables with zero mean and unit variance.

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

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

$$B_i = \beta_{0a} + \beta_{1a} A_i + u_i,$$

$$D_i = \beta_{0b} + \beta_{1b} C_i + u_i,$$

$$C_i - D_i = \beta_{0c} + \beta_{1c}\,(A_i - B_i) + u_i,$$

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

4.5 Hypothesis Testing on βs

Consider the following regression

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

We would like to test whether β_1 equals zero.

Suppose we find that $\widehat{\beta}_1 = 0.34$ from the sample. After the estimation, we may perform hypothesis testing. 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, X cannot explain Y and the regression model will be useless. When we test this hypothesis, we need a test statistic and find its distribution. In the context of regression models, the random elements are u_i. Note that we have not yet specified the distribution of u_i. Thus far, we have only assumed that u_i 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: $u_i \sim N(0, \sigma^2)$.

This assumption is not needed as far as estimation is concerned. It is called for when we would like 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

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

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

$$Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

from Example 4.1. Since u_i 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 t 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_i in the true model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

Now after the OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have been obtained, the estimated residual is

$$\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$$

and we define

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

We use s^2 to estimate σ^2. The reason why we have to use $(n-2)$ is because s^2 is an unbiased estimator of σ^2. This number should be equal to the number of $\beta's$ in the regression. If we have a multiple regression with k $\beta's$, then it should be $(n-k)$ at the bottom. The test will have a t-distribution with degrees of freedom $(n-2)$.

Exercise 4.12: Consider the sample period from 1/9/14-30/9/14. Let
Y=Daily closing price of the call warrant [25453];
X=Price of [2628] China Life ;
i) Plot (X,Y).
ii) Run the following regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

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 warrant price affected by the price of the underlying stock?

Exercise 4.13: Let Z_1, Z_2 be independent $N(0,1)$ random variables. Suppose we draw a sample size n from the above distributions of Z_1 and Z_2. In a linear regression model $Z_{2i}^2 = \beta_0 + \beta_1 Z_{1i}^2 + u_i$, what will $\widehat{\beta}_1$ converge to?

Exercise 4.14: 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) $Z = XY$.

Suppose we draw a sample size n from the above distributions of X, Y and Z, and run the following regressions:
(i) $Y_i = \beta_0 + \beta_1 X_i + u_i$,
(ii) $Z_i = \beta_0 + \beta_1 X_i + u_i$,
(iii) $Z_i = \beta_0 + \beta_1 Y_i + u_i$.

When n goes to infinity, what are the values of $\widehat{\beta}_0$, $\widehat{\beta}_1$ in each of the possible cases ?

Exercise 4.15: Find the closing (i.e., unadjusted closing) price of [572] CHINA PACKAGING from September1-September 30, 2014. Extract your data from Yahoo Finance. Let P_t be the price and $r_t = \ln P_t - \ln P_{t-1}$ be the daily return of GOME on day t. Assume that $r_t \sim N(\mu, \sigma^2)$. Consider a sample of r_t from 2/9/14 to 30/9/14.

(a) Find $\overline{r}$ and s^2.

(b) Use t-test to test the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ at $\alpha = 5\%$.

(c) Construct a 95% confidence interval for the population mean μ.

(d) Let HSI be the Hang Seng Index of the same period, estimate the following regression model

$$P_t = \beta_0 + \beta_1 HSI_t + u_t,$$

(e) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$. Is the price of CHINA PACKAGING affected by Hang Seng Index?

4.6 Multiple Regression

In many situations, a single explanatory variable is not sufficient to explain the variation of Y. We may regress Y on some more other explanatory variables. A multiple regression is of the following form:

$$Y_i = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i.$$

The OLS estimated model is:

$$\widehat{Y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + ... + \widehat{\beta}_k X_{ki}.$$

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 or the predicted value of Y is beyond its possible range. $\widehat{\beta}_k$ is interpreted as the increase in the value of $\widehat{Y}$ if X_k is increased by 1 unit, holding all other X's constant. Sometimes, the sign of $\widehat{\beta}$ may be 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.

Example 4.4: Consider a regression model

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

We have the following data

i	y_i	x_{1i}	x_{2i}
1	2	3	1
2	1	1	2
3	4	2	3
4	5	0	4

Define

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

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \\ 5 \end{pmatrix}, \quad U = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix},$$

and $Y = X\beta + U$. The least square method is to find β to minimize $\sum u_i^2 = \min_\beta U'U = \min_\beta (Y - X\beta)' (Y - X\beta)$. The first-order condition is

$$2(-X)'(Y - X\beta) = 0$$

and we solve that

$$\widehat{\beta} = (X'X)^{-1} X'Y.$$

We need to find the inverse of $X'X$

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

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

$$\widehat{\beta} = \begin{pmatrix} \frac{299}{36} & -\frac{35}{18} & -\frac{37}{18} \\ -\frac{35}{18} & \frac{5}{9} & \frac{4}{9} \\ -\frac{37}{18} & \frac{4}{9} & \frac{5}{9} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 3 & 1 & 2 & 0 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 4 \\ 5 \end{pmatrix} = \begin{pmatrix} -\frac{7}{2} \\ 1 \\ 2 \end{pmatrix}.$$

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 goes to infinity, R^2 should approach 1. However, 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, the R^2 may still increase. 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 in $\overline{R}^2$ 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 substantially, and the indirect effect will dominate the direct effect, ending up with an drop in $\overline{R}^2$. However, if the additional variable is a garbage variable, R^2 will only increase much. 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 across models. Note that when $\overline{R}^2$ is maximized, the absolute value of the t statistics of all the slope coefficient estimates will be greater than one.

Exercise 4.16: True/False. Explain.

(a) The more explanatory variables we have, the higher the $\overline{R}^2$.

(b). The $\overline{R}^2$ cannot be negative.

(c) When the sample size increases, the R^2 must be higher.

4.7 Simple Hypothesis Testing

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

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.

$$\begin{aligned}
\widehat{sd}\left(\widehat{\beta}_i\right) &= \sqrt{s^2 c_{i+1,i+1}}, \\
s^2 &= \frac{\sum\limits_{i=1}^{n} \widehat{u}_i^2}{n-k-1}, \\
\widehat{u}_i &= Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i} - \ldots - \widehat{\beta}_k X_{ki},
\end{aligned}$$

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

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

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

Example 4.5: Consider the following data

	$i=1$	$i=2$	$i=3$	$i=4$
X_{1i}	3	1	2	0
X_{2i}	1	2	3	4
Y_i	2	1	4	5

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

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

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

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

4.8 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 X_2, X_3 and X_4 do not affect Y.

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 just interested in the linear relationship among $\beta's$ rather than testing if the $\beta's$ equal some prespecified values. For instance, we may like 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:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i.$$

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_5 X_{5i} + ... + \beta_k X_{ki} + u_i.$$

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

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 4.6: Consider the following demand function for chicken.

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \beta_4 \ln X_{4i} + u_i.$$

Suppose we run an OLS regression and obtain

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

$i = 1, 2, ..., 30.$

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 would like 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

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

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

Solution: Given

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \beta_4 \ln X_{4i} + u_i.$$

(a)

$$\begin{aligned}\beta_i &= \frac{\partial \ln Y_i}{\partial \ln X_{it}} = \frac{\partial \ln Y_i}{\partial Y_i}\frac{\partial Y_i}{\partial X_{it}}\frac{\partial X_{it}}{\partial \ln Xi_i} = \frac{\partial Y_i / Y}{\partial X_{it}/X_{it}} \\ &= \text{elasticity of } Y \text{ with respect to } X_i \text{ for } i = 1,2,3,4 \\ &\quad (i.e., \text{ when } X_i \text{ increases } 1\%, Y \text{ will increase } \beta_i\%)\end{aligned}$$

Thus,

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

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 $(30 - 5) = 25$.

$$t = 2.06.$$

The observed 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.
Using $R^2 = 1 - \dfrac{ESS}{TSS}$, we have

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

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 4.17: A model of death tolls due to heart disease is estimated as follows:

$$\begin{aligned}
\widehat{CHD}_i &= 139.68 + 10.71 CIG_i + 3.38 EDFAT_i + 26.75 SPIRITS_i - 4.13 BEER_i \\
n &= \text{Sample size} = 34 \\
k &= 4 = \text{Number of explanatory variables excluding the constant term} \\
ESS &= \sum_{i=1}^{34} \left(CHD_i - \widehat{CHD}_i \right)^2 = 2122 \\
\overline{R}^2 &= 1 - \frac{ESS/(n-k-1)}{TSS/(n-1)} = 0.672
\end{aligned}$$

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 ($TSS = \sum_{i=1}^{34} \left(CHD_i - \overline{CHD} \right)^2$) and the Regression Sum of Squares (RSS) in the above model.

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

$$CHD_i = \beta_0 + u_i.$$

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

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

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 $F = \dfrac{(ESS_r - ESS_u)\,/\,(df_r - df_u)}{ESS_u/df_u}$.

Exercise 4.18: Suppose we have 4 observations of a trivariate model.

	$i=1$	$i=2$	$i=3$	$i=4$
X_{1i}	3	1	2	0
X_{2i}	1	2	3	4
Y_i	2	1	4	5

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

(b) Find $\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i} \qquad$ for $i = 1, 2, 3, 4$;

(c) Find $s^2 = \dfrac{\sum_{i=1}^{n} \widehat{u}_i^2}{n-2-1}$;

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

(e) Test

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

for $i = 0, 1, 2$.

Exercise 4.19: Consider the model:

$$PRICE_i = \beta_0 + \beta_1 SQFT_i + \beta_2 BEDROOM_i + u_i,$$

$i = 1, 2, ..., 19$.
where
$PRICE_i$ is the price of house i (thousands of dollars)
$SQFT_i$ is the living areas of house i. (square feet)
$BEDROOM_i$ is the number of bedrooms in house i
Suppose we estimate the model and obtain

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

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 ($TSS = \sum_{i=1}^{19} \left(PRICE_i - \overline{PRICE} \right)^2$) and the Regression Sum of Squares ($RSS = TSS - ESS$) in the above model.

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

$$PRICE_i = \beta_0 + u_i.$$

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

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

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

Exercise 4.20: If the true model has X_1, but we estimate a model with X_1 and X_2. If $S_{y2} = 0$, then β_1 will be over-estimated. True/False/Uncertain. Explain.

Exercise 4.21: Consider the following production function for gross national product at time t.

$$\ln Y_t = \beta_0 + \beta_1 \ln K_t + \beta_2 \ln L_t + u_t.$$

Suppose we run an OLS and get

$$\begin{aligned} \widehat{\ln Y_t} &= \underset{(6.94)}{1.18} + \underset{(3.13)}{0.25} \ln K_t + \underset{(2.42)}{0.46} \ln L_t, \qquad t = 1, 2, ..., 30; \\ R^2 &= 0.93; \end{aligned}$$

where
Y_t=GDP at time t in constant dollars;
L_t=Total employment at time t;
K_t=Capital stock at time t in constant dollars;
and the figures in parentheses are t-ratios.

Define an F-statistic

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

where df_R and df_U are the degrees of freedom of the restricted and unrestricted models respectively; ESS_R and ESS_U are the error sum of squares of the restricted and unrestricted models respectively.

(a) Use the definition $R^2 = 1 - \dfrac{ESS}{TSS}$, show that the F-test can be rewritten as

$$F = \frac{(R^2_U - R^2_R) \ / \ (df_R - df_U)}{(1 - R^2_U) \ / \ df_U}.$$

(b) Suppose we want to test $H_0 : \beta_1 = \beta_2 = 0$ at $\alpha = 5\%$.What is restricted model? Show that the $R^2 = 0$ in this restricted model.

(c) Compute the value of F in part (b) under $H_0 : \beta_1 = \beta_2 = 0$.

Exercise 4.22: Consider the sample period from 1/9/14-30/9/14. Let

Y=Daily closing price of the call warrant [25453];

X_1=Price of [2628] China Life;

X_2=The square of the price range of [2628] China Life in the previous trading day, i.e, $(P_{\max,t-1} - P_{\min,t-1})^2$.

i) Run the following regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,$$

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 warrant price affected by the price of China Life?

iii) Test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ at $\alpha = 0.05$. Is the warrant price affected by the volatility of China Life?

iv) Compare your results with those from the simple regression. What are the differences in terms of the estimated values of the coefficients, test result for $H_0 : \beta_1 = 0$, R^2 and the adjusted R^2.

4.9 Multivariate Multiple Regression

Multivariate regression is a technique that estimates a regression model with more than one outcome variable. Mathematically speaking, one would like to model the relationship between m responses $Y_1, Y_2, ..., Y_m$ and a single set of predictor variables $z_1, z_2, ..., z_r$. Each response is assumed to follow its own regression model, so that for $j = 1, 2, ..., n$

$$Y_{j1} = \beta_{01} + \beta_{11} z_{j1} + ... + \beta_{r1} z_{jr} + \varepsilon_{j1}$$

$$Y_{j2} = \beta_{02} + \beta_{12} z_{j1} + ... + \beta_{r2} z_{jr} + \varepsilon_{j2}$$

$$\vdots$$

$$Y_{jm} = \beta_{0m} + \beta_{1m} z_{j1} + ... + \beta_{rm} z_{jr} + \varepsilon_{jm}$$

where $z_{j1}, ..., z_{jr}$ denote the values of the predictor variables for the j^{th} observation.

For example, one may like to examine how the three measures of health of individual j, namely, cholesterol (Y_{j1}), blood pressure (Y_{j2}), and weight (Y_{j3}) are affected by his/her eating habits such as how many ounces of red meat (z_{j1}), fish (z_{j2}), dairy products (z_{j3}), and chocolate (z_{j4}) consumed per day.

In matrix notation,

$$\underset{n\times(r+1)}{\mathbf{Z}} = \begin{pmatrix} 1 & z_{11} & \cdots & z_{1r} \\ 1 & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & \cdots & z_{nr} \end{pmatrix}, \qquad \underset{n\times m}{\mathbf{Y}} = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nm} \end{pmatrix}$$

$$\underset{n\times m}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{pmatrix}, \qquad \underset{(r+1)\times m}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rm} \end{pmatrix}$$

The multivariate linear regression model is

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The estimator is for β

$$\widehat{\beta}= \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}.$$

Example 4.7: Consider the following model for $j = 1, 2, ..., 5$.

$$Y_{j1} = \beta_{01} + \beta_{11} z_{j1} + \varepsilon_{j1}.$$

$$Y_{j2} = \beta_{02} + \beta_{12} z_{j1} + \varepsilon_{j2.}$$

The data are given as follows:

z_1	0	1	2	3	4
Y_1	1	4	3	8	9
Y_2	-1	-1	2	3	2

$$\underset{5\times 2}{\mathbf{Z}} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} \qquad \underset{5\times 2}{\mathbf{Y}} = \begin{pmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{pmatrix}$$

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y} \\
&= \left(\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{pmatrix} \\
&= \begin{pmatrix} 5 & 10 \\ 10 & 30 \end{pmatrix}^{-1} \begin{pmatrix} 25 & 5 \\ 70 & 20 \end{pmatrix} \\
&= \frac{1}{5 \times 30 - 10^2} \begin{pmatrix} 30 & -10 \\ -10 & 5 \end{pmatrix} \begin{pmatrix} 25 & 5 \\ 70 & 20 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix}.
\end{aligned}$$

The fitted values are generated from

$$\widehat{Y}_{j1} = 1 + 2z_{j1}.$$

$$\widehat{Y}_{j2} = -1 + z_{j1}.$$

$$\widehat{\mathbf{Y}} = \mathbf{Z}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{pmatrix}.$$

The residual matrix is

$$\widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \begin{pmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{pmatrix} - \begin{pmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & -1 \\ -2 & 1 \\ 1 & 1 \\ 0 & -1 \end{pmatrix}.$$

Note that the sum of residual terms in each column is zero.

$$\widehat{\boldsymbol{\varepsilon}}'\widehat{\mathbf{Y}} = \begin{pmatrix} 0 & 1 & -2 & 1 & 0 \\ 0 & -1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

$$\mathbf{Y}'\mathbf{Y} = \begin{pmatrix} 1 & 4 & 3 & 8 & 9 \\ -1 & -1 & 2 & 3 & 2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{pmatrix} = \begin{pmatrix} 171 & 43 \\ 43 & 19 \end{pmatrix}.$$

$$\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} = \begin{pmatrix} 1 & 3 & 5 & 7 & 9 \\ -1 & 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{pmatrix} = \begin{pmatrix} 165 & 45 \\ 45 & 15 \end{pmatrix}.$$

$$\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = \begin{pmatrix} 0 & 1 & -2 & 1 & 0 \\ 0 & -1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & -1 \\ -2 & 1 \\ 1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 6 & -2 \\ -2 & 4 \end{pmatrix}.$$

Note that

$$\mathbf{Y}'\mathbf{Y} = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}.$$

Exercise 4.23: Consider the model

$$Y_{j1} = \beta_{01} + \beta_{11} z_{j1} + \varepsilon_{j1}$$

$$Y_{j2} = \beta_{02} + \beta_{12} z_{j1} + \varepsilon_{j2}$$

The data are given as follows:

z_1	-2	-1	0	1	2
Y_1	5	3	4	2	1
Y_2	-3	-1	-1	2	3

(a) Solve $\widehat{\beta}_{01}, \widehat{\beta}_{11}, \widehat{\beta}_{02}, \widehat{\beta}_{12}$.

(b) Find $\widehat{\mathbf{Y}}$.

(c) Verify that $\mathbf{Y}'\mathbf{Y} = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}$.

(d) Repeat (a), (b), (c) if the data are given as follows:

z_1	3	1	0	2
Y_1	3	5	6	4
Y_2	1	1	1	1

Chapter 5

Principal Components Analysis

Principal components analysis (PCA) aims to transform a set of correlated response variables into a smaller set of uncorrelated variables called principal components. The objectives of a PCA are (1) to reduce the dimensionality of the data set and (2) to identify new meaningful underlying variables. If the data are plotted in a p-dimensional space, will the data take up all p dimensions? If not, the original variables can be replaced by a smaller number of underlying variables without losing any information. Note that we cannot guarantee that the new variables, called principal components, will be meaningful. The principal components have the following properties:

(1) They are uncorrelated;

(2) The first principal component accounts for much of the variability in the data as possible;

(3) Each succeeding component accounts for as much of the remaining variability as is possible.

5.1 The Two-Variable Case

Let the random vector $\mathbf{X}' = (X_1, X_2)$ have the covariance matrix

$$\mathbf{\Omega} = \begin{pmatrix} Var\left(X_1\right) & Cov\left(X_1, X_2\right) \\ Cov\left(X_2, X_1\right) & Var\left(X_2\right) \end{pmatrix},$$

with eigenvalues $\lambda_1 \geq \lambda_2 \geq 0$. We make two linear combinations of X variables, called them Y variables. Note that X may be correlated, but Y must be uncorrelated. If one of the X is uncorrelated with other X, then this X will become one of our Y, i.e., the weight associated with other X is zero. In the extreme case, where all X are uncorrelated, then Y will just be X. Mathematically speaking, consider the linear combination

$$Y_1 = a_{11}X_1 + a_{12}X_2 = \mathbf{a}_1'\mathbf{X},$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 = \mathbf{a}_2'\mathbf{X}.$$

$$Var\left(Y_1\right) = Var\left(\mathbf{a}_1'\mathbf{X}\right) = \mathbf{a}_1' Var\left(\mathbf{X}\right)\mathbf{a}_1 = \mathbf{a}_1'\mathbf{\Omega}\mathbf{a}_1,$$

$$Var\left(Y_2\right) = Var\left(\mathbf{a}_2'\mathbf{X}\right) = \mathbf{a}_2' Var\left(\mathbf{X}\right)\mathbf{a}_2 = \mathbf{a}_2'\mathbf{\Omega}\mathbf{a}_2,$$

$$Cov\left(Y_1, Y_2\right) = \mathbf{a}_1'\mathbf{\Omega}\mathbf{a}_2.$$

The first principal component=linear combination of $\mathbf{a}_1'\mathbf{X}$ that maximizes

$$Var\left(\mathbf{a}_1'\mathbf{X}\right)$$

subject to

$$\mathbf{a}_1'\mathbf{a}_1 = \sum_{j=1}^{p} a_{1j}^2 = 1.$$

The second principal component=linear combination of $\mathbf{a}_2'\mathbf{X}$ that maximizes $Var\left(\mathbf{a}_2'\mathbf{X}\right)$ subject to $\mathbf{a}_2'\mathbf{a}_2 = \sum_{j=1}^{p} a_{2j}^2 = 1$ and $\mathrm{Cov}(\mathbf{a}_2'\mathbf{X}, \mathbf{a}_1'\mathbf{X}) = 0$.

What values of the vector **a** will satisfy the above condition? Here, we recall the eigenvalues and eigenvectors that we have learned in previous classes. In general, if

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 \qquad \text{for } i = 1, 2,$$

where $\mathbf{e}_i = (e_{i1}, e_{i2})'$ is the eigenvector of $\mathbf{\Omega}$ associated with the i^{th} eigenvalue λ_i, then the above condition will be satisfied.

Note that since $\mathbf{\Omega}$ is a covariance matrix, it is a positive definite matrix and its spectral decomposition can be expressed as

$$\mathbf{\Omega} = \sum_{i=1}^{2} \lambda_i \mathbf{e}_i \mathbf{e}_i',$$

where λ_i is the i^{th} eigenvalue and e_i is the i^{th} eigenvector. We can rewrite the decomposition in matrix form such that

$$\underset{p\times p}{\mathbf{\Omega}} = \sum_{i=1}^{2} \lambda_i \mathbf{e}_i \mathbf{e}_i' = \begin{pmatrix} \mathbf{e}_1, & \mathbf{e}_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \mathbf{e}_1' \\ \mathbf{e}_2' \end{pmatrix} = \mathbf{P\Lambda P}',$$

where

$$\underset{2\times 2}{\mathbf{\Lambda}} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix},$$

P is a matrix collecting the eigenvectors

$$\underset{2\times 2}{\mathbf{P}} = \begin{pmatrix} \mathbf{e}_1, & \mathbf{e}_2 \end{pmatrix} = \begin{pmatrix} e_{11} & e_{21} \\ e_{12} & e_{22} \end{pmatrix}.$$

Using the properties that $\mathbf{e}_i'\mathbf{e}_i = 1$ and $\mathbf{e}_i'\mathbf{e}_j = 0$ for $i \neq j$, we have

$$\begin{aligned}
Var\,(Y_1) &= Var\,(\mathbf{e}_1'\mathbf{X}) = \mathbf{e}_1' Var\,(\mathbf{X})\,\mathbf{e}_1 = \mathbf{e}_1'\boldsymbol{\Omega}\mathbf{e}_1 = \mathbf{e}_1'\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'\mathbf{e}_1 \\
&= \mathbf{e}_1' \begin{pmatrix} \mathbf{e}_1, & \mathbf{e}_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \mathbf{e}_1' \\ \mathbf{e}_2' \end{pmatrix} \mathbf{e}_1 \\
&= \begin{pmatrix} \mathbf{e}_1'\mathbf{e}_1, & \mathbf{e}_1'\mathbf{e}_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \mathbf{e}_1'\mathbf{e}_1 \\ \mathbf{e}_2'\mathbf{e}_1 \end{pmatrix} \\
&= \begin{pmatrix} 1, & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \lambda_1.
\end{aligned}$$

Similarly, Var$(Y_2) = \lambda_2$ and

$$\begin{aligned}
Cov\,(Y_1, Y_2) &= \mathbf{e}_1'\boldsymbol{\Omega}\mathbf{e}_2 = \mathbf{e}_1'\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'\mathbf{e}_2 \\
&= \mathbf{e}_1' \begin{pmatrix} \mathbf{e}_1, & \mathbf{e}_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \mathbf{e}_1' \\ \mathbf{e}_2' \end{pmatrix} \mathbf{e}_2 \\
&= \begin{pmatrix} \mathbf{e}_1'\mathbf{e}_1, & \mathbf{e}_1'\mathbf{e}_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \mathbf{e}_1'\mathbf{e}_2 \\ \mathbf{e}_2'\mathbf{e}_2 \end{pmatrix} \\
&= \begin{pmatrix} 1, & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0.
\end{aligned}$$

The proportion of total population variance due to first principal component $= \dfrac{\lambda_1}{\lambda_1 + \lambda_2}$.

Example 5.1: Consider the covariance matrix

$$\Omega = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}.$$

(a) Determine the population components Y_1 and Y_2.

(b) Calculate the proportion of the total population variance explained by the first principal component.

Solution: Recall from Chapter 2 that for a 2 by 2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the eigenvalues are

$$\lambda_1 = \frac{1}{2}\left(a + d + \sqrt{(a-d)^2 + 4bc}\right),$$

$$\lambda_2 = \frac{1}{2}\left(a + d - \sqrt{(a-d)^2 + 4bc}\right).$$

Thus, we have

$$\lambda_1 = 100.16, \qquad \mathbf{e}_1 = \begin{pmatrix} 0.04034 \\ 0.99998 \end{pmatrix},$$

$$\lambda_2 = 0.83865, \qquad \mathbf{e}_2 = \begin{pmatrix} 0.99998 \\ -0.04034 \end{pmatrix}.$$

$$Y_1 = \mathbf{e}_1'\mathbf{X} = 0.04034X_1 + 0.99998X_2,$$

$$Y_2 = \mathbf{e}_2'\mathbf{X} = 0.99998X_1 - 0.04034X_2.$$

Note that the first principal component attaches a very large weight to X_2, since X_2 has a large variance (This large variance may be due to the unit of measurement used).

$$\begin{aligned} Var\,(Y_1) &= Var\,(0.04034X_1 + 0.99998X_2) \\ &= (0.04034)^2\, Var\,(X_1) + (0.99998)^2\, Var\,(X_2) \\ &\quad +2\,(0.04034)\,(0.99998)\, Cov(X_1, X_2) \\ &= (0.04034)^2\,(1) + (0.99998)^2\,(100) + 2\,(0.04034)\,(0.99998)\,(4) \\ &= 100.16 \\ &= \lambda_1. \end{aligned}$$

Similarly, we can show that $Var\,(Y_2) = 0.83865 = \lambda_2$. Note that

$$
\begin{aligned}
Cov\left(Y_1, Y_2\right) &= Cov\left(0.04034X_1 + 0.99998X_2, 0.99998X_1 - 0.04034X_2\right) \\
&= 0.04034\left(0.99998\right)Var\left(X_1\right) - \left(\left(0.99998\right)^2 - \left(0.04034\right)^2\right)Cov\left(X_1, X_2\right) \\
&\quad -0.04034\left(0.99998\right)Var\left(X_2\right) \\
&= \left(0.04034\right)\left(0.99998\right) + \left(\left(0.99998\right)^2 - \left(0.04034\right)^2\right)\left(4\right) - \left(0.04034\right)\left(0.99998\right)(1 \\
&= 0.
\end{aligned}
$$

Therefore, the proportion of total population variance due to first principal component $=\dfrac{\lambda_1}{\lambda_1+\lambda_2} = \dfrac{100.16}{100.16+0.83865} = 0.99.$

Exercise 5.1: The two Eigen values of a 2 by 2 square matrix can be equal to each other. True/ False

Exercise 5.2: The smallest Eigen values of a 2 by 2 square matrix can be equal to zero. True/ False

Exercise 5.3. Determine the population components Y_1 and Y_2. and calculate the proportion of the total population variance explained by first principal component for the covariance matrix $\mathbf{\Omega} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

Exercise 5.4. Determine the population components Y_1 and Y_2. and calculate the proportion of the total population variance explained by first principal component for the covariance matrix $\mathbf{\Omega} = \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}$.

Exercise 5.5: True/False. For the correlation matrix $\boldsymbol{\rho}= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$,

(a). The corresponding covariance matrix can be $\mathbf{\Omega} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$.

(b). The corresponding covariance matrix should also be $\mathbf{\Omega} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

5.2 The General Case

Let the random vector $\mathbf{X}' = (X_1, X_2, ..., X_p)$ have the covariance matrix

$$\mathbf{\Omega} = \begin{pmatrix} Var\,(X_1) & Cov\,(X_1, X_2) & \cdots & Cov\,(X_1, X_p) \\ Cov\,(X_2, X_1) & Var\,(X_2) & \cdots & Cov\,(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\,(X_p, X_1) & Cov\,(X_p, X_2) & \cdots & Var\,(X_p) \end{pmatrix},$$

with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. We make p linear combinations of X variables, called them Y variables. Consider the linear combination

$$Y_1 = \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p,$$

$$Y_2 = \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p,$$

$$\vdots$$

$$Y_p = \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p.$$

Note that

$$Var\,(Y_i) = \mathbf{a}_i'\mathbf{\Omega}\mathbf{a}_i,$$

$$Cov\,(Y_i, Y_k) = \mathbf{a}_i'\mathbf{\Omega}\mathbf{a}_k.$$

First principal component=linear combination of $\mathbf{a}_1'\mathbf{X}$ that maximizes

$$Var\,(\mathbf{a}_1'\mathbf{X})$$

subject to

$$\mathbf{a}_1'\mathbf{a}_1 = \sum_{j=1}^{p} a_{1j}^2 = 1.$$

The second principal component=linear combination of $\mathbf{a}_2'\mathbf{X}$ that maximizes $Var\left(\mathbf{a}_2'\mathbf{X}\right)$ subject to $\mathbf{a}_2'\mathbf{a}_2 = \sum_{j=1}^{p} a_{2j}^2 = 1$ and $\text{Cov}(\mathbf{a}_2'\mathbf{X}, \mathbf{a}_1'\mathbf{X}) = 0$.

The i^{th} principal component=linear combination of $\mathbf{a}_i'\mathbf{X}$ that maximizes $Var\left(\mathbf{a}_i'\mathbf{X}\right)$ subject to $\mathbf{a}_i'\mathbf{a}_i = \sum_{j=1}^{p} a_{ij}^2 = 1$ and $\text{Cov}(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_k'\mathbf{X}) = 0$ for $k < i$.

What values of the vector $\mathbf{a}$ will satisfy the above condition? In general, if

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p, \qquad (i = 1, 2, ..., p),$$

where $\mathbf{e}_i = (e_{i1}, e_{i2}, ..., e_{ip})'$ is the eigenvector associated with the i^{th} eigenvalue λ_i, then the above condition will be satisfied. Note that Ω is a positive definite matrix with the spectral decomposition of a p by p symmetric matrix X can be expressed as

$$\mathbf{\Omega} = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i',$$

where λ_i is the i^{th} eigenvalue and e_i is the i^{th} eigenvector. We can rewrite the decomposition in matrix form as

$$\underset{p\times p}{\mathbf{\Omega}} = \mathbf{P\Lambda P}',$$

where

$$\underset{p\times p}{\mathbf{\Lambda}} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix},$$

$\mathbf{P}$ is a matrix collecting the eigenvectors

$$\underset{p\times p}{\mathbf{P}} = \begin{pmatrix} \mathbf{e}_1, & \mathbf{e}_2, & \cdots & ,\mathbf{e}_p \end{pmatrix}.$$

For $i = 1, 2, ..., p$,

$$\begin{aligned}
& Var\left(Y_i\right) \\
= \; & Var\left(\mathbf{e}_i'\mathbf{X}\right) = \mathbf{e}_i' Var\left(\mathbf{X}\right)\mathbf{e}_i = \mathbf{e}_i'\boldsymbol{\Omega}\mathbf{e}_i = \mathbf{e}_i'\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'\mathbf{e}_i \\
= \; & \begin{pmatrix} \mathbf{e}_i'\mathbf{e}_1, & \mathbf{e}_i'\mathbf{e}_2, & \cdots & \mathbf{e}_i'\mathbf{e}_i, & \cdots & ,\mathbf{e}_i'\mathbf{e}_p \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \lambda_i & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & \lambda_p \end{pmatrix} \begin{pmatrix} \mathbf{e}_1'\mathbf{e}_i \\ \mathbf{e}_2'\mathbf{e}_i \\ \vdots \\ \mathbf{e}_i'\mathbf{e}_i \\ \vdots \\ \mathbf{e}_p'\mathbf{e}_i \end{pmatrix} \\
= \; & \begin{pmatrix} 0, & 0, & \cdots & 1, & \cdots & ,0 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \lambda_i & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & \lambda_p \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}. \\
= \; & \lambda_i.
\end{aligned}$$

Note that

$$\begin{aligned}
Var\left(Y_1\right) + Var\left(Y_2\right) + ... + Var\left(Y_p\right) &= \lambda_1 + \lambda_2 + ... + \lambda_p \\
&= Trace\left(\boldsymbol{\Omega}\right) \\
&= Var\left(X_1\right) + Var\left(X_2\right) + ... + Var\left(X_p\right)
\end{aligned}$$

and

$$Cov\left(Y_i, Y_k\right) = \mathbf{e}_i'\boldsymbol{\Omega}\mathbf{e}_k = \mathbf{e}_i'\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'\mathbf{e}_k = 0 \qquad \text{for } i \neq k.$$

The proportion of total population variance due to k^{th} principal component is $\dfrac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$. Note that a good Y variable should have large variation, since we need the variation of Y to reflect the variation in X. Although we have p principal components Y, not all of them are useful. For example, if it just happens that one of the Y variables (say, Y_p) has no variation at all, i.e., for all the X observations we have, Y_p has the same value. In this case, Y_p contains no information of X and can be dropped, so the number of variables is reduced from p to $p-1$. The principal components analysis will be extremely useful if we can reduce a very large value of p (say, 50) to just a few useful variables (say, 3).

Example 5.2: If the covariance matrix of X_1, X_2 and X_3 is

$$\Omega = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

We can show that

$$\lambda_1 = 5.83, \qquad \mathbf{e}_1 = \begin{pmatrix} 0.383 \\ -0.924 \\ 0 \end{pmatrix},$$

$$\lambda_2 = 2, \qquad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

$$\lambda_3 = 0.17, \qquad \mathbf{e}_3 = \begin{pmatrix} 0.924 \\ 0.383 \\ 0 \end{pmatrix}.$$

$$Y_1 = \mathbf{e}_1'\mathbf{X} = 0.383X_1 - 0.924X_2,$$

$$Y_2 = \mathbf{e}_2'\mathbf{X} = X_3,$$

$$Y_3 = \mathbf{e}_3'\mathbf{X} = 0.924X_1 + 0.383X_2.$$

$$\begin{aligned}
Var\left(Y_1\right) &= Var\left(0.383X_1 - 0.924X_2\right) \\
&= \left(0.383\right)^2 Var\left(X_1\right) + \left(-0.924\right)^2 Var\left(X_2\right) + 2\left(0.383\right)\left(-0.924\right) Cov(X_1, X_2) \\
&= \left(0.383\right)^2 \left(1\right) + \left(-0.924\right)^2 \left(5\right) + 2\left(0.383\right)\left(-0.924\right)\left(-2\right) \\
&= 5.83 = \lambda_1.
\end{aligned}$$

$$\begin{aligned}
Cov\left(Y_1, Y_2\right) &= Cov\left(0.383X_1 - 0.924X_2, X_3\right) \\
&= 0.383 Cov\left(X_1, X_3\right) - 0.924 Cov\left(X_2, X_3\right) \\
&= 0.383\left(0\right) - 0.924\left(0\right) \\
&= 0.
\end{aligned}$$

Similarly, we can show that $Var\left(Y_2\right) = 2$ and $Var\left(Y_3\right) = 0.17$.

Therefore, the proportion of total population variance due to first principal component $= \dfrac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \dfrac{5.83}{5.83 + 2 + 0.17} = 0.73$.

The proportion of total population variance due to second principal component $= \dfrac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = \dfrac{2}{5.83 + 2 + 0.17} = 0.25$.

Thus, the first two components account for 98% of the population variance. In this case, the component Y_3 can be dropped.

5.3 Principal Components Obtained From Correlation Matrices

Since the covariance matrix will be affected by the unit of measurement, sometimes it is better to standardize the variable and use the correlation matrix. Principal components obtained from covariance and correlation matrices are different.

Example 5.3: Note from the previous example that the first principal component attaches a very large weight to X_2, since X_2 has a large variance. This large variance may be due to the unit of measurement used. The problem can be solved by using the correlation matrix. Consider the covariance matrix of the previous example

$$\boldsymbol{\Omega} = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}.$$

(a) Convert the covariance matrix into a correlation matrix.

(b) Determine the population components Y_1 and Y_2 from the correlation matrix.

(c) Calculate the proportion of the total population variance explained by the first principal component.

Solution: We first perform a standardization of X

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{Var\,(X_1)}},$$

$$Z_2 = \frac{X_2 - \mu_2}{\sqrt{Var\,(X_2)}}.$$

The corresponding correlation matrix is

$$\boldsymbol{\rho} = \begin{pmatrix} Var\,(Z_1) & Cov\,(Z_1, Z_2) \\ Cov\,(Z_2, Z_1) & Var\,(Z_2) \end{pmatrix} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}.$$

$$\lambda_1 = 1.4, \qquad \mathbf{e}_1 = \begin{pmatrix} 0.707 \\ 0.707 \end{pmatrix},$$

$$\lambda_2 = 0.6, \qquad \mathbf{e}_2 = \begin{pmatrix} 0.707 \\ -0.707 \end{pmatrix}.$$

$$\begin{aligned}
Y_1 &= 0.707Z_1 + 0.707Z_2 \\
&= 0.707\frac{X_1-\mu_1}{\sqrt{Var\left(X_1\right)}} + 0.707\frac{X_2-\mu_2}{\sqrt{Var\left(X_2\right)}} \\
&= 0.707\frac{X_1-\mu_1}{1} + 0.707\frac{X_2-\mu_2}{10} \\
&= 0.707\left(X_1-\mu_1\right) + 0.0707\left(X_2-\mu_2\right).
\end{aligned}$$

Similarly

$$Y_2 = 0.707\left(X_1-\mu_1\right) - 0.0707\left(X_2-\mu_2\right).$$

Note that

$$\begin{aligned}
Var\left(Y_1\right) &= Var\left(0.707Z_1 + 0.707Z_2\right) \\
&= 0.707^2 Var\left(Z_1\right) + 0.707^2 Var\left(Z_2\right) + 2\left(0.707\right)\left(0.707\right)Cov\left(Z_1,Z_2\right) \\
&= 0.707^2\left(1\right) + 0.707^2\left(1\right) + 2\left(0.707\right)\left(0.707\right)\left(0.4\right) \\
&= 1.4 = \lambda_1.
\end{aligned}$$

Similarly

$$Var\left(Y_2\right) = 0.6.$$

$$\begin{aligned}
Cov\left(Y_1,Y_2\right) &= Cov\left(0.707Z_1 + 0.707Z_2, 0.707Z_1 - 0.707Z_2\right) \\
&= 0.707^2[Cov\left(Z_1,Z_1\right) - Cov\left(Z_1,Z_2\right) + Cov\left(Z_2,Z_1\right) - Cov\left(Z_2,Z_2\right)] \\
&= 0.707^2[Cov\left(Z_1,Z_1\right) - Cov\left(Z_2,Z_2\right)] \\
&= 0.707^2\left[Var\left(Z_1\right) - Var\left(Z_2\right)\right] \\
&= 0.707^2\left[1-1\right] \\
&= 0.
\end{aligned}$$

Therefore, the proportion of total population variance due to first principal component $=\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{1.4}{1.4+0.6} = 0.7$. Note that this proportion is much lower than the case of the previous example when the variables are not standardized.

Exercise 5.6: For the covariance matrix $\boldsymbol{\Omega} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$,

(a) Determine the population components Y_1 and Y_2.

(b) Calculate the proportion of the total population variance explained by Y_1.

(c) Convert the covariance matrix to a correlation matrix. Repeat (a) and (b).

(d) Compare the components in (a) and (c), Are they the same?

Exercise 5.7: For the covariance matrix $\boldsymbol{\Omega} = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$,

(a) Show that the corresponding correlation matrix is

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & -\frac{2}{\sqrt{5}} & 0 \\ -\frac{2}{\sqrt{5}} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

(b) Show that the eigenvalues are $\lambda_1 = 1.89443$, $\lambda_2 = 1$, $\lambda_3 = 0.10557$. find the corresponding eigenvectors.

(c) Calculate the proportion of the total population variance explained by Y_1.

Exercise 5.8: Find the daily return $r_t = \ln P_t - \ln P_{t-1}$ of the six stocks of Hang Seng Index Property sector [1], [12], [16], [83], [101] and [688] for October 3 to October 31, 2014.

(a) Construct the sample covariance matrix S, and find the sample principal components.

(b) Determine the proportion of the total sample variance explained by the first three principal components.

5.4 Covariance Matrices with Special Structures

$$\mathbf{\Omega} = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}.$$

Setting

$$\lambda_1 = 5, \lambda_2 = 2.$$

$$\mathbf{e}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

$$Y_1 = \mathbf{e}_1'\mathbf{X} = X_1.$$

$$Y_2 = \mathbf{e}_2'\mathbf{X} = X_2.$$

Thus, the set of principal components is just the original set of uncorrelated variables, and nothing is gained by extracting the principal components.

In general, if we have a set of p uncorrelated variable with covariance matrix

$$\mathbf{\Omega} = \begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{pmatrix}.$$

with $\sigma_{11} \geq \sigma_{22} \geq ... \geq \sigma_{pp}$. Setting

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, ..., \mathbf{e}_p = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

$$\lambda_1 = \sigma_{11}, \lambda_2 = \sigma_{22}, ..., \lambda_p = \sigma_{pp}.$$

We will have $Y_i = X_i$ for all i.

Exercise 5.9: For the covariance matrix

$$\mathbf{\Omega} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

(a) Determine the population principal components Y_1, Y_2 and Y_3.

(b) Calculate the proportion of the total population variance explained by the first principal component.

5.5 Equicorrelation Matrix

Consider the 3 by 3 covariance matrix

$$\Omega = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.$$

For $0 < \rho \leq 1$.

The corresponding correlation matrix is

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.$$

It can be shown that the greatest eigenvalue of this matrix is

$$\lambda_1 = 1 + (3-1)\,\rho = 1 + 2\rho$$

and its normalized eigenvector is

$$\mathbf{e}_1' = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right).$$

The remaining two eigenvalues are all equal, i.e., $\lambda_2 = \lambda_3$. Since $\lambda_1 + \lambda_2 + \lambda_3$ is the dimension of the correlation matrix (=3), we have

$$\lambda_2 = \lambda_3 = \frac{3 - \lambda_1}{2} = \frac{3 - (1 + 2\rho)}{2} = 1 - \rho.$$

We can also show that

$$\mathbf{e}_2 = \begin{pmatrix} \frac{1}{\sqrt{1\times2}} \\ \frac{-1}{\sqrt{1\times2}} \\ 0 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} \frac{1}{\sqrt{2\times3}} \\ \frac{1}{\sqrt{2\times3}} \\ \frac{-2}{\sqrt{2\times3}} \end{pmatrix}.$$

The first principal component is

$$Y_1 = \mathbf{e}_1'\mathbf{X} = \frac{1}{\sqrt{3}}X_1 + \frac{1}{\sqrt{3}}X_2 + \frac{1}{\sqrt{3}}X_3,$$

which accounts for $\dfrac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \dfrac{1 + 2\rho}{1 + 2\rho + 1 - \rho + 1 - \rho} = (1 + 2\rho)/3$ of the total variance. Note that the higher the value of ρ, the higher the importance of the first principal component. It is proportional to the sum of the three original variables, which might be regarded as an "index" with equal weights.

Example 5.4: Let

$$\mathbf{\Omega} = 3\begin{pmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 1 \end{pmatrix}.$$

(a) Determine the population components Y_1 to Y_3.

(b) Calculate the proportion of the total population variance explained by Y_1.

Solution: It can be shown that the greatest eigenvalue of this matrix is

$$\lambda_1 = 3\left[1 + 2\left(0.6\right)\right]$$

and its normalized eigenvector is

$$\mathbf{e}_1' = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right).$$

$$\lambda_2 = \lambda_3 = 3\,(1 - 0.6) = 1.2.$$

$$\mathbf{e}_2 = \begin{pmatrix} \frac{1}{\sqrt{1\times 2}} \\ \frac{-1}{\sqrt{1\times 2}} \\ 0 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} \frac{1}{\sqrt{2\times 3}} \\ \frac{1}{\sqrt{2\times 3}} \\ \frac{-2}{\sqrt{2\times 3}} \end{pmatrix}.$$

The first principal component is

$$Y_1 = \frac{1}{\sqrt{3}} X_1 + \frac{1}{\sqrt{3}} X_2 + \frac{1}{\sqrt{3}} X_3,$$

$$Y_2 = \frac{1}{\sqrt{2}} X_1 - \frac{1}{\sqrt{2}} X_2,$$

$$Y_3 = \frac{1}{\sqrt{6}} X_1 + \frac{1}{\sqrt{6}} X_2 - \frac{2}{\sqrt{6}} X_3,$$

which accounts for $[1 + 2\,(0.6)]\,/3 = 0.7333$ (or 73.33 percent) of the total variance.

In general, consider the p by p covariance matrix

$$\mathbf{\Omega} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

For $0 < \rho \leq 1$.

Note: Please do not mix up p and ρ.

The greatest eigenvalue of this matrix is

$$\lambda_1 = \sigma^2 \left[1 + (p - 1)\,\rho \right]$$

and its normalized eigenvector is

$$\mathbf{e}_1' = \left(\frac{1}{\sqrt{p}}, ..., \frac{1}{\sqrt{p}}\right).$$

The first principal component is

$$Y_1 = \mathbf{e}_1'\mathbf{X} = \frac{1}{\sqrt{p}}X_1 + \frac{1}{\sqrt{p}}X_2 + ... + \frac{1}{\sqrt{p}}X_p,$$

which accounts for

$$\left[1 + (p-1)\,\rho\right]/p$$

of the total variance. The remaining $p-1$ eigenvalues are all equal to

$$\lambda_2 = \lambda_3 = ... = \lambda_p = \sigma^2\left[1-\rho\right].$$

The remaining $p-1$ eigenvectors are

$$\mathbf{e}_2 = \begin{pmatrix} \frac{1}{\sqrt{1\times 2}} \\ \frac{-1}{\sqrt{1\times 2}} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} \frac{1}{\sqrt{2\times 3}} \\ \frac{1}{\sqrt{2\times 3}} \\ \frac{-2}{\sqrt{2\times 3}} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, ..., \mathbf{e}_i = \begin{pmatrix} \frac{1}{\sqrt{(i-1)i}} \\ \vdots \\ \frac{1}{\sqrt{(i-1)i}} \\ \frac{-(i-1)}{\sqrt{(i-1)i}} \\ \vdots \\ 0 \end{pmatrix}, ..., \mathbf{e}_p = \begin{pmatrix} \frac{1}{\sqrt{(p-1)p}} \\ \vdots \\ \vdots \\ \vdots \\ \frac{1}{\sqrt{(p-1)p}} \\ \frac{-(p-1)}{\sqrt{(p-1)p}} \end{pmatrix}.$$

Exercise 5.10: Find the eigenvalues of the correlation matrix

$$\mathbf{\Omega} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Exercise 5.11: Let

$$\mathbf{\Omega} = 3\begin{pmatrix} 1 & 0.6 & \cdots & 0.6 \\ 0.6 & 1 & \cdots & 0.6 \\ \vdots & \vdots & \ddots & \vdots \\ 0.6 & 0.6 & \cdots & 1 \end{pmatrix}.$$

(a) Determine the population components Y_1 to Y_p.

(b) Calculate the proportion of the total population variance explained by Y_1.

5.6 Sample Principal Components

Let the sample covariance matrix of the random vector $\mathbf{X}' = (X_1, X_2, ..., X_p)$ be

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}.$$

The i^{th} sample principal component is given by

$$\widehat{Y}_i = \widehat{\mathbf{e}}_i'\mathbf{X} = \widehat{e}_{i1}X_1 + \widehat{e}_{i2}X_2 + ... + \widehat{e}_{ip}X_p.$$

Sample variance of Y

$$Var\left(\widehat{Y}_i\right) = \widehat{\lambda}_i, \qquad i = 1, 2, ..., p.$$

Sample covariance

$$Cov\left(\widehat{Y}_i, \widehat{Y}_k\right) = 0 \qquad \text{for } i \neq k.$$

Total sample variance of $Y = s_{11} + s_{22} + .. + s_{pp} = \widehat{\lambda}_1 + \widehat{\lambda}_2 + ... + \widehat{\lambda}_p$.

5.7 Standardizing the Sample Principal Components

Let the standardized observations be

$$\underset{n\times p}{\mathbf{Z}} = \begin{pmatrix} \dfrac{x_{11}-\overline{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{12}-\overline{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{1p}-\overline{x}_p}{\sqrt{s_{pp}}} \\ \dfrac{x_{21}-\overline{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{22}-\overline{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{2p}-\overline{x}_p}{\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{x_{n1}-\overline{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{n2}-\overline{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{np}-\overline{x}_p}{\sqrt{s_{pp}}} \end{pmatrix}.$$

The sample mean vector is

$$\overline{\mathbf{Z}} = \begin{pmatrix} \sum_{j=1}^{n} \dfrac{x_{j1}-\overline{x}_1}{\sqrt{s_{11}}} \\ \sum_{j=1}^{n} \dfrac{x_{j2}-\overline{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \sum_{j=1}^{n} \dfrac{x_{jp}-\overline{x}_p}{\sqrt{s_{pp}}} \end{pmatrix} = \mathbf{0}.$$

The i^{th} sample principal component is given by

$$\widehat{Y}_i = \widehat{e}_{i1}Z_1 + \widehat{e}_{i2}Z_2 + ... + \widehat{e}_{ip}Z_p = \widehat{\mathbf{e}}_i'\mathbf{Z}.$$

Sample variance of y

$$Var\left(\widehat{Y}_i\right) = \widehat{\lambda}_i \qquad\qquad i = 1, 2, ..., p.$$

Sample covariance

$$Cov\left(\widehat{Y}_i, \widehat{Y}_k\right) = 0 \qquad \text{for } i \neq k.$$

Total sample variance of $Y = \widehat{\lambda}_1 + \widehat{\lambda}_2 + ... + \widehat{\lambda}_p = p.$

Example 5.5: Let $X_1, ..., X_5$ denote observed weekly rates of return for Allied Chemical, du pont, Union Carbide, Exxon, and Texaco, respectively. Suppose we have

$$\overline{\mathbf{X}} = \begin{pmatrix} 0.0054 \\ 0.0048 \\ 0.0057 \\ 0.0063 \\ 0.0037 \end{pmatrix},$$

and the sample correlation matrix is

$$\mathbf{R} = \begin{pmatrix} 1 & 0.577 & 0.509 & 0.387 & 0.462 \\ 0.577 & 1 & 0.599 & 0.389 & 0.322 \\ 0.509 & 0.599 & 1 & 0.436 & 0.426 \\ 0.387 & 0.389 & 0.436 & 1 & 0.523 \\ 0.462 & 0.322 & 0.426 & 0.523 & 1 \end{pmatrix}.$$

The eigenvalues and the corresponding normalized eigenvectors of R are $\widehat{\lambda}_1 = 2.857$, $\widehat{\lambda}_2 = 0.809$, $\widehat{\lambda}_3 = 0.540$, $\widehat{\lambda}_4 = 0.452$, $\widehat{\lambda}_5 = 0.343$ and

$$\begin{aligned} \mathbf{e}_1 &= \begin{pmatrix} 0.464 \\ 0.457 \\ 0.470 \\ 0.421 \\ 0.421 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0.240 \\ 0.509 \\ 0.269 \\ -0.526 \\ -0.582 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} -0.612 \\ 0.178 \\ 0.335 \\ 0.541 \\ -0.435 \end{pmatrix}, \\ \mathbf{e}_4 &= \begin{pmatrix} 0.387 \\ 0.206 \\ -0.662 \\ 0.472 \\ -0.382 \end{pmatrix}, \mathbf{e}_5 = \begin{pmatrix} -0.451 \\ 0.676 \\ -.400 \\ -0.176 \\ 0.385 \end{pmatrix}. \end{aligned}$$

$$\widehat{Y}_1 = \widehat{\mathbf{e}}_1'\mathbf{Z} = 0.464Z_1 + 0.457Z_2 + 0.470Z_3 + 0.421Z_4 + 0.421Z_5,$$

$$\widehat{Y}_2 = \widehat{\mathbf{e}}_2'\mathbf{Z} = 0.240Z_1 + 0.509Z_2 + 0.269Z_3 - 0.526Z_4 - 0.582Z_5.$$

The first two components account for $\frac{2.857 + 0.809}{5} = 73\%$ of the total standardized sample variance. Note that $\widehat{Y}_1 \approx 0.45\,(Z_1 + Z_2 + Z_3 + Z_4 + Z_5) = 0.45\left(5\overline{Z}\right) = 2.25\overline{Z}$. Therefore, the first component is a roughly proportion to the sample average, which can be perceived as a general stock-market component. The second component represents a contrast between the chemical stocks (Allied Chemical, du Pont, and Union Carbide) and oil stocks (Exxon and Texaco). It might be called an industry component. Thus, most of the variation in these stock returns is due to market activity and uncorrelated industry activity. The remaining components are hard to interpret. They may be variation specific to each stock.

5.8 Determining the Number of Principal Components

Note that some of the Y variables have little variation, so we may drop them without much loss of information. But what is the rule for dropping Y? There are two methods to determine the number of principal components. Both are based on the eigenvalues of covariance matrix Ω. One is to drop those Y with eigenvalue less than one. Another useful rule to determining an appropriate number of principal components is a scree plot, with the eigenvalues ordered from the largest to smallest. For example, if $p = 6$, and the eigenvalues are 2, 0.9, 0.7, 0.24, 0.22, 0.19, then the first three Y should be used. In the previous example, with $\widehat{\lambda}_1 = 2.857$, $\widehat{\lambda}_2 = 0.809$, $\widehat{\lambda}_3 = 0.540$, $\widehat{\lambda}_4 = 0.452$, $\widehat{\lambda}_5 = 0.34$, if we use the first rule, then number of principal components should be one. If we use a scree plot, we may retain the first two principal components.

Exercise 5.12: Find the unadjusted daily closing price from Yahoo Finance for the following Hong Kong stocks from 30/9/2014 to 31/10/2014: [1], [5], [11], [12], [16].

(a) Calculate the daily returns $r_t = \ln P_t - \ln P_{t-1}$ for these stocks from 3/10/2014 to 31/10/2014 using the log difference of price.

(b) Standardized the returns and calculate the sample correlation matrix $\mathbf{R}$ for the standardized daily returns of these 5 stocks.

(c) Based on the sample correlation matrix $\mathbf{R}$, find the sample principal components.

(d) Determine the proportion of the total sample variance explained by the first two principal components.

Chapter 6

Factor Analysis

Suppose variables can be grouped by their correlations. i.e., all variables within a particular group are highly correlated among themselves, but they have relatively small correlations with variables in a different group. Then it is conceivable that each group of variables represents a single underlying construct, or factor, that is unobservable but is responsible for the observed correlations. Factor analysis can be considered as an extension of principal components analysis. Principal components analysis is concerned with explaining the variance in the variables while factor analysis is concerned with explaining the covariances.

Factor analysis is an interdependence technique in which all variables are simultaneously considered, each related to all others. The factor model can be written as

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + ... + l_{1m}F_m + \varepsilon_1,$$

$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + ... + l_{2m}F_m + \varepsilon_2,$$

$$\vdots$$

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + ... + l_{pm}F_m + \varepsilon_p.$$

where

$X_1, X_{2,...,}X_p$ are observed variables;

$F_1, F_2, ..., F_m$, are unobserved common factors, with $m \leq p$;

$\varepsilon_1, \varepsilon_{2,...,}\varepsilon_p$ are the error terms, or can be considered as specific factor.

In matrix notation, we have

$$\underset{(p\times 1)}{\mathbf{X}-\boldsymbol{\mu}} = \underset{(p\times m)}{\mathbf{L}}\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}}.$$

The coefficient l_{ij} is called the loading of the i^{th} variable on the j^{th} factor

$\mathbf{L}$ is the matrix of factor loadings.

$$\mathbf{L} = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{pmatrix}.$$

Note that $\mathbf{F}$ is unobservable, so factor model is different from regression model. We assume that

$$E(\mathbf{F}) = \mathbf{0},$$

$$Cov(\mathbf{F}) = E(\mathbf{F}\mathbf{F}') = \mathbf{I},$$

$$E(\boldsymbol{\varepsilon}) = \mathbf{0},$$

$$Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Psi} = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix},$$

$$Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{0}.$$

These assumptions constitute the orthogonal factor model. The orthogonal factor model implies a covariance structure for $\mathbf{X}$.

$$\begin{aligned}\boldsymbol{\Omega} &= Cov\left(\mathbf{X}\right)=E\left(\mathbf{X}-\boldsymbol{\mu}\right)\left(\mathbf{X}-\boldsymbol{\mu}\right)'\\ &= E\left(\mathbf{LF}+\varepsilon\right)\left(\mathbf{LF}+\varepsilon\right)'\\ &= \mathbf{L}E\left(\mathbf{FF}'\right)\mathbf{L}'+\mathbf{L}E\left(\mathbf{F}\varepsilon'\right)+\left(\varepsilon\mathbf{F}'\right)\mathbf{L}'+E\left(\varepsilon\varepsilon'\right)\\ &= \mathbf{LIL}'+\mathbf{L0}+\mathbf{0L}'+\boldsymbol{\Psi}\\ &= \mathbf{LL}'+\boldsymbol{\Psi}.\end{aligned}$$

In factor analysis, the covariance matrix is partitioned into two parts: that due to the common factors and that due to the unique factors. Any covariance (correlation) not explained by the common factors are associated with the mutual uncorrelated unique (residual) factors. In principal component analysis, there is no residual variance, all variance is explained by the components.

$$\begin{aligned}Cov\left(\mathbf{X},\mathbf{F}\right) &= E\left(\left(\mathbf{X}-\boldsymbol{\mu}\right)\mathbf{F}'\right)\\ &= E\left(\left(\mathbf{LF}+\varepsilon\right)\mathbf{F}'\right)\\ &= \mathbf{L}E\left(\mathbf{FF}'\right)+E\left(\varepsilon\mathbf{F}'\right)\\ &= \mathbf{LI}+\mathbf{0}\\ &= \mathbf{L}.\end{aligned}$$

Thus, we have

$$Cov\left(X_i,F_j\right)=l_{ij}.$$

The portion of variance of the i^{th} variable contributed by the m common factors is called the i^{th} communality, denoted by

$$h_i^2=l_{i1}^2+l_{i2}^2+...+l_{im}^2.$$

The portion of $Var(X_i)$ due to the specific factor is called the uniqueness, or specific variance ψ_i.

$$Var(X_i)=\sigma_{ii}=h_i^2+\psi_i,$$

$i = 1, 2, ..., p$.

Note that what we can observe are the X variables and their covariance structure. We would like to derive the loading matrix.

Example 6.1: Consider the covariance matrix

$$\mathbf{\Omega} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

Suppose there is one factor, i.e., $m = 1$, we can decompose the matrix as

$$\begin{aligned} \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} &= \begin{pmatrix} l_{11} \\ l_{21} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix} \\ &= \begin{pmatrix} l_{11}^2 & l_{11}l_{21} \\ l_{11}l_{21} & l_{21}^2 \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix} \\ &= \begin{pmatrix} l_{11}^2 + \psi_1 & l_{11}l_{21} \\ l_{11}l_{21} & l_{21}^2 + \psi_2 \end{pmatrix} \\ &= \mathbf{LL}' + \mathbf{\Psi}. \end{aligned}$$

We have

$$l_{11}^2 + \psi_1 = 1,$$

$$l_{11}l_{21} = \frac{1}{2},$$

$$l_{21}^2 + \psi_2 = 1.$$

Note that there is no unique solution in this case. One solution is $l_{11} = l_{21} = \sqrt{\frac{1}{2}}$, and $\psi_1 = \psi_2 = \frac{1}{2}$. The portion of variance of the first variable contributed by the single common factor, i.e., the communality of X_1 is

$$h_1^2 = l_{11}^2 = \frac{1}{2}.$$

and the variance of X_1 can be decomposed as

$$Var(X_1) = \sigma_{11} = h_1^2 + \psi_1 = \underbrace{\frac{1}{2}}_{\text{communality}} + \underbrace{\frac{1}{2}}_{\text{specific variance}}.$$

Example 6.2: Consider the covariance matrix

$$\Omega = \begin{pmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{pmatrix}.$$

We can decompose the matrix as

$$\begin{aligned} \begin{pmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{pmatrix} &= \begin{pmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{pmatrix} \begin{pmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{pmatrix} + \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix} \\ &= \mathbf{LL}' + \mathbf{\Psi}. \end{aligned}$$

The portion of variance of the first variable contributed by the 2 common factors, i.e., the communality of X_1 is

$$\begin{aligned} h_1^2 &= l_{11}^2 + l_{12}^2 \\ &= 4^2 + 1^2 \\ &= 17. \end{aligned}$$

and the variance of X_1 can be decomposed as

$$Var(X_1) = \sigma_{11} = 19 = \underbrace{17}_{\text{communality}} + \underbrace{2}_{\text{specific variance}}.$$

A similar breakdown occurs for other variables.

When $m > 1$, there is always some inherent ambiguity associated with the factor model. Let Γ be an $m \times m$ orthogonal matrix such that $\mathbf{\Gamma\Gamma}' = \mathbf{\Gamma}'\mathbf{\Gamma} = \mathbf{I}$.

$$\mathbf{X}-\boldsymbol{\mu}=\mathbf{LF}+\boldsymbol{\varepsilon}=\mathbf{L\Gamma\Gamma'F}+\boldsymbol{\varepsilon}=\mathbf{L}^{*}\mathbf{F}^{*}+\boldsymbol{\varepsilon}.$$

The factors $\mathbf{F}$ and $\mathbf{F}^{*}$ have the same statistical properties, with

$$E\left(\mathbf{F}^{*}\right)=E\left(\mathbf{\Gamma'F}\right)=\mathbf{\Gamma'}E\left(\mathbf{F}\right)=\mathbf{0}.$$

$$Cov\left(\mathbf{F}^{*}\right)=E\left(\mathbf{\Gamma'FF'\Gamma}\right)=\mathbf{\Gamma'}E\left(\mathbf{FF'}\right)\mathbf{\Gamma}=\mathbf{\Gamma'\Gamma}=\mathbf{I}.$$

The loadings $\mathbf{L}^{*}$ are also different from the loadings $\mathbf{L}$

$$\mathbf{\Omega}=\mathbf{LL'}+\mathbf{\Psi}=\mathbf{L}\left(\mathbf{\Gamma\Gamma'}\right)\mathbf{L'}+\mathbf{\Psi}=\left(\mathbf{L}^{*}\right)\left(\mathbf{L}^{*}\right)'+\mathbf{\Psi}.$$

Note that principal component analysis is merely a transformation of the data. No assumptions are made about the form of covariance matrix from which data comes. On the other hand, factor analysis assumes that the data comes from a well-defined model, where underlying factors satisfy the above assumptions. Also, in principal component analysis the emphasis is on a transformation from the observed variables to the principal components, whereas in factor analysis the emphasis is on a transformation from the underlying factors to the observed variables.

Exercise 6.1: Show that the covariance matrix

$$\boldsymbol{\rho}=\begin{pmatrix}1 & 0.63 & 0.45\\ 0.63 & 1 & 0.35\\ 0.45 & 0.35 & 1\end{pmatrix}$$

for standardized random variables Z_1, Z_2 and Z_3 can be generated by the following factor model:

$$Z_1=0.9F_1+\varepsilon_1,$$

$$Z_3=0.7F_1+\varepsilon_2,$$

$$Z_3 = 0.5F_1 + \varepsilon_3.$$

where $Var\left(F_1\right) = 1$, $Cov\left(\boldsymbol{\varepsilon}, F_1\right) = \mathbf{0}$, and

$$\boldsymbol{\Psi} = Cov\left(\boldsymbol{\varepsilon}\right) = \begin{pmatrix} 0.19 & 0 & 0 \\ 0 & 0.51 & 0 \\ 0 & 0 & 0.75 \end{pmatrix}.$$

Exercise 6.2: Suppose the test score of a student depends on its intelligence (an unobservable common factor),

$$Chinese = l_{11}F_1 + \varepsilon_1,$$

$$English = l_{21}F_1 + \varepsilon_2,$$

$$Maths = l_{31}F_1 + \varepsilon_3.$$

and suppose the correlation of the test score is

	Chinese	*English*	*Maths*
Chinese	1	0.4	0.9
English	0.4	1	0.7
Maths	0.9	0.7	1

Show that there is a unique choice of $\mathbf{L}$ and $\boldsymbol{\Psi}$ with $\boldsymbol{\Omega} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$, but that $\psi_3 < 0$, so the choice is not admissible.

6.1 Methods of Estimation

6.1.1 The Principal Component Method

Let $\boldsymbol{\Omega}$ have eigenvalue-eigenvector pairs $(\lambda_i, \mathbf{e}_i)$ with $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p \geqslant 0$ and $m = p$. Then

$$\underset{p\times p}{\mathbf{\Omega}} = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i', = \left(\sqrt{\lambda_1}\mathbf{e}_1, \quad \sqrt{\lambda_2}\mathbf{e}_2, \quad \cdots \quad , \sqrt{\lambda_p}\mathbf{e}_p \right) \begin{pmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \sqrt{\lambda_2}\mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_p}\mathbf{e}_p' \end{pmatrix} = \underset{(p\times p)}{\mathbf{L}}\underset{(p\times p)}{\mathbf{L}}'.$$

In this case, if all the p factors are used, we have

$$\mathbf{\Psi} = \mathbf{0}.$$

Note that since not all factors are used, if we just use m factors $(m < p)$, then

$$\begin{aligned} \mathbf{\Omega} &\approx \sum_{i=1}^{m} \lambda_i \mathbf{e}_i \mathbf{e}_i', \\ &= \left(\sqrt{\lambda_1}\mathbf{e}_1, \quad \sqrt{\lambda_2}\mathbf{e}_2, \quad \cdots \quad , \sqrt{\lambda_m}\mathbf{e}_m \right) \begin{pmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \sqrt{\lambda_2}\mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_m}\mathbf{e}_m' \end{pmatrix} \\ &= \underset{(p\times m)}{\mathbf{L}}\underset{(m\times p)}{\mathbf{L}}'. \end{aligned}$$

Allowing for specific factors, the approximation becomes

$$\begin{aligned} \mathbf{\Omega} &\approx \mathbf{L}\mathbf{L}' + \mathbf{\Psi} \\ &= \left(\sqrt{\lambda_1}\mathbf{e}_1, \quad \sqrt{\lambda_2}\mathbf{e}_2, \quad \cdots \quad , \sqrt{\lambda_m}\mathbf{e}_m \right) \begin{pmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \sqrt{\lambda_2}\mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_m}\mathbf{e}_m' \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}, \end{aligned}$$

where

$$\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2.$$

6.1.2 Maximum Likelihood Method

If we assume F and ε to be jointly normal, the observations X are then normal. For each observation $\mathbf{x}_j = (x_{j1}, x_{j2}, ..., x_{jp})'$. The joint density of $\mathbf{x}_j$ will be

$$f\left(x_{j1}, x_{j2}, ..., x_{jp}\right) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Omega}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)' \boldsymbol{\Omega}^{-1}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)\right).$$

This is the joint density for one point of observation of p variables. If we have n points of observations in our sample, and if each observation is obtained independently, the overall joint density will be

$$\begin{aligned}
& \Pi_{j=1}^{n}\frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Omega}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)' \boldsymbol{\Omega}^{-1}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)\right) \\
= \; & \frac{1}{(2\pi)^{np/2}\left|\boldsymbol{\Omega}\right|^{n/2}}\Pi_{j=1}^{n} \exp\left(-\frac{1}{2}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)' \boldsymbol{\Omega}^{-1}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)\right) \\
= \; & \frac{1}{(2\pi)^{np/2}\left|\boldsymbol{\Omega}\right|^{n/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)' \boldsymbol{\Omega}^{-1}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)\right) \\
= \; & \frac{1}{(2\pi)^{np/2}\left|\mathbf{L}\mathbf{L}'+\boldsymbol{\Psi}\right|^{n/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)' \left(\mathbf{L}\mathbf{L}'+\boldsymbol{\Psi}\right)^{-1}\left(\mathbf{x}_j - \boldsymbol{\mu}\right)\right).
\end{aligned}$$

This joint density function is a function of $\mathbf{X}$. Given our data $\mathbf{X}$, we can also consider it as a function of $\mathbf{L}$ and $\boldsymbol{\Psi}$, we call this the likelihood function. The maximum likelihood method is to choose the values in $\mathbf{L}$ and $\boldsymbol{\Psi}$ to maximize the above function. We can solve for the initial loadings and $\boldsymbol{\Psi}$ after proper constraints are imposed.

6.2 Factor Rotation

When a set of factors are derived, they are not always easy to interpret. Do not try to interpret underlying factors until you have performed a factor rotation. Most rotation procedures try to make as many factor loadings as possible near zero and to maximize as many of the others as possible.

Since factors are independent, it would be nice if response variables were not loaded heavily on more than one factor. Consider the rotation in the two factor cases. Let $\widehat{\mathbf{L}}$ be the original unrotated loadings, the rotated loading is given by

$$\underset{(p\times 2)}{\widehat{\mathbf{L}}}^{*} = \underset{(p\times 2)}{\widehat{\mathbf{L}}}\underset{(2\times 2)}{\mathbf{\Gamma}},$$

where

$$\mathbf{\Gamma} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \qquad \text{clockwise rotation;}$$

$$\mathbf{\Gamma} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \qquad \text{counterclockwise rotation.}$$

Example 6.3: Consider a simple case where $p = 2$ and

$$\widehat{\mathbf{L}} = \begin{pmatrix} 0.56 & 0.82 \\ 0.78 & -0.52 \end{pmatrix},$$

what is the new coordinate if the axes are rotated clockwise / counterclockwise by 45^o?

Solution: For clockwise rotation

$$\begin{aligned} \underset{(2\times 2)}{\widehat{\mathbf{L}}}^{*} &= \underset{(2\times 2)(2\times 2)}{\widehat{\mathbf{L}}\ \mathbf{\Gamma}} = \begin{pmatrix} 0.56 & 0.82 \\ 0.78 & -0.52 \end{pmatrix}\begin{pmatrix} \cos 45^o & \sin 45^o \\ -\sin 45^o & \cos 45^o \end{pmatrix} \\ &= \begin{pmatrix} 0.56 & 0.82 \\ 0.78 & -0.52 \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} -0.1838 & 0.9758 \\ 0.9192 & 0.1838 \end{pmatrix}. \end{aligned}$$

For counterclockwise rotation

$$\begin{aligned}
\underset{(2\times 2)}{\widehat{\mathbf{L}}^{*}} &= \underset{(2\times 2)}{\widehat{\mathbf{L}}}\,\underset{(2\times 2)}{\mathbf{\Gamma}} \\
&= \begin{pmatrix} 0.56 & 0.82 \\ 0.78 & -0.52 \end{pmatrix} \begin{pmatrix} \cos 45^o & -\sin 45^o \\ \sin 45^o & \cos 45^o \end{pmatrix} \\
&= \begin{pmatrix} 0.56 & 0.82 \\ 0.78 & -0.52 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \\
&= \begin{pmatrix} 0.9758 & 0.1838 \\ 0.1838 & -0.9192 \end{pmatrix}.
\end{aligned}$$

Example 6.2: Consider the following correlation matrix on test scores on 100 students

	Chinese	*English*	*History*	*Maths*	*A. Maths*	*Physics*
Chinese	1	.439	.410	.288	.329	.248
English	.439	1	.351	.354	.320	.329
History	.410	.351	1	.164	.190	.181
Maths	.288	.354	.164	1	.595	.470
A. Maths	.329	.320	.190	.595	1	.464
Physics	.248	.329	.181	.470	.464	1

The maximum likelihood solution is

	Unrotated	*Loadings*	*Rotated*	*Loadings*	$(\phi \simeq 20^o)$	
					Communalities	*Specific*
	F_1	F_2	F_1^*	F_2^*	$\widehat{h}_i^{*2} = l_{i1}^{*2} + l_{i2}^{*2}$	*Variance*
					$= l_{i1}^2 + l_{i2}^2 = \widehat{h}_i^2$	$\widehat{\psi}_i = 1 - \widehat{h}_i^2$
1.*Chinese*	.553	.429	.369	**.594**	.490	.510
2.*English*	.568	.288	.433	**.467**	.406	.594
3.*History*	.392	.450	.211	**.558**	.356	.644
4.*Maths*	.740	−.273	**.789**	.001	.623	.377
5.*A. Maths*	.724	−.211	**.752**	.054	.568	.432
6.*Physics*	.595	−.132	**.752**	.083	.372	.628

Note that half of the original loadings are positive and another half of them are negative for the second factor. A factor with this pattern of loading is called a bipolar factor. We rotate the original factor by about 20 degrees. This angle is chosen so that one of the new axes passes through the fourth point (0.740, -.273). Note that all values are positive now after the notation, and the two distinct clusters of variables are more clearly revealed. The first factor might be called a mathematical-ability factor, while the second factor might be labeled a verbal-ability factor.

6.3 Varimax Rotation Method

Sometimes, it may not be possible to rotate the factors just by visual inspection, especially when we are dealing with a higher dimensional space. Let $\widehat{l}^*_{i1}, \widehat{l}^*_{i2}, ..., \widehat{l}^*_{im}$ be the estimated rotated loadings with the estimated communality $\widehat{h}^{*2}_i = \widehat{l}^{*2}_{i1} + \widehat{l}^{*2}_{i2} + ... + \widehat{l}^{*2}_{im}$. Let

$$\widetilde{l}^{*2}_{ij} = \frac{\widehat{l}^{*2}_{ij}}{\widehat{h}^{*2}_i}.$$

The varimax procedure selects the orthogonal transformation $\mathbf{\Gamma}$ that maximizes

$$V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\widetilde{l}^{*4}_{ij} - \frac{1}{p}\left(\sum_{i=1}^{p}\widetilde{l}^{*2}_{ij}\right)^2\right].$$

It can be rewritten as

$$V = \sum_{j=1}^{m}\left[\frac{1}{p}\sum_{i=1}^{p}\left(\widetilde{l}^{*2}_{ij} - \overline{\widetilde{l}^{*2}_{j}}\right)^2\right].$$

where

$$\overline{\widetilde{l}^{*2}_{j}} = \frac{1}{p}\sum_{i=1}^{p}\widetilde{l}^{*2}_{ij},$$

V can be considered as the sum of variance of squares of scaled loadings for the j^{th} factor for all j. Since the squared loadings are all between 0 and 1, trying to maximize the variance of the squared loadings within a column is somewhat equivalent to trying to spread out the squared loadings within a column, i.e., forcing as many of the loadings as possible towards 0 and forcing the others towards 1. After solving $\widetilde{l}^*_{ij}$, we can solve $\widehat{l}^*_{ij} = \widehat{h}^*_i \widetilde{l}^*_{ij}$.

Example 6.3: Consider the rotated loadings in Example 6.2. Calculate the value of V.

Solution:

Note that $m = 2$ and $p = 6$ in this case, we have

$$\widetilde{l}^{*2}_{ij} = \frac{\widehat{l}^{*2}_{ij}}{\widehat{l}^{*2}_{i1} + \widehat{l}^{*2}_{i2}}.$$

$$\widetilde{l}^{*2}_{11} = \frac{\widehat{l}^{*2}_{11}}{\widehat{l}^{*2}_{11} + \widehat{l}^{*2}_{12}} = \frac{.369^2}{.369^2 + .594^2} = 0.27845,$$

$$\widetilde{l}^{*2}_{21} = \frac{\widehat{l}^{*2}_{21}}{\widehat{l}^{*2}_{21} + \widehat{l}^{*2}_{22}} = \frac{.433^2}{.433^2 + .467^2} = 0.46228,$$

$$\widetilde{l}^{*2}_{31} = \frac{\widehat{l}^{*2}_{31}}{\widehat{l}^{*2}_{31} + \widehat{l}^{*2}_{32}} = \frac{.211^2}{.211^2 + .558^2} = 0.12510,$$

$$\widetilde{l}^{*2}_{41} = \frac{\widehat{l}^{*2}_{41}}{\widehat{l}^{*2}_{41} + \widehat{l}^{*2}_{42}} = \frac{.789^2}{.789^2 + .001^2} = 1.00000,$$

$$\widetilde{l}^{*2}_{51} = \frac{\widehat{l}^{*2}_{51}}{\widehat{l}^{*2}_{51} + \widehat{l}^{*2}_{52}} = \frac{.752^2}{.752^2 + .054^2} = 0.99487,$$

$$\widetilde{l}^{*2}_{61} = \frac{\widehat{l}^{*2}_{61}}{\widehat{l}^{*2}_{61} + \widehat{l}^{*2}_{62}} = \frac{.752^2}{.752^2 + .083^2} = 0.98796,$$

$$\overline{\widetilde{l}^{*2}_1} = \frac{0.27845 + 0.46228 + 0.12510 + 1.00000 + 0.99487 + 0.98796}{6} = 0.64144.$$

$$\widetilde{l}_{12}^{*2} = \frac{\widehat{l}_{12}^{*2}}{\widehat{l}_{11}^{*2} + \widehat{l}_{12}^{*2}} = \frac{.594^2}{.369^2 + .594^2} = 0.72155,$$

$$\widetilde{l}_{22}^{*2} = \frac{\widehat{l}_{22}^{*2}}{\widehat{l}_{21}^{*2} + \widehat{l}_{22}^{*2}} = \frac{.467^2}{.433^2 + .467^2} = 0.53772,$$

$$\widetilde{l}_{32}^{*2} = \frac{\widehat{l}_{32}^{*2}}{\widehat{l}_{31}^{*2} + \widehat{l}_{32}^{*2}} = \frac{.558^2}{.211^2 + .558^2} = 0.87490,$$

$$\widetilde{l}_{42}^{*2} = \frac{\widehat{l}_{42}^{*2}}{\widehat{l}_{41}^{*2} + \widehat{l}_{42}^{*2}} = \frac{.001^2}{.789^2 + .001^2} = 0.00000,$$

$$\widetilde{l}_{52}^{*2} = \frac{\widehat{l}_{52}^{*2}}{\widehat{l}_{51}^{*2} + \widehat{l}_{52}^{*2}} = \frac{.054^2}{.752^2 + .054^2} = 0.00513,$$

$$\widetilde{l}_{62}^{*2} = \frac{\widehat{l}_{62}^{*2}}{\widehat{l}_{61}^{*2} + \widehat{l}_{62}^{*2}} = \frac{.083^2}{.752^2 + .083^2} = 0.01204,$$

$$\overline{\widetilde{l}_{1}^{*2}} = \frac{0.721\,55 + 0.53772 + 0.87490 + 0.00000 + 0.00513 + 0.01204}{6} = 0.35856.$$

$$\begin{aligned}
V &= \sum_{j=1}^{2} \left[\frac{1}{6} \sum_{i=1}^{6} \left(\widetilde{l}_{ij}^{*2} - \overline{\widetilde{l}_{j}^{*2}} \right)^2 \right] \\
&= \frac{1}{6} \left[\begin{array}{c} (0.27845 - 0.64144)^2 + (0.46228 - 0.64144)^2 + (0.12510 - 0.64144)^2 \\ + (1.00000 - 0.64144)^2 + (0.99487 - 0.64144)^2 + (0.98796 - 0.64144)^2 \end{array} \right] \\
& \quad \frac{1}{6} \left[\begin{array}{c} (0.72155 - 0.35856)^2 + (0.53772 - 0.35856)^2 + (0.87490 - 0.35856)^2 \\ + (0.00000 - 0.35856)^2 + (0.00513 - 0.35856)^2 + (0.01204 - 0.35856)^2 \end{array} \right] \\
&= \frac{1}{6}(0.804) + \frac{1}{6}(0.804) \\
&= 0.268.
\end{aligned}$$

Exercise 6.3: Repeat the calculation of Example 6.3 using the unrotated loadings in Example 6.2. Compare the value of V in both cases.

Exercise 6.4:

(a) Show that

$$\frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\widetilde{l}_{ij}^{*4}-\frac{1}{p}\left(\sum_{i=1}^{p}\widetilde{l}_{ij}^{*2}\right)^2\right]=\sum_{j=1}^{m}\left[\frac{1}{p}\sum_{i=1}^{p}\left(\widetilde{l}_{ij}^{*2}-\overline{\widetilde{l}_{j}^{*2}}\right)^2\right].$$

(b) When $m=2$, show that

$$\sum_{i=1}^{p}\left(\widetilde{l}_{i1}^{*2}-\overline{\widetilde{l}_{1}^{*2}}\right)^2=\sum_{i=1}^{p}\left(\widetilde{l}_{i2}^{*2}-\overline{\widetilde{l}_{2}^{*2}}\right)^2.$$

Exercise 6.5: Find the daily closing price of the following Hong Kong stocks from 3/10/2014 to 31/10/2014: [1], [2], [3], [16], [823].

(a) Calculate the daily returns $r_t = \ln P_t - \ln P_{t-1}$ for these stocks from 3/10/2014 to 31/10/2014 .

(b) Standardized the returns and calculate the sample correlation matrix $\mathbf{R}$ for the standardized daily returns of these 5 stocks.

(c) Based on the sample correlation matrix, perform a factor analysis assuming there are 2 factors. Solve the factor model using the principal component method. Find the communalities and the proportion of variance explained by each factor.

(d) Find the residual matrix $\mathbf{R}-\widehat{\mathbf{L}}\widehat{\mathbf{L}}'-\widehat{\mathbf{\Psi}}$.

(e) Perform a Varimax rotation.

Exercise 6.6: True/False.

(a) The portion of variance contributed by the i^{th} factor is called the i^{th} communality.

(b) Six factors can be obtain from five variables.

(c) Consider the estimated loadings in the two factor case, with $\widehat{\mathbf{L}} = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}$. The new loading matrix is $\widehat{\mathbf{L}}^* = \begin{pmatrix} 3 & 4 \\ 4 & 3 \end{pmatrix}$ if the axes are rotated clockwise by 90^o.

(d) Most rotation procedures try to make the factor loadings as close to each other as possible.

Exercise 6.7: Consider the estimated loadings in the two factor case, with

$$\widehat{\mathbf{L}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

(a) What is the new coordinate if the axes are rotated clockwise by 45^o?

(b) What is the new coordinate if the axes are rotated counterclockwise by 45^o?

(c) Repeat (a) and (b) if

$$\widehat{\mathbf{L}} = \begin{pmatrix} 0.5 & 0.8 \\ 0.7 & -0.5 \end{pmatrix}.$$

Chapter 7

Discrimination and Classification

7.1 Introduction

Discrimination and classification are multivariate techniques concerned with separating distinct sets of or observations and with allocating new observations to previously defined groups. A good classification procedure should avoid misclassification. In other words, the probability of misclassification should be small. Consider a very simple example, suppose we have two groups of population π_1 and π_2. For population 1, we have

$$Pr(x=0)=0.25, \qquad Pr(x=1)=0.5, \qquad Pr(x=2)=0.25.$$

For population 2, we have

$$Pr(x=1)=0.25, \qquad Pr(x=2)=0.5, \qquad Pr(x=3)=0.25.$$

If we have an observation with value $x_0 = 1$, should we classify this observation as population 1 or population 2? Suppose each population has the same size, and there is no misclassification cost, we should classify this observation as population 1, since it has a probability of 0.5, which is higher than the probability that this observation is coming from population 2.

However, if we know that the size of population 2 is much larger than population 1, for example, let p_1 be the prior probability of π_1 and p_2 be the prior probability of π_2, where $p_1 + p_2 = 1$. If $p_1 = 0.01$, and $p_2 = 0.99$, it may be more reasonable to classify an observation as population 2. Therefore, an optimal classification rule should take these "prior probability of occurrence" into account. An empirical example is that there tend to be more financially sound firms than bankrupt firms. If we really believe that the (prior) probability of a financially distressed and ultimately bankrupted firm is very small, then one should classify a randomly selected firm as non-bankrupt unless the data overwhelmingly favors bankruptcy.

Another consideration in classifying observations is the cost of misclassification. In general, the cost of the two type of misclassification are not equal. Sometimes, classifying a π_1 observation as belonging to π_2 represents a more serious error than classifying a π_2 observation as belonging to π_1.

In the previous example, suppose the sizes of the two population are the same, but the costs of misclassification are different. For example, if the cost of misclassifying π_2 observation as belonging to π_1 is 1000 HK dollars, but the cost of misclassifying π_1 observation as belonging to π_2 is only 1 HK dollar. Then you may have a second thought when you would like to classify the observation as π_1 in the previous example. In reality, for example, failing to diagnose a potentially fatal illness is substantially more "costly" than concluding that disease is present when it is not. Therefore, an optimal classification procedure should also account for the costs associated with misclassification.

7.2 Expected cost of misclassification (ECM)

Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be the probability density function associated with the $p \times 1$ vector random variable $\mathbf{X}$ for the population π_1 and π_2 respectively.

An observation with associated measurement $\mathbf{x}$ must be assigned to either π_1 or π_2. Let R_1 be the set of $\mathbf{x}$ values for which we classify objects as π_1 and R_2 be the remaining values for which we classify objects as π_2.

The conditional probability of classifying an observation from π_1 as π_2 is

$$P\left(2|1\right) = P\left(\mathbf{X} \in R_2|\pi_1\right).$$

The conditional probability of classifying an observation from π_2 as π_1 is

$$P\left(1|2\right) = P\left(\mathbf{X} \in R_1|\pi_2\right).$$

Let p_1 be the prior probability of π_1 and p_2 be the prior probability of π_2, where $p_1 + p_2 = 1$. We have

P(observation is correctly classified as π_1)
$= P$(observation comes from π_1 and is correctly classified as π_1)
$= P\left(\mathbf{X} \in R_1|\pi_1\right) P\left(\pi_1\right) = P\left(1|1\right) p_1$.

P(observation is misclassified as π_1)
$= P$(observation comes from π_2 and is misclassified as π_1)
$= P\left(\mathbf{X} \in R_1|\pi_2\right) P\left(\pi_2\right) = P\left(1|2\right) p_2$.

P(observation is correctly classified as π_2)
$= P$(observation comes from π_2 and is correctly classified as π_2)
$= P\left(\mathbf{X} \in R_2|\pi_2\right) P\left(\pi_2\right) = P\left(2|2\right) p_2$.

$= P$(observation is misclassified as π_2)
$= P$(observation comes from π_1 and is misclassified as π_2)
$= P\left(\mathbf{X} \in R_2|\pi_1\right) P\left(\pi_1\right) = P\left(2|1\right) p_1$.

The costs of misclassification can be defined by a cost matrix

		Classify	as
		π_1	π_2
True	π_1	0	$c\,(2\vert 1)$
population	π_2	$c(1\vert 2)$	0

We define the expected cost of misclassification (ECM) as

$$\begin{aligned} ECM &= c\,(2|1)\,P(\text{observation is misclassified as }\pi_2) \\ &\quad +c(1|2)P(\text{observation is misclassified as }\pi_1) \\ &= c\,(2|1)\,P\,(2|1)\,p_1 + c(1|2)P\,(1|2)\,p_2. \end{aligned}$$

It can be proved (difficult) that the regions R_1 and R_2 that minimize the ECM are defined by the values of $\mathbf{x}$ for which the following inequalities hold

$$R_1 : \frac{f_1\,(\mathbf{x})}{f_2\,(\mathbf{x})} \geq \frac{c(1|2)}{c\,(2|1)}\frac{p_2}{p_1},$$

$$R_2 : \frac{f_1\,(\mathbf{x})}{f_2\,(\mathbf{x})} < \frac{c(1|2)}{c\,(2|1)}\frac{p_2}{p_1}.$$

In other words, we compare the values of

$$c\,(2|1)\,f_1\,(\mathbf{x})\,p_1$$

and

$$c\,(1|2)\,f_2\,(\mathbf{x})\,p_2.$$

We allocate $\mathbf{x}_0$ to π_1 if

$$c\,(1|2)\,f_2\,(\mathbf{x}_0)\,p_2 < c\,(2|1)\,f_1\,(\mathbf{x}_0)\,p_1.$$

7.3 Special cases

1. If $p_1 = p_2$

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)},$$

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)}.$$

2. If $c(2|1) = c(1|2)$

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1},$$

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}.$$

3. If $\dfrac{c(1|2)}{c(2|1)}\dfrac{p_2}{p_1} = 1$

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1,$$

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1.$$

Example 7.1: Consider the case of one X variable. Suppose the first group of X is normally distributed with $N(0,1)$, and the second group of X is normally distributed with $N(2,1)$. Consider a point $x_0 = 0.5$, which group does this point belong to if $\dfrac{c(1|2)}{c(2|1)}\dfrac{p_2}{p_1} = 1$?

Solution:

$$\begin{aligned}
\frac{f_1(x)}{f_2(x)} &= \frac{\dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{(0.5-0)^2}{2}\right)}{\dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{(0.5-2)^2}{2}\right)} \\
&= \exp\left(\frac{(0.5-2)^2}{2} - \frac{(0.5-0)^2}{2}\right) \\
&= \exp(1) \simeq 2.71828 \geq 1.
\end{aligned}$$

So $x_0 = 0.5 \in R_1$ and we should classify $x_0 = 0.5$ to group 1.

7.4 Classification of normal population when $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$

Suppose that the joint density of $\mathbf{X} = (X_1,X_2,...,X_p)'$ for population π_1 and π_2 are given by

$$f_1(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{\Omega}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right),$$

$$f_2(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{\Omega}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right).$$

Here, we assume $\mathbf{\Omega}_1 = \mathbf{\Omega}_2 = \mathbf{\Omega}$. Using the fact that the product of the matrices $\mathbf{a}'\mathbf{B}\mathbf{c} = \mathbf{c}'\mathbf{B}\mathbf{a}$ if $\mathbf{a}'\mathbf{B}\mathbf{c}$ is a 1 by 1 scalar, we have

$$\begin{aligned}
\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{\frac{1}{(2\pi)^{p/2}|\mathbf{\Omega}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{\Omega}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right)}{\frac{1}{(2\pi)^{p/2}|\mathbf{\Omega}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{\Omega}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right)} \\
&= \exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{\Omega}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{\Omega}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right) \\
&= \exp\begin{pmatrix} \frac{1}{2}\mathbf{x}'\mathbf{\Omega}^{-1}\mathbf{x} - \boldsymbol{\mu}_2'\mathbf{\Omega}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_2'\mathbf{\Omega}^{-1}\boldsymbol{\mu}_2 \\ -\frac{1}{2}\mathbf{x}'\mathbf{\Omega}^{-1}\mathbf{x} + \boldsymbol{\mu}_1'\mathbf{\Omega}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_1'\mathbf{\Omega}^{-1}\boldsymbol{\mu}_1 \end{pmatrix} \\
&= \exp\left((\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\mathbf{\Omega}^{-1}\mathbf{x} - \frac{1}{2}\left(\boldsymbol{\mu}_1'\mathbf{\Omega}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\mathbf{\Omega}^{-1}\boldsymbol{\mu}_2\right)\right) \\
&= \exp\left((\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\mathbf{\Omega}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\mathbf{\Omega}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right).
\end{aligned}$$

The regions R_1 and R_2 that minimize the expected cost of misclassification (ECM) are defined by the values of $\mathbf{x}$ for which the following inequalities hold

$$R_1 : \exp\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Omega}^{-1}\mathbf{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Omega}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right) \geq \frac{c(1|2)}{c\,(2|1)} \frac{p_2}{p_1},$$

$$R_2 : \exp\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Omega}^{-1}\mathbf{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Omega}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right) < \frac{c(1|2)}{c\,(2|1)} \frac{p_2}{p_1}.$$

Thus, we allocate a point $\mathbf{x}_0$ to population 1 if

$$\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Omega}^{-1}\mathbf{x}_0 - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Omega}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right) \geq \ln\left(\frac{c(1|2)}{c\,(2|1)} \frac{p_2}{p_1} \right).$$

The above is based on the assumption that $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\mathbf{\Omega}$ are known. In an empirical sample, we have to replace $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ by $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$ respectively. How about the sample variance? The two sample variance $\mathbf{S}_1$ and $\mathbf{S}_2$ will generally be different. Under the assumption that $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$, we pool the two sample variances together and let

$$\mathbf{S}_{pooled} = \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} \mathbf{S}_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} \mathbf{S}_2.$$

Therefore, in an observed sample, we allocate a point $\mathbf{x}_0$ to population 1 if

$$\left((\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1}\mathbf{x}_0 - \frac{1}{2} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2) \right) \geq \ln\left(\frac{c(1|2)}{c\,(2|1)} \frac{p_2}{p_1} \right).$$

If $\dfrac{c(1|2)}{c\,(2|1)} \dfrac{p_2}{p_1} = 1$, we allocate a point $\mathbf{x}_0$ to π_1 if

$$\left((\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1}\mathbf{x}_0 - \frac{1}{2} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2) \right) \geq \ln(1) = 0.$$

or equivalently

$$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} \mathbf{x}_0 \geq (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} \left(\frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2} \right).$$

Therefore, if $\frac{c(1|2)}{c\,(2|1)} \frac{p_2}{p_1} = 1$, we can define the linear discriminant function as

$$\widehat{y} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} \mathbf{x} = \widehat{\mathbf{a}}' \mathbf{x}.$$

Evaluate $\widehat{y}$ at $\mathbf{x}_0$ and compare $\widehat{y}_0$ to

$$\widehat{m} = \frac{\overline{y}_1 + \overline{y}_2}{2},$$

where

$$\overline{y}_1 = \widehat{\mathbf{a}}' \overline{\mathbf{x}}_1 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} \overline{\mathbf{x}}_1,$$

$$\overline{y}_2 = \widehat{\mathbf{a}}' \overline{\mathbf{x}}_2 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} \overline{\mathbf{x}}_2.$$

Intuitively speaking, if there is only one X variable, and assume $\overline{x}_1 - \overline{x}_2 > 0$, we allocate a point x_0 to population 1 if $x_0 \geq \frac{\overline{x}_1 + \overline{x}_2}{2}$, i.e., if the observation x_0 is above the mid-point of the two sample mean, or equivalently if x_0 is closer to the bigger mean $\overline{x}_1$, we allocate it to population 1. If there are more than one X variables, we transform the set of X variables into a scalar value $\widehat{y}$ and compare $\widehat{y}_0$ with $\frac{\overline{y}_1 + \overline{y}_2}{2}$.

Example 7.2: Consider the following mean vectors

$$\overline{\mathbf{x}}_1 = \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix},$$

$$\overline{\mathbf{x}}_2 = \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix},$$

$$\mathbf{S}_{pooled}^{-1} = \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix}.$$

Should the point $\mathbf{x}_0 = \begin{pmatrix} -0.210 \\ -0.044 \end{pmatrix}$ be classified as population 1 or 2 if $\dfrac{c(1|2)}{c\,(2|1)}\dfrac{p_2}{p_1} = 1$?

Solution: The linear discriminant function is

$$\begin{aligned}
\widehat{y} &= \widehat{\mathbf{a}}'\mathbf{x} \\
&= (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\,\mathbf{S}_{pooled}^{-1}\mathbf{x} \\
&= \left(\begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix} - \begin{pmatrix} -0.2483 \\ -0.0262 \end{pmatrix}\right)' \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= \begin{pmatrix} 0.2418 & -0.0652 \end{pmatrix} \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= 37.61x_1 - 28.92x_2.
\end{aligned}$$

$$\overline{y}_1 = \widehat{\mathbf{a}}'\overline{\mathbf{x}}_1 = \begin{pmatrix} 37.61 & -28.92 \end{pmatrix} \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix} = 0.88.$$

$$\overline{y}_2 = \widehat{\mathbf{a}}'\overline{\mathbf{x}}_2 = \begin{pmatrix} 37.61 & -28.92 \end{pmatrix} \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix} = -10.10.$$

$$\widehat{m} = \frac{1}{2}\left(\overline{y}_1 + \overline{y}_2\right) = \frac{1}{2}\left(0.88 - 10.10\right) = -4.61.$$

$$\begin{aligned}
\widehat{y}_0 &= 37.61\,(-0.210) - 28.92\,(-0.044) \\
&= -6.62 \\
&< -4.61 \\
&= \widehat{m}.
\end{aligned}$$

Therefore, we classify $\mathbf{x}_0 = \begin{pmatrix} -0.210 \\ -0.044 \end{pmatrix}$ as π_2.

Exercise 7.1: Consider the following data sets

$$\mathbf{X}_1 = \begin{pmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{pmatrix}, \qquad \mathbf{X}_2 = \begin{pmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{pmatrix}$$

$$\overline{\mathbf{x}}_1 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \qquad \overline{\mathbf{x}}_2 = \begin{pmatrix} 5 \\ 8 \end{pmatrix}, \qquad \mathbf{S}_{pooled} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

(a) Calculate the linear discriminant function $\widehat{y} = \widehat{\mathbf{a}}'\mathbf{x}$.

(b) Should the point $\mathbf{x}_0 = \begin{pmatrix} 2 \\ 7 \end{pmatrix}$ be classified as population 1 or 2 if $\dfrac{c(1|2)}{c\,(2|1)}\dfrac{p_2}{p_1} = 1$?

Example 7.3: The following table shows the survey results for the evaluation of a new model of mobile phone. Evaluation are made on a 10-point scale (1=very poor to 10=excellent).

Group based on purchase intention	x_1(Durability)	x_2(Performance)	x_3(Style)
Group 1: Would purchase			
Subject 1	8	9	6
Subject 2	6	7	5
Subject 3	10	6	3
Subject 4	9	4	4
Subject 5	4	8	2
Group mean	7.4	6.8	4.0
Group 2: Would not purchase			
Subject 6	5	4	7
Subject 7	3	7	2
Subject 8	4	5	5
Subject 9	2	4	3
Subject 10	2	2	2
Group mean	3.2	4.4	3.8
Difference between group mean	4.2	2.4	0.2

Group\Discriminant function	$\widehat{y} = x_1$	$\widehat{y} = x_1 + x_2$	$\widehat{y} = -4.53 + 0.476x_1 + 0.359x_2$
Group 1: Would purchase			
Subject 1	8	17	2.51
Subject 2	6	13	0.84
Subject 3	10	16	2.38
Subject 4	9	13	1.19
Subject 5	4	12	0.25
Group 2: Would not purchase			
Subject 6	5	9	−0.71
Subject 7	3	10	−0.59
Subject 8	4	9	−0.83
Subject 9	2	6	−2.14
Subject 10	2	4	−2.86
Cutting score	5.3	10.9	−0.32

Classification accuracy for $\widehat{y} = x_1$, using the cutting score of 5.3:

	Predicted	group
Actual group	1	2
1: Would purchase	4	1
2: Would not purchase	0	5

Classification accuracy for $\widehat{y} = x_1 + x_2$, using the cutting score of 10.9:

	Predicted	group
Actual group	1	2
1: Would purchase	5	0
2: Would not purchase	0	5

Classification accuracy for $\widehat{y} = -4.53 + 0.476x_1 + 0.359x_2$, using the cutting score of -0.32:

	Predicted	group
Actual group	1	2
1: Would purchase	5	0
2: Would not purchase	0	5

Exercise 7.2: Suppose we would like to classify stocks into Hang Seng Index Constituent Stocks and non-Constituent Stocks. As of 31/10/2014, we obtain the following financial information from the efinet website at http://www.finet.hk/mainsite/index.htm.

Stock code	**Company name**	**Total market capitalization (billions)**	**PE Ratio**	**HSI Constituent Stock**
[1]	Cheung Kong	318.70	9.041	*Yes*
[16]	Sun Hung Kai Properties	315.45	9.285	*Yes*
[66]	MTR Corporation	183.91	14.044	*Yes*
[11]	Hang Seng Bank	251.22	9.419	*Yes*
[388]	HK Exchanges and Clearing	200.78	43.519	*Yes*
[8]	PCCW	36.74	18.976	*No*
[10]	Hang Lung Group	52.84	11.538	*No*
[20]	Wheelock	75.89	4.478	*No*
[54]	Hopewell Holdings	23.96	17.628	*No*
[823]	The Link	104.457	6.065	*No*

We can summarize the data as the following matrices:

$$\mathbf{X}_1 = \begin{pmatrix} 318.70 & 9.041 \\ 315.45 & 9.285 \\ 183.91 & 14.044 \\ 251.22 & 9.419 \\ 200.78 & 43.519 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 36.74 & 18.976 \\ 52.84 & 11.538 \\ 75.89 & 4.478 \\ 23.96 & 17.628 \\ 104.457 & 6.065 \end{pmatrix}.$$

(a) Find the mean vectors $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$.

(b) Assume the variance covariance matrices are the same for the two populations, find the sample pooled variance matrix

$$\mathbf{S}_{pooled} = \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\mathbf{S}_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\mathbf{S}_2.$$

(c) Assume joint normality of the two populations and suppose $\frac{c(1|2)}{c\,(2|1)}\frac{p_2}{p_1} = 1$, find the linear discriminant function

$$\widehat{y} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\,\mathbf{S}_{pooled}^{-1}\mathbf{x} = \widehat{\mathbf{a}}'\mathbf{x}.$$

(d) Define the cutting score to be

$$\widehat{m} = \widehat{\mathbf{a}}'\left(\frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right).$$

Fill in the following Table

Group\Discriminant function	$\widehat{y} = x_1$	$\widehat{y} = x_2$	$\widehat{y} = \widehat{\mathbf{a}}'\mathbf{x}$
Group 1: HSI Constituent Stock			
Cheung Kong	318.70	9.041	?
Sun Hung Kai Properties	315.45	9.285	?
MTR Corporation	183.91	14.044	?
Hang Seng Bank	251.22	9.419	?
Hong Kong Exchanges and Clearing	200.78	43.519	?
Group 2: non-HSI Constituent Stock			
PCCW	36.74	18.976	?
Hang Lung Group	52.84	11.538	?
Wheelock and Company	75.89	4.478	?
Hopewell Holdings	23.96	17.628	?
The Link	104.457	6.065	?
Cutting score	$\overline{x}_1 =?$	$\overline{x}_2 =?$	?

Classification accuracy for $\widehat{y} = x_1$:

	Predicted	group
Actual group	1	2
1: Constituent Stock	?	?
2: Non-Constituent Stock	?	?

Classification accuracy for $\widehat{y} = x_2$:

	Predicted	group
Actual group	1	2
1: Constituent Stock	?	?
2: Non-Constituent Stock	?	?

Classification accuracy for $\widehat{y} = \widehat{\mathbf{a}}'\mathbf{x}$:

	Predicted	group
Actual group	1	2
1: Constituent Stock	?	?
2: Non-Constituent Stock	?	?

7.5 Scaling

The coefficient vectors $\widehat{\mathbf{a}} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\,\mathbf{S}_{pooled}^{-1}$ is unique only up to a multiplicative constant. Thus, for $c \neq 0$, any vector $c\widehat{\mathbf{a}}$ will also serve as discriminant coefficients. The vector $\widehat{\mathbf{a}}$ is frequently scaled or normalized to ease the interpretation of its elements. A commonly employed normalizations is

$$\widehat{\mathbf{a}}^* = \frac{\widehat{\mathbf{a}}}{\sqrt{\widehat{\mathbf{a}}'\widehat{\mathbf{a}}}},$$

so that $\widehat{\mathbf{a}}^*$ has unit length and its elements all lie in $[-1, 1]$. Another normalization is to scale the first element to 1, i.e.,

$$\widetilde{\mathbf{a}} = \frac{\widehat{\mathbf{a}}}{\widehat{a}_1}.$$

Normalization is recommended only if the X variables have been standardized.

Example 7.4: In Example 7.2, $\widehat{\mathbf{a}} = \begin{pmatrix} 37.61 \\ -28.92 \end{pmatrix}$, we have

$$\begin{aligned}\widehat{\mathbf{a}}^* &= \frac{\widehat{\mathbf{a}}}{\sqrt{\widehat{\mathbf{a}}'\widehat{\mathbf{a}}}} = \frac{1}{\sqrt{\begin{pmatrix} 37.61 & -28.92 \end{pmatrix}\begin{pmatrix} 37.61 \\ -28.92 \end{pmatrix}}}\begin{pmatrix} 37.61 \\ -28.92 \end{pmatrix} \\ &= \frac{1}{\sqrt{2251}}\begin{pmatrix} 37.61 \\ -28.92 \end{pmatrix} = \begin{pmatrix} 0.7927 \\ -0.6096 \end{pmatrix}.\end{aligned}$$

$$\widetilde{\mathbf{a}} = \frac{\widehat{\mathbf{a}}}{\widehat{a}_1} = \frac{1}{37.61}\begin{pmatrix} 37.61 \\ -28.92 \end{pmatrix} = \begin{pmatrix} 1 \\ -0.7689 \end{pmatrix}.$$

7.6 Classification with three populations

Let p_i be the prior probability of population π_i for $i = 1, 2, 3$ with $p_1+p_2+p_3 = 1$. We have

P(observation is misclassified as π_1)
$= P$(observation comes from π_2 and is misclassified as π_1)
$+P$(observation comes from π_3 and is misclassified as π_1)
$= P(1|2)\,p_2 + P(1|3)\,p_3$.

P(observation is misclassified as π_2)
$= P$(observation comes from π_1 and is misclassified as π_2)
$+P$(observation comes from π_3 and is misclassified as π_2)
$= P(2|1)\,p_1 + P(2|3)\,p_3$.

P(observation is misclassified as π_3)
$= P$(observation comes from π_1 and is misclassified as π_3)
$+P$(observation comes from π_2 and is misclassified as π_3)
$= P(3|1)\,p_1 + P(3|2)\,p_2$.

The costs of misclassification can be defined by a cost matrix

		Classify	*as*	
		π_1	π_2	π_3
True	π_1	0	$c\,(2\mid 1)$	$c\,(3\mid 1)$
population	π_2	$c(1\mid 2)$	0	$c\,(3\mid 2)$
	π_3	$c\,(1\mid 3)$	$c\,(2\mid 3)$	0

The expected cost of misclassification (ECM)

$$\begin{aligned} ECM &= c\,(1|2)\,P\,(1|2)\,p_2 + c\,(1|3)\,P\,(1|3)\,p_3 \\ &\quad + c\,(2|1)\,P\,(2|1)\,p_1 + c\,(2|3)\,P\,(2|3)\,p_3 \\ &\quad + c\,(3|1)\,P\,(3|1)\,p_1 + c\,(3|2)\,P\,(3|2)\,p_2. \end{aligned}$$

Recall that in the two-group case, we allocate $\mathbf{x}_0$ to π_1 if

$$c\,(1|2)\,f_2\,(\mathbf{x}_0)\,p_2 < c\,(2|1)\,f_1\,(\mathbf{x}_0)\,p_1.$$

In the three-group case, we compare

$$c\,(1|2)\,f_2\,(\mathbf{x}_0)\,p_2 + c\,(1|3)\,f_3\,(\mathbf{x}_0)\,p_3,$$

$$c\,(2|1)\,f_1\,(\mathbf{x}_0)\,p_1 + c\,(2|3)\,f_3\,(\mathbf{x}_0)\,p_3,$$

nd

$$c\,(3|1)\,f_1\,(\mathbf{x}_0)\,p_1 + c\,(3|2)\,f_2\,(\mathbf{x}_0)\,p_2.$$

We allocate $\mathbf{x}_0$ to π_1 if

$$c\,(1|2)\,f_2\,(\mathbf{x}_0)\,p_2 + c\,(1|3)\,f_3\,(\mathbf{x}_0)\,p_3$$

is the **smallest** among the three;

We allocate $\mathbf{x}_0$ to π_2 if

$$c\,(2|1)\,f_1\,(\mathbf{x}_0)\,p_1 + c\,(2|3)\,f_3\,(\mathbf{x}_0)\,p_3$$

is the **smallest** among the three;

We allocate $\mathbf{x}_0$ to π_3 if

$$c\,(3|1)\,f_1\,(\mathbf{x}_0)\,p_1 + c\,(3|2)\,f_2\,(\mathbf{x}_0)\,p_2$$

is the **smallest** among the three.

If all the misclassification costs are equal, it can be shown that we should allocate $\mathbf{x}_0$ to π_k if

$$f_k\,(\mathbf{x}_0)\,p_k$$

is the **biggest** among the three, $k = 1, 2, 3$.

Example 7.5: Consider the following case,

		Classify	as	
		π_1	π_2	π_3
True	π_1	$c(1\|1) = 0$	$c\,(2\|1) = 10$	$c\,(3\|1) = 50$
population	π_2	$c(1\|2) = 500$	$c(2\|2) = 0$	$c\,(3\|2) = 200$
	π_3	$c\,(1\|3) = 100$	$c\,(2\|3) = 50$	$c(3\|3) = 0$
Prior probability		$p_1 = 0.05$	$p_2 = 0.60$	$p_3 = 0.35$
Densities at $\mathbf{x}_0$		$f_1\,(\mathbf{x}_0) = 0.01$	$f_2\,(\mathbf{x}_0) = 0.85$	$f_3\,(\mathbf{x}_0) = 2$

(a) Should the point $\mathbf{x}_0$ be classified as π_1, π_2 or π_3 using the minimum ECM procedure?

(b) If all misclassification costs are the same, should the point $\mathbf{x}_0$ be classified as π_1, π_2 or π_3?

Solution:

(a)

$$\begin{aligned} & c(1|2)\, f_2(\mathbf{x}_0)\, p_2 + c(1|3)\, f_3(\mathbf{x}_0)\, p_3 \\ = \; & 500\,(0.85)\,(0.60) + 100\,(2)\,(0.35) \\ = \; & 325, \end{aligned}$$

$$\begin{aligned} & c(2|1)\, f_1(\mathbf{x}_0)\, p_1 + c(2|3)\, f_3(\mathbf{x}_0)\, p_3 \\ = \; & 10\,(0.01)\,(0.05) + 50\,(2)\,(0.35) \\ = \; & 35, \end{aligned}$$

$$\begin{aligned} & c(3|1)\, f_1(\mathbf{x}_0)\, p_1 + c(3|2)\, f_2(\mathbf{x}_0)\, p_2 \\ = \; & 50\,(0.01)\,(0.05) + 200\,(0.85)\,(0.60) \\ = \; & 102. \end{aligned}$$

Thus, we allocate $\mathbf{x}_0$ to π_2 since $c(2|1)\, f_1(\mathbf{x}_0)\, p_1 + c(2|3)\, f_3(\mathbf{x}_0)\, p_3$ is the **smallest** among the three;

(b) If all misclassification are the same, we have

$$f_1(\mathbf{x}_0)\, p_1 = (0.01)\,(0.05) = 0.000\,5,$$

$$f_2(\mathbf{x}_0)\, p_2 = (0.85)\,(0.60) = 0.51,$$

$$f_3(\mathbf{x}_0)\, p_3 = (2)\,(0.35) = 0.7.$$

We should allocate $\mathbf{x}_0$ to π_3 since $f_3(\mathbf{x}_0)\, p_3$ is the **biggest** among the three.

7.7 Classification with normal population

An important special case occurs when the density is multivariate normal with p-dimensions, with

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Omega}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\Omega}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)\right), \qquad i = 1, 2, 3.$$

To simplify the analysis, assume all the costs of misclassification are the same and equal 1, and the covariance matrices are equal. We compare

$$\begin{aligned} \ln(f_i(\mathbf{x})\, p_i) &= \ln p_i + \ln f_i(\mathbf{x}) \\ &= \ln p_i + \ln\left[\frac{1}{(2\pi)^{p/2} |\boldsymbol{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\Omega}^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)\right)\right] \\ &= \ln p_i - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Omega}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\Omega}^{-1} (\mathbf{x}-\boldsymbol{\mu}_i) \end{aligned}$$

In practice, the mean and variance matrices are unknown, we replace them by their sample estimates. Further, since the term $\frac{p}{2}\ln(2\pi)$ and $\frac{1}{2}\ln|\boldsymbol{\Omega}|$ are the same for all i, we can skip them and define

$$D_i^2(\mathbf{x}) = (\mathbf{x}-\overline{\mathbf{x}}_i)' \mathbf{S}_{pooled}^{-1} (\mathbf{x}-\overline{\mathbf{x}}_i).$$

We should allocate $\mathbf{x}_0$ to π_i if

$$\ln p_i - \frac{1}{2} D_i^2(\mathbf{x}_0)$$

is the **biggest** among the three. If all the prior probability p$_i$ are the same, then we allocate $\mathbf{x}_0$ to π_i if

$$\frac{1}{2} D_i^2(\mathbf{x}_0)$$

is the **smallest** among the three.

Example 7.6: Consider three groups of populations, and two bivariate normal X variables. Assume $p_1 = p_2 = 0.25$, and $p_3 = 0.5$. Suppose we draw a sample of three observations from each group and obtain

$$\mathbf{X}_1 = \begin{pmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{pmatrix} \quad \mathbf{X}_3 = \begin{pmatrix} 1 & -2 \\ 0 & 0 \\ -1 & 4 \end{pmatrix}$$

$$\overline{\mathbf{x}}_1 = \begin{pmatrix} -1 \\ 3 \end{pmatrix}, \quad \overline{\mathbf{x}}_2 = \begin{pmatrix} 1 \\ 4 \end{pmatrix} \quad \overline{\mathbf{x}}_3 = \begin{pmatrix} 0 \\ -2 \end{pmatrix},$$

$$\mathbf{S}_1 = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}, \quad \mathbf{S}_2 = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} \quad \mathbf{S}_3 = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}.$$

Which group does the point $\mathbf{x}_0 = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$ belong to?

Solution:

$$\begin{aligned}
\mathbf{S}_{pooled} &= \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}\mathbf{S}_1 \\
&\quad + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}\mathbf{S}_2 \\
&\quad + \frac{n_3 - 1}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}\mathbf{S}_3 \\
&= \frac{3 - 1}{(3 - 1) + (3 - 1) + (3 - 1)}\begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} \\
&\quad + \frac{3 - 1}{(3 - 1) + (3 - 1) + (3 - 1)}\begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} \\
&\quad + \frac{3 - 1}{(3 - 1) + (3 - 1) + (3 - 1)}\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \\
&= \frac{2}{6}\begin{pmatrix} 3 & -1 \\ -1 & 12 \end{pmatrix} \\
&= \begin{pmatrix} 1 & -\frac{1}{3} \\ -\frac{1}{3} & 4 \end{pmatrix}.
\end{aligned}$$

$$|\mathbf{S}| = \frac{35}{9}.$$

$$\mathbf{S}_{pooled}^{-1} = \begin{pmatrix} 1 & -\frac{1}{3} \\ -\frac{1}{3} & 4 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{36}{35} & \frac{3}{35} \\ \frac{3}{35} & \frac{9}{35} \end{pmatrix}.$$

$$D_i^2\left(\mathbf{x}\right) = \left(\mathbf{x} - \overline{\mathbf{x}}_i\right)' \mathbf{S}_{pooled}^{-1} \left(\mathbf{x} - \overline{\mathbf{x}}_i\right).$$

$$\begin{aligned}
\ln p_1 - \frac{1}{2} D_1^2\left(\mathbf{x}_0\right) &= \ln p_1 - \frac{1}{2}\left(\mathbf{x}_0 - \overline{\mathbf{x}}_1\right)' \mathbf{S}_{pooled}^{-1} \left(\mathbf{x}_0 - \overline{\mathbf{x}}_1\right) \\
&= \ln 0.25 - \frac{1}{2}\left(\begin{pmatrix} -2 \\ -1 \end{pmatrix} - \begin{pmatrix} -1 \\ 3 \end{pmatrix}\right)' \begin{pmatrix} \frac{36}{35} & \frac{3}{35} \\ \frac{3}{35} & \frac{9}{35} \end{pmatrix} \left(\begin{pmatrix} -2 \\ -1 \end{pmatrix} - \begin{pmatrix} -1 \\ 3 \end{pmatrix}\right) \\
&= \ln 0.25 - \frac{1}{2}\begin{pmatrix} -1 & -4 \end{pmatrix} \begin{pmatrix} \frac{36}{35} & \frac{3}{35} \\ \frac{3}{35} & \frac{9}{35} \end{pmatrix} \begin{pmatrix} -1 \\ -4 \end{pmatrix} \\
&= -4.30.
\end{aligned}$$

$$\begin{aligned}
\ln p_2 - \frac{1}{2} D_2^2\left(\mathbf{x}_0\right) &= \ln p_2 - \frac{1}{2}\left(\mathbf{x}_0 - \overline{\mathbf{x}}_2\right)' \mathbf{S}_{pooled}^{-1} \left(\mathbf{x}_0 - \overline{\mathbf{x}}_2\right) \\
&= \ln 0.25 - \frac{1}{2}\left(\begin{pmatrix} -2 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 \\ 4 \end{pmatrix}\right)' \begin{pmatrix} \frac{36}{35} & \frac{3}{35} \\ \frac{3}{35} & \frac{9}{35} \end{pmatrix} \left(\begin{pmatrix} -2 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 \\ 4 \end{pmatrix}\right) \\
&= \ln 0.25 - \frac{1}{2}\begin{pmatrix} -3 & -5 \end{pmatrix} \begin{pmatrix} \frac{36}{35} & \frac{3}{35} \\ \frac{3}{35} & \frac{9}{35} \end{pmatrix} \begin{pmatrix} -3 \\ -5 \end{pmatrix} \\
&= -10.51.
\end{aligned}$$

$$\begin{aligned}
\ln p_3 - \frac{1}{2} D_3^2\left(\mathbf{x}_0\right) &= \ln p_3 - \frac{1}{2}\left(\mathbf{x}_0 - \overline{\mathbf{x}}_3\right)' \mathbf{S}_{pooled}^{-1} \left(\mathbf{x}_0 - \overline{\mathbf{x}}_3\right) \\
&= \ln 0.5 - \frac{1}{2}\left(\begin{pmatrix} -2 \\ -1 \end{pmatrix} - \begin{pmatrix} 0 \\ -2 \end{pmatrix}\right)' \begin{pmatrix} \frac{36}{35} & \frac{3}{35} \\ \frac{3}{35} & \frac{9}{35} \end{pmatrix} \left(\begin{pmatrix} -2 \\ -1 \end{pmatrix} - \begin{pmatrix} 0 \\ -2 \end{pmatrix}\right) \\
&= \ln 0.5 - \frac{1}{2}\begin{pmatrix} -2 & 1 \end{pmatrix} \begin{pmatrix} \frac{36}{35} & \frac{3}{35} \\ \frac{3}{35} & \frac{9}{35} \end{pmatrix} \begin{pmatrix} -2 \\ 1 \end{pmatrix} \\
&= -2.707.
\end{aligned}$$

Thus, we allocate $\mathbf{x}_0$ to π_3 since $\ln p_3 - \frac{1}{2}D_3^2(\mathbf{x}_0)$ is the **biggest** among the three.

Exercise 7.3: Consider three groups of students applying for the MBA program of CUHK. Let $x_1 = GPA$ score, $x_2 = GMAT$ score of the applicants. Group 1 students are admitted to the program, group 2 students are not admitted, and group 3 is marginal. Assume the proportion of each population is the same, i.e., $p_1 = p_2 = p_3 = \frac{1}{3}$. Suppose we have a sample of 31 admitted students, 28 not admitted, and 26 students are marginal, i.e., $n_1 = 31$, $n_2 = 28$, $n_3 = 26$. The mean score of each group are

$$\overline{\mathbf{x}}_1 = \begin{pmatrix} 3.40 \\ 561.23 \end{pmatrix}, \quad \overline{\mathbf{x}}_2 = \begin{pmatrix} 2.48 \\ 447.07 \end{pmatrix} \quad \overline{\mathbf{x}}_3 = \begin{pmatrix} 2.99 \\ 446.23 \end{pmatrix},$$

$$\mathbf{S}_{pooled} = \begin{pmatrix} 0.0361 & -2.0188 \\ -2.0188 & 3655.9011 \end{pmatrix}.$$

Suppose you would like to apply for the MBA program of CUHK. Your GPA and GMAT score are $\mathbf{x}_0 = \begin{pmatrix} 3.21 \\ 497 \end{pmatrix}$. Will you be admitted?

Exercise 7.4: Consider the case of one X variable. Suppose the first group of X is normally distributed with $N(0,1)$, and the second group of X is normally distributed with $N(1,1)$. Consider a point $x_0 = 0$, which group does this point belong to if $\frac{c(1|2)\, p_2}{c(2|1)\, p_1} = 1$?

Exercise 7.5: Consider the following data sets

$$\mathbf{X}_1 = \begin{pmatrix} 5 & 2 \\ 7 & 3 \\ 6 & 1 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 0 & 4 \\ 1 & 5 \\ 2 & 6 \end{pmatrix}, \quad \overline{\mathbf{x}}_1 = \begin{pmatrix} 6 \\ 2 \end{pmatrix}, \quad \overline{\mathbf{x}}_2 = \begin{pmatrix} 1 \\ 5 \end{pmatrix}$$

(a) Find $\mathbf{S}_{pooled}$ and $\mathbf{S}_{pooled}^{-1}$.

(b) Calculate the linear discriminant function $\widehat{y} = \widehat{\mathbf{a}}'\mathbf{x}$.

(c) Should the point $\mathbf{x}_0 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$ be classified as population 1 or 2 if $\dfrac{c(1|2)}{c\,(2|1)}\dfrac{p_2}{p_1} = 1$?

Exercise 7.6: True/False. Let p_k be the prior probability of population k, $k = 1, 2, 3$.

(a) If all the misclassification costs are equal, then we should allocate $\mathbf{x}_0$ to population k if p_k is the smallest of the three.

(b) P(observation is misclassified as population 1)$= 1 - P$(observation is classified as population 1).

(c) If all the misclassification costs are equal, the we should allocate $\mathbf{x}_0$ to population k if $f_k\,(\mathbf{x}_0)\,p_k$ is the smallest among the three.

Exercise 7.7: Consider the following data sets

$$\mathbf{X}_1 = \begin{pmatrix} 0 & 4 \\ 1 & 5 \\ 2 & 6 \end{pmatrix}, \qquad \mathbf{X}_2 = \begin{pmatrix} 10 & 8 \\ 11 & 5 \\ 12 & 8 \end{pmatrix}$$

$$\overline{\mathbf{x}}_1 = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \qquad \overline{\mathbf{x}}_2 = \begin{pmatrix} 11 \\ 7 \end{pmatrix}.$$

(a) Find $\mathbf{S}_{pooled}$ and $\mathbf{S}_{pooled}^{-1}$.

(b) Calculate the linear discriminant function $\widehat{y} = \widehat{\mathbf{a}}'\mathbf{x}$.

(c) Should the point $\mathbf{x}_0 = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$ be classified as population 1 or 2 if $\dfrac{c(1|2)}{c\,(2|1)}\dfrac{p_2}{p_1} = 1$?

Exercise 7.8: Consider the following data sets

$$\mathbf{X}_1 = \begin{pmatrix} 3 & 6 \\ 2 & 4 \\ 4 & 5 \end{pmatrix}, \qquad \mathbf{X}_2 = \begin{pmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{pmatrix}$$

$$\overline{\mathbf{x}}_1 = \begin{pmatrix} 3 \\ 5 \end{pmatrix}, \qquad \overline{\mathbf{x}}_2 = \begin{pmatrix} 5 \\ 8 \end{pmatrix},$$

(a) Calculate $\mathbf{S}_{pooled}^{-1}$.

(b) Calculate the linear discriminant function $\widehat{y} = \widehat{\mathbf{a}}'\mathbf{x}$.

(c) Should the point $\mathbf{x}_0 = \begin{pmatrix} 7 \\ 2 \end{pmatrix}$ be classified as population 1 or 2 if $\frac{c(1|2)}{c\,(2|1)} \frac{p_2}{p_1} = 1$?

Exercise 7.9: Suppose there are two groups of individuals. Each individual can be characterized by a single value x, which follows an extreme value distribution, with

$$f\,(x) = \exp\,(-x) \exp\,(-\exp\,(-x)) \qquad \text{for} -\infty < x < \infty$$

Suppose $f\,(x)$ is the same for both groups. For $i = 1, 2$ and $j = 1, 2$, let p_i be the prior probability of group i, and $c(i|j)$ be the cost if an individual from group j is misclassified into group i. Suppose we would like to minimize the expected cost of missclassification. Consider a point $x_0 = 3$, which group does this point belong to if

(a) $\frac{c(1|2)}{c\,(2|1)} > \frac{p_1}{p_2}$?

(b) $\frac{c(1|2)}{c\,(2|1)} = \frac{p_1}{p_2}$?

Chapter 8

Cluster Analysis

Cluster analysis involves techniques that produce classifications from data that are initially unclassified, and must not be confused with discriminant analysis, in which one initially knows how many distinct groups exist and also has data that are known to come from each of these distinct groups. To perform a cluster analysis, one must first be able to measure the similarity or dissimilarity between two clusters of observations.

8.1 Similarity Measures

Let x_{ij} be the score (1 or 0) of the j^{th} binary variable on the i^{th} item and x_{kj} be the score (1 or 0) of the j^{th} binary variable on the k^{th} item, $j = 1, 2, ..., p$.

$$\begin{aligned} (x_{ij} - x_{kj})^2 &= 0 \quad \text{if } x_{ij} = x_{kj} = 1 \text{ or } x_{ij} = x_{kj} = 0 \\ &= 1 \quad \text{if } x_{ij} \neq x_{kj}. \end{aligned}$$

The square Euclidean distance

$$\sum_{j=1}^{p} (x_{ij} - x_{kj})^2$$

provides a count of the number of mismatches. A large distance corresponds to many mismatches. Let us arrange the frequencies of matches and mismatches for items i and k in the form of a contingency table:

		Item	k	
		1	0	Totals
Item i	1	a	b	$a+b$
	0	c	d	$c+d$
	Totals	$a+c$	$b+d$	$p=a+b+c+d$

where a represents the frequency of 1-1 matches and so on.

However, the measure suffers from weighting the 1-1 and 0-0 matches equally. In some cases, a 1-1 match is a stronger indication of similarity than a 0-0 match. For instance, in grouping people, the evidence that two persons both are the president of the United States is stronger evidence of similarity than the absence of this position. Thus, it might be reasonable to discount the 0-0 matches. We define some similarity coefficients for clustering items as follows:

	Coefficient	Rationale
1	$\dfrac{a+d}{a+b+c+d}$	Equal weights for 1-1 matches and 0-0 matches.
2	$\dfrac{2(a+d)}{2(a+d)+b+c}$	Double weights for 1-1 matches and 0-0 matches.
3	$\dfrac{a+d}{a+d+2(b+c)}$	Double weights for unmatched pairs.
4	$\dfrac{a}{a+b+c+d}$	No 0-0 matches in numerator.
5	$\dfrac{a}{a+b+c}$	No 0-0 matches in numerator or denominator. (The 0-0 matches are treated as irrelevant.)
6	$\dfrac{2a}{2a+b+c}$	No 0-0 matches in numerator or denominator, d=0. (Double weights for 1-1 matches)
7	$\dfrac{a+d}{a+2(b+c)}$	No 0-0 matches in numerator or denominator, d=0. (Double weights for unmatched pairs.)
8	$\dfrac{a}{b+c}$	Ratio of matches to mismatches with 0-0 matches excluded.

Example 8.1: Suppose five individuals possess the following characteristics:

	Height (inch)	Weight (lb)	Eye Color	Hair Color	Handedness	Gender
Individual 1	68	140	Green	Blond	Right	Female
Individual 2	72	185	Brown	Brown	Right	Male
Individual 3	67	165	Blue	Blond	Right	Male
Individual 4	64	120	Brown	Brown	Right	Female
Individual 5	76	210	Brown	Brown	Left	Male

Define six binary variables $X_1, X_2, X_3, X_4, X_5, X_6$ as

$$\begin{aligned} X_1 &= 1 \quad \text{if height} \geqslant 72 \text{ in.} \\ &= 0 \quad \text{if height} < 72 \text{ in.} \end{aligned}$$

$$\begin{aligned} X_2 &= 1 \quad \text{if weight} \geqslant 150 \text{ lb.} \\ &= 0 \quad \text{if weight} < 150 \text{ lb.} \end{aligned}$$

$$\begin{aligned} X_3 &= 1 \quad \text{if brown eyes.} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

$$\begin{aligned} X_4 &= 1 \quad \text{if blond hair.} \\ &= 0 \quad \text{if not blond hair.} \end{aligned}$$

$$\begin{aligned} X_5 &= 1 \quad \text{if right handed.} \\ &= 0 \quad \text{if left handed.} \end{aligned}$$

$$\begin{aligned} X_6 &= 1 \quad \text{if female.} \\ &= 0 \quad \text{if male.} \end{aligned}$$

The scores for individuals 1 and 2 on these 6 variables are

	X_1	X_2	X_3	X_4	X_5	X_6
Individual 1	0	0	0	1	1	1
Individual 2	1	1	1	0	1	0

and the number of matches and mismatches are indicated in the two-way array

		Individual	2	
		1	**0**	**Totals**
Individual 1	**1**	1	2	3
	0	3	0	3
	Totals	4	2	6

Employing the first similarity coefficient, which gives equal weight to matches, we have

$$\frac{a+d}{a+b+c+d} = \frac{1+0}{1+2+3+0} = \frac{1}{6},$$

we have

				Individual		
		1	**2**	**3**	**4**	**5**
	1	1				
	2	$\frac{1}{6}$	1			
Individual	**3**	$\frac{4}{6}$	$\frac{3}{6}$	1		
	4	$\frac{4}{6}$	$\frac{3}{6}$	$\frac{2}{6}$	1	
	5	0	$\frac{5}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	1

Based on the magnitude of the similarity coefficient, we should conclude that individuals 2 and 5 are most similar and individuals 1 and 5 are least similar. Other pairs fall between these extremes. If we were to divide individuals into two relatively homogeneous subgroups, we might form the subgroups (1, 3, 4) and (2, 5).

Example 8.2: The following table gives the first 10 numbers in eleven languages. Use the first letters of the numbers to split the languages into different groups.

Eng.	**Nor**	**Dan**	**Dutch**	**Ger.**	**Fren.**	**Span.**	**Italian**	**Polish**	**Hung.**	**Finnish**
one	*en*	*en*	*een*	*eins*	*un*	*uno*	*uno*	*jeden*	*egy*	*yksi*
two	*to*	*to*	*twee*	*zwei*	*deux*	*dos*	*due*	*dwa*	*ketto*	*kaksi*
three	*tre*	*tre*	*drie*	*drei*	*trois*	*tres*	*tre*	*trzy*	*harom*	*kolme*
four	*fire*	*fire*	*vier*	*vier*	*quatre*	*cuatro*	*quattro*	*cztery*	*negy*	*neua*
five	*fem*	*fem*	*vijf*	*funf*	*cinq*	*cinco*	*cinque*	*piec*	*ot*	*viisi*
six	*seks*	*seks*	*zes*	*sechs*	*six*	*seis*	*sei*	*szesc*	*hat*	*kuusi*
seven	*sju*	*syv*	*zeven*	*sieben*	*sept*	*siete*	*sette*	*siedem*	*het*	*seitseman*
eight	*atte*	*otte*	*acht*	*acht*	*huit*	*ocho*	*otto*	*osiem*	*nyolc*	*kahdeksan*
nine	*ni*	*ni*	*negen*	*neun*	*neuf*	*nueve*	*nove*	*dziewiec*	*kilenc*	*yhdeksan*
ten	*ti*	*ti*	*tien*	*zehn*	*dix*	*diez*	*dieci*	*dziesiec*	*tiz*	*kymmenen*

From the following table, we see that English and Norwegian have the same first letter for 8 of the 10 word pairs. The remaining frequencies are calculated in the same manner.

	Eng.	**Nor**	**Dan**	**Dutch**	**Ger.**	**Fren.**	**Span.**	**Ital.**	**Polish**	**Hung.**	**Fin.**
English	10										
Norwegian	8	10									
Danish	8	9	10								
Dutch	3	5	4	10							
German	4	6	5	5	10						
French	4	4	4	1	3	10					
Spanish	4	4	5	1	3	8	10				
Italian	4	4	5	1	3	9	9	10			
Polish	3	3	4	0	2	5	7	6	10		
Hungarian	1	2	2	2	1	0	0	0	0	10	
Finnish	1	1	1	1	1	1	1	1	1	2	10

From the table, English, Norwegian, Danish, Dutch and German seem to form a group. French, Spanish, Italian and Polish might be grouped together, whereas Hungarian and Finnish appear to stand alone.

Exercise 8.1: Consider the following ten Hong Kong stocks as of 31/10/2014.:

Company name	**Total market capitalization (billions)**	**PE Ratio**	**HSI Constituent Stock**	**Sector**
Cheung Kong	318.70	9.041	*Yes*	Property Development
Sun Hung Kai Properties	315.45	9.285	*Yes*	Property Development
MTR Corporation	183.91	14.044	*Yes*	Traffic
Hang Seng Bank	251.22	9.419	*Yes*	Bank
HKExchanges and Clearing	200.78	43.519	*Yes*	Exchanges
PCCW	36.74	18.976	*No*	Telecommunications
Hang Lung Group	52.84	11.538	*No*	Property Development
Wheelock	75.89	4.478	*No*	Property Development
Hopewell Holdings	23.96	17.628	*No*	Consolidated Enterprises
The Link	104.457	6.065	*No*	REIT

Define four binary variables X_1, X_2, X_3, X_4 as

$$\begin{aligned} X_1 &= 1 \quad \text{if total market capitalization >200 billions} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

$$\begin{aligned} X_2 &= 1 \quad \text{if PE >10} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

$$\begin{aligned} X_3 &= 1 \quad \text{if HSI Constituent stock} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

$$\begin{aligned} X_4 &= 1 \quad \text{if from Property Development Sector} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

(a) Calculate the coefficient $\dfrac{a+d}{a+b+c+d}$ for pairs of stocks.

(b) How would you classify the stocks into two clusters? How would you classify the stocks into three clusters?

Exercise 8.2 Consider the following table for the US presidents.

President	**Birthplace**	**Elected First Term**	**Party**	**Congressman**	**Vice President**
R. Reagan	*Midwest*	*Yes*	*Republican*	*No*	*No*
J.Carter	*South*	*Yes*	*Democrat*	*No*	*No*
G. Ford	*Midwest*	*No*	*Republican*	*Yes*	*Yes*
R. Nixon	*West*	*Yes*	*Republican*	*Yes*	*Yes*
L. Johnson	*South*	*No*	*Democrat*	*Yes*	*Yes*
J. Kennedy	*East*	*Yes*	*Democrat*	*Yes*	*No*

Define five binary variables X_1, X_2, X_3, X_4, X_5 as

$$\begin{aligned} X_1 &= 1 \quad \text{if birthplace is South.} \\ &= 0 \quad \text{if birthplace is non-South.} \end{aligned}$$

$$\begin{aligned} X_2 &= 1 \quad \text{if elected first term.} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

$$\begin{aligned} X_3 &= 1 \quad \text{if Republican.} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

$$\begin{aligned} X_4 &= 1 \quad \text{if Congressman.} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

$$\begin{aligned} X_5 &= 1 \quad \text{if served as vice president.} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

(a) Calculate the coefficient $\dfrac{a+d}{a+b+c+d}$ for pairs of presidents.
(b) How would you put the presidents into clusters?

8.2 Agglomerative hierarchical clustering method

When the first cluster is formed, we need to measure the distance between this cluster and other clusters/objects. Two commonly used methods are the single linkage method and the complete linkage method.

8.2.1 Single linkage (nearest-neighbor) method

Consider the hypothetical distances between pairs of five objects as follows:

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{cccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{1} & 0 & & & & \\ \mathbf{2} & 9 & 0 & & & \\ \mathbf{3} & 3 & 7 & 0 & & \\ \mathbf{4} & 6 & 5 & 9 & 0 & \\ \mathbf{5} & 11 & 10 & \mathbf{2} & 8 & 0 \end{array}$$

First, we merge the two closet items. Since $d_{35} = 2$ is the smallest, objects 3 and 5 are merged to form the cluster (35). Next, we calculate the distance between this new cluster (35) and the remaining objects. The nearest neighbor distances are

$$d_{(35)1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3.$$

$$d_{(35)2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7.$$

$$d_{(35)4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8.$$

The new distance matrix becomes

$$\begin{array}{ccccc} & \mathbf{(35)} & \mathbf{1} & \mathbf{2} & \mathbf{4} \\ \mathbf{(35)} & 0 & & & \\ \mathbf{1} & \mathbf{3} & 0 & & \\ \mathbf{2} & 7 & 9 & 0 & \\ \mathbf{4} & 8 & 6 & 5 & 0 \end{array}$$

Since $d_{(35)1}$ is the smallest, object 1 and cluster (35) and are merged to form the cluster (135). The nearest neighbor distances between the new cluster (135) and the remaining objects are

$$d_{(135)2} = \min\left\{d_{(35)2}, d_{12}\right\} = \min\{7, 9\} = 7.$$

$$d_{(135)4} = \min\left\{d_{(35)4}, d_{14}\right\} = \min\{8, 6\} = 6.$$

The new distance matrix becomes

$$\begin{array}{cccc} & \mathbf{(135)} & \mathbf{2} & \mathbf{4} \\ \mathbf{(135)} & 0 & & \\ \mathbf{2} & 7 & 0 & \\ \mathbf{4} & 6 & \mathbf{5} & 0 \end{array}$$

Since $d_{(42)} = 5$ is the smallest, objects 2 and 4 are merged to form the cluster (24) . At this point we have 2 clusters, their nearest neighbor distance is

$$d_{(135)(24)} = \min \left\{ d_{(135)2}, d_{(135)4} \right\} = \min \left\{ 7, 6 \right\} = 6.$$

The final distance matrix becomes

$$\begin{array}{ccc} & \mathbf{(135)} & \mathbf{(24)} \\ \mathbf{(135)} & 0 & \\ \mathbf{(24)} & 6 & 0 \end{array}$$

How to cluster the objects depends on how many cluster we would like to have. If we would like to have two cluster, then the two clusters are (135) and (24). If we need three cluster, then we have (135), 2 and 4.

Example 8.3: Consider the clustering of 11 languages in the previous example, the matrix of distances is as follows:

	1	2	3	4	5	6	7	8	9	10	11
	Eng.	**Nor**	**Dan**	**Dutch**	**Ger.**	**Fren.**	**Span.**	**Ital.**	**Polish**	**Hung.**	**Fin.**
English	0										
Nor.	2	0									
Danish	2	**1**	0								
Dutch	7	5	6	0							
German	6	4	5	5	0						
French	6	6	6	9	7	0					
Spanish	6	6	5	9	7	2	0				
Italian	6	6	5	9	7	**1**	**1**	0			
Polish	7	7	6	10	8	5	3	4	0		
Hung.	9	8	8	8	9	10	10	10	10	0	
Finnish	9	9	9	9	9	9	9	9	9	8	0

We first search for the minimum distance between pairs of languages (clusters). The minimum distance is 1, which occurs between Danish and Norwegian, Italian and French, and Italian and Spanish. Numbering the languages in the order in which they appear across the top of the array, we have

$$d_{23} = 1.$$

$$d_{68} = 1.$$

$$d_{78} = 1.$$

Note that 6, 7, 8 cannot be merged at this stage since $d_{67} = 2 > 1$. We first merge 6 and 8. Next, we calculate the distance between the two clusters (23), (68), and the remaining objects. The nearest neighbor distances are

$$d_{(23)1} = \min\{d_{21}, d_{31}\} = \min\{2, 2\} = 2.$$

$$d_{(23)4} = \min\{d_{24}, d_{34}\} = \min\{5, 6\} = 5.$$

$$d_{(23)5} = \min\{d_{25}, d_{35}\} = \min\{4, 5\} = 4.$$

$$d_{(23)7} = \min\{d_{27}, d_{37}\} = \min\{6, 5\} = 5.$$

$$d_{(23)9} = \min\{d_{29}, d_{39}\} = \min\{7, 6\} = 6.$$

$$d_{(23)10} = \min\{d_{2,10}, d_{3,10}\} = \min\{8, 8\} = 8.$$

$$d_{(23)11} = \min\{d_{2,11}, d_{3,11}\} = \min\{9, 9\} = 9.$$

$$d_{(68)1} = \min\{d_{61}, d_{81}\} = \min\{6, 6\} = 6.$$

$$d_{(68)4} = \min\{d_{64}, d_{84}\} = \min\{9, 9\} = 9.$$

$$d_{(68)5} = \min\{d_{65}, d_{85}\} = \min\{7, 7\} = 7.$$

$$d_{(68)7} = \min\{d_{67}, d_{87}\} = \min\{2, 1\} = 1.$$

$$d_{(68)9} = \min\{d_{69}, d_{89}\} = \min\{5, 4\} = 4.$$

$$d_{(68)10} = \min\{d_{6,10}, d_{8,10}\} = \min\{10, 10\} = 10.$$

$$d_{(68)11} = \min\{d_{6,11}, d_{8,11}\} = \min\{9, 9\} = 9.$$

$$d_{(68)(23)} = \min\{d_{62}, d_{63}, d_{82}, d_{83}\} = \min\{6, 6, 6, 5\} = 5.$$

Now, the new distance matrix becomes

	$(2,3)$	$(6,8)$	1	4	5	7	9	10	11
	Nor, **Dan**	**French**, **Ital**.	**Eng**.	**Dutch**	**Ger**.	**Span**.	**Polish**	**Hung**.	**Fin**.
Norwegian, **Danish**	0								
French, **Italian**	5	0							
English	2	6	0						
Dutch	5	9	7	0					
German	4	7	6	5	0				
Spanish	5	**1**	6	9	7	0			
Polish	6	4	7	10	8	3	0		
Hungarian	8	10	9	8	9	10	10	0	
Finnish	9	9	9	9	9	9	9	8	0

The nearest neighbor distances between (678) and the remaining objects are

$$d_{(678)1} = \min\left\{d_{(68)1}, d_{71}\right\} = \min\left\{6, 6\right\} = 6.$$

$$d_{(678)4} = \min\left\{d_{(68)4}, d_{74}\right\} = \min\left\{9, 9\right\} = 9.$$

$$d_{(678)5} = \min\left\{d_{(68)5}, d_{75}\right\} = \min\left\{7, 7\right\} = 7.$$

$$d_{(678)9} = \min\left\{d_{(68)9}, d_{79}\right\} = \min\left\{4, 3\right\} = 3.$$

$$d_{(678)10} = \min\left\{d_{(68),10}, d_{7,10}\right\} = \min\left\{10, 10\right\} = 10.$$

$$d_{(678)11} = \min\left\{d_{(68)11}, d_{7,11}\right\} = \min\left\{9, 9\right\} = 9.$$

$$d_{(678)(23)} = \min\left\{d_{(68)(23)}, d_{(23),7}\right\} = \min\{5, 5\} = 5.$$

	$(2,3)$	$(6,7,8)$	1	4	5	9	10	11
	Nor, **Dan**	**French**, **Span**., **Ital**.	**Eng**.	**Dutch**	**Ger**.	**Polish**	**Hung**.	**Fin**.
Norwegian, **Danish**	0							
French, **Spanish**, **Italian**	5	0						
English	**2**	6	0					
Dutch	5	9	7	0				
German	4	7	6	5	0			
Polish	6	3	7	10	8	0		
Hungarian	8	10	9	8	9	10	0	
Finnish	9	9	9	9	9	9	8	0

Since $d_{(23)1}$ is the smallest, object 1 and cluster (23) and are merged to form the cluster (123). The nearest neighbor distances between (123) and the remaining objects are

$$d_{(123)4} = \min\left\{d_{14}, d_{(23)4}\right\} = \min\{7, 5\} = 5.$$

$$d_{(123)5} = \min\left\{d_{15}, d_{(23)5}\right\} = \min\{6, 4\} = 4.$$

$$d_{(123)9} = \min\left\{d_{19}, d_{(23)9}\right\} = \min\{7, 6\} = 6.$$

$$d_{(123)10} = \min\left\{d_{1,10}, d_{(23)10}\right\} = \min\{9, 8\} = 8.$$

$$d_{(123)11} = \min\left\{d_{1,11}, d_{(23)11}\right\} = \min\{9, 9\} = 9.$$

$$d_{(123)(678)} = \min\left\{d_{(678)1}, d_{(678)(23)}\right\} = \min\{6, 5\} = 5.$$

	$(1,2,3)$	$(6,7,8)$	4	5	9	10	11
	Eng., **Nor**, **Dan**	**French**, **Span.**, **Ital.**	**Dutch**	**Ger.**	**Polish**	**Hung.**	**Fin.**
English, **Norwegian**, **Danish**	0						
French, **Spanish**, **Italian**	5	0					
Dutch	5	9	0				
German	4	7	5	0			
Polish	6	**3**	10	8	0		
Hungarian	8	10	8	9	10	0	
Finnish	9	9	9	9	9	8	0

Since $d_{(678)9} = 3$ is the smallest, object 9 and cluster (678) and are merged to form the cluster (6789). The nearest neighbor distances between (6789) and the remaining objects are

$$d_{(6789)4} = \min\left\{d_{(678)4}, d_{94}\right\} = \min\{9, 10\} = 9.$$

$$d_{(6789)5} = \min\left\{d_{(678)5}, d_{95}\right\} = \min\{7, 8\} = 7.$$

$$d_{(6789)10} = \min\left\{d_{(678)10}, d_{9,10}\right\} = \min\{10, 10\} = 10.$$

$$d_{(6789)11} = \min\left\{d_{(678)11}, d_{9,11}\right\} = \min\{9, 9\} = 9.$$

$$d_{(123)(6789)} = \min\left\{d_{(123)(678)}, d_{(123),9}\right\} = \min\{5, 6\} = 5.$$

	$(1,2,3)$	$(6,7,8,9)$	4	5	10	11
	Eng., Nor, Dan	**French, Span., Ital., Polish**	**Dutch**	**Ger.**	**Hung.**	**Fin.**
English, Norwegian, Danish	0					
French, Spanish, Italian, Polish	5	0				
Dutch	5	9	0			
German	**4**	7	5	0		
Hungarian	8	10	8	9	0	
Finnish	9	9	9	9	8	0

Since $d_{(123)5}$ is the smallest, object 5 and cluster (123) and are merged to form the cluster (1235). The nearest neighbor distances between (1235) and the remaining objects are

$$d_{(1235)4} = \min\left\{d_{(123)4}, d_{54}\right\} = \min\{5,5\} = 5.$$

$$d_{(1235)10} = \min\left\{d_{(123)10}, d_{5,10}\right\} = \min\{8,9\} = 8.$$

$$d_{(1235)(6789)} = \min\left\{d_{(123)(6789)}, d_{5(6789)}\right\} = \min\{5,7\} = 5.$$

	$(1,2,3,5)$	$(6,7,8,9)$	4	10	11
	Eng., **Nor**, **Dan**, **Ger.**	**French**, **Span.**, **Italian**, **Polish**	**Dutch**	**Hung.**	**Fin.**
English, **Norwegian**, **Danish**, **German**	0				
French, **Spanish**, **Italian**, **Polish**	**5**	0			
Dutch	**5**	9	0		
Hungarian	8	10	8	0	
Finnish	9	9	9	8	0

Note that $d_{(1235)(6789)} = d_{(1235)4} = 5$, we can group them to form the cluster (123456789). The nearest neighbor distances between (123456789) and the remaining objects are

$$d_{(123456789)10} = \min\left\{d_{(1235)10}, d_{(6789)10}, d_{4,10}\right\} = \min\{8, 10, 8\} = 8.$$

$$d_{(123456789)11} = \min\left\{d_{(1235)11}, d_{(6789)11}, d_{4,11}\right\} = \min\{9, 9, 9\} = 9.$$

	$(1,2,3,4,5,6,7,8,9)$	10	11
	Eng., **Nor**, **Dan**, **Dutch**, **Ger.French**, **Span.**, **Italian**, **Polish**	**Hung.**	**Fin.**
English, **Norwegian**, **Danish**, **Dutch**, **German**, **French**, **Spanish**, **Italian**, **Polish**	0		
Hungarian	8	0	
Finnish	9	8	0

Note that $d_{(123456789)10} = d_{10,11} = 8$, are the smallest, but $d_{(123456789)11} = 9 > 8$, we cannot group (123456789) and 10, but we can group 10 and 11 to form the cluster (10,11). The minimum distances between (123456789) and $(10, 11)$ is

$$d_{(123456789)(10,11)} = \min\left\{d_{(123456789)10}, d_{(123456789)11}\right\} = \min\{8, 9\} = 8.$$

	$(1,2,3,4,5,6,7,8,9)$ **Eng.**, **Nor**, **Dan**, **Dutch**, **Ger.**, **French**, **Span.**, **Italian**, **Polish**	$(10,11)$ **Hung.**, **Fin.**
English, **Norwegian**, **Danish**, **Dutch**, **German**, **French**, **Spanish**, **Italian**, **Polish**	0	
Hungarian, **Finnish**	8	0

8.2.2 Complete linkage (Farthest-neighbor) method

The single linkage has a shortcoming that, as long as a new object is close to one of the objects in the cluster, it will be assigned to this cluster even if it is very different from other objects in the cluster. For example, consider a cluster that contains 1000 African people and one Chinese, then a Chinese not in this cluster will be assigned to it since there is a single linkage (Chinese-Chinese). Because of this shortcoming, we need another clustering method. One method that can avoid the aforementioned shortcoming is called the complete linkage method. Complete linkage clustering is different from single linkage clustering in that at each stage, the distance between clusters is the maximum distance between two elements from each cluster. In the above example, a Chinese who is not in this cluster will not be assigned to the cluster.

Example 8.4: Consider again the hypothetical distances between pairs of five objects as follows:

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{cccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{1} & 0 & & & & \\ \mathbf{2} & 9 & 0 & & & \\ \mathbf{3} & 3 & 7 & 0 & & \\ \mathbf{4} & 6 & 5 & 9 & 0 & \\ \mathbf{5} & 11 & 10 & \mathbf{2} & 8 & 0 \end{array}$$

At the first stage, we merge the two closet items. Since $d_{35} = 2$ is the smallest, objects 3 and 5 are merged to form the cluster (35).

At stage 2, we calculate the maximum distance between this new cluster (35) and the remaining objects. The maximum distances are

$$d_{(35)1} = \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11.$$

$$d_{(35)2} = \max\{d_{32}, d_{52}\} = \max\{7, 10\} = 10.$$

$$d_{(35)4} = \max\{d_{34}, d_{54}\} = \max\{9, 8\} = 9.$$

Now, the new distance matrix becomes

$$\begin{array}{ccccc} & \mathbf{(35)} & \mathbf{1} & \mathbf{2} & \mathbf{4} \\ \mathbf{(35)} & 0 & & & \\ \mathbf{1} & 11 & 0 & & \\ \mathbf{2} & 10 & 9 & 0 & \\ \mathbf{4} & 9 & 6 & \mathbf{5} & 0 \end{array}$$

The next merger occurs between the most similar groups, 2 and 4, to form cluster (24).

At stage 3, we have

$$d_{(24)(35)} = \max\left\{d_{(35)2}, d_{(35)4}\right\} = \max\{10, 9\} = 10.$$

$$d_{(24)1} = \max\{d_{21}, d_{41}\} = \max\{9, 6\} = 9.$$

Now, the new distance matrix becomes

$$\begin{array}{cccc} & (\mathbf{35}) & (\mathbf{24}) & \mathbf{1} \\ (\mathbf{35}) & 0 & & \\ (\mathbf{24}) & 10 & 0 & \\ \mathbf{1} & 11 & \mathbf{9} & 0 \end{array}$$

Repeat the merging procedure again. Since $d_{(24)1} = 9$ is the smallest, cluster (24) and 1 are merged to form the cluster (124) .

At the final stage, the groups (35) and (124) are merged as the single cluster (12345), with

$$d_{(124)(35)} = \max\left\{d_{(35)1}, d_{(35)(24)}\right\} = \max\{11, 10\} = 11.$$

The final distance matrix becomes

$$\begin{array}{ccc} & (\mathbf{124}) & (\mathbf{35}) \\ (\mathbf{124}) & 0 & \\ (\mathbf{35}) & 11 & 0 \end{array}$$

Note that object one is grouped with 2 and 4 under the complete linkage, while it is grouped with 3 and 5 in the single linkage case.

Example 8.5: Consider the clustering of 11 language in the previous example, The first two clusters are (23), (68). We find the maximum distances between (23), (68), and the remaining objects. The maximum distances are

$$d_{(23)1} = \max\{d_{21}, d_{31}\} = \max\{2, 2\} = 2.$$

$$d_{(23)4} = \max\{d_{24}, d_{34}\} = \max\{5, 6\} = 6.$$

$$d_{(23)5} = \max\{d_{25}, d_{35}\} = \max\{4, 5\} = 5.$$

$$d_{(23)7} = \max\{d_{27}, d_{37}\} = \max\{6, 5\} = 6.$$

$$d_{(23)9} = \max\{d_{29}, d_{39}\} = \max\{7, 6\} = 7.$$

$$d_{(23)10} = \max\{d_{2,10}, d_{3,10}\} = \max\{8, 8\} = 8.$$

$$d_{(23)11} = \max\{d_{2,11}, d_{3,11}\} = \max\{9, 9\} = 9.$$

$$d_{(68)1} = \max\{d_{61}, d_{81}\} = \max\{6, 6\} = 6.$$

$$d_{(68)4} = \max\{d_{64}, d_{84}\} = \max\{9, 9\} = 9.$$

$$d_{(68)5} = \max\{d_{65}, d_{85}\} = \max\{7, 7\} = 7.$$

$$d_{(68)7} = \max\{d_{67}, d_{87}\} = \max\{2, 1\} = 2.$$

$$d_{(68)9} = \max\{d_{69}, d_{89}\} = \max\{5, 4\} = 5.$$

$$d_{(68)10} = \max\{d_{6,10}, d_{8,10}\} = \max\{10, 10\} = 10.$$

$$d_{(68)11} = \max\{d_{6,11}, d_{8,11}\} = \max\{9, 9\} = 9.$$

$$d_{(68)(23)} = \max\{d_{62}, d_{63}, d_{82}, d_{83}\} = \max\{6, 6, 6, 5\} = 6.$$

Now, the new distance matrix becomes

	$(2,3)$	$(6,8)$	1	4	5	7	9	10	11
	Nor, **Dan**	**Fren.**, **Ital.**	**Eng.**	**Dutch**	**Ger.**	**Span.**	**Polish**	**Hung.**	**Fin.**
Nor., **Danish**	0								
French, **Italian**	6	0							
English	**2**	6	0						
Dutch	6	9	7	0					
German	5	7	6	5	0				
Spanish	6	**2**	6	9	7	0			
Polish	7	5	7	10	8	3	0		
Hung.	8	10	9	8	9	10	10	0	
Finnish	9	9	9	9	9	9	9	8	0

Since $d_{(23)1}$ $d_{(68)7}$ are the smallest, object 1 and cluster (23) and are merged to form the cluster (123). Object 7 and cluster (68) and are merged to form the cluster (678). The maximum distances between (123), (678) and the remaining objects are

$$d_{(123)4} = \max\left\{d_{14}, d_{(23)4}\right\} = \max\{7,6\} = 7.$$

$$d_{(123)5} = \max\left\{d_{15}, d_{(23)5}\right\} = \max\{6,5\} = 6.$$

$$d_{(123)9} = \max\left\{d_{19}, d_{(23)9}\right\} = \max\{7,6\} = 7.$$

$$d_{(123)10} = \max\left\{d_{110}, d_{(23)10}\right\} = \max\{9,8\} = 9.$$

$$d_{(123)11} = \max\left\{d_{1,11}, d_{(23)11}\right\} = \max\{9,9\} = 9.$$

$$d_{(678)1} = \max\left\{d_{(68)1}, d_{71}\right\} = \max\{6,6\} = 6.$$

$$d_{(678)4} = \max\left\{d_{(68)4}, d_{74}\right\} = \max\{9, 9\} = 9.$$

$$d_{(678)5} = \max\left\{d_{(68)5}, d_{75}\right\} = \max\{7, 7\} = 7.$$

$$d_{(678)9} = \max\left\{d_{(68)9}, d_{79}\right\} = \max\{5, 3\} = 5.$$

$$d_{(678)10} = \max\left\{d_{(68)10}, d_{7,10}\right\} = \max\{10, 10\} = 10.$$

$$d_{(678)11} = \max\left\{d_{(68)11}, d_{7,11}\right\} = \max\{9, 9\} = 9.$$

$$d_{(123)(678)} = \max\left\{d_{1(68)}, d_{(23)(68)}, d_{17}, d_{(23)7}\right\} = \max\{6, 6, 6, 6\} = 6.$$

	$(1,2,3)$	$(6,7,8)$	4	5	9	10	11
	Eng.,	**French**,					
	Nor,	**Span**.,	**Dutch**	**Ger**.	**Polish**	**Hung**.	**Fin**.
	Dan	**Ital**.					
English, **Norwegian**, **Danish**	0						
French, **Spanish**, **Italian**	6	0					
Dutch	7	9	0				
German	6	7	**5**	0			
Polish	7	**5**	10	8	0		
Hungarian	9	10	8	9	10	0	
Finnish	9	9	9	9	9	8	0

Since d_{45} and $d_{(678)9}$ is the smallest, objects 4 and 5 and are merged to form the cluster (45). Object 9 and cluster (678) and are merged to form (6789). The maximum distances between $(45), (6789)$ and the remaining objects are

$$d_{(45)10} = \max\{d_{4,10}, d_{5,10}\} = \max\{8, 9\} = 9.$$

$$d_{(45)11} = \max\{d_{4,11}, d_{5,11}\} = \max\{9, 9\} = 9.$$

$$d_{(45)(123)} = \max\left\{d_{(123)4}, d_{(123)5}\right\} = \max\{7, 6\} = 7.$$

$$d_{(6789)(123)} = \max\left\{d_{(678)(123)}, d_{9(123)}\right\} = \max\{6, 7\} = 7.$$

$$d_{(6789)10} = \max\left\{d_{(678)10}, d_{910}\right\} = \max\{10, 10\} = 10.$$

$$d_{(6789)11} = \max\left\{d_{(678)11}, d_{9,11}\right\} = \max\{9, 9\} = 9.$$

$$d_{(45)(6789)} = \max\left\{d_{(6789)4}, d_{(6789)5}\right\} = \max\{9, 9\} = 9.$$

	$(1, 2, 3)$	$(6, 7, 8, 9)$	$4, 5$	10	11
	Eng., **Nor**, **Dan**	**French**, **Span.**, **Italian**, **Polish**	**Dutch**, **German**	**Hung.**	**Fin.**
English, **Norwegian**, **Danish**	0				
French, **Spanish**, **Italian**, **Polish**	**7**	0			
Dutch, **German**	**7**	9	0		
Hungarian	9	10	9	0	
Finnish	9	9	9	8	0

Note that $d_{(123)(6789)} = d_{(123)(45)} = 7$, but $d_{(6789)(45)} = 9 > 7$, we cannot group (6789) and (45) at this stage, but we can group (123) and (6789) to form the cluster (1236789). The maximum distances between (1236789) and the remaining objects are

$$d_{(1236789)10} = \max\left\{d_{(123)10}, d_{(6789),10}\right\} = \max\{9, 10\} = 10.$$

$$d_{(1236789)11} = \max\left\{d_{(123)11}, d_{(6789),11}\right\} = \max\{9, 9\} = 9.$$

$$d_{(1236789)(45)} = \max\left\{d_{(123)(45)}, d_{(6789)(45)}\right\} = \max\{7, 9\} = 9.$$

	$(1,2,3,6,7,8,9)$	$(4,5)$	10	11
	Eng., **Nor**, **Dan**, **French**, **Span.**, **Italian**, **Polish**	**Dutch**, **German**	**Hung.**	**Fin.**
English, **Norwegian**, **Danish**, **French**, **Spanish**, **Italian**, **Polish**	0			
Dutch, **German**	9	0		
Hungarian	10	9	0	
Finnish	9	9	**8**	0

Since $d_{10,11}$ is the smallest, objects 10 and 11 and are merged to form the cluster (10,11). The maximum distances between $(10, 11)$ and the remaining objects are

$$d_{(1236789)(10,11)} = \max\left\{d_{(1236789)10}, d_{(1236789),11}\right\} = \max\{10, 9\} = 10.$$

$$d_{(10,11)(45)} = \max\left\{d_{(45)10}, d_{(45)11}\right\} = \max\{9, 9\} = 9.$$

	$(1,2,3,6,7,8,9)$ **Eng., Nor, Dan,** **French, Span.,** **Italian, Polish**	$(4,5)$ **Dutch,** **German**	$(10,11)$ **Hung.,** **Finnish**
English, Norwegian, Danish, **French, Spanish,** **Italian, Polish**	0		
Dutch, German	9	0	
Hungarian, Finnish	10	9	0

Exercise 8.3: For the following dissimilarity matrix

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{ccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \mathbf{1} & 0 & & & \\ \mathbf{2} & 9 & 0 & & \\ \mathbf{3} & 7 & 6 & 0 & \\ \mathbf{4} & 7 & 10 & 7 & 0 \end{array}$$

Cluster the five items using each of the following procedures.

(a) Single linkage hierarchical procedure.

(b) Complete linkage hierarchical procedure.

(c) Draw the dendrograms and compare the results in (a) and (b).

(d) Repeat (a) to (c) if

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{cccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{1} & 0 & & & & \\ \mathbf{2} & 2 & 0 & & & \\ \mathbf{3} & 4 & 8 & 0 & & \\ \mathbf{4} & 7 & 9 & 3 & 0 & \\ \mathbf{5} & 9 & 8 & 7 & 5 & 0 \end{array}$$

8.3 Non-hierarchical clustering method

8.3.1 K-means method

Non-hierarchical methods start from an initial partition of items into groups, then assign an item to the cluster whose centroid (mean) is nearest.

Example 8.6: Suppose we measure two variables X_1 and X_2 for each of the four items A, B, C and D. The data are given in the following table:

Item\Observations	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

The objective is to divide these items into K=2 clusters such that the items within a cluster are closer to one other than they are to the items in different clusters. First, we arbitrarily partition the items into two clusters, such as (AB) and (CD), and compute the coordinates of the cluster centroid (mean), $(\overline{x}_1, \overline{x}_2)$. We have

Cluster\Centroid	$\overline{x}_1$	$\overline{x}_2$
(**AB**)	$\frac{5+(-1)}{2} = 2$	$\frac{3+1}{2} = 2$
(**CD**)	$\frac{1+(-3)}{2} - 1$	$\frac{-2+(-2)}{2} = -2$

Next, we compute the Euclidean distance of each item from the group centroids and reassign each item to the nearest group. Note that the cluster centroids must be updated before proceeding. We compute the squared distances

$$d^2\left(A, (AB)\right) = (5-2)^2 + (3-2)^2 = 10.$$

$$d^2\left(A, (CD)\right) = (5+1)^2 + (3+2)^2 = 61.$$

Since A is closer to cluster (AB) than to cluster (CD), it is not reassigned. Next, we check

$$d^2(B,(AB)) = (-1-2)^2 + (1-2)^2 = 10$$

$$d^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9.$$

Now, we need to reassign B to cluster (CD), giving cluster (BCD). We need to update the coordinates of the centroid to

Cluster\Centroid	$\overline{x}_1$	$\overline{x}_2$
A	5	3
(**BCD**)	$\frac{-1+1+(-3)}{3} = -1$	$\frac{1+(-2)+(-2)}{3} = -1$

Each item is checked for reassignment. Computing the squared distances gives the following table:

	squared	**distances to**	**group**	**centroid**
Cluster\Item	**A**	**B**	**C**	**D**
A	0	40	41	89
(**BCD**)	52	4	5	5

Since the items B, C and D is closer to the centroid of the cluster (BCD) than to A, the final K=2 clusters are A and (BCD).

Exercise 8.4: Suppose we measure two variables X_1 and X_2 for each of the four items A, B, C and D. The data are given as follows:

Item\Observations	x_1	x_2
A	5	4
B	1	-2
C	-1	1
D	3	1

Use the K-means clustering technique to divide the items into K=2 clusters. Start with the initial groups (AB) and (CD).

Exercise 8.5: Suppose we measure two variables X_1 and X_2 for each of the four items A, B, C and D. The data are given as follows:

Item\Observations	x_1	x_2
A	-2	0
B	2	0
C	0	4
D	0	-4

Use the K-means clustering technique to divide the items into K=2 clusters.

(i) Start with the initial groups (AB) and (CD).

(ii) Start with the initial groups (AD) and (BC).

Exercise 8.6: True/ False.

(a). The complete linkage clustering is a hierarchical clustering method.

(b). The solutions of the single linkage and complete linkage procedures can be the same.

(c). The solution of the single linkage hierarchical procedure is unique.

(d). The single linkage clustering is a hierarchical clustering method.

(e). In the complete linkage clustering, the distance between clusters is the maximum distance between two elements from each cluster.

Exercise 8.7: Suppose we measure two variables X_1 and X_2 for each of the four items A, B, C and D. The data are given as follows:

Item\Observations	x_1	x_2
A	-2	2
B	2	10
C	0	15
D	0	1

Use the K-means clustering technique to divide the items into K=2 clusters.

(i) Start with the initial groups (AB) and (CD).

(ii) Start with the initial groups (AD) and (BC).

Exercise 8.8: For the following dissimilarity matrix $\mathbf{D} = \{d_{ik}\} = \begin{array}{ccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \mathbf{1} & 0 & & & \\ \mathbf{2} & 8 & 0 & & \\ \mathbf{3} & 5 & 1 & 0 & \\ \mathbf{4} & 6 & 10 & 7 & 0 \end{array}$

Cluster the four items using each of the following procedures.

(a) Single linkage hierarchical procedure.

(b) Complete linkage hierarchical procedure.

(c) Draw the dendrograms and compare the results in (a) and (b).

Exercise 8.9: For the following dissimilarity matrix

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{cccccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \\ \mathbf{1} & 0 & & & & \\ \mathbf{2} & 2 & 0 & & & \\ \mathbf{3} & 4 & 8 & 0 & & \\ \mathbf{4} & 6 & 9 & 3 & 0 & \\ \mathbf{5} & 10 & 1 & 7 & 5 & 0 \end{array}$$

Cluster the five items using each of the following procedures.

(a) Single linkage hierarchical procedure.

(b) Complete linkage hierarchical procedure.

(c) Draw the dendrograms and compare the results in (a) and (b).

Exercise 8.10: Suppose we measure two variables X_1 and X_2 for each of the four items A, B, C and D. The data are given as follows:

Item\Observations	x_1	x_2
A	1	1
B	2	-2
C	-3	1
D	5	4
E	0	-1
F	-2	0

Use the K-means clustering technique to divide the items into K=2 clusters. Start with the initial groups (ABC) and (DEF).

Chapter 9

Binary and Multinomial Dependent Variable Models

In empirical studies, we often encounter variables which are qualitative rather than quantitative. For example, we may be interested in whether people participate in the labor force or not; whether people get married or not; whether people buy a car or not, etc., all these yes-no decisions are not quantifiable. In the case where the variable of interest belongs to one of the two categories, we normally give it a value of 1 if it falls into one category, and assign a value of 0 to it if it falls into another category.

9.1 Linear Probability Model

Consider a simple binary regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

Note very carefully that we cannot simply assume u_i to be $i.i.d.\,(0, \sigma^2)$, as Y_i cannot be treated as a predicted value in a regression line plus an arbitrary residual. This is because Y_i only takes either 0 or 1, so the residuals also take only two possible values for a given value of X_i. First, note that

$$E(Y_i) = 1 \times \Pr(Y_i = 1) + 0 \times \Pr(Y_i = 0) = \Pr(Y_i = 1).$$

Further, if $Y_i = 1$, then $u_i = 1-\beta_0-\beta_1 X_i$, and if $Y_i = 0$, $u_i = -\beta_0-\beta_1 X_i$.

$$\begin{aligned} E(u_i) &= (1-\beta_0-\beta_1 X_i)\Pr(Y_i=1)+(-\beta_0-\beta_1 X_i)\Pr(Y_i=0) \\ &= (1-\beta_0-\beta_1 X_i)\Pr(Y_i=1)+(-\beta_0-\beta_1 X_i)(1-\Pr(Y_i=1)) \\ &= \Pr(Y_i=1)-\beta_0-\beta_1 X_i. \end{aligned}$$

We can still assume $E(u_i) = 0$ in order to obtain an unbiased estimator. This will imply

$$\Pr(Y_i=1)-\beta_0-\beta_1 X_i = 0,$$

or

$$\Pr(Y_i=1) = \beta_0+\beta_1 X_i.$$

We call this a linear probability model, where β_1 can be interpreted as the marginal effect of X_i on the probability of getting $Y_i = 1$. To give a concrete example, suppose we have data on two groups of people, one group purchases sport car while the other purchases family car.

We define $Y_i = 1$ if a family car is purchased and $Y_i = 0$ if a sport car is purchased. Suppose X_i is the family size. Then β_1 can be interpreted as: if there is one more member in the family, by how much will the chance of buying a family car increase?

The advantage of using the linear probability model is that it is very simple, and the parameters are easily interpretable. We just need to run a regression and 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 $Var(u_i)$ to be a constant in this framework. To see why, note that

$$
\begin{aligned}
Var\left(u_i\right) &= E\left(u_i^2\right) - E^2\left(u_i\right) = E\left(u_i^2\right) \\
&= \left(1-\beta_0-\beta_1 X_i\right)^2 \Pr\left(Y_i=1\right) + \left(-\beta_0-\beta_1 X_i\right)^2 \Pr\left(Y_i=0\right) \\
&= \left(1-\beta_0-\beta_1 X_i\right)^2 \Pr\left(Y_i=1\right) + \left(\beta_0+\beta_1 X_i\right)^2 \Pr\left(Y_i=0\right) \\
&= \left(1-\Pr\left(Y_i=1\right)\right)^2 \Pr\left(Y_i=1\right) + \Pr\left(Y_i=1\right)^2 \Pr\left(Y_i=0\right) \\
&= \Pr\left(Y_i=0\right)^2 \Pr\left(Y_i=1\right) + \Pr\left(Y_i=1\right)^2 \Pr\left(Y_i=0\right) \\
&= \Pr\left(Y_i=0\right) \Pr\left(Y_i=1\right)\left[\Pr\left(Y_i=0\right)+\Pr\left(Y_i=1\right)\right] \\
&= \Pr\left(Y_i=0\right) \Pr\left(Y_i=1\right) \\
&= \left(1-\beta_0-\beta_1 X_i\right)\left(\beta_0+\beta_1 X_i\right),
\end{aligned}
$$

which is not a constant and will vary with X_i. Further, it may even be negative. Thus, we have the problem of heteroskedasticity, and the estimators will be inefficient.

Non-normality of the disturbances

Another problem is that the error distribution is not normal. This is because given the value of X_i, the disturbance u_i only takes 2 values, namely, $u_i = 1 - \beta_0 - \beta_1 X_i$ or $u_i = -\beta_0 - \beta_1 X_i$. We cannot apply the classical statistical tests to the estimates when the sample is small, since the tests depend on the normality of the errors. However, as sample size increases, it can be shown that the OLS estimators tend to be normally distributed. Therefore, in large samples, statistical inference of the LPM can be carried out as usual.

Low value of R^2

The conventional R^2 tends to be low in the binary regression model. Since all the Y values will either lie along the X axis or along the line corresponding to 1, no linear regression line will fit the data well. As a result, the conventional R^2 is likely to be much lower than 1 for such models. In most cases, the R^2 ranges from 0.2 to 0.6.

Nonfulfillment of $\mathbf{0} < \Pr\widehat{(Y_i = 1)} < \mathbf{1}$.

The other problem is on prediction. Since

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i = \Pr\widehat{(Y_i = 1)}$$

is the predicted probability of Y_i being equal to 1 given X_i, 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_i. An obvious solution for this problem is to set extreme predictions equal to 1 or 0, thereby constraining predicted probabilities within the zero-one interval.

This solution is not very satisfying either, as it suggests that we might have a predicted probability of 1 when it is entirely possible that an event may not occur, or we might have a predicted probability 0 when an event may actually occur. While the estimation procedure might yield unbiased estimates, the predictions obtained from the estimation process are clearly biased.

Example 9.1: Consider the following linear probability model:

$$Y_i = \beta_0 + \beta_1 INCOME_i + \beta_2 MARRIED_i + u_i,$$

where

$Y_i = 1$ if individual i purchased a car in the year of the survey and $Y_i = 0$ if not.

$INCOME_i$ =monthly income of individual i (in dollars).

$MARRIED_i = 1$ if individual i is married and $MARRIED_i = 0$ if not.

a) Show that $E(Y_i) = \Pr(Y_i = 1)$.

b) Show that $E(u_i) = 0$ implies

$$\Pr(Y_i = 1) = \beta_0 + \beta_1 INCOME_i + \beta_2 MARRIED_i.$$

c) Show that $\text{Var}(u_i) = \Pr(Y_i = 1)\Pr(Y_i = 0)$.

d) Suppose we estimate the model by OLS and obtain:

$$\widehat{Y}_i = -.1 + 0.0001INCOME_i + 0.3MARRIED_i.$$

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)

$$E(Y_i) = 0 \times \Pr(Y_i = 0) + 1 \times \Pr(Y_i = 1) = \Pr(Y_i = 1).$$

(b)

$$E(u_i) = 0 \Rightarrow E(Y_i) = \beta_0 + \beta_1 INCOME_i + \beta_2 MARRIED_i.$$

By using the result of part (a), i.e., $E(Y_i) = \Pr(Y_i = 1)$, we have

$$\Pr(Y_i = 1) = \beta_0 + \beta_1 INCOME_i + \beta_2 MARRIED_i.$$

(c) When $Y_i = 1$,

$$\begin{aligned} u_i &= 1 - \beta_0 - \beta_1 INCOME_i - \beta_2 MARRIED_i \\ &= 1 - \Pr(Y_i = 1) \\ &= \Pr(Y_i = 0). \end{aligned}$$

When $Y_i = 0$,

$$\begin{aligned} u_i &= 0 - \beta_0 - \beta_1 INCOME_i - \beta_2 MARRIED_i \\ &= -\Pr(Y_i = 1). \end{aligned}$$

Now,

$$\begin{aligned} Var(u_i) &= E\left(u_i^2\right) \text{ since } E(u_i) = 0 \\ &= \Pr(Y_i = 0)^2 \times \Pr(Y_i = 1) + (-\Pr(Y_i = 1))^2 \times \Pr(Y_i = 0) \\ &= \Pr(Y_i = 1)\Pr(Y_i = 0)\left[\Pr(Y_i = 0) + \Pr(Y_i = 1)\right] \\ &= \Pr(Y_i = 1)\Pr(Y_i = 0). \end{aligned}$$

(d)

$\beta_1 =$ Marginal Effect of change in monthly income on the probability of $Y_i = 1$.

$\beta_2 =$ Marginal Effect of change in marriage on the probability of $Y_i = 1$.

$\beta_0 =$ Effect on the probability of $Y_i = 1$ when the other variables are zero.

(e)

(i)

$$\widehat{Y} = -0.1 + (0.0001)(5000) + (0.3)(1) = 0.7.$$

(ii)

$$\widehat{Y} = -0.1 + (0.0001)(10000) + (0.3)(1) = 1.2.$$

(iii)

$$\widehat{Y} = -0.1 + (0.0001)(1000) + (0.3)(0) = 0.$$

(f) Advantage : It is convenient to carry out. Disadvantage : $0 < \widehat{Y}_i < 1$ may not be satisfied.

9.2 Logistic Regression

Since a linear probability model may yield a predicted value that is outside the [0,1] range, it is not a good model as far as prediction is concerned. To improve the linear probability model, one can modify the dependent variable a little bit. Suppose for each distinct value of X, we have many observations of Y, some are equal to 1, and some are equal to zero. For example, for a given value of X_i, we have N_i observations of Y_i, and n_i $(0 < n_i < N_i)$ of these Y_i are 1, and $N_i - n_i$ are 0. We let $P_i = \frac{n_i}{N_i}$ be the observed probability of observing $Y_i = 1$ given the value of X_i. We take a transformation of and let $Z_i = \ln \dfrac{P_i}{1-P_i}$, then Z_i can take any real value. We can run the following regression:

$$Z_i = \beta_0 + \beta_1 X_i + u_i.$$

Note that

$$\exp(Z_i) = \exp\left(\ln \frac{P_i}{1-P_i}\right) = \frac{P_i}{1-P_i}.$$

Thus, we have

$$\exp(-Z_i) = \frac{1}{P_i} - 1$$

and

$$P_i = \frac{1}{1+\exp(-Z_i)} = \frac{1}{1+\exp(-(\beta_0 + \beta_1 X_i))}.$$

Given the estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$from the Z regression , the predicted values of the probability that event will occur is

$$\widehat{P}_i = \frac{1}{1+\exp\left(-\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\right)},$$

which lies between 0 and 1. We call this method the Logistic regression, since $\dfrac{1}{1+\exp(-Z_i)}$ is the distribution function of a Logistic distribution.

9.3 Nonlinear Regression Approach

The linear probability model and the logistic regression model are linear regressions, in that all the $\beta's$ in the model are linearly related. To ensure a realistic predicted value, an alternative approach is to re-estimate the parameters subject to the constraint that the predicted value is bounded between zero and one. Since predicted value is the value in a regression curve, we can find a nonlinear function $\widehat{Y_i} = g\left(X_i, \beta\right)$ such that $0 \leq g\left(X_i, \beta\right) \leq 1$ for all β and X_i. Clearly $g\left(X_i, \beta\right)$ cannot be linear in either β or X, i.e., $g\left(X_i, \beta\right) = \beta_0 + \beta_1 X_i$ 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? For example, the cumulative normal distribution has an increasing, S-shaped CDF bounded between zero and one. Another example is the logistic distribution, i.e.,

$$g\left(X_i, \beta\right) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 X_i\right)\right]}.$$

Note that as $\beta_1 X_i \to -\infty$, $g\left(X_i, \beta\right) \to 0$, and as $\beta_1 X_i \to \infty$, $g\left(X_i, \beta\right) \to 1$. Since $g\left(X_i, \beta\right)$ is not linear in β, we cannot use the linear least squares method. Instead, we should run a nonlinear regression

$$Y_i = \frac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 X_i\right)\right]} + u_i.$$

i.e., we find β_0 and β_1 to minimize $\sum_{i=1}^{n}\left(Y_i - \dfrac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 X_i\right)\right]}\right)^2$. Or we can assume $u_i = Y_i - \dfrac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 X_i\right)\right]}$ to have a certain distribution and apply the Maximum Likelihood method to estimate β_0 and β_1.

Exercise 9.1: For the Logistic distribution function $F(x) = \dfrac{1}{1 + \exp\left(-x\right)}$, find the density function $f\left(x\right)$. Is $f\left(x\right)$ a symmetric density?

9.4 Random Utility Model

Suppose you have to make a decision on two alternatives, say, whether to buy a sport car or family car. Given the characteristics X_i of individual i , for example, his/her family size, income, etc. Let

$$U_{i1} = \alpha_0 + \alpha_1 X_i + \varepsilon_{i1},$$

$$U_{i2} = \gamma_0 + \gamma_1 X_i + \varepsilon_{i2},$$

where U_{i1} is the utility derived from a family car, and U_{i2} is the utility derived from a sport car. The individual will buy a family car if $U_{i1} > U_{i2}$, or $U_{i1} - U_{i2} > 0$. Subtracting the second equation from the first equation gives

$$U_{i1} - U_{i2} = \alpha_0 - \gamma_0 + (\alpha_1 - \gamma_1) X_i + \varepsilon_{i1} - \varepsilon_{i2}.$$

Suppose we define $Y_i^* = U_{i1} - U_{i2}$, $\beta_0 = \alpha_0 - \gamma_0$, $\beta_1 = \alpha_1 - \gamma_1$, $u_i = \varepsilon_{i1} - \varepsilon_{i2}$. We can rewrite the model as

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i.$$

However, we cannot observe the exact value of Y_i^*, what we observe is whether the individual buy a family car or not. That is, we only observe whether $Y_i^* > 0$ or $Y_i^* < 0$. If $Y_i^* > 0$, the individual will buy a family car, we assign a value $Y_i = 1$ for this observation, and assign $Y_i = 0$ otherwise.

Denote the density function and distribution function of u_i by $f(\cdot)$ and $F(\cdot)$ respectively, and suppose it is symmetric about zero, i.e., $f(u_i) = f(-u_i)$, and $F(u_i) = 1 - F(-u_i)$. We then have:

$$\begin{aligned}
\Pr(Y_i = 1) &= \Pr(Y_i^* > 0) \\
&= \Pr(\beta_0 + \beta_1 X_i + u_i > 0) \\
&= \Pr(-u_i < \beta_0 + \beta_1 X_i) \\
&= \Pr(u_i < \beta_0 + \beta_1 X_i) \qquad \text{since } u_i \text{ is symmetrically distributed about zero,} \\
&= F(\beta_0 + \beta_1 X_i),
\end{aligned}$$

and

$$\Pr(Y_i = 0) = 1 - \Pr(Y_i = 1) = 1 - F(\beta_0 + \beta_1 X_i).$$

Note that the marginal effects of an increase in X_i in the probability is nonlinear in $\beta' s$, in particular,

$$\frac{\partial \Pr(Y = 0)}{\partial X_i} = -f(\beta_0 + \beta_1 X_i)\beta_1,$$

$$\frac{\partial \Pr(Y_i = 1)}{\partial X_i} = f(\beta_0 + \beta_1 X_i)\beta_1.$$

Consider the case where $\beta_1 > 0$, since $f(\cdot) > 0$, we have

$$\begin{aligned}
\frac{\partial \Pr(Y_i = 0)}{\partial X_i} &< 0 \\
\frac{\partial \Pr(Y_i = 1)}{\partial X_i} &> 0.
\end{aligned}$$

9.5 Maximum Likelihood Estimation

The principle of maximum likelihood provides a mean of choosing an asymptotically efficient estimator for a set of parameters. Let $\{y_i\}_{i=1}^n$ be i.i.d. random variable with joint density $f(y_1, y_{2...}, y_n; \theta)$, where $\theta = (\theta_1, \theta_2, ...\theta_K)'$. Since the sample values have been observed and therefore fixed number, we regard $f(y_i; \theta)$ as a function of θ. Let $y = (y_1, y_{2...}, y_n)'$, we defined the **likelihood function** as

$$L(y;\theta) = f(y_1, y_2, ..., y_n; \theta) = \prod_{i=1}^{n} f(y_i;\theta).$$

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

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

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

If the distribution is correctly specified, then the Maximum Likelihood estimator is unbiased and is asymptotically more efficient than any estimators. If the variable is discrete, the density function can be replaced by the probability that each discrete value will take.

Example 9.2: **:** Consider a random sample of 10 observations from a Normal distribution $y_1, y_2, ..., y_{10}$. The density of y_i is

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

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

(a) Find the log-likelihood function.

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

Solution:

$$\begin{aligned}
L\left(y;\mu,\sigma^2\right) &= f\left(y_1,y_2,...,y_{10};\mu,\sigma^2\right) \\
&= \prod_{i=1}^{10} f\left(y_i;\mu,\sigma^2\right). \\
&= \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(y_i-\mu\right)^2}{2\sigma^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{10}\exp\left(-\frac{\sum_{i=1}^{10}\left(y_i-\mu\right)^2}{2\sigma^2}\right) \\
&= \left(2\pi\sigma^2\right)^{-5}\exp\left(-\frac{\sum_{i=1}^{10}\left(y_i-\mu\right)^2}{2\sigma^2}\right).
\end{aligned}$$

$$\begin{aligned}
\ln L\left(y;\mu,\sigma^2\right) &= \ln\left[\left(2\pi\sigma^2\right)^{-5}\exp\left(-\frac{\sum_{i=1}^{10}\left(y_i-\mu\right)^2}{2\sigma^2}\right)\right] \\
&= -5\ln\left(2\pi\right)-5\ln\left(\sigma^2\right)-\frac{\sum_{i=1}^{10}\left(y_i-\mu\right)^2}{2\sigma^2}.
\end{aligned}$$

$$\widehat{\theta}_{ML} = \arg\max\left(\ln L\left(y;\mu,\sigma^2\right)\right).$$

First-order condition,

$$\frac{\partial}{\partial\mu}\ln L\left(y;\mu,\sigma^2\right) = \frac{\sum_{i=1}^{10}\left(y_i-\mu\right)}{\sigma^2} = 0.$$

$$\frac{\partial}{\partial\sigma^2}\ln L\left(y;\mu,\sigma^2\right) = -\frac{5}{\sigma^2}+\frac{\sum_{i=1}^{10}\left(y_i-\mu\right)^2}{2\sigma^4} = 0.$$

$$\widehat{\mu}_{ML} = \frac{\sum_{i=1}^{10} y_i}{10} = \overline{y}.$$

Plug $\widehat{\mu}_{ML} = \overline{y}$ into the second equation, we have

$$\begin{aligned}
-\frac{5}{\sigma^2}+\frac{\sum_{i=1}^{10}\left(y_i-\overline{y}\right)^2}{2\sigma^4} &= 0 \\
-1+\frac{\sum_{i=1}^{10}\left(y_i-\overline{y}\right)^2}{10\sigma^2} &= 0
\end{aligned}$$

$$\widehat{\sigma}_{ML}^{2}=\frac{\sum_{i=1}^{10}\left(y_{i}-\overline{y}\right)^{2}}{10}.$$

Example 9.3: Consider a random sample of 10 observations from a Poisson distribution $y_1, y_2, ..., y_{10}$. The probability of each observation is

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

with

$$E\left(y_i\right)=\theta,$$

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

$$\begin{aligned}L\left(y;\theta\right) &= f\left(y_1,y_2,...,y_{10};\theta\right)\\ &= \underset{i=1}{\overset{10}{\Pi}} f\left(y_i;\theta\right).\\ &= \underset{i=1}{\overset{10}{\Pi}}\frac{\theta^{y_i}\exp\left(-\theta\right)}{y_i!}\\ &= \frac{\theta^{y_1+y_2...+y_{10}}\exp\left(-10\theta\right)}{\underset{i=1}{\overset{10}{\Pi}} y_i!}.\end{aligned}$$

$$\ln L\left(y;\theta\right)=\ln\frac{\theta^{y_1+y_2...+y_{10}}\exp\left(-10\theta\right)}{\underset{i=1}{\overset{10}{\Pi}} y_i!}=\left(\sum_{i=1}^{10}y_i\right)\ln\theta-10\theta-\ln\left(\underset{i=1}{\overset{10}{\Pi}} y_i!\right)$$

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

First-order condition,

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

$$\widehat{\theta}_{ML} = \frac{\sum_{i=1}^{10} y_i}{10}.$$

Exercise 9.2: Consider a random sample of 10 observations from a Normal distribution $y_1, y_2, ..., y_{10}$. The density of y_i is

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

where θ_1, θ_2 are unknown parameters.

(a) Find the log-likelihood function.

(b) Now let the observations be

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

Find the values of ML estimates for θ_1 and θ_2.

Exercise 9.3: Consider the following density function of a random variable X.

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

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

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

ii) Find the joint density of X_1 and X_2.

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

Exercise 9.4: Suppose the random variable $y_i \sim N\left(\exp\left(\theta\right), 1\right)$, $i = 1, 2, ..., 100$, y_i and y_j are independent for all $i \neq j$. Thus,

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

a) Derive the log-likelihood function $\ln L\left(y;\theta\right)$.

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

Exercise 9.5: Given the data $y=\left(y_1,y_2,...,y_n\right)'$. y_i is an i.i.d. random variable with density function

$$f\left(y_i;\theta\right)=\frac{1}{\theta}e^{-\frac{y_i}{\theta}},\qquad 0<y_i<\infty.$$

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

b) Find the ML estimator of θ.

Exercise 9.6: Suppose the span of human life follows a uniform distribution $U\left(0,\theta\right)$, with $\theta<\infty$. Suppose we have a sample of n observations $y_1,y_2,...,y_n$ on people's life span.

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

b) Find the ML estimator of θ by solving the first-order condition. Does your estimator depend on the data?

c) Suggest another ML estimator that uses the information of the data and is based on the maximum of the log-likelihood function.

9.6 Maximum Likelihood Estimation of the Probit and Logit Models

Let $L(y_1, y_2, ..., y_n; \beta)$ be the joint probability density of the sample observations when the true parameter is β. This is a function of $y_1, y_2, ..., y_n$ and β. As a function of the sample observation it is called a joint probability density function of $y_1, y_2, ..., y_n$. 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 $L(y_1, y_2, ..., y_n; \beta)$.

Intuitively speaking, if we have several values of β, each of which might be the true value, we would like to find a value of β which gives the sample we actually observe the highest probability. Suppose we have n observations of Y and X, where Y takes the value zero or one. The probability of observing such data is

$$\begin{aligned} L &= \Pr(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) \\ &= \Pr(Y_1 = y_1)\Pr(Y_2 = y_2)...\Pr(Y_n = y_n) \end{aligned}$$

by the independence of u_i.

Since y_i only takes either zero or one, we can assign them to two groups. The likelihood function is

$$\begin{aligned} L &= \prod_{y_i=1} \Pr(Y_i = 1) \prod_{y_i=0} \Pr(Y_i = 0) \\ &= \prod_{y_i=1} F(\beta_0 + \beta_1 X_i) \prod_{y_i=0} [1 - F(\beta_0 + \beta_1 X_i)] \\ &= \prod_{i=1}^{n} [F(\beta_0 + \beta_1 X_i)]^{Y_i} [1 - F(\beta_0 + \beta_1 X_i)]^{1-Y_i}. \end{aligned}$$

$$\begin{aligned}
\ln L &= \ln\left\{\prod_{i=1}^{n}\left[F\left(\beta_0+\beta_1X_i\right)\right]^{Y_i}\left[1-F\left(\beta_0+\beta_1X_i\right)\right]^{1-Y_i}\right\} \\
&= \sum_{i=1}^{n}\ln\left\{\left[F\left(\beta_0+\beta_1X_i\right)\right]^{Y_i}\left[1-F\left(\beta_0+\beta_1X_i\right)\right]^{1-Y_i}\right\} \\
&= \sum_{i=1}^{n}Y_i\ln F\left(\beta_0+\beta_1X_i\right)+\sum_{i=1}^{n}\left(1-Y_i\right)\ln\left[1-F\left(\beta_0+\beta_1X_i\right)\right].
\end{aligned}$$

We would like to maximize L, or equivalently, maximize $\ln L$ since $\ln(\cdot)$ is a monotonic increasing function. The first-order conditions are

$$\frac{\partial\ln L}{\partial\beta_0}=\sum_{i=1}^{n}Y_i\frac{f\left(\beta_0+\beta_1X_i\right)}{F\left(\beta_0+\beta_1X_i\right)}-\sum_{i=1}^{n}\left(1-Y_i\right)\frac{f\left(\beta_0+\beta_1X_i\right)}{1-F\left(\beta_0+\beta_1X_i\right)}=0,$$

$$\frac{\partial\ln L}{\partial\beta_1}=\sum_{i=1}^{n}Y_iX_i\frac{f\left(\beta_0+\beta_1X_i\right)}{F\left(\beta_0+\beta_1X_i\right)}-\sum_{i=1}^{n}\left(1-Y_i\right)X_i\frac{f\left(\beta_0+\beta_1X_i\right)}{1-F\left(\beta_0+\beta_1X_i\right)}=0.$$

These two equations can be solved to obtain estimators for $\beta's$. However, as $\ln L$ is a highly nonlinear function of $\beta's$, we cannot easily obtain the estimator of $\beta's$ by simple substitutions. We may use the grid-search method and a computer algorithm to solve the problem.

The MLE procedure has a number of desirable properties. When the sample size is large, all estimators are consistent and efficient if there is no misspecification on the probability distribution. In addition, all parameters are normally distributed when sample size is large.

If we **assume** u_i to be **normally** distributed $N\left(0,\sigma^2\right)$, i.e.,

$$f\left(\beta_0+\beta_1X_i\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(\beta_0+\beta_1X_i\right)^2}{2\sigma^2}\right),$$

$$F\left(\beta_0+\beta_1 X_i\right)=\int_{-\infty}^{\beta_0+\beta_1 X_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{i^2}{2\sigma^2}\right) dt,$$

then we have the **Probit Model**.

The first-order condition can be simplified to

$$\frac{\partial \ln L}{\partial \beta_0}=\sum_{Y_i=1} \frac{\exp\left(-\frac{\left(\beta_0+\beta_1 X_i\right)^2}{2\sigma^2}\right)}{\int_{-\infty}^{\beta_0+\beta_1 X_i} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv}-\sum_{Y_i=0} \frac{\exp\left(-\frac{\left(\beta_0+\beta_1 X_i\right)^2}{2\sigma^2}\right)}{\int_{\beta_0+\beta_1 X_i}^{\infty} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv}=0,$$

$$\frac{\partial \ln L}{\partial \beta_1}=\sum_{Y_i=1} \frac{X_i \exp\left(-\frac{\left(\beta_0+\beta_1 X_i\right)^2}{2\sigma^2}\right)}{\int_{-\infty}^{\beta_0+\beta_1 X_i} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv}-\sum_{Y_i=0} \frac{X_i \exp\left(-\frac{\left(\beta_0+\beta_1 X_i\right)^2}{2\sigma^2}\right)}{\int_{\beta_0+\beta_1 X_i}^{\infty} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv}=0.$$

Although the normal distribution is a commonly used distribution, its distribution function is not a closed form function of u_i. 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:

$$\begin{aligned} f\left(\beta_0+\beta_1 X_i\right) &= \frac{\exp\left(\beta_0+\beta_1 X_i\right)}{\left(1+\exp\left(\beta_0+\beta_1 X_i\right)\right)^2}, \\ F\left(\beta_0+\beta_1 X_i\right) &= \frac{\exp\left(\beta_0+\beta_1 X_i\right)}{1+\exp\left(\beta_0+\beta_1 X_i\right)}. \end{aligned}$$

If we assume u_i to have a logistic distribution, then we have the **Logit Model**. The first-order condition can be simplified to

$$\frac{\partial \ln L}{\partial \beta_0}=\sum_{Y_i=1} \frac{1}{1+\exp\left(\beta_0+\beta_1 X_i\right)}-\sum_{Y_i=0} \frac{1}{1+\exp\left(-\beta_0-\beta_1 X_i\right)}=0,$$

$$\frac{\partial \ln L}{\partial \beta_1}=\sum_{Y_i=1}\frac{X_i}{1+\exp\left(\beta_0+\beta_1 X_i\right)}-\sum_{Y_i=0}\frac{X_i}{1+\exp\left(-\beta_0-\beta_1 X_i\right)}=0.$$

Exercise 9.7: True/False.

(a) A Probit model assumes that the error term has a uniform distribution.

(b)A Probit model assumes that the error term has an F distribution.

Exercise 9.8: Consider the Probit model

$$Y_i^*=\beta_0+\beta_1 X_i+u_i.$$

Suppose we can only observe the sign of Y_i^*. If $Y_i^*>0$, we assign a value $Y_i=1$ for this observation, and assign $Y_i=0$ otherwise. Denote the density function and distribution function of u_i by $f(\cdot)$ and $F(\cdot)$ respectively, where

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

$$F\left(u_i\right)=\int_{-\infty}^{u_i}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{v^2}{2\sigma^2}\right)dv,$$

(a) Show that

$$\Pr\left(Y_i=1\right)=F\left(\beta_0+\beta_1 X_i\right),$$

and

$$\Pr\left(Y_i=0\right)=1-F\left(\beta_0+\beta_1 X_i\right).$$

(b) Suppose we have n observations of Y and X, where Y takes the value zero or one. Assume u_i to be independent, show that the log-likelihood function can be simplified to

$$\begin{aligned}\ln L &= \sum_{i=1}^{n} Y_i \ln \int_{-\infty}^{\beta_0+\beta_1 X_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv \\ &\quad + \sum_{i=1}^{n} (1-Y_i) \ln \left[\int_{\beta_0+\beta_1 X_i}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{v^2}{2\sigma^2}\right) dv\right].\end{aligned}$$

(c) Let $w = \frac{v}{\sigma}$, show that

$$\begin{aligned}\ln L &= \sum_{i=1}^{n} Y_i \ln \int_{-\infty}^{\frac{\beta_0}{\sigma}+\frac{\beta_1}{\sigma} X_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) dw \\ &\quad + \sum_{i=1}^{n} (1-Y_i) \ln \left[\int_{\frac{\beta_0}{\sigma}+\frac{\beta_1}{\sigma} X_i}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) dw\right].\end{aligned}$$

(d) Given the data $\{X_i, Y_i\}_{i=1}^{n}$, suppose $\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}\right) = (1, 2, 3)$ maximizes the log-likelihood function, will $\left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}\right) = (2, 4, 6)$ also maximize the log-likelihood function? Discuss the identifiability of β_0 and β_1.

Exercise 9.9: Consider the following linear probability model:

$$\begin{aligned}DIVORCE_i &= \beta_0 + \beta_1 INCOME_i + \beta_2 YEARMARRIED_i + \beta_3 AFFAIR_i \\ &\quad + \beta_4 CHILDREN_i + u_i,\end{aligned}$$

where

$DIVORCE_i = 1$ if couple i got divorce in the year of the survey, and $DIVORCE_i = 0$ if not.

$INCOME_i =$ family's monthly income of couple i (in dollars).

$YEARMARRIED_i =$ years of marriage of couple i.

$AFFAIR_i = 1$ if the husband or the wife (or both) has an extramarital affair, and $AFFAIR_i = 0$ if not.

$CHILDREN_i =$ number of children of couple i.

a) Show that $E\left(DIVORCE_i\right) = \Pr\left(DIVORCE_i = 1\right)$.

b) Interpret each of the above coefficients $\beta_0, ..., \beta_4$.

c) Show that $E(u_i) = 0$ implies

$$\begin{aligned}\Pr(DIVORCE_i = 1) &= \beta_0 + \beta_1 INCOME_i + \beta_2 YEARMARRIED_i + \beta_3 AFFAIR_i \\ &\quad + \beta_4 CHILDREN_i.\end{aligned}$$

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 with 6 years of marriage, 2 children, family's monthly income of 1000 dollars, and no extramarital affair.

ii) a couple with 1 year of marriage, no children, family's monthly income of 2000 dollars, and the husband has an extramarital affair.

iii) a couple with 30 years of marriage, 3 children, family's monthly income of 4000 dollars, and the wife has an extramarital affair.

f) State an advantage and a shortcoming of the linear probability model.

Exercise 9.10: Consider the following linear probability model:

$$\begin{aligned}AFFAIR_i &= \beta_0 + \beta_1 INCOME_i + \beta_2 SPOUSEINCOME_i + \beta_3 YEARMARRIED_i \\ &\quad + \beta_4 CHILDREN_i + \beta_5 HRTOGETHER_i + \beta_6 SEX_i + u_i,\end{aligned}$$

where

$AFFAIR_i = 1$ if individual i has an extramarital affair, and $= 0$ if not,

$INCOME_i$ = monthly income of individual i (in dollars),

$SPOUSEINCOME_i$ = monthly income of the spouse of individual i,

$YEARMARRIED_i$ =years of marriage of individual i,

$CHILDREN_i$ = number of children of individual i,

$HRTOGETHER_i$ =number of hours per week that individual i spends with his/her spouse.

$SEX_i = 1$ if individual i is a male, and $= 0$ otherwise.

(a) Interpret each of the above coefficients $\beta_1, ..., \beta_6$, what are their expected signs? Explain.

(b) Show that $E(u_i) = 0$ implies

$$\begin{aligned}\Pr(AFFAIR_i = 1) &= \beta_0 + \beta_1 INCOME_i + \beta_2 SPOUSEINCOME_i \\ &\quad + \beta_3 YEARMARRIED_i + \beta_4 CHILDREN_i \\ &\quad + \beta_5 HRTOGETHER_i + \beta_6 SEX_i.\end{aligned}$$

(c) Show that $\mathrm{Var}(u_i) = \Pr(AFFAIR_i = 1)\Pr(AFFAIR_i = 0)$.

(d) Suggest a method to fix the problem of heteroskedasticity in part (c). What is the advantage and shortcoming of your method?

(e) Suppose the we estimate the model by OLS and obtain:

$$\begin{aligned}\widehat{AFFAIR}_i &= .5 + .008 INCOME_i - .009 SPOUSEINCOME_i \\ &\quad -.015 YEARMARRIED_i - .03 CHILDREN_i \\ &\quad -.004 HRTOGETHER_i + .007 SEX_i.\end{aligned}$$

What is the chance of having an extramarital affair for:

i) a man with 6 years of marriage, 2 children, monthly income of 1000 dollars, wife's income is 800 and he spends 100 hours per week with his wife.

ii) a woman with 1 years of marriage, 1 child, monthly income of 1000 dollars, husband's income is 900 and she spends 56 hours per week with his husband.

iii) a man with 30 years of marriage, 3 children, monthly income of 700 dollars, wife's income is 500 and he spends 120 hours per week with his wife.

9.7 The Multinomial Logit Model

Suppose there are n individuals and J categories, e.g., Occupational choice. Define $Y_{ij} = 1$ if individual i chooses category j, and $Y_{ij} = 0$ otherwise. Thus, $\sum_{j=1}^{J} Y_{ij} = 1$ for all i.

For example, let $J = 3$. Suppose that an individual i whose utilities associated with three alternatives are given by

$$U_{ij} = X_i' \beta_j + \varepsilon_{ij}, \qquad j = 1, 2, 3.$$

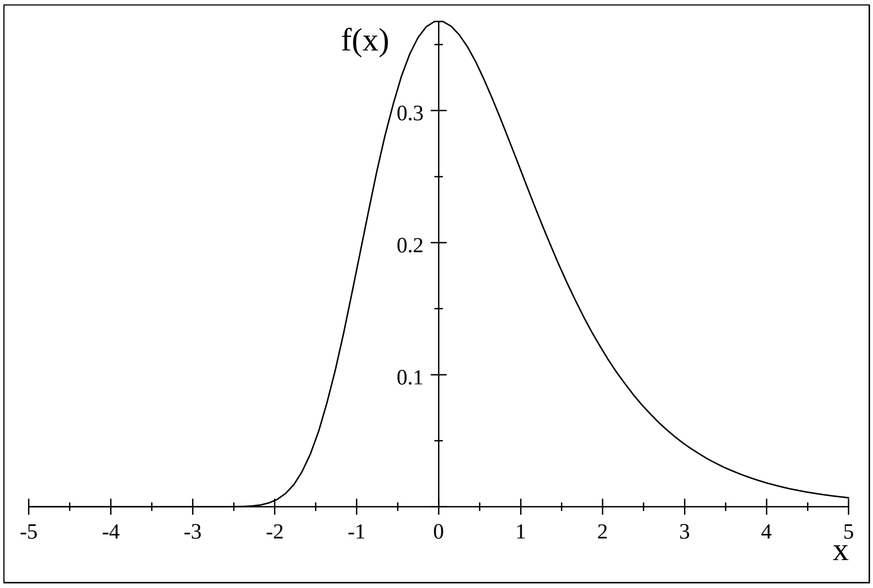
where X and β are vectors.

Assume that ε_{ij} are independent and identically distributed, each with the *extreme value* distribution

$$F(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij})),$$

$$f(\varepsilon_{ij}) = \exp(-\varepsilon_{ij}) \exp(-\exp(-\varepsilon_{ij})).$$

The density is shown in the following diagram:



The density function of an extreme value distribution

Now, if there are 3 categories, category 1, 2 and 3. The probability that individual i will choose category 2 is

$$\begin{aligned}
& \Pr\left(Y_{i2}=1\right) \\
= & \Pr\left(U_{i2}>U_{i1} \text{ and } U_{i2}>U_{i3}\right) \\
= & \Pr\left(X_i'\beta_2+\varepsilon_{i2}>X_i'\beta_1+\varepsilon_{i1} \text{ and } X_i'\beta_2+\varepsilon_{i2}>X_i'\beta_3+\varepsilon_{i3}\right) \\
= & \Pr\left(\varepsilon_{i1}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_1\right) \text{ and } \varepsilon_{i3}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_3\right)\right) \\
= & \int_{-\infty}^{\infty} f\left(\varepsilon_{i2}\right)\Pr\left(\varepsilon_{i1}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_1\right) \text{ and } \varepsilon_{i3}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_3\right)|\varepsilon_{i2}\right)d\varepsilon_{i2} \\
= & \int_{-\infty}^{\infty} f\left(\varepsilon_{i2}\right)\Pr\left(\varepsilon_{i1}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_1\right)|\varepsilon_{i2}\right)\Pr\left(\varepsilon_{i3}<\varepsilon_{i2}+X_i'\left(\beta_2-\beta_3\right)|\varepsilon_{i2}\right)d\varepsilon_{i2} \\
= & \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\varepsilon_{i2}+X_i'(\beta_2-\beta_1)} f\left(\varepsilon_{i1}\right)d\varepsilon_{i1}\right]\left[\int_{-\infty}^{\varepsilon_{i2}+X_i'(\beta_2-\beta_3)} f\left(\varepsilon_{i3}\right)d\varepsilon_{i3}\right]dF\left(\varepsilon_{i2}\right)
\end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \exp\left[-\exp\left(-\varepsilon_{i2}\right)\exp\left(X_i'\left(\beta_1-\beta_2\right)\right)\right]\exp\left[-\exp\left(-\varepsilon_{i2}\right)\exp\left(X_i'\left(\beta_3-\beta_2\right)\right)\right]dF\left(\varepsilon_{i2}\right)\\
&= \int_{-\infty}^{\infty} F\left(\varepsilon_{i2}\right)^{\exp\left(X_i'(\beta_1-\beta_2)\right)} F\left(\varepsilon_{i2}\right)^{\exp\left(X_i'(\beta_3-\beta_2)\right)} dF\left(\varepsilon_{i2}\right)\\
&= \int_{-\infty}^{\infty} F\left(\varepsilon_{i2}\right)^{\exp\left(X_i'(\beta_1-\beta_2)\right)+\exp\left(X_i'(\beta_3-\beta_2)\right)} dF\left(\varepsilon_{i2}\right)\\
&= \left[\frac{F\left(\varepsilon_{i2}\right)^{1+\exp\left(X_i'(\beta_1-\beta_2)\right)+\exp\left(X_i'(\beta_3-\beta_2)\right)}}{1+\exp\left(X_i'\left(\beta_1-\beta_2\right)\right)+\exp\left(X_i'\left(\beta_3-\beta_2\right)\right)}\right]_{-\infty}^{\infty}\\
&= \frac{1}{1+\exp\left(X_i'\left(\beta_1-\beta_2\right)\right)+\exp\left(X_i'\left(\beta_3-\beta_2\right)\right)}\\
&= \frac{\exp\left(X_i'\beta_2\right)}{\exp\left(X_i'\beta_1\right)+\exp\left(X_i'\beta_2\right)+\exp\left(X_i'\beta_3\right)}.
\end{aligned}$$

Therefore, if there are J categories, the probability that individual i will choose the jth category will be

$$\Pr\left(Y_{ij}=1\right)=\frac{\exp\left(X_i'\beta_j\right)}{\sum_{k=1}^{J}\exp\left(X_i'\beta_k\right)}.$$

One problem arises here, the β_j here cannot be identified as if we change all the β to $\beta+c$, where c is a vector of any constant, $\Pr\left(Y_{ij}=1\right)$ will still be the same since

$$\frac{\exp\left(X_i'\left(\beta_j+c\right)\right)}{\sum_{k=1}^{J}\exp\left(X_i'\left(\beta_k+c\right)\right)}=\frac{\exp\left(X_i'c\right)\exp\left(X_i'\left(\beta_j+c\right)\right)}{\exp\left(X_i'c\right)\sum_{k=1}^{J}\exp\left(X_i'\left(\beta_k+c\right)\right)}=\frac{\exp\left(X_i'\beta_j\right)}{\sum_{k=1}^{J}\exp\left(X_i'\beta_k\right)}.$$

Therefore, for the parameter to be identified, we must impose some restrictions on β. We can simply let $\beta_1=0$, so that

$$\Pr\left(Y_{i1}=1\right)=\frac{1}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)},$$

$$\Pr\left(Y_{ij}=1\right)=\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}\qquad j=2,3,...,J.$$

The likelihood function is

$$L=\prod_{i=1}^{n}\prod_{j=1}^{J}\Pr\left(Y_{ij}=1\right)^{Y_{ij}}=\prod_{i=1}^{n}\prod_{j=1}^{J}\left[\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}\right]^{Y_{ij}}$$

By using the conditions that $\beta_1=0$ and $\sum_{j=1}^{J}Y_{ij}=1$, we have

$$\begin{aligned}
\ln L &= \sum_{i=1}^{n}\sum_{j=1}^{J}Y_{ij}\ln\left(\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{J}Y_{ij}\left(X_i'\beta_j-\ln\left[1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)\right]\right)\\
&= \sum_{i=1}^{n}\left(\sum_{j=1}^{J}Y_{ij}X_i'\beta_j-\left(\sum_{j=1}^{J}Y_{ij}\right)\ln\left[1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)\right]\right)\\
&= \sum_{i=1}^{n}\left(\sum_{j=2}^{J}Y_{ij}X_i'\beta_j-\ln\left[1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)\right]\right).
\end{aligned}$$

$$\frac{\partial\ln L}{\partial\beta_j}=\sum_{i=1}^{n}\left(Y_{ij}X_i'-\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}X_i'\right)=\sum_{i=1}^{n}\left(Y_{ij}-\frac{\exp\left(X_i'\beta_j\right)}{1+\sum_{k=2}^{J}\exp\left(X_i'\beta_k\right)}\right)X_i'.$$

Exercise 9.11: Find $E\left(X\right)$ and $Var\left(X\right)$ of the random variable X with

$$F\left(x\right)=\exp\left(-\exp\left(-x\right)\right),$$

$$f\left(x\right)=\exp\left(-x\right)\exp\left(-\exp\left(-x\right)\right).$$

9.8 Ordered Data

Some multinomial-choice variables are inherently ordered, e.g., Bond ratings, opinion surveys, employment (unemployed, part time, or full time). Consider the model

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i.$$

where Y_i^* is unobserved. What we observe is

$$\begin{aligned} Y_i &= 1 \quad \text{if } \mu_0 < Y_i^* \le \mu_1, \\ &= 2 \quad \text{if } \mu_1 < Y_i^* \le \mu_2, \\ &= 3 \quad \text{if } \mu_2 < Y_i^* \le \mu_3, \\ &\vdots \\ &= J \quad \text{if } \mu_{J-1} < Y_i^* \le \mu_J, \end{aligned}$$

where $\mu_0 = -\infty$ and $\mu_J = \infty$. Other $\mu's$ are unknown parameters to be estimated with $\beta's$.

$$\begin{aligned} \Pr(Y_i = j) &= \Pr\left(\mu_{j-1} < Y_i^* \le \mu_j\right) \\ &= \Pr\left(\mu_{j-1} < \beta_0 + \beta_1 X_i + u_i \le \mu_j\right) \\ &= \Pr\left(u_i \le \mu_j - \beta_0 - \beta_1 X_i\right) - \Pr\left(u_i \le \mu_{j-1} - \beta_0 - \beta_1 X_i\right) \\ &= F\left(\mu_j - \beta_0 - \beta_1 X_i\right) - F\left(\mu_{j-1} - \beta_0 - \beta_1 X_i\right), \end{aligned}$$

We can either assume that u_i is normally distributed, or has a logistic distribution.

Suppose we have n observations of Y and X, where Y takes the value $1, 2, ..., J$. The probability of getting such observations is

$$L = \Pr(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) = \Pr(Y_1 = y_1)\Pr(Y_2 = y_2)...\Pr(Y_n = y_n)$$

by the independence of u_i.

The likelihood function is

$$\begin{aligned} L &= \prod_{y_i=1} \Pr(Y_i = 1) \prod_{y_i=2} \Pr(Y_i = 2) ... \prod_{y_i=J} \Pr(Y_i = J). \\ &= \prod_{i=1}^{n}\prod_{j=1}^{J} \left[F\left(\mu_j - \beta_0 - \beta_1 X_i\right) - F\left(\mu_{j-1} - \beta_0 - \beta_1 X_i\right)\right]^{d_j} \end{aligned}$$

where $d_j = 1$ if $Y_i = j$ and $d_j = 0$ otherwise.

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{J} d_j \ln \left\{ \left[F\left(\mu_j - \beta_0 - \beta_1 X_i\right) - F\left(\mu_{j-1} - \beta_0 - \beta_1 X_i\right) \right] \right\}.$$

Example 9.3: Suppose there are only 3 ordered categories, then

$$\begin{aligned}
\Pr(Y_i = 1) &= F(\mu_1 - \beta_0 - \beta_1 X_i), \\
\Pr(Y_i = 2) &= F(\mu_2 - \beta_0 - \beta_1 X_i) - F(\mu_1 - \beta_0 - \beta_1 X_i), \\
\Pr(Y_i = 3) &= 1 - F(\mu_2 - \beta_0 - \beta_1 X_i).
\end{aligned}$$

Consider the case where $\beta_1 > 0$. For the three probabilities, the marginal effects of changes in the regressors are

$$\begin{aligned}
\frac{\partial \Pr(Y_i = 1)}{\partial X_i} &= -f(\mu_1 - \beta_0 - \beta_1 X_i)\,\beta_1 < 0, \\
\frac{\partial \Pr(Y_i = 2)}{\partial X_i} &= [f(\mu_2 - \beta_0 - \beta_1 X_i) - f(\mu_1 - \beta_0 - \beta_1 X_i)]\,\beta_1 =?, \\
\frac{\partial \Pr(Y_i = 3)}{\partial X_i} &= f(\mu_2 - \beta_0 - \beta_1 X_i)\,\beta_1 > 0.
\end{aligned}$$

Thus, in the general case, given the signs of the coefficients, only the signs of the changes in $\Pr(Y_i = 1)$ and $\Pr(Y_i = J)$ are unambiguous. What happens to the middle cell is unknown.

9.9 Truncation of data

Sometimes we cannot perfectly observe the actual value of the dependent variable. If we only observe a subpopulation such as individuals with income above a certain level, then the data is said to be lower-truncated, in the sense that we cannot observe people with income below that level in the sample.

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 $Y > a$.

$$\Pr(Y < c|Y > a) = \frac{\Pr(Y < c \cap Y > a)}{P(Y > a)} = \frac{\int_a^c f(Y)\,dY}{\int_a^\infty f(Y)\,dY},$$

$$f(Y = c|Y > a) = \frac{d\Pr(Y < c|Y > a)}{dc} = \frac{f(c)}{\int_a^\infty f(Y)\,dY}.$$

Example 9.4: Suppose Y is uniformly distributed in the $[0, 1]$ interval. Since $f(Y) = 1$ and $F(Y) = Y$, it is easy to find the unconditional probability $\Pr(Y > 3/4) = 1/4$. Suppose now we know that Y must be greater than $1/2$, how will this affect our prediction for $\Pr(Y > 3/4)$?

Solution: Using the above rule

$$\Pr\left(Y > \frac{3}{4}\middle| Y > \frac{1}{2}\right) = \frac{\Pr\left(Y > \frac{3}{4} \cap Y > \frac{1}{2}\right)}{\Pr\left(Y > \frac{1}{2}\right)} = \frac{\Pr\left(Y > \frac{3}{4}\right)}{\Pr\left(Y > \frac{1}{2}\right)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

9.9.1 Moments of Truncated Distributions

Note that $E(Y)$ is a weighted average of $E(Y|Y < a)$ and $E(Y|Y > a)$ since

$$\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} Yf(Y)\,dY \\
&= \int_{-\infty}^{a} Yf(Y)\,dY + \int_{a}^{\infty} Yf(Y)\,dY \\
&= \int_{-\infty}^{a} Y\frac{f(Y)}{\Pr(Y<a)}dY\Pr(Y<a) + \int_{a}^{\infty} Y\frac{f(Y)}{\Pr(Y>a)}dY\Pr(Y>a) \\
&= \int_{-\infty}^{a} Yf(Y|Y<a)\,dY\Pr(Y<a) + \int_{a}^{\infty} Yf(Y|Y>a)\,dY\Pr(Y>a) \\
&= E(Y|Y<a)\Pr(Y<a) + E(Y|Y>a)\Pr(Y>a).
\end{aligned}$$

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

$$\begin{aligned}
E(Y|Y\geq a) &= \int_{a}^{\infty} Yf(Y|Y\geq a)\,dY \geq E(Y), \\
E(Y|Y<a) &= \int_{-\infty}^{a} Yf(Y|Y<a)\,dY \leq E(Y).
\end{aligned}$$

If the truncation is from below, the mean of the truncated variable is greater than the mean of the original one. If the truncation is from above, the mean of the truncated variable is smaller than the mean of the original one.

Example 9.5: Find $E(u|u>1)$ and $Var(u|u>1)$ if $f(u)=\exp(-u)$, $u>0$, and compare them to their unconditional mean and variance.

Solution:

$$\begin{aligned}
E(u \mid u>1) &= \int_1^\infty uf(u \mid u>1)\,du \\
&= \frac{1}{1-F(1)}\int_1^\infty uf(u)\,du \\
&= \frac{1}{e^{-1}}\int_1^\infty u\exp(-u)\,du \\
&= \frac{1}{e^{-1}}\left\{[-u\exp(-u)]_1^\infty + \int_1^\infty \exp(-u)\,du\right\} \\
&= \frac{e^{-1}}{e^{-1}} + \frac{1-F(1)}{1-F(1)} \\
&= 2 > E(u) = 1.
\end{aligned}$$

$$\begin{aligned}
Var(u \mid u>1) &= E\left(u^2 \mid u>1\right) - [E(u \mid u>1)]^2 \\
&= \int_1^\infty u^2 f(u \mid u>1)\,du - 4 \\
&= \frac{1}{1-F(1)}\int_1^\infty u^2 f(u)\,du - 4 \\
&= e\int_1^\infty u^2 f(u)\,du - 4 \\
&= e\int_1^\infty u^2\exp(-u)\,du - 4 \\
&= e\left[\left[-u^2\exp(-u)\right]_1^\infty + 2\int_1^\infty u\exp(-u)\,du\right] - 4 \\
&= e\left[e^{-1} + 2\times 2e^{-1}\right] - 4 \\
&= 1 = Var(u)\,. \blacksquare
\end{aligned}$$

9.9.2 Maximum Likelihood Estimation of the Truncated Model

Consider the simple model

$$Y_i = \beta_0 + \beta_1 X_i + u_i > a.$$

$$\Pr\left(Y_i > a\right) = \Pr\left(\beta_0 + \beta_1 X_i + u_i > a\right) = \Pr\left(u_i > a - \beta_0 - \beta_1 X_i\right) = 1 - F\left(a - \beta_0 - \beta_1 X_i\right).$$

The Likelihood function is

$$\begin{aligned} L &= f\left(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n | Y_1 > a, Y_2 > a, ..., Y_n > a\right) \\ &= f\left(y_1 - \beta_0 - \beta_1 X_1 | Y_1 > a\right) f\left(y_2 - \beta_0 - \beta_1 X_2 | Y_2 > a\right) ... f\left(y_n - \beta_0 - \beta_1 X_n | Y_n > a\right) \end{aligned}$$

The Log-likelihood function is

$$\begin{aligned} \ln L &= \ln\left[f\left(y_1 - \beta_0 - \beta_1 X_1 | Y_1 > a\right) f\left(y_2 - \beta_0 - \beta_1 X_2 | Y_2 > a\right) ... f\left(y_n - \beta_0 - \beta_1 X_n | Y_n > \right.\right. \\ &= \sum_{i=1}^{n} \ln f\left(y_i - \beta_0 - \beta_1 X_i | Y_i > a\right) = \sum_{i=1}^{n} \ln \frac{f\left(y_i - \beta_0 - \beta_1 X_i\right)}{\Pr\left(Y_i > a\right)} \\ &= \sum_{i=1}^{n} \ln f\left(y_i - \beta_0 - \beta_1 X_i\right) - \sum_{i=1}^{n} \ln\left[1 - F\left(a - \beta_0 - \beta_1 X_i\right)\right]. \end{aligned}$$

First order conditions:

$$\frac{\partial \ln L}{\partial \beta_0} = -\sum_{i=1}^{n} \frac{f'\left(y_i - \beta_0 - \beta_1 X_i\right)}{f\left(y_i - \beta_0 - \beta_1 X_i\right)} - \sum_{i=1}^{n} \frac{f\left(a - \beta_0 - \beta_1 X_i\right)}{1 - F\left(a - \beta_0 - \beta_1 X_i\right)} = 0,$$

$$\frac{\partial \ln L}{\partial \beta_1} = -\sum_{i=1}^{n} X_i \frac{f'\left(y_i - \beta_0 - \beta_1 X_i\right)}{f\left(y_i - \beta_0 - \beta_1 X_i\right)} - \sum_{i=1}^{n} X_i \frac{f\left(a - \beta_0 - \beta_1 X_i\right)}{1 - F\left(a - \beta_0 - \beta_1 X_i\right)} = 0.$$

Exercise 9.12: Consider the truncated model

$$Y_i = \beta_0 + \beta_1 X_i + u_i > a,$$

where u_i are i.i.d. with density function and distribution function

$$f\left(u_i\right) = \exp\left(-u_i\right)$$

and

$$F(u_i) = 1 - \exp(-u_i)$$

respectively.

(a) Show that $\Pr(Y_i > a) = \exp(\beta_0 + \beta_1 X_i - a)$.

(b) Suppose we have n observations of Y and X, find the log-likelihood function.

(c) Find $\frac{\partial \ln L}{\partial \beta_0}$ and $\frac{\partial \ln L}{\partial \beta_1}$. Discuss the identifiability of β_0 and β_1.

Exercise 9.13: Find $E(u|u > 1)$ and $Var(u|u > 1)$ if $u \sim N(0,1)$, and compare them to their unconditional mean and variance.

9.10 Maximum Likelihood Estimation of the Tobit Model

Sometimes data are **censored** rather than truncated. When the dependent variable is censored, values in a certain range are all reported as a single value. Suppose we are interested in the accommodation demand for a certain hotel. If the demand is higher than the hotel's capacity, we will never know the value of actual demand, and the over-demand values are reported as the maximum capacity of this hotel. We may also observe people either work for a certain hour or not work at all. If people do not work at all, their optimal working hour may be negative. However, we will never observe a negative working hour, we observe zero working hour instead. Suppose the data is lower-censored at zero.

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 X_i + u_i, \\ Y_i &= 0 \qquad \text{if } Y_i^* \leq 0, \\ Y_i &= Y_i^* \qquad \text{if } Y_i{}^* > 0. \end{aligned}$$

Y_i^* is not observable, we can only observe Y_i and X_i. To fully utilize the information, if the observation is not censored, we calculate the density value at that point of observation $f(Y_i - \beta_0 - \beta_1 X_i)$. If the observation is censored, we use the probability of observing a censored value $\Pr(Y_i = 0)$. Note that:

$$\begin{aligned}\Pr(Y_i = 0) &= \Pr(\beta_0 + \beta_1 X_i + u_i \leq 0) \\ &= \Pr(u_i \leq -\beta_0 - \beta_1 X_i) \\ &= 1 - F(\beta_0 + \beta_1 X_i).\end{aligned}$$

The likelihood function is

$$L = \prod_{Y_i > 0} f(Y_i - \beta_0 - \beta_1 X_i) \prod_{Y_i = 0} \Pr(Y_i = 0).$$

The log-likelihood function is

$$\begin{aligned}\ln L &= \ln \left[\prod_{Y_i > 0} f(Y_i - \beta_0 - \beta_1 X_i) \prod_{Y_i = 0} \Pr(Y_i = 0) \right] \\ &= \sum_{Y_i > 0} \ln f(Y_i - \beta_0 - \beta_1 X_i) + \sum_{Y_i = 0} \ln \left[1 - F(\beta_0 + \beta_1 X_i) \right].\end{aligned}$$

First-order condition:

$$\frac{\partial \ln L}{\partial \beta_0} = -\sum_{Y_i > 0} \frac{f'(Y_i - \beta_0 - \beta_1 X_i)}{f(Y_i - \beta_0 - \beta_1 X_i)} - \sum_{Y_i = 0} \frac{f(\beta_0 + \beta_1 X_i)}{1 - F(\beta_0 + \beta_1 X_i)} = 0,$$

$$\frac{\partial \ln L}{\partial \beta_1} = -\sum_{Y_i > 0} X_i \frac{f'(Y_i - \beta_0 - \beta_1 X_i)}{f(Y_i - \beta_0 - \beta_1 X_i)} - \sum_{Y_i = 0} X_i \frac{f(\beta_0 + \beta_1 X_i)}{1 - F(\beta_0 + \beta_1 X_i)} = 0.$$

If $u_i \sim N(0, \sigma^2)$, and let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution functions of an $N(0, 1)$ respectively.

$$f(Y_i - \beta_0 - \beta_1 X_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right) = \frac{1}{\sigma} \phi\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma} \right).$$

$$f'\left(Y_i - \beta_0 - \beta_1 X_i\right) = \frac{1}{\sigma^2}\phi'\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right).$$

$$f\left(\beta_0 + \beta_1 X_i\right) = \frac{1}{\sigma}\phi\left(\frac{\beta_0 + \beta_1 X_i}{\sigma}\right).$$

$$F\left(\beta_0 + \beta_1 X_i\right) = \Phi\left(\frac{\beta_0 + \beta_1 X_i}{\sigma}\right).$$

Then the log-likelihood can be rewritten as

$$\ln L = \sum_{Y_i > 0} \ln \frac{1}{\sigma}\phi\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right) + \sum_{Y_i = 0} \ln\left[1 - \Phi\left(\frac{\beta_0 + \beta_1 X_i}{\sigma}\right)\right].$$

Example 9.6: Consider the model $Y_i = \beta_0 + \beta_1 X_i + u_i$. 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

$$\begin{aligned} L &= \prod_{Y_i > c_2} f\left(Y_i - \beta_0 - \beta_1 X_i \mid Y_i < c_1\right) \prod_{Y_i = c_2} \Pr\left(Y_i = c_2 \mid Y_i < c_1\right) \\ &= \prod_{Y_i > c_2} \frac{f\left(Y_i - \beta_0 - \beta_1 X_i\right)}{\Pr\left(Y_i < c_1\right)} \prod_{Y_i = c_2} \frac{\Pr\left(Y_i = c_2\right)}{\Pr\left(Y_i < c_1\right)}. \end{aligned}$$

where

$$\begin{aligned} \Pr\left(Y_i = c_2\right) &= \Pr\left(\beta_0 + \beta_1 X_i + u_i < c_2\right) \\ &= \Pr\left(u_i < c_2 - \beta_0 - \beta_1 X_i\right) \\ &= F\left(c_2 - \beta_0 - \beta_1 X_i\right) \\ \text{and } \Pr\left(Y_i < c_1\right) &= \Pr\left(\beta_0 + \beta_1 X_i + u_i < c_1\right) \\ &= F\left(c_1 - \beta_0 - \beta_1 X_i\right). \end{aligned}$$

The log-likelihood function is given by

$$\begin{aligned}\ln L &= \sum_{Y_i>c_2} \ln \frac{f\left(Y_i-\beta_0-\beta_1 X_i\right)}{\Pr\left(Y_i<c_1\right)}+\sum_{Y_i=c_2} \ln \frac{\Pr\left(Y_i=c_2\right)}{\Pr\left(Y_i<c_1\right)} \\ &= \sum_{Y_i>c_2} \ln \frac{f\left(Y_i-\beta_0-\beta_1 X_i\right)}{F\left(c_1-\beta_0-\beta_1 X_i\right)}+\sum_{Y_i=c_2} \ln \frac{F\left(c_2-\beta_0-\beta_1 X_i\right)}{F\left(c_1-\beta_0-\beta_1 X_i\right)}. \blacksquare\end{aligned}$$

Exercise 9.14: True/False. Let X be a random variable, and c be a constant, then

(a) $Var\left(X\right) \geq Var\left(X|X=c\right)$.

(b) $Var(X|X<c)<Var(X)$.

Exercise 9.15: True/False/Uncertain.

(a) If we only observe a subpopulation such as individuals with income above a certain level, then the data is said to be lower-truncated.

(b) If we only observe a subpopulation, such as individuals with income above a certain level, then the data are said to be lower-censored.

(c) When the dependent variable is censored, values in a certain range are all reported as single value.

(d) When the dependent variable is truncated, values in a certain range are all reported as a single value.

(e) If X is a random variable which has an extreme value distribution with density $f\left(x\right)=\exp\left(-x\right)\exp\left(-\exp\left(-x\right)\right)$ for $-\infty<x<\infty$. Let $Y=\exp\left(-X\right)$, then $E\left(Y\right)=1$.

(f). An extreme value distribution has the distribution function $F\left(x\right)=1-\exp\left(-\exp\left(-x\right)\right)$ for $-\infty<x<\infty$.

(g). For a random variable X, we can have $E\left(X|X\leq 0\right)>0$.