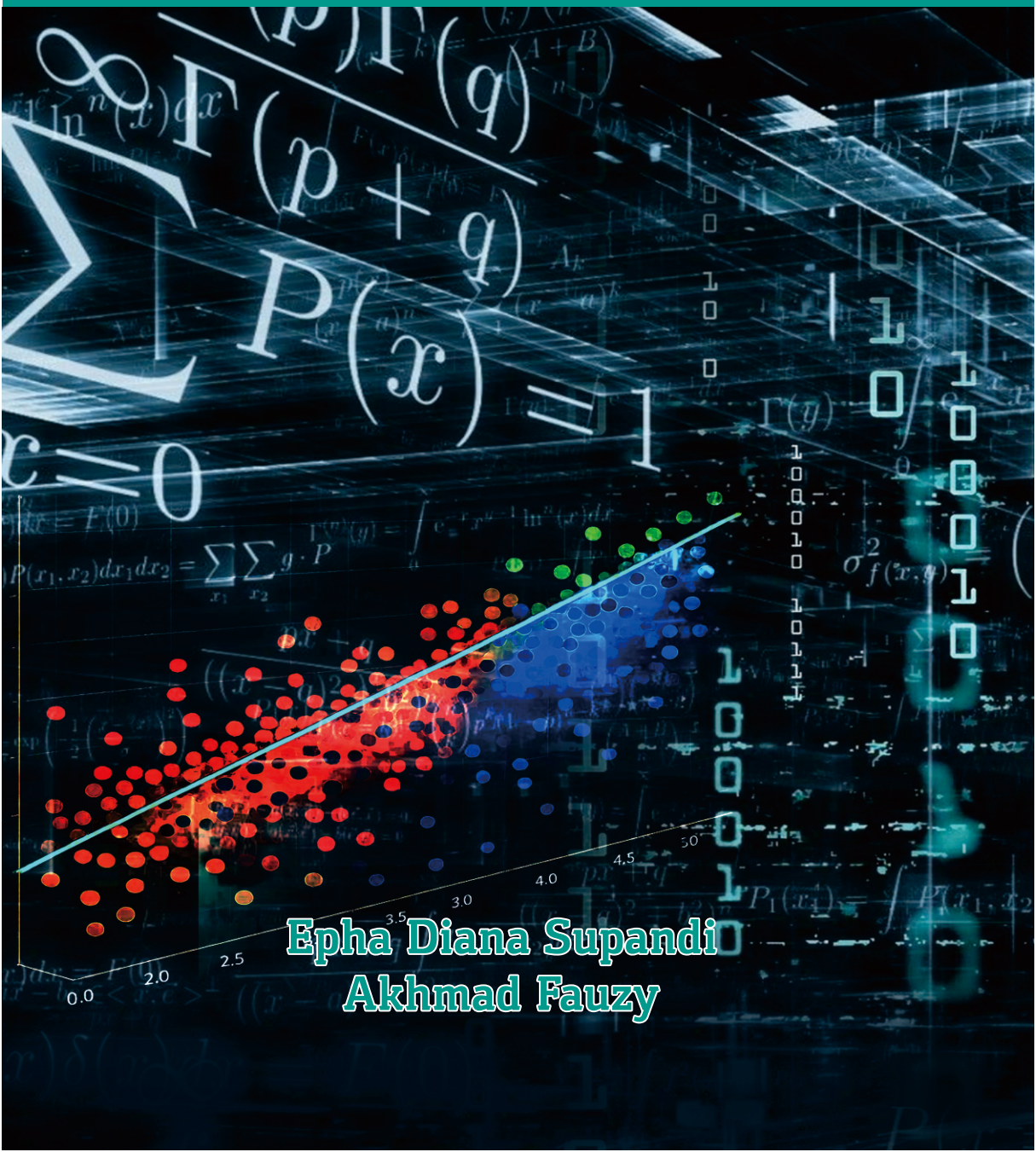
ANALISIS MULTIVARIAT DENGAN R

Konsep dan Implementasi



Epha Diana Supandi
Akhmad Fauzy

ANALISIS MULTIVARIAT DENGAN R

Konsep dan Implementasi

Epha Diana Supandi lahir di Tasikmalaya pada 12 September 1975. Ia menempuh pendidikan S1 Statistika di Institut Pertanian Bogor, kemudian melanjutkan studi Magister di Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, serta meraih gelar Doktor bidang Matematika dengan minat Statistika dari FMIPA Universitas Gadjah Mada pada tahun 2012. Sejak 2008 ia menjadi dosen tetap di Program Studi Matematika, Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta, dengan bidang pengajaran antara lain metode statistika, analisis data, dan analisis multivariat. Selain aktif dalam pengajaran dan penelitian, ia juga terlibat dalam pengembangan sistem penjaminan mutu akademik serta berpengalaman sebagai asesor BAN-PT, LAMDIK, dan LAMSAMA.

Akhmad Fauzy lahir di Tegal, 8 Juli 1970. Ia menempuh pendidikan Statistika di Universitas Gadjah Mada (S1), Institut Pertanian Bogor (S2), dan Universiti Putra Malaysia (S3). Sejak 1995, ia menjadi dosen tetap Program Studi Statistika FMIPA Universitas Islam Indonesia (UII) Yogyakarta dengan bidang pengajaran meliputi statistika nonparametrik, metode sampling, statistika industri, dan metodologi penelitian. Selain aktif mengajar dan meneliti, ia juga produktif menulis buku-buku di bidang statistika dan metode penelitian, di antaranya *Statistik Industri, Metode Sampling*, dan *Pengantar Statistika Nonparametrik*. Ia pernah menjabat sebagai Dekan FMIPA UII, Direktur DPPM UII, serta anggota Dewan Eksekutif BAN-PT.

ANALISIS MULTIVARIAT DENGAN R

Konsep dan Implementasi

Dr. Epha Diana Supandi, S.Si., M.Sc.
Prof. Akhmad Fauzy, S.Si., M.Si., Ph.D.

SUKA PRESS
Sunan Kalijaga State Islamic University

Analisis Multivariat dengan R
Konsep dan Implementasi

Cetakan Pertama, Mei 2026
xii + 324 Halaman; 15,5 x 23 cm
ISBN: 978-634-7179-10-4

Penulis	: Epha Diana Supandi Akhmad Fauzy
Editor	: Bernando J. Sujibto
Layout	: Effendi Chairi
Desain Sampul	: Mohammad Ali Tsabit

Diterbitkan oleh

SUKA PRESS
Sunan Kalijaga State Islamic University

Suka Press (Anggota IKAPI)
Jl. Laksda Adisucipto, Depok, Sleman,
Daerah Istimewa Yogyakarta 55281
Surel: suka.press@uin-suka.ac.id
Website: https://sukapress.uin-suka.ac.id
WhatsApp: +62 823-2675-5046
Instagram: sukapress
Facebook: Suka Press

Kata Pengantar

Alhamdulillah, segala puji syukur penulis panjatkan kepada Allah SWT yang telah melimpahkan rahmat dan hidayahNya sehingga buku yang berjudul *Analisis Multivariat dengan R: Konsep dan Implementasi* dapat diselesaikan.

Analisis multivariat merupakan salah satu teknik statistik yang digunakan untuk memahami struktur data dalam dimensi tinggi. Analisis multivariat memiliki kemampuan untuk memahami hubungan kompleks antar banyak variabel secara bersamaan. Kemampuan ini sangat relevan dalam dunia nyata karena masalah yang dihadapi sering kali melibatkan lebih dari dua variabel dan saling berhubungan. Oleh karena itu, mempelajari analisis multivariat adalah kunci untuk mengelola dan menafsirkan data yang kompleks dan membuat keputusan yang cerdas, menemukan peluang, serta mengatasi masalah di berbagai sektor

Dalam analisis multivariat diperlukan suatu pengolah data yang *powerfull* karena sering melibatkan banyak variabel dan hubungan yang kompleks. Salah satu *software* statistik yang banyak digunakan saat ini adalah *software* R.

Software R adalah salah satu *software open source* yang mampu menyelesaikan dengan mudah berbagai teknik analisis data mulai dari sistem grafik yang menarik (*visualitation data*), metode statistik klasik, metode statistik modern sampai perkembangan sains data terkini.

Buku ini memberikan pengetahuan berbagai metode dalam analisis multivariat baik secara konsep dan implementasinya pada berbagai kasus dengan menggunakan bantuan *software* R. Oleh karena itu, untuk memahami buku ini, pengetahuan dasar *software* R sangat diperlukan. Pengenalan dasar *software* R dijelaskan pada Bab I, namun disarankan untuk membaca buku

pengantar *software* R yang lebih lengkap lainnya.

Pada Bab 2 berisi pengantar analisis multivariat beserta struktur data, sementara Bab 3 memuat distribusi normal multivariat dimana distribusi ini sering dijadikan asumsi dalam berbagai teknik analisis multivariat. Selanjutnya inferensi vektor rata-rata dan *Multivariate analisis of variance* (MANOVA) diuraikan pada Bab 4 dan Bab 5. Analisis regresi Multivariat, Analisis Korelasi Kanonik, Analisis Diskriminan dapat dipelajari pada Bab 6 sampai Bab 8. Teknik pengelompokan ada pada Bab 9, sementara Bab 10 yang berisi Analisis Klaster dan Penskalaan Mutltidimensi (*Multidimensional Scalling*). Adapun Analisis Komponen Utama, Analisis Faktor, Analisis Jalur dan Model Persamaan Struktural dapat dipelajari pada Bab 11 hingga Bab 14.

Penulis menyadari bahwa buku ini belum sempurna, oleh karena itu saran dan kritik yang membangun sangat diharapkan untuk perbaikan buku ini di masa yang akan datang. Kritik dan saran yang membangun dapat dikirim ke email: epha.supandi@uin-suka.ac.id. Semoga buku ini dapat memberikan manfaat, khususnya bagi perkembangan ilmu statistik di Indonesia dan secara umum bagi para pembaca.

Sleman, Maret 2026

Penulis

Daftar Isi

BAB

I

PENGENALAN DASAR SOFTWARE R

Software R termasuk ke dalam kelompok *software* statistik yang *open source,* artinya *software* ini tidak memerlukan lisensi atau gratis. R dapat diperoleh secara gratis di alamat http ://cran.r-project.org. R sebenarnya bukan bahasa pemrograman yang baru, versi awal dibuat oleh Ross Ihaka dan Robe Gentleman pada tahun 1992 di Universitas Auckland, New Zealand.

Dengan berkembangnya era *data analysis* atau dikenal dengan era big data, maka *software* R ini semakin berkembang bukan hanya untuk analisis statistik saja, tetapi digunakan untuk berbagai kebutuhan analisis bidang lainnya.

Tidak seperti *software* SPSS, Minitab atau *software* lain yang berbasis *graphical user interface* (GUI), pada *software* R biasanya digunakan dengan mengetikkan perintah secara interaktif pada jendela konsol (*R-console*). Namun demikian, R juga memiliki antarmuka grafis dengan menginstal pustaka (*library*) tambahan.

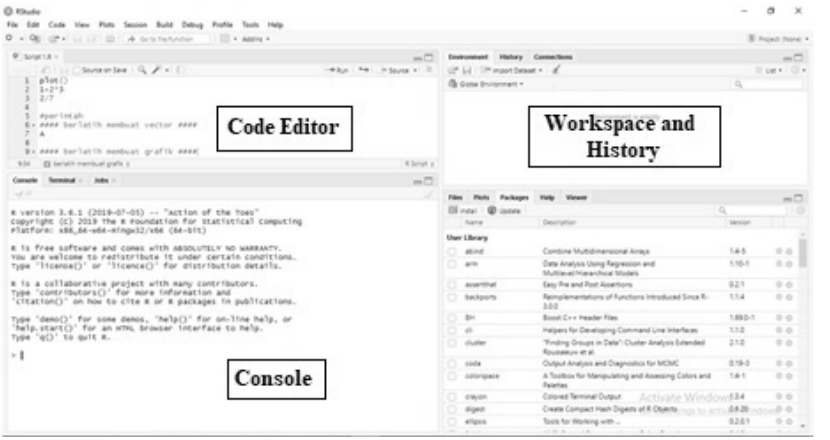
Pada dasarnya perintah dasar dalam R berbasis teks atau command line sehingga pengguna harus mengetikkan perintah-perintah tertentu dan harus hafal perintah-perintahnya. Selain itu, R juga menyediakan *package* (paket) yang berisi kumpulan perintah atau fungsi yang dapat digunakan untuk melakukan analisis tertentu. Paket dasar yang tersedia diantaranya **base, MASS, stats** dan lain lain.

1.1 Memulai R

Untuk memulai menggunakan R, langkah pertama adalah menginstal R terlebih dahulu. Instalasi R dalam sistem operasi Windows adalah mengunduh file **R-4.2.2-win.exe** (versi terbaru per Februari 2023) yang dapat diperoleh di *http ://cran.rproject.org.*

Proses instalasi R, dapat dilakukan dengan cara klik dua kali (*double click*) pada file **R-4.2.2-win.exe** yang terdapat pada direktori yang telah disediakan. Lanjutkan jalannya proses instalasi dengan mengikuti *Wizard* dan menggunakan pilihan-pilihan *default* instalasi. Setelah R terinstal maka *icon* dan *menu* R akan muncul pada program *list*. Jika instalasi berlangsung dengan baik, maka jendela program R akan terbuka seperti yang terlihat pada

Gambar 1.



Gambar 1. Tampilan Jendela pada R Studio

Pada gambar 1. muncul 4 (empat) jendela yaitu :

1. **Code Editor**: jendela yang berisi editor teks yang memungkinkan untuk bekerja dengan file *script*. Pada jendela ini dapat memasukkan beberapa baris kode, kemudian *script* disimpan dan melakukan tugas-tugas pada *script* lainnya.
2. **Console**: jendela untuk melakukan semua pekerjaan interaktif dengan R.
3. **Workspace and History**: jendela ini untuk melihat semua objek, perintah/kode serta nilai yang telah dihasilkan dan disimpan dalam *worksheet*. Bagian ini sangat membantu untuk mengeksplore *workspace* dan seluruh isinya.
3. **File, plot,** package **dan help**: berada di jendela sudut kanan bawah yang berfungsi untuk :
    a. **File:** menelusuri *folder* dan *file*;
    b. **Plot:** menampilkan plot (diagram atau grafik)
    c. Package: melihat daftar *package* yang telah diinstal
    d. **Help**: menelusuri sistem *help* di R

R. merupakan *object oriented programming*, dimana untuk memasukkan satu atau lebih nilai kepada sebuah objek dimulai dengan perintah "<-" atau "=". Contoh seperti di bawah ini.

```
> X = 10                    #objek X bernilai 10
> X #perintah mencetak X
[1] 10
> A = c(1,2,3,4,5)          # objek A adalah vektor
> print(A)                  #perintah untuk mencetak A
[1] 1 2 3 4 5
```

Setiap mulai bekerja dengan R, sebaiknya menggunakan direktori yang berbeda untuk setiap tugas yang dikerjakan. Kegiatan ini bertujuan untuk memudahkan dalam melihat *history* dan objek yang digunakan dalam tugas tersebut.

Langkah-langkah membuat direktori khusus dengan menggunakan R:

1. Buatlah direktori baru misalnya C :/R_Kerja, selanjutnya buat sub direktori dengan nama Praktikum, sehingga diperoleh direktori untuk menyimpan file kerja praktikum, yaitu C :/R_Kerja/Praktikum
2. Ketikkan pada jendela konsol perintah

```
> setwd("C :/R_Kerja/Praktikum") #perintah
untuk mengubah direktori kerja
```

3. Untuk mengetahui pada direktori mana R bekerja, maka ketikkan perintah

```
> getwd() #perintah untuk mengetahui direktori
kerja [1] "C :/R_Kerja/Praktikum"
```

Agar dapat menggunakan R dengan secara lebih baik, pengetahuan mengenai beberapa perintah perlu diketahui. Ada beberapa cara yang digunakan untuk mencari bantuan atau *help* terhadap suatu perintah atau fungsi di R, diantaranya adalah:

```
help(nama_fungsi) atau ? nama_fungsi
```

Setiap perintah help(nama_fungsi) atau ?(nama_fungsi) dari suatu perintah akan memuat keterangan-keterangan berikut ini.

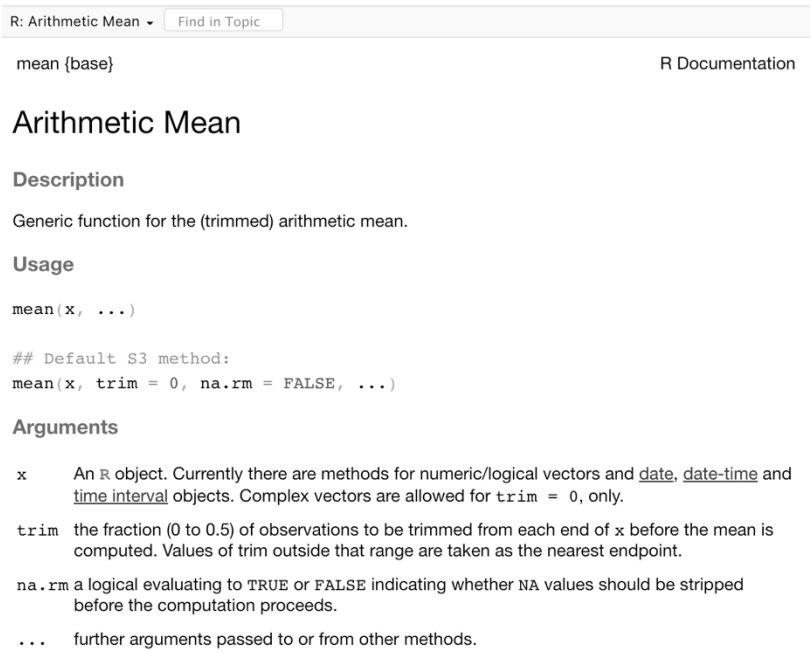
- *Title*: nama perintah.

- `Description`**:** deskripsi singkat tentang perintah.
- `Usage`: menampilkan sintaks perintah untuk penggunaan perintah tersebut.
- `Arguments`: keterangan mengenai *argument/input* yang diperlukan pada perintah tersebut.
- `Details`: keterangan lebih lengkap lengkap tentang perintah tersebut.
- `Value`: keterangan tentang *output* suatu perintah dapat diperoleh pada bagian ini.
- `Author(s)`: memberikan keterangan tentang *Author* dari perintah tersebut.
- `References`: seringkali referensi yang dapat digunakan untuk memperoleh keterangan lebih lanjut terhadap suatu perintah ditampilkan pada bagian ini.
- `See also`: bagian ini berisikan daftar perintah/fungsi yang berhubungan erat dengan perintah tersebut.
- `Example`: berisikan contoh-contoh penggunaan perintah tersebut.

Misalkan ingin mengetahui bagaimana cara menuliskan perintah untuk menghitung rata-rata suatu vektor. Caranya ketik pada *command line* di jendela konsol perintah berikut ini.

```
> help(mean)
atau
> ?mean
```

Setelah salah satu dari perintah tersebut dijalankan maka pada jendela *help* akan muncul keterangan sebagai berikut.



Gambar 2. Tampilan dari fungsi help(means) atau ?means

Sofware R berisi perintah-perintah dasar baik untuk memanipulasi maupun menganalisis statistik. Salah satu kelebihan R adalah tersedianya fungsi-fungsi lain dalam bentuk *bundle* yang disebut "*package*". R *package* berisi tidak hanya sekumpulan fungsi-fungsi, namun dapat berisi sekumpulan data.

R *package* ini digunakan untuk mengimplementasikan tidak hanya metode statistik tingkat lanjut, visualisasi, analisis media sosial tetapi digunakan juga sampai pada proses *parallel computing*. Hal ini menjadi kelebihan dari R dan menyebabkan R menjadi *software* yang berkembang sangat pesat dan banyak digunakan sampai saat ini.

R *package* dapat dibuat oleh siapa saja baik akademisi maupun praktisi. Setiap *package* memiliki tujuan yang berbeda-beda, daftar seluruh R *package* dapat dilihat pada laman: https ://cran.r-project.org/web/*package*s/available_*package*s_by_name.html

1.2 Operasi Perhitungan

Di dalam R banyak sekali perintah atau fungsi yang berkaitan dengan operasi-operasi perhitungan matematika. Beberapa fungsi yang penting dan sering digunakan disajikan berikut ini (disertai dengan # komentar).

```
> 10 + 25          #penjumlahan
[1] 35
> 10 * 25          #perkalian
[1] 250
> 10 - 25          #pengurangan
[1] -10
> 10/25            #pembagian
[1] 0.4
> log(10)          #logaritma natural
 [1] 2.302585
> log10(10)        #logaritma basis 10
[1] 1
> sqrt(10)         #akar 10
[1] 3.16227 8
> 10^2             #10 dikuadratkan
[1] 100
> rep(2,5)         #mengulang 2 sebanyak 5 kali
[1] 2 2 2 2 2
> rep(5,2)         #mengulang 5 sebanyak 2 kali
[1] 5 5
> seq(1,10)  #membuat barisan bilangan dari 1 s.d 10
[1] 1 2 3 4 5 6 7 8 9 10
> seq(1,10,by=2)   #membuat barisan bilangan dari 1
s.d 10 dengan kenaikan 2
[1] 1 3 5 7 9
```

Tentu saja ada ratusan fungsi R yang lain. Namun, kita hanya menggunakan yang perlu dan relevan dengan fokus kajian saja.

1.3 Penamaan Variabel

Dalam R pemberian nama variabel bersifat sensitive, artinya huruf kecil dan huruf besar dibedakan. Contoh nama variabel ipk, IPK dan Ipk adalah

berbeda. Setiap variabel dapat ditugaskan dengan suatu nilai atau variabel lainnya dengan menggunakan simbol "<-" atau " =". Misalkan :

```
> ipk <- 3.5 #ipk diberi nilai 20
> ipk              #print ipk
[1] 3.5
> ipk = 2.75 #ipk diberi nilai 2.75
> ipk              #print ipk
[1] 2.75
```

Apabila diketik seperti ini, maka akan muncul keterangan sebagai berikut.

```
> IPK              #print IPK
error: object 'IPK' not found
> Ipk              #print Ipk
error: object 'Ipk' not found
```

Pada *output* di atas ditulis IPK dan Ipk maka akan terjadi *error*, karena pendefinisian menggunakan nama ipk menggunakan huruf kecil semua. Oleh karena itu, ketika diketik tidak sesuai dengan pendefinisian awal maka akan terjadi *error*.

1.4 Data Vektor

Dalam melakukan analisis data biasanya dalam bentuk vektor atau matriks. Pada bagian ini akan dijelaskan hal-hal yang terkait dengan vektor dan matriks.

Vektor merupakan suatu *array* atau himpunan bilangan, *character* atau *string*, *logical value*, dan merupakan objek paling dasar yang dikenal dalam R. Vektor dibuat dengan menggunakan fungsi *concatenate* c(), seperti yang pada contoh berikut ini.

```
> X = c(1 :10)
> X                 #print X
[1] 1 1 2 3 4 5 6 7 8 9 10
> Y = c("Ani", "Siti", "Ahmad", "Soni")
> Y                 #print Y
[1] "Ani", "Siti", "Ahmad", "Soni"
> length(X)         #menampilkan panjang vektor X
[1] 10
> length(Y)         #mengetahui panjang vektor Y
[1] 4
```

Ekstraksi sebagian data vektor dapat dilakukan dengan berbagai cara. Berikut ini adalah beberapa contoh hasil ekstraksi dari suatu data vektor.

```
> Y[2]              #menampilkan elemen ke-2 dari
vektor Y
[1] "Siti"
> X[X>6]            #menampilkan elemen pada vektor X
yang lebih besar dari 6
[1] 7 8 9 10
> X[-c(1,8)]#menampilkan semua elemen vektor X
kecuali elemen 1 dan 8
[1] 2 3 4 5 6 7 9 10
```

1.5 Data Matriks

Matriks adalah salah satu tipe data yang banyak digunakan dalam pemrograman statistik. Sebagian besar fungsi-fungsi yang ada dalam R banyak menggunakan data dalam bentuk matriks.

Cara pembuatan matriks dapat menggunakan fungsi matrix(), *argument* yang digunakan dalam fungsi ini adalah :

matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE, dimnames = NULL)

Misalkan: $A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$

Cara membuat matriks A dengan menggunakan R adalah :

```
> A = matrix(c(1 :9), nrow=3, ncol=3)
> A #print matriks A
 [,1] [,2] [,3]
 [1,] 1 4 7
 [2,] 2 5 8
 [3,] 3 6 9
```

Untuk mengetahui dimensi dari suatu matriks, kita dapat menggunakan fungsi dim().

```
> dim(A) #menampilkan dimensi matriks A
[1] 3 3
```

Cara untuk menyeleksi baris atau kolom dalam dimensi dapat dilakukan sama seperti dalam vektor. Berikut adalah contoh penerapannya.

```
> A[2,] #menampilkan baris ke-2 dari matriks A
[1] 2 5 8
> A[ ,3] #menampilkan kolom ke-3 dari matriks A
[1] 7 8 9
> A[2,3] #menampilkan elemen pada baris ke-2 dan kolom
ke-3 dari matriks A
[1] 8
```

Operasi matematika pada matriks dapat menjadi lebih kompleks dibanding pada vektor.

Berikut beberapa operasi yang sering digunakan dalam matriks.

Tabel 1. Operasi pada Matriks

Operator	Keterangan
*	Perkalian elemen pada matriks
%*%	Perkalian matriks
solve	Invers matriks
t	Transpose matriks

Contoh penggunaannya.

```
> A * A      #menampilkan perkalian elemen matriks A
dengan matriks A
      [,1] [,2] [,3]
[1,]    1   16   49
[2,]    4   25   64
[3,]    9   36   81

> A %*% A    #menampilkan perkalian matriks A dengan
matriks A
      [,1] [,2] [,3]
[1,]   30   66   102
[2,]   36   81   126
[3,]   42   96   150

> t(A)       #menampilkan transpose matriks A
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

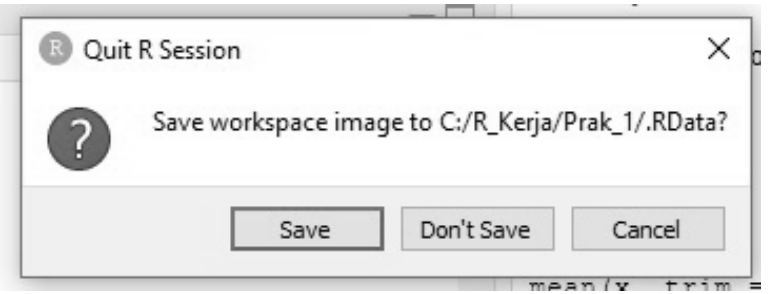
Untuk mencari invers suatu matriks, misalkan diketahui matriks B yaitu:

$$B = \begin{bmatrix} 1 & 4 & 7 \\ 10 & 5 & 3 \\ 2 & 6 & 9 \end{bmatrix}$$

Dibuat dalam R menjadi :

```
> B = matrix(c(1,10,2,4,5,6,7,3,9), nrow=3, ncol=3)
> B #print matriks B
      [,1]  [,2]  [,3]
[1,]    1     4     7
[2,]   10     5     3
[3,]    2     6     9

> solve(B) #menampilkan invers matriks B
       [,1]             [,2]           [,3]
[1,]  0.6585366   0.14634146  -0.5609756
[2,] -2.0487805  -0.12195122   1.6341463
[3,]  1.2195122   0.04878049  -0.8536585
```

Untuk mengakhiri sesi penggunaan R maka dapat dilakukan dengan cara mengeklik menu *File* kemudian pilih *Quit Session*, sehingga akan muncul tampilan berikut ini.



Gambar 3. Tampilan kotak dialog untuk keluar dari sesi R

Jangan lupa untuk selalu menyimpan sesi R dengan memilih *Save*.

1.6 Import Data

Memasukkan data pada R dapat dilakukan dengan beberapa metode. Secara umum proses *importing* data pada R dapat dilakukan dengan dua cara, yaitu menggunakan perintah di *command line* (*R Console*) dan menggunakan fasilitas GUI R-CMDR.

Ingat untuk memastikan direktori kerja R, gunakan perintah setwd(). Misalkan data dibuat dengan *Notepad* dengan nama file Penjualan. File

disimpan di direktori C :/R_Kerja/Praktikum/Penjualan.txt

```
 83  67 108 112  56 78  39  60  48  71
 28  48  27 136  83 82  72  42  39  29
 28 100  73  48 103 78 120  96  78  64
 43  62  42  73  64 96 102  43  72 118
 39  48  71  63  64 38  26  43  33  74
```

Data di atas dapat dibuat dalam *Notepad* dan termasuk ke dalam file ASCII. Suatu file ASCII terdiri dari sekumpulan data yang dipisahkan oleh spasi, tab, tanda akhir baris atau tanda baris baru serta pembatas lainnya.

Cara membaca data dalam bentuk file ASCII dapat dilakukan dengan menggunakan fungsi scan. Fungsi scan ini digunakan untuk data dalam jumlah yang besar dan tidak memiliki kepala (*header*) dan data tersimpan dalam suatu berkas teks (ekstension.txt).

Misalkan file disimpan dalam folder Praktikum, maka perintahnya dilakukan seperti di bawah ini.

```
> setwd("C :/R_Kerja/Praktikum") #mengubah direktori
kerja untuk Praktikum
> Penjualan=scan("Penjualan.txt") #membaca file
Penjualan.txt
Read 50 items #Keterangan terbaca 50 item
> Penjualan #Print variabel Penjualan
[1] 83 67 108 112 56 78 39 60 48 71 28 48 27 136 83
82 72 42 39 29
[21]28 100 73 48 103 78 120 96 78 64 43 62 42 73 64
96 102 43 72 118
[41]39 48 71 63 64 38 26 43 33 74
```

Apabila data di atas ingin dibuat matrik dengan dimensi (ukuran) = 5 x 10, maka lakukan perintah berikut ini.

```
> A = matrix(Penjualan, 5,10)
> A
       [,1]  [,2]  [,3]  [,4]  [,5]
 [1,]   83    28    28    43    39
 [2,]   67    48   100    62    48
 [3,]  108    27    73    42    71
 [4,]  112   136    48    73    63
 [5,]   56    83   103    64    64
 [6,]   78    82    78    96    38
 [7,]   39    72   120   102    26
 [8,]   60    42    96    43    43
 [9,]   48    39    78    72    33
[10,]   71    29    64   118    74
```

Selanjutnya, apabila data dibuat di Excell dengan ektenson. Xls atau .xlxs dapat diimpor dengan dengan fasilitas *command line*. Untuk dapat dieksekusi di R, file tersebut perlu diubah ke dalam format .TXT (*Text Tax Delimited*) atau format CSV (*comma delimited*). Setelah itu, data dapat diimpor dengan menggunakan perintah read.tabel atau read.csv.

Misalkan data sensus penduduk dibuat di Excell dengan format CSV seperti dalam gambar berikut ini.

	A	B	C	D	E	F
1	Nama	Umur	JK	BB	TB	
2	Zahra	16	P	55	165	
3	Fadhia	14	P	50	155	
4	Ghina	18	P	58	158	
5	Azka	10	L	45	140	
6	Nauval	12	L	44	143	
7	Zauzah	15	P	54	160	
8						
9						
10						

Gambar 4. Data penjualan dalam format CSV atau TXT

Misalkan file tersebut disimpan di folder C :/R_Kerja/Praktikum/Sensus. CSV. Perintah untuk mengimpor file tersebut di R adalah:

```
> Data1 = read.csv("Sensus.CSV", header=TRUE, sep=";")
> Data1
     Nama  Umur  JK   BB   TB
1   Zahra   16    P   55  165
2  Fadhia   14    P   50  155
3   Ghina   18    P   58  158
4    Azka   10    L   45  140
5  Nauval   12    L   44  143
6  Zauzah   15    P   54  160
```

Argumen optional `header = TRUE` digunakan apabila baris pertama dalam file tersebut adalah *header* atau nama variabel. Sedangkan optional `sep=`";" adalah pemisah atau pembatas antarkolom (variabel).

Perintah impor data untuk file yang ektensi .TXT adalah :

```
> Data2=read.table("Sensus.txt", header=TRUE)
> Data2
     Nama  Umur  JK   BB   TB
1   Zahra   16    P   55  165
2  Fadhia   14    P   50  155
3   Ghina   18    P   58  158
4    Azka   10    L   45  140
5  Nauval   12    L   44  143
6  Zauzah   15    P   54  160
```

Kedua Data 1 dan Data 2 menghasilkan data yang sama. Perintah untuk mengimpor data selain yang diuraikan di sini, perintah yang lainnya dapat dilihat dengan mengetik ?read pada *command line*. Daftar atau *list* nanti akan muncul pada jendela *Help*, juga ada beberapa cara *import* data yang tidak dijelaskan dalam buku ini misalnya read.delim dan lain-lain.

1.7 Ringkasan Numerik Data

Pembahasan pada bagian ini difokuskan untuk membuat statistik deskriptif, khususnya mengenai pembuatan ringkasan (*summary*) data. Ringkasan statistik deskriptif dilakukan dengan menggunakan perintah :

```
summary(nama_file).
```

Ringkasan ini akan menampilkan beberapa besaran statistic, yaitu `Mean, Min, Max, Quartil1, Median` dan `Quartil3`.

Nilai-nilai tersebut memberikan gambaran data berupa ukuran pusat, letak data, dan ukuran dispersi dalam bentuk numerik. Selain itu, ringkasan numerik juga diperlukan sebagai estimasi dari nilai-nilai karakteristik data.

Misalkan ingin menghitung ringkasan numerik data Penjualan, maka perintahnya :

```
> summary(Penjualan)
  Min.  1st Qu.  Median   Mean   3rd Qu.    Max.
  26.0    43.0     64.0   66.2      81.0   136.0
```

Hasil di atas menunjukkan bahwa minimal penjualan sebesar 26, kuartil 1 = 43, Median = 64, rata-rata = 66,2, kuartil 3 = 81 dan maksimum = 136.

Untuk mencari ringkasan numerik dapat dilakukan dengan menggunakan perintah berikut ini.

```
> mean(Penjualan)          #menampilkan rata – rata
Penjualan
[1] 66.2
> median(Penjualan)        #menampilkan median
Penjualan
[1] 64
> var(Penjualan)           #menampilkan variansi data
Penjualan
[1] 759.5102
> sd(Penjualan)
[1] 27.55921
> min(Penjualan)           #menampilkan nilai minimum
data Penjualan
[1] 26
```

```
> max(Penjualan)          #menampilkan nilai maksimum
data Penjualan
[1] 136
> quantile(Penjualan)     #menampilan nilai kuantil
data penjualan
 0% 25% 50% 75% 100%
 26 43 64 81 136
```

R menyediakan dua macam cara untuk menampilkan ringkasan numerik, yaitu menampilkan ringkasan numerik untuk satu variabel tertentu saja dan menampilkan ringkasan numerik untuk semua variabel.

Untuk menjelaskan ringkasan numerik untuk semua variabel dan variabel tertentu saja dilakukan dengan menggunakan data yang ada di R. Software R menyediakan data yang bisa digunakan untuk lebih memperjelas cara kerja setiap fungsi.

Pada menu **`help(summary)`**, pada bagian `example` terlihat data yang digunakan adalah data `attenu`. Data attenu berisi tentang informasi gempa bumi yang diukur dari stasiun yang ada di California. Berikut keterangan mengenai file `attenu`:

`A data frame with 182 observations on 5 variables.`

[,1]	*event*	*numeric*	*Event Number*
[,2]	*mag*	*numeric*	*Moment Magnitude*
[,3]	*station*	*factor*	*Station Number*
[,4]	*dist*	*numeric*	*Station-hypocenter distance (km)*
[,5]	*accel*	*numeric*	*Peak acceleration (g)*

Untuk melihat isi dari data attenu, ketik seperti di bawah ini.

```
> attenu            #menampilkan isi data attenu
      event   mag   station   dist   accel
1         1   7.0       117   12.0   0.359
2         2   7.4      1083  148.0   0.014
3         2   7.4      1095   42.0   0.196
4         2   7.4       283   85.0   0.135
5         2   7.4       135  107.0   0.062
…         …     …         …      …       …
180      23   5.3      5069   47.7   0.033
181      23   5.3      5073   49.2   0.017
182      23   5.3      5072   53.1   0.022
```

Output di atas hanya menampilkan data `attenu` untuk pengamatan 5 teratas dan 3 terbawah. Misalkan ringkasan numerik hanya dilakukan untuk variabel `mag`,`dist`, dan `accel`.

Langkah pertama adalah mengekstrak data attenu untuk mengambil variabel yang diinginkan.

```
> mag=attenu[,2]      #mengambil data pada kolom ke-2
dari attenu
> distance=attenu[,4] #mengambil data pada kolom ke-2
dari attenu
> accel=attenu[,5]    #mengambil data pada kolom ke-2
dari attenu
> data3=cbind(mag,distance,accel)#menggabungkan
mag,distance, accel
> data3                    #print data3
       mag  distance   accel
 [1,]  7.0      12.0   0.359
 [2,]  7.4     148.0   0.014
 [3,]  7.4      42.0   0.196
 [4,]  7.4      85.0   0.135
 [5,]  7.4     107.0   0.062
    …   …         …       …
[178,] 5.3      46.1    0.070
[179,] 5.3      47.1    0.080
[180,] 5.3      47.7    0.033
```

```
[181,] 5.3      49.2    0.017
[182,] 5.3      53.1    0.022
> summary(data3)    #menampilkan ringkasan numerik data3
     mag              distance           accel
 Min.    :5.000    Min.   : 0.50    Min.    :0.00300
 1st Qu. :5.300    1st Qu.: 11.32   1st Qu. :0.04425
 Median  :6.100    Median : 23.40   Median  :0.11300
 Mean    :6.084    Mean   : 45.60   Mean    :0.15422
 3rd Qu. :6.600    3rd Qu.: 47.55   3rd Qu. :0.21925
 Max.    :7.700    Max.   :370.00   Max.    :0.81000
```

Terlihat bahwa perintah *summary* dapat dilakukan untuk menghitung statistik deskriptif secara serentak untuk semua variabel yang diinginkan.

Hasil di atas menunjukkan ringkasan numerik untuk variabel `mag`, `distance`, dan `accel`.

1 Pada variabel mag (*moment magnitude*): nilai minimum = 5,00; kuartil 1 (Q1) = 5,30; Median (Q2) = 6,10; rata-rata = 6,084, kuartil 3 (Q3)= 6,60 dan nilai maksimum = 7,70;

2 Pada variabel distance (*Station-hypocenter distance* (km)): nilai minimum = 0,50; kuartil 1 (Q1) = 11,32; Median (Q2) = 23,40; rata-rata = 45,60, kuartil 3 (Q3) = 47,55 dan nilai maksimum = 370,0;

3 Pada variabel accel (*Peak acceleration* (g)): nilai minimum = 0,00300; kuartil 1 (Q1) = 0,04425; Median (Q2) = 0,11300; rata-rata = 0,15422, kuartil 3 (Q3) = 0,21925 dan nilai maksimum = 0,81000.

1.8 Latihan

1 Buatlah vektor dengan elemen 1-20! Beri nama vektor tersebut X!

2 Buatlah vektor dengan elemen 1-100 dengan kenaikan setiap 5! Beri nama vektor tersebut Y!

3 Diketahui matriks $\boldsymbol{A}$ dan $\boldsymbol{B}$ berikut ini.

$$A = \begin{bmatrix} 2 & 3 & 7 \\ 1 & 5 & 3 \\ 2 & 6 & 4 \end{bmatrix} \quad \text{dan} \quad B = \begin{bmatrix} -1 & 3 \\ 2 & 5 \\ 2 & 6 \end{bmatrix}$$

Berdasarkan matriks di atas :

a. Buatkah matriks A dan B dengan R!
b. Carilah transpose kedua matriks tersebut!
c. Carilah invers matriks A!
d. Lakukan perkalian matriks A dengan B!

4. Diketahui matriks dimana kolom pertama Nama, kolom kedua Umur, kolom ketiga Jenis Kelamin dan kolom keempat Umur dengan elemen-elemen berikut ini.

$$C = \begin{bmatrix} \text{Yani} & 27 & \text{Perempuan} & 50 \\ \text{Didi} & 25 & \text{Laki-laki} & 66 \\ \text{Sari} & 30 & \text{Perempuan} & 50 \\ \text{Maman} & 22 & \text{Laki-laki} & 65 \end{bmatrix}$$

a. Buatlah matriks ***C*** dengan R!
b. Lakukan ekstrasi matriks ***C***, dimana matriks baru hanya berisi variabel Nama dan Jenis Kelamin!
c. Lakukan ekstrakdi matriks ***C***, dimana matriks baru hanya berisi elemen yang berjenis kelamin Perempuan!

5. Berdasarkan pengamatan Badan Meteorologi, Klimatologi dan Geofisika (BMKG) suhu udara 16 wilayah di Provinsi Jawa Tengah sebagai berikut.

31, 27, 31, 30, 30, 33, 32, 32, 32, 30, 30, 31, 30, 31, 34, 10.

a. Hitunglah ringkasan statistik untuk data di atas!
b. Carilah variansi dan standar deviasi!

6. Di bawah ini disajikan data penjualan sepeda motor PT ”A” dan PT ”B” selama tahun 2019.

Pengamatan	1	2	3	4	5	6	7	8	9	10
A	20	25	30	31	30	20	25	33	31	30
B	35	40	37	42	43	50	51	49	40	43

a. Hitunglah ringkasan statistik untuk data di atas!
b. Carilah variansi dan standar deviasi!

7. Terlampir adalah data ekspor Migas dan Non Migas sejak tahun 1975 sampai dengan 2019 yang diunduh di www.bps.go.id.

Tahun	Migas	Non Migas	Tahun	Migas	Non Migas	Tahun	**Migas**	**Non Migas**
1975	4769,8	4561,3	1990	11659,7	40456,8	2005	17457,7	40243,2
1976	5673,1	5235,4	1991	14021,5	46062,3	2006	18962,9	42102,6
1977	732	5498,3	1992	13806,7	49489,4	2007	21932,8	52540,6
1978	579,7	6110,7	1993	2170,5	26157,3	2008	30552,9	98644,4
1979	793,3	6409,0	1994	2367,2	29621,4	2009	18980,7	77848,5
1980	1744,0	9090,4	1995	2910,8	37743,3	2010	27412,7	108250,6
1981	1721,3	11550,8	1996	3589,7	39338,9	2011	40701,6	136734,1
1982	3544,8	13314,1	1997	3924,1	37755,7	2012	42564,4	149126,6
1983	4144,8	12207,0	1998	2653,7	24683,2	2013	45266,4	141362,3
1984	2696,8	11185,3	1999	3681,1	20322,2	2014	43459,9	134718,9
1985	1275,6	8983,5	2000	6019,5	27495,3	2015	24613,1	118081,4
1986	1086,4	9632,0	2001	5471,8	25490,3	2016	18739,4	116913,4
1987	1067,9	11302,4	2002	6525,8	24763,1	2017	24316,2	132669,3
1988	909	12339,5	2003	7610,9	24939,8	2018	29868,8	158842,4
1989	1195,2	15164,4	2004	11732,0	34792,5	2019	21885,3	148842,1

a. Carilah ringkasan (*summary*) statistik untuk data Migas dan Non Migas! Analisis hasilnya!
b. Hitung variansi dan standar deviasi untuk data Migas dan Non Migas! Bandingkan kedua data tersebut!

BAB

II

PENGANTAR ANALISIS MULTIVARIAT

2.1 Pendahuluan

Sebagian besar fenomena yang terjadi dalam kehidupan nyata, baik di bidang sosial, ekonomi, kesehatan, maupun ilmu alam, umumnya dipengaruhi oleh lebih dari satu variabel. Variabel-variabel tersebut sering kali saling berhubungan dan membentuk pola hubungan yang kompleks. Oleh karena itu, diperlukan suatu teknik analisis statistik yang mampu mengkaji hubungan antarvariabel secara simultan, agar informasi yang diperoleh dapat menggambarkan fenomena yang diteliti secara lebih utuh dan akurat.

Berdasarkan jumlah variabel yang dianalisis, metode analisis statistik dapat dikelompokkan menjadi tiga jenis, yaitu analisis univariat, analisis bivariat, dan analisis multivariat. Analisis univariat merupakan analisis yang hanya melibatkan satu variabel, dengan tujuan utama untuk mendeskripsikan karakteristik data, seperti nilai rata-rata, median, modus, serta ukuran penyebaran data. Sebagai contoh, penelitian yang bertujuan untuk mengetahui gambaran tingkat pendapatan masyarakat di Provinsi Daerah Istimewa Yogyakarta termasuk dalam analisis univariat.

Analisis bivariat melibatkan dua variabel dan umumnya digunakan untuk mengkaji hubungan atau pengaruh antara satu variabel terhadap variabel lainnya. Misalnya, penelitian yang menganalisis pengaruh tingkat pendidikan terhadap pendapatan masyarakat merupakan contoh analisis bivariat. Teknik statistik yang sering digunakan dalam analisis bivariat antara lain korelasi, regresi sederhana, dan uji perbedaan dua kelompok.

Selanjutnya, analisis multivariat digunakan ketika penelitian melibatkan lebih dari dua variabel yang dianalisis secara bersamaan. Sebagai contoh, penelitian yang mengkaji pengaruh tingkat pendidikan, umur, dan lama bekerja terhadap pendapatan termasuk dalam analisis multivariat. Pada analisis ini, hubungan antarvariabel tidak hanya dilihat secara terpisah, tetapi dipertimbangkan secara simultan sehingga dapat memberikan gambaran yang lebih komprehensif mengenai fenomena yang diteliti.

Analisis multivariat merupakan teknik statistik yang digunakan untuk mengolah dan menganalisis data berdimensi tinggi, yaitu data yang terdiri atas banyak variabel yang umumnya saling berkorelasi. Perlu dibedakan

antara istilah *multivariabel* dan *multivariat*. Multivariabel mengacu pada analisis yang melibatkan lebih dari satu variabel, sedangkan multivariat secara khusus merujuk pada analisis multivariabel yang memperhitungkan adanya hubungan atau korelasi antarvariabel tersebut. Dengan demikian, analisis multivariat selalu bersifat multivariabel, tetapi tidak semua analisis multivariabel dapat disebut sebagai analisis multivariat.

Dalam analisis multivariat, satu unit penelitian dapat diamati dari berbagai dimensi atau sudut pandang. Sebagai contoh, dalam penelitian mengenai kesejahteraan masyarakat, unit analisis yang digunakan dapat berupa rumah tangga. Setiap rumah tangga dapat diukur melalui berbagai variabel, seperti pendapatan, pengeluaran, konsumsi pangan atau protein, jumlah anggota rumah tangga, tingkat pendidikan kepala rumah tangga, dan karakteristik sosial ekonomi lainnya. Banyaknya variabel yang diamati tersebut mencerminkan kompleksitas kondisi kesejahteraan rumah tangga. Oleh karena itu, pendekatan analisis multivariat sangat tepat digunakan untuk mengidentifikasi pola, hubungan, dan struktur data yang tidak dapat dijelaskan secara memadai melalui analisis univariat atau bivariat saja.

Misalkan dimiliki sekumpulan data terdiri dari p variabel bebas dan n observasi, $\{x_i\}_{i=1}^{n}$, dimana setiap pengamatan x_i *memiliki* p dimensi dinyatakan dengan :

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{in})$$

merupakan elemen dari vektor variabel $X \in \Re^p$, sehingga terdiri dari variabel random

$$X = (X_1, X_2, \cdots, X_p)$$

Pada sekumpulan data, mungkin saja tertarik untuk mengetahui hal-hal berikut ini.

1. Apakah ada komponen X yang lebih menyebar daripada yang lain?
2. Apakah ada beberapa komponen yang membentuk suatu kelompok baru?
3. Apakah ada nilai-nilai ektsrem dalam sebuah komponen?
4. Bagaimana distribusi datanya?
5. Apakah mungkin dilakukan pengurangan dimensi sehingga

memungkinkan membuat dimensi yang lebih kecil?

Berdasarkan pertanyaan-pertanyaan di atas memunculkan berbagai analisis multivariat seperti analisis komponen utama, klaster, diskriminan dan lain-lain.

1. Menurut Johnson and Wichern (2007), tujuan melakukan analisis multivariat diantaranya adalah :
2. Mereduksi data menjadi lebih sederhana (*data reduction*). Tujuan dari mereduksi data adalah membuat data menjadi lebih sederhana (dimensi menjadi lebih kecil) sehingga mudah diinterpretasikan;
3. Mengelompokkan objek berdasarkan variabel (*grouping*). Tujuan dari mengelompokkan yaitu membuat objek-objek yang memiliki kemiripan karakteristik (variabel) yang sama akan berada dalam suatu kelompok;
4. Menyelidiki keterkaitan antara beberapa variabel (*investigate of dependence among several variables*). Analisis mengenai keterkaitan antara variabel adalah untuk menyelidiki adakah hubungan antarvariabel yang diteliti;
5. Meramalkan (*prediction*). Tujuan dari meramalkan adalah memprediksi suatu variabel jika variabel lainnya diketahui;
6. Menguji suatu hipotesis. Uji hipotesis dilakukan untuk menguji kebenaran suatu dugaan/hipotesa tertentu.

Klasifikasi teknik analisis data di dalam analisis multivariat dapat dibedakan menjadi:

1. *Dependence Method*. Metode di mana suatu variabel atau kumpulan variabel yang diketahui sebagai variabel tak bebas diprediksi atau dijelaskan oleh variabel-variabel yang lain, disebut sebagai variabel bebas. Beberapa metode dependensi diantaranya analisis regresi multivariat, *multivariate analysis of variance* (MANOVA), analisis diskriminan, analisis kanonik, analisis jalur (*path analysis*), model persamaan sturktural (*structural equation modelling*/SEM) dan lain-lain.
1. *Interdependence Method*. Metode dimana tidak ada satu atau sekelompok variabel yang didefinisikan sebagai variabel bebas

ataupun variabel tak bebas. Teknik- teknik yang termasuk golongan metode interdependensi diantaranya adalah *cluster analysis*, *factor analysis* dan *multidimensional scaling* (MDS).

Berikut penjelasan beberapa metode analisis multivariat.

Tabel 2. Jenis-jenis Metode Analisis Multivariat

No	Metode	Tujuan	Model
1	*Principal Component Analysis*	Mereduksi dimensi data dengan cara membangkitkan variabel baru (komponen utama) yang merupakan kombinasi linear dari variabel asal sedemikan hingga varians komponen utama menjadi maksimum dan antar komponen utama bersifat saling bebas	$\mathbf{Y}_j = \mathbf{a}' \mathbf{X}$ maks var ($\mathbf{Y}_i$) dan corr ($\mathbf{Y}_i$, $\mathbf{Y}_j$) = 0
2	*Factor Analysis*	Mereduksi dimensi data dengan cara menyatakan variabel asal sebagai kombinasi linear sejumlah faktor, sedemikian hingga sejumlah faktor tersebut mampu menjelaskan sebesar mungkin keragaman data yang dijelaskan oleh variabel asal	$\mathbf{X} = \mathbf{CF} + \varepsilon$ maks var (**CF**)
3	*Multivariate Regression*	Memodelkan hubungan antara kelompok variabel respon (**Y**) dengan kelompok variabel (**X**) yang diduga mempengaruhi variabel respon	$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$
4	*Multivariate Analysis of Variance*	Menganalisis hubungan antara vektor variabel respon (**Y**) yang diduga dipengaruhi oleh beberapa perlakuan (*treatment*).	$Y_{ijk} = \mu_k + \tau_k + \varepsilon_{ijk}$ i=1,...,t j=1,...,n_i k=1,...,p

No	Metode	Tujuan	Model
5	*Discriminant Analysis*	Membentuk fungsi yang memisahkan antar kelompok berdasarkan variabel pembeda, fungsi tersebut disusun sedemikian nisbah keragaman data antar dan kelompok maksimum.	$D = b0 + b_1X_1 + ... + b_kX_k$
6	*Cluster Analysis*	Mengelompokkan data ke dalam beberapa kelompok sedemikian hingga data yang berada di dalam kelompok yang sama cenderung mempunyai sifat yang lebih homogen daripada data yang berada di kelompok yang berbeda	Teknik Hirarki dan Teknik Non hirarki
7	*Canonicl Correlation Analysis*	Mengidentifikasi dan mengukur hubungan linier, yang melibatkan beberapa variabel dependen dan independen	$\boldsymbol{U = a' X}$ dan $\boldsymbol{v = b' Y}$
8	Multidimensional Scalling	Memetakan atau mencari konfigurasi sejumlah objek dalam ruang dimensi rendah, misalnya dua atau tiga, yang diharapkan dapat merefleksikan sebaik mungkin ukuran ketidakmiripan yang diketahui antar objek tersebut.	Peta spatial
9	*Path Analysis*	Mengidentifikasi dan menguji hubungan kausal antara berbagai variabel dalam suatu model. Pemodelan analisis jalur biasanya ditampilkan dalam bentuk diagram jalur (*path diagram*) yang menggambarkan hubungan antara variabel-variabel dengan menggunakan panah atau jalur yang menghubungkan satu variabel ke variabel lainnya.	Model regresi linear berganda, Model Mediasi

No	Metode	Tujuan	Model
10	*Structural Equation Modeling*	Menganalisis hubungan kausal antara variabel, baik yang teramati (terukur) maupun yang tidak teramati (laten). SEM menggabungkan prinsip analisis faktor dan analisis jalur untuk menguji hubungan yang kompleks dan memodelkan hubungan sebab akibat	$\boldsymbol{Y = \lambda_y \eta + \xi}$ $\boldsymbol{X = \lambda_y \xi + \delta}$

2.2 Struktur Data Analisis Multivariat

Analisis multivariat merupakan teknik statistik yang berkaitan dengan pengukuran sejumlah besar variabel yang diamati secara simultan pada setiap objek penelitian dalam satu atau lebih sampel. Dengan kata lain, analisis multivariat mempelajari hubungan bersama (simultan) antar beberapa variabel, yaitu lebih dari dua variabel, yang umumnya saling berkorelasi. Pendekatan ini memungkinkan peneliti untuk memahami pola, struktur, serta keterkaitan antarvariabel secara menyeluruh, bukan secara parsial.

Untuk memahami konsep analisis multivariat, maka perlu diketahui bagaimana penyusunan/pengorganisasian data di dalam analisis multivariat. Misalkan, data yang merupakan hasil pengukuran *n* objek atas *p*, maka x_{ij} adalah variabel acak yang menunjukkan pengamatan ke-*i* dan variabel ke-*j* dengan

$i = 1, 2, ..., n$ dan

$j = 1, 2, ..., p$

Data multivariat tersebut dapat disajikan dalam bentuk tabel atau matriks sebagai berikut.

Tabel 3. Tabulasi Data Multivariat

	Variabel 1	Variabel 2	...	Variabel k	...	Variabel p
Obs 1	X_{11}	X_{12}	...	X_{1k}	...	X_{1p}
Obs 2	X_{21}	X_{22}	...	X_{2k}	...	X_{2p}
...			...		...	

Obs *j*	X_{j1}	X_{j2}	...	X_{jk}	...	X_{jp}
...			...		...	
Obs *n*	X_{n1}	X_{n2}	...	X_{nk}	...	X_{np}

Secara matematis, struktur data tersebut dapat dinyatakan dalam bentuk matriks data berukuran *n* x *p*, yaitu

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} \ldots & x_{1p} \\ x_{21} & x_{22} \ldots & x_{2p} \\ \vdots & \vdots \ddots & \ldots \\ x_{n1} & x_{n2} \ldots & x_{np} \end{bmatrix}$$

Berdasarkan matriks di atas dapat dihitung beberapa ukuran numerik data, diantaranya :

vektor rata-rata sampel yaitu $\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$

dimana $\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$

Matriks variansi-kovariansi adalah: $\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} \ldots & s_{1p} \\ s_{21} & s_{22} \ldots & s_{2p} \\ \vdots & \vdots \ddots & \ldots \\ s_{p1} & s_{p2} \ldots & s_{pp} \end{bmatrix}$

dengan:

$$s_{ii} = s_i^2 = \frac{\sum_{j=1}^{n} \left(x_{ji} - \bar{x}_i\right)^2}{n-1} \text{ dan } s_{ik} = \frac{\sum_{j=1}^{n} \left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right)}{n-1}$$

Sedangkan matriks korelasi sampel adalah

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} \ldots & r_{1p} \\ r_{21} & 1 \ \ldots & r_{2p} \\ \vdots & \vdots \ddots & \ldots \\ r_{p1} & r_{p2} \ \ldots & 1 \end{bmatrix}$$

Contoh

Data berikut ini adalah data pendapatan perusahaan Jaya dalam empat tahun.

Tabel 4. Data Pendapatan Perusahaan PT Jaya

	Tahun ke-1	Tahun ke-2	Tahun ke-3	Tahun ke-4
Pendapatan (x 100 juta)	45	50	43	42
Banyaknya karyawan	4	5	4	3

Matriks data adalah: $\boldsymbol{X} = \begin{bmatrix} 45 & 4 \\ 50 & 5 \\ 43 & 4 \\ 42 & 3 \end{bmatrix}$

1. Vektor Rata-rata, Matriks Variansi-Kovariansi dan Matriks Korelasi

Ukuran-ukuran statistik deskriptif yang penting dipelajari diantaranya adalah vektor rata-rata, matriks variansi-kovariansi dan matriks korelasi.

Vektor Rata-rata Populasi

Diketahui sekumpulan data populasi yaitu $x_{ij} \in \boldsymbol{X}$ dengan $i = 1, 2, ..., N$ dan $j = 1, 2, ..., N$. maka rata-rata populasi untuk variabel ke-j adalah :

$$\mu_j = \frac{\sum_{i=1}^{N} x_{ij}}{N};\ \ j = 1,2, \dots, p$$

Selanjutnya dapat dibentuk vektor rata-rata poulasi adalah :

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \qquad (2.1)$$

Vektor Rata-rata Sampel

Diketahui sekumpulan data sampel yaitu $x_{ij} \in \boldsymbol{X}$ dengan $i = 1, 2, ..., n$ dan $j = 1, 2, ..., p$. maka rata-rata sampel untuk variabel ke-j adalah :

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n};\ \ j = 1,2,\dots,p$$

Misalkan ingin menghitung rata-rata sampel pada variabel pertama, maka rumusnya adalah :

$$\bar{x}_1 = \frac{\sum_{i=1}^{n} x_{i1}}{n} = \frac{x_{11} + x_{21} + \cdots + x_{n1}}{n}$$

Selanjutnya dapat dibentuk vektor rata-rata sampel sebagai berikut :

$$\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad (2.2)$$

Contoh

Hitunglah vektor rata-rata pada data pendapatan (Tabel 4)

Rata-rata untuk setiap variabel dapat dihitung yaitu :

$$\bar{x}_1 = \frac{45 + 50 + 43 + 42}{4} = \frac{180}{4} = 45$$

$$\bar{x}_2 = \frac{4 + 5 + 4 + 3}{4} = \frac{16}{4} = 4$$

Maka, vektor rata-rata sampel adalah :

$$\bar{\boldsymbol{x}} = \begin{bmatrix} 45 \\ 4 \end{bmatrix}$$

Matriks Variansi-Kovariansi Populasi

Variansi populasi variabel ke-*j* dapat dihitung menggunakan rumus berikut ini.

$$\sigma_{jj} = \frac{\sum_{i=1}^{N} \left(x_{ij} - \mu_j\right)^2}{N};\ \ j = 1,2,\dots,p$$

Sedangkan kovariansi populasi variabel ke-k dengan variabel ke-j adalah:

$$\sigma_{jk} = \frac{\sum_{i=1}^{N}(x_{ik} - \mu_k)(x_{ij} - \mu_j)}{N}; \; j \neq k \; dan \; j, k = 1,2, \dots, p$$

Matriks variansi-kovariansi sampel dapat dibentuk menjadi :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} \cdots & \sigma_{2p} \\ \vdots & \vdots \ddots & \cdots \\ \sigma_{p1} & \sigma_{p2} \cdots & \sigma_{pp} \end{bmatrix} \tag{2.3}$$

Matriks Variansi-Kovariansi Sampel

Variansi sampel variabel ke-j dapat dihitung menggunakan rumus berikut ini.

$$s_{jj} = \frac{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}{n-1}; \; j = 1,2, \dots, p$$

Misalkan ingin menghitung variansi sampel pada variabel pertama, maka rumusnya adalah :

$$s_{11} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}{n-1}$$

$$= \frac{(x_{11} - \bar{x}_1)^2 + (x_{21} - \bar{x}_1)^2 + \cdots + (x_{n1} - \bar{x}_1)^2}{n-1}$$

Sedangkan kovariansi sampel variabel ke-k dengan variabel ke-j adalah :

$$s_{kj} = \frac{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j)}{n-1}; \; j \neq k \; dan \; j, k = 1,2, \dots, p$$

Misalkan ingin menghitung kovariansi sampel pada variabel pertama dengan variabel kedua, maka rumusnya adalah :

$$s_{12} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{n-1};$$

$$= \frac{(x_{11} - \bar{x}_1)(x_{12} - \bar{x}_2) + (x_{21} - \bar{x}_1)(x_{22} - \bar{x}_2) + \cdots + (x_{n1} - \bar{x}_1)(x_{n2} - \bar{x}_2)}{n-1}$$

Matriks variansi-kovariansi sampel dapat dibentuk menjadi :

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} \cdots & s_{1p} \\ s_{21} & s_{22} \cdots & s_{2p} \\ \vdots & \vdots \ddots & \cdots \\ s_{p1} & s_{p2} \cdots & s_{pp} \end{bmatrix} \qquad (2.4)$$

Contoh

Carilah variansi dan kovariansi pada data pendapatan (Tabel 4) dan tuliskan matriks kovariansinya!

Jawab

Variansi variabel pertama adalah :

$$s_{11} = \frac{(45-45)^2 + (50-45)^2 + (43-45)^2 + (42-45)^2}{4-1}$$

$$= \frac{(0)^2 + (5)^2 + (-2)^2 + (-3)^2}{3} = 12{,}67$$

Variansi variabel kedua adalah :

$$s_{22} = \frac{(4-4)^2 + (5-4)^2 + (4-4)^2 + (3-4)^2}{4-1}$$

$$= \frac{(0)^2 + (1)^2 + (0)^2 + (-1)^2}{3} = 0{,}67$$

Kovariansi variabel pertama dengan kedua adalah :

$s_{12} = s_{21}$

$$= \frac{(45-45)(4-4)+(50-45)(5-4)+(43-45)(4-4)+(42-45)(3-4)}{4-1}$$

$$= \frac{(0)(0)+(5)(1)+(-2)(0)+(-3)(-1)}{3} = 2{,}67$$

sehingga, matriks variansi kovariansi adalah :

$$\boldsymbol{S} = \begin{bmatrix} 12{,}67 & 2{,}67 \\ 2{,}67 & 0{,}67 \end{bmatrix}$$

Matriks Korelasi Populasi

Statistik deskriptif yang sering dipakai dalam analisis multivariat adalah korelasi. Korelasi populasi antara variabel ke-*k* dan variabel ke-*j* adalah :

$$\rho_{kj} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\ \sqrt{\sigma_{kk}}}$$

Matriks korelasi populasi dengan *p*-variabel adalah :

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \cdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \qquad (2.5)$$

Matriks Korelasi Sampel

Statistik deskriptif yang sering dipakai dalam analisis multivariat adalah korelasi. Korelasi sampel antara variabel ke-*k* dan variabel ke-*j* adalah :

$$r_{kj} = \frac{s_{jk}}{\sqrt{s_{jj}}\ \sqrt{s_{kk}}}$$

Misalkan ingin dicari korelasi antara variabel ke-1 dan ke-2 yaitu

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}}\ \sqrt{s_{22}}}$$

Matriks korelasi untuk data dengan *p*-variabel adalah :

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} \dots & r_{1p} \\ r_{21} & 1 \dots & r_{2p} \\ \vdots & \vdots \ddots & \dots \\ r_{p1} & r_{p2} \dots & 1 \end{bmatrix} \tag{2.6}$$

Contoh

Buatlah matrik korelasi untuk data pendapatan perusahaan Jaya (lihat Tabel 4)!

Berdasarkan pada perhitungan matriks variansi-kovariansi diketahui bahwa:

s_{12} = 2,67; s_{11} = 12,67 dan s_{22} = 0,67

sehingga, korelasi antara variabel pertama dan kedua dapat dihitung yaitu:

$$r_{12} = \frac{2{,}67}{\sqrt{12{,}67}\ \sqrt{0{,}67}} = 0{,}9164$$

Matriks korelasi sampel adalah :

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0{,}9164 \\ 0{,}9164 & 1 \end{bmatrix}$$

2.3 Tabulasi Data dengan Menggunakan Software R

Berikut data yang terdiri dari 3 variabel dengan 5 pengamatan yaitu:

X_1	X_2	X_3
9	12	3
2	8	4
5	6	0
6	4	2
8	10	1

Berdasarkan data di atas, carilah vektor rata-rata, matriks kovariansi, dan matriks korelasinya!

Jawab

Tabel di atas disimpan di excell dengan nama ContohData.csv. Untuk memulai maka pertama panggil data (import data) dari Excell ke R dengan menggunakan perintah :

```
> X=read.csv2("ContohData.csv")
> X
   X1  X2  X3
1   9  12   3
2   2   8   4
3   5   6   0
4   6   4   2
5   8  10   1
```

Data sudah ada di R dengan nama X, selanjutnya dicari vektor rata-rata yaitu :

```
> colMeans(X)

X1 X2 X3
 6 8 2
```

sehingga $\bar{\boldsymbol{X}} = \begin{bmatrix} 6 \\ 8 \\ 2 \end{bmatrix}$

Menghitung matriks kovariansi yaitu :

```
> var(X)
       X1     X2     X3
X1   7.50    4.5  -1.25
X2   4.50   10.0   1.50
X3  -1.25    1.5   2.50
```

sehingga $\boldsymbol{S} = \begin{bmatrix} 7{,}50 & 4{,}50 & -1{,}25 \\ 4{,}50 & 10{,}0 & 1{,}50 \\ -1{,}25 & 1{,}50 & 2{,}50 \end{bmatrix}$

Menghitung matriks korelasi dengan menggunakan perintah berikut ini.

```
> cor(X)
            X1          X2          X3
X1   1.0000000   0.5196152  -0.2886751
X2   0.5196152   1.0000000   0.3000000
X3  -0.2886751   0.3000000   1.0000000
```

sehingga $\boldsymbol{R} = \begin{bmatrix} 1{,}0000 & 0{,}5196 & -0{,}2887 \\ 0{,}5196 & 1{,}0000 & 0{,}3000 \\ -0{,}2887 & 0{,}3000 & 1{,}0000 \end{bmatrix}$

2.4 Latihan

1. Buktikanlah bahwa $s_{kj} = s_{jk}$!
2. Buktikan bahwa $r_{ii} = 1$!
3. Berikut ini adalah data pengunjung ke perpustakaan UIN Sunan Kalijaga Yogyakarta.

	Hari ke-				
	1	2	3	4	5
Jumlah pengunjung (X_1)	30	35	42	38	45
Banyaknya buku yang dipinjam (X_2)	10	15	20	12	28

Berdasarkan data di atas

a. Buatlah vektor rata-rata!
b. Buatlah matriks variansi-kovariansi!
c. Buatlah matriks korelasi!

4. Berikut ini adalah data penduduk kota/kabupaten di provinsi DI Yogyakarta.

	Yogyakarta	Sleman	Kulonprogo	Bantul	Gunungkidul
Jumlah penduduk (x 10.000)	40	120	40	95	73
Rata-rata pengeluaran per kapita (x 10.000)	280	245	190	240	160
Rata-rata lama sekolah (tahun)	12	11	8	9	7

Berdasarkan data di atas, gunakan *software* untuk mencari nilai-nilai berikut ini.

a. Buatlah vektor rata-rata!
b. Buatlah matriks variansi-kovariansi!
c. Buatlah matriks korelasi!

BAB

III

DISTRIBUSI NORMAL MULTIVARIAT

3.1 Pendahuluan

Distribusi probabilitas memegang peranan penting dalam statistika, khususnya dalam pengembangan metode inferensial. Pada analisis univariat, distribusi normal merupakan salah satu distribusi yang paling banyak digunakan karena sifat-sifat matematisnya yang sederhana dan kemampuannya dalam merepresentasikan berbagai fenomena alam maupun sosial. Namun, dalam konteks analisis multivariat, objek penelitian umumnya diukur berdasarkan lebih dari satu variabel yang saling berkorelasi, sehingga pendekatan distribusi univariat tidak lagi memadai untuk menggambarkan karakteristik data secara menyeluruh.

Distribusi normal multivariat merupakan perluasan dari distribusi normal univariat ke dalam dimensi yang lebih tinggi. Distribusi ini digunakan untuk memodelkan vektor perubahan secara acak yang terdiri atas beberapa variabel yang diamati secara simultan dan memiliki keterkaitan satu sama lain. Karakteristik utama distribusi normal multivariat ditentukan oleh dua parameter, yaitu vektor rata-rata dan matriks kovariansi , yang secara bersama-sama menggambarkan kecenderungan pusat data serta pola keragaman dan hubungan antarvariabel.

Keberadaan distribusi normal multivariat sangat fundamental dalam analisis multivariat, karena banyak metode statistik multivariat inferensial didasarkan pada asumsi bahwa data mengikuti distribusi normal multivariat. Metode seperti analisis diskriminan, MANOVA, regresi multivariat, dan pengujian hipotesis vektor rata-rata maupun matriks kovariansi memerlukan asumsi normalitas multivariat agar hasil analisis yang diperoleh bersifat sahih dan dapat diinterpretasikan secara tepat.

Selain itu, distribusi normal multivariat memiliki sejumlah sifat matematis yang penting, seperti kestabilan terhadap transformasi linier dan keterkaitannya dengan distribusi marginal dan kondisional. Sifat-sifat ini menjadikan distribusi normal multivariat sebagai landasan teoritis yang kuat dalam pengembangan dan penerapan berbagai teknik analisis multivariat.

Oleh karena itu, bab ini membahas konsep dasar distribusi normal multivariat, meliputi definisi, fungsi kepadatan peluang, parameter-parameter distribusi, serta sifat-sifat pentingnya. Pemahaman terhadap distribusi normal

multivariat diharapkan dapat memberikan dasar teoritis yang kuat dalam mempelajari dan menerapkan metode-metode analisis multivariat pada bab-bab selanjutnya.

3.2 Model Distribusi Normal Multivariat

Distribusi normal multivariat adalah bentuk generalisasi dari distribusi normal univariat. Suatu variabel acak yang memiliki distribusi normal univariat dengan rata- rata μ dan σ^2 variansi memiliki fungsi peluang seperti di bawah ini.

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{\left(\frac{x-\mu}{\sigma}\right)^2}{2}\right]; -\infty \leq x \leq +\infty \qquad (3.1)$$

Suatu variabel acak yang berdimensi p yaitu $\boldsymbol{X}$ = ($X1$, $X2$, ..., Xp)' berdistribusi normal multivariat dengan parameter rata-rata $\boldsymbol{\mu}$ = (μ_1, μ_1, ..., μ_p,)' dan matriks variansi-kovariansi $\boldsymbol{\Sigma}$ memiliki fungsi peluang normal multivariat sebagai berikut.

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right]; -\infty \leq x_i \leq +\infty \qquad (3.2)$$

Suatu variabel acak memiliki distribusi normal multivariat dapat dituliskan dalam bentuk $\boldsymbol{X} \sim N_p$ ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$).

Keterangan :

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \text{ dan } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Contoh

Misalkan suatu variabel acak berdistribusi bivariat dengan parameter sebagai berikut. $E(X_1) = \mu_1$; $E(X_2) = \mu_2$; $Var\ (X_1) = \sigma\ 11$; $Var\ (X_2) = \sigma\ 22$; $Cov\ (X_1 X_2) = \sigma\ 12 = \sigma\ 21$ dan $Corr\ (X_1 X_2) = \rho_{12}$ sehingga, vektor rata-rata dan

matriks variansi-kovariansi untuk data bivariat adalah :

$$\boldsymbol{\mu} = \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} \text{ dan } \boldsymbol{\Sigma} = \begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}$$

Buatlah fungsi distribusi normal bivariate!

Jawab

Ingat kembali rumus korelasi variabel pertama dan kedua adalah :

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$$
$$\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \qquad (3.3)$$

Selanjutnya, dicari matriks invers dari $\boldsymbol{\Sigma}$ yaitu

$$\boldsymbol{\Sigma^{-1}} = \frac{1}{\sigma_{11}\sigma_{22} - (\sigma_{12})^2}\begin{bmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{21} & \sigma_{11}\end{bmatrix}$$

Dengan menggunakan rumus korelasi (2.3), matriks invers di atas dapat diubah menjadi :

$$\boldsymbol{\Sigma^{-1}} = \frac{1}{\sigma_{11}\sigma_{22} - (\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}})^2}\begin{bmatrix}\sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11}\end{bmatrix}$$

$$= \frac{1}{\sigma_{11}\sigma_{22} - {\rho_{12}}^2\sigma_{11}\sigma_{22}}\begin{bmatrix}\sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11}\end{bmatrix}$$

$$= \frac{1}{\sigma_{11}\sigma_{22}(1 - {\rho_{12}}^2)}\begin{bmatrix}\sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11}\end{bmatrix} \qquad (3.4)$$

Perhatikan bentuk persamaan di bawah ini.

$$(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma^{-1}}(\boldsymbol{x} - \boldsymbol{\mu}) = (x_1 - \mu_1, x_2 - \mu_2)\boldsymbol{\Sigma^{-1}}\begin{pmatrix}x_1 - \mu_1\\ x_2 - \mu_2\end{pmatrix} \qquad (3.5)$$

Substitusikan persamaan (3.4) ke dalam persamaan (3.5) diperoleh:

$$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \frac{1}{1-{\rho_{12}}^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right] \quad (3.6)$$

Ingat bahwa determinan dari matrik $\boldsymbol{\Sigma}$ ($p = 2$) adalah :

$$|\boldsymbol{\Sigma}| = \sigma_{11}\sigma_{22}(1-{\rho_{12}}^2) \quad (3.6)$$

Substitusikan persamaan (3.6) dan (3.7) ke dalam persamaan (3.2), maka fungsi distribusi normal bivariat adalah :

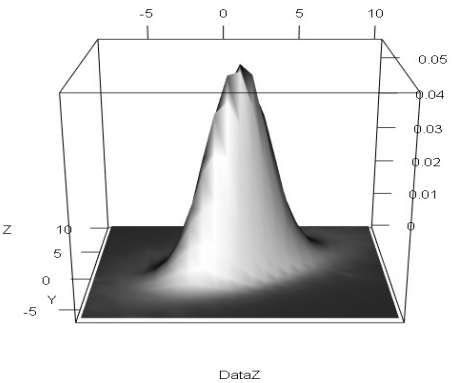
$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-{\rho_{12}}^2)}} \times \exp\left\{-\frac{1}{2(1-{\rho_{12}}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\}$$

Persamaan di atas merupakan perluasan dari persamaan (3.2) untuk data bivariat.

Perhatikan bahwa apabila tidak ada korelasi antara variabel X_1 dan X_2, artinya $\rho_{12} = 0$. Maka fungsi distribusi normal bivariat adalah perkalian antara fungsi normal univariat variabel X_1 dan X_2 yaitu:

$$f(x_1,x_2) = f(x_1)f(x_2) \text{ , jika } X_1 \text{ dan } X_2 \text{ independen}$$

Bentuk dari fungsi distribusi normal multivariate dengan dua variabel dapat digambarkan seperti di bawah ini.



Gambar 5. Distribusi Normal Multivariat (p = 2)

3.3 Sifat-Sifat Distribusi Normal Multivariat

Misalkan $X_1, X_2, ..., X_p$ adalah sampel acak berukuran n yang berasal dari distribusi normal multivariat dengan rata-rata $\boldsymbol{\mu}$ dan matriks variansi-kovariansi $\boldsymbol{\Sigma}$. maka :

1. $\overline{\boldsymbol{X}}$ berdistribusi normal multivariat dengan rata-rata $\boldsymbol{\mu}$ dan matriks variansi-kovariansi $\frac{1}{n}\boldsymbol{\Sigma}$ atau $\overline{\boldsymbol{X}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$;
2. $(n-1)\boldsymbol{S}$ berdistribusi Wishart dengan derajat bebas $(n-1)$;
3. $\overline{\boldsymbol{X}}$ dan $\boldsymbol{S}$ saling independen (bebas).

Teorema Limit Pusat

Diberikan sampel acak berukuran besar yaitu $X_1, X_2, ..., X_p$ yang berasal dari sembarang distribusi dengan rata-rata $\boldsymbol{\mu}$ dan matriks variansi-kovariansi $\boldsymbol{\Sigma}$. Maka $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ akan menghampiri distribusi normal multivariat dengan rata-rata $\mathbf{0}$ dan matriks variansi-kovariansi $\boldsymbol{\Sigma}$ atau $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$.

Lebih lanjut, menurut Johnson dan Wichern (2007), sifat-sifat distribusi sampling dari data yang berasal dari distribusi normal multivariat seperti di bawah ini.

- Teorema

  Jika $\boldsymbol{X}$ berdistribusi normal multivariat N_p $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ maka kombinasi linear dari variabel $\boldsymbol{a'X} = a_1X_1 + a_2X_2 + ... + a_pX_p$ akan berdistribusi

normal N_p ($\boldsymbol{\alpha\mu}$, $\boldsymbol{\alpha'\Sigma a}$).

- Teorema
  Jika berdistribusi normal multivariat $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ maka $(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ akan berdistribusi khi kuadrat X^2dengan p derajat bebas.
- Teorema
  Jika diambil sampel acak $X_1, X_2, ..., X_p$ dari distribusi normal multivariat dengan rata- rata $\boldsymbol{\mu}$ dan matriks variansi-kovariansi $\boldsymbol{\Sigma}$, maka penduga maksimum likelihood untuk $\boldsymbol{\mu}$ dan $\boldsymbol{\Sigma}$ berturut-turut adalah:
  $$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} \text{ dan } \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})' = \frac{n-1}{n}\boldsymbol{S}$$

3.4 Prosedur Analisis Data Berdistribusi Normal Multivariat dengan R

Pada sub bab ini akan dibahas mengenai cara pengolahan data distribusi normal multivariat dengan menggunakan R.

1. Menguji Data Normal Multivariat

Dalam analisis multivariat, uji normalitas multivariat digunakan untuk menilai apakah vektor variabel mengikuti distribusi normal multivariat. Berbeda dengan uji normalitas univariat, uji ini mempertimbangkan hubungan antarvariabel sekaligus.

Berikut jenis-jenis uji distribusi normal multivariat yang umum digunakan:

a. Uji Mardia

Uji Mardia merupakan salah satu uji normalitas multivariat yang paling banyak digunakan. Uji ini didasarkan pada dua ukuran, yaitu skewness multivariat dan kurtosis multivariat.

Uji skewness multivariat mengukur tingkat kemencengan distribusi data secara simultan, sedangkan uji kurtosis multivariat mengukur tingkat keruncingan distribusi. Di bawah hipotesis nol bahwa data berdistribusi normal multivariat, statistik uji skewness mengikuti distribusi X^2, sedangkan statistik uji kurtosis mengikuti distribusi normal baku.

1) Statistik Uji Skewnes Multivariat

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(x_i - \bar{x})^T S^{-1}(x_i - \bar{x})]^3$$

2) Statistik Uji Kurtosis Multivariat

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}[(x_i - \bar{x})^T S^{-1}(x_i - \bar{x})]^3$$

b. Uji Henze–Zirkler

Uji Henze–Zirkler merupakan uji normalitas multivariat yang relatif modern dan memiliki kekuatan uji yang baik untuk berbagai ukuran sampel. Statistik uji Henze–Zirkler (HZ) dihitung berdasarkan kombinasi jarak Mahalanobis antara pengamatan dan jarak pengamatan terhadap rataan data.

Statistik uji :

$$HZ = n\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} exp\left(-\frac{\beta^2}{2}D_{ij}^2\right) - 2(1+\beta^2)^{-p/2}\frac{1}{n}\sum_{i=1}^{n} exp\left(-\frac{\beta^2}{2(1+\beta^2)}D_i^2\right) + (1+\beta^2)^{-p/2}\right]$$

dengan :

- $D_i^2 = (x_i - \overline{x})^T S^{-1}(x_i - \overline{x})$ adalah jarak Mahalanobis terhadap rataan,
- p adalah banyaknya variabel,
- n adalah ukuran sampel,
- β adalah parameter pemulus (*smoothing parameter*) yang umumnya ditentukan sebagai :

Uji Henze–Zirkler sering direkomendasikan sebagai alternatif uji Mardia karena lebih stabil terhadap pencilan dan tidak terlalu sensitif terhadap ukuran sampel.

c. Uji Royston

Uji Royston merupakan pengembangan dari uji Shapiro–Wilk untuk kasus multivariat. Uji ini mengombinasikan statistik uji Shapiro–Wilk dari masing-masing variabel dengan memperhitungkan korelasi antarvariabel.

Statistik uji Royston dinyatakan dalam bentuk :

$$H = \sum_{j=1}^{p} Z_j^2$$

dengan :

- Z_i adalah skor baku hasil transformasi dari statistik Shapiro–Wilk variabel ke-*j*,
- *n* adalah jumlah variabel.

Uji Royston cocok digunakan pada data dengan dimensi relatif kecil dan ukuran sampel terbatas.

d. Uji Doornik–Hansen

Uji Doornik–Hansen merupakan perluasan dari uji Jarque–Bera ke dalam konteks multivariat. Uji ini menguji skewness dan kurtosis secara simultan dan banyak digunakan dalam bidang ekonometrika multivariat.

$$H = \sum_{j=1}^{p} \left(Z_{1j}^2 + Z_{2j}^2\right)$$

Dengan

- Statistik *skewness* terstandarisasi (Z_{1j}^2)
- Statistik kurtosis terstandarisasi (Z_{2j}^2)

Semua uji tersebut memiliki rumusan hipotesis sebagai berikut

- H_0: Data berdistribusi normal multivariat
- H_1: Data tidak berdistribusi normal multivariat.

Keputusan uji ditentukan berdasarkan nilai *p-value* :

- Jika *p-value* > α (misalnya 0,05), maka H_0 gagal ditolak, sehingga data dapat dianggap berdistribusi normal multivariat.
- Jika *p-value* ≤ α, maka H_0 ditolak, yang berarti data tidak berdistribusi normal multivariat

Selain uji formal, normalitas multivariat juga dapat dievaluasi secara grafis, antara lain melalui :

1) **Plot Q–Q Chi-Square**, yaitu plot antara jarak Mahalanobis dan kuantil distribusi X^2. Pola titik yang mengikuti garis lurus

mengindikasikan normalitas multivariat.

2) **Scatterplot matrix**, di mana pola sebaran berbentuk elips menunjukkan indikasi normal multivariat.

Pendekatan grafis bersifat pendukung dan sebaiknya digunakan bersama uji statistik formal.

Uji normalitas multivariat di R Studio tersedia pada menggunakan perintah `mvn()` yang berada pada `MVN package`. Package MVN menyediakan berbagai uji (Mardia, Royston, Henze-Zirkler, Doornik-Hansen) dalam satu fungsi.

Contoh

Berikut data penjualan suatu barang selama 20 periode di perusahaan Prima Jaya.

Tabel 5. Data Penjualan Perusahaan Prima Jaya

Periode	Penjualan	Biaya	Promosi
1	247	33	179
2	246	33	180
3	247	34	180
4	247	34	182
5	246	34	181
6	246	35	180
7	247	35	180
8	247	35	180
9	247	36	181
10	246	35	181
11	246	34	180
12	246	36	178
13	245	35	180
14	248	35	179
15	246	34	178
16	244	33	181
17	245	33	179
18	246	33	178
19	246	35	180

Periode	Penjualan	Biaya	Promosi
20	245	32	180

Misalkan data di atas disimpan dalam file dengan nama Jual.csv. Berikut cara melakukan uji Mardia dan Henze–Zirkler (HZ) untuk menguji data normal multivariat.

```
> install.packages("MVN")#instal package MVN jika belum ada
> library(MVN) #panggil package MVN
> Jual=read.csv2("Jual.csv")
> # Uji Normalitas Mardia
> result = mvn(Jual, mvn_test = "mardia")
> # Menampilkan hasil
> result$multivariate_normality
             Test Statistic  p.value      Method      MVN
1 Mardia Skewness     2.676    0.988  asymptotic ✓ Normal
2 Mardia Kurtosis    -0.715    0.474  asymptotic ✓ Normal
```

Berdasarkan hasil uji Mardia, diperoleh dua statistik pengujian, yaitu skewness multivariat dan kurtosis multivariat. Untuk uji skewness, nilai statistik sebesar 2,676 dengan *p-value* 0,988, sedangkan untuk uji kurtosis diperoleh nilai statistik –0,715 dengan *p-value* 0,474.

Karena kedua *p-value* lebih besar dari tingkat signifikansi $\alpha = 0{,}05$, maka hipotesis nol gagal ditolak. Dengan demikian, data dapat dinyatakan berdistribusi normal multivariat baik dari sisi kemencengan maupun keruncingan. Hasil ini menunjukkan bahwa asumsi normalitas multivariat terpenuhi.

Sedangkan uji normalitas multivariat dengan menggunakan uji Henze–Zirkler (HZ) dilakukan dengan perintah sebagai berikut :

```
> # Uji Normalitas Henze-Zirkler (HZ)
> result = mvn(Jual, mvn_test = "hz")
> # Menampilkan hasil
> result$multivariate_normality
           Test  Statistic  p.value       Method      MVN
1 Henze-Zirkler      0.396     0.92   asymptotic  ✓ Normal
```

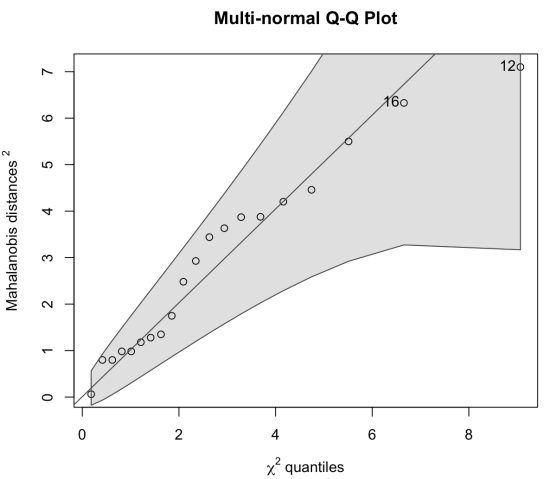
Berdasarkan hasil uji Henze–Zirkler, diperoleh nilai statistik uji sebesar 0,396 dengan *p-value* 0,92. Karena nilai *p-value* tersebut lebih besar dari

tingkat signifikansi $\alpha = 0{,}05$, maka hipotesis nol gagal ditolak.

Dengan demikian, dapat disimpulkan bahwa data memenuhi asumsi normalitas multivariat. Hasil ini menunjukkan bahwa distribusi bersama dari variabel-variabel dalam data tidak berbeda secara signifikan dari distribusi normal multivariat teoritis.

Untuk membuat Q–Q plot normalitas multivariat di R menggunakan paket RVAideMemoire, gunakan fungsi `mqqnorm()`. Fungsi ini digunakan untuk memvisualisasikan apakah data mengikuti distribusi normal multivariat dengan cara memplot jarak Mahalanobis yang telah diurutkan terhadap kuantil distribusi chi-kuadrat.

```
> install.packages("RVAideMemoire")
> library(RVAideMemoire)
> mqqnorm(Jual, main = "Multi-normal Q-Q Plot")
[1] 12 16
```



Hasil menunjukkan bahwa pengamatan ke-12 dan ke-16 teridentifikasi sebagai titik yang paling menyimpang dari pola garis lurus pada Q–Q plot. Namun demikian, interpretasi Q–Q plot tidak hanya bergantung pada adanya beberapa titik ekstrem, melainkan pada pola keseluruhan. Jika sebagian besar titik mengikuti garis lurus dan hanya sedikit pengamatan yang menyimpang

ringan, maka data masih dapat dianggap mendekati normal multivariat.

Dalam konteks ini, keberadaan observasi ke-12 dan ke-16 tidak langsung mengindikasikan pelanggaran serius terhadap normalitas multivariat, terutama karena :

1) Uji formal Mardia dan Henze–Zirkler sebelumnya menunjukkan *p-value* > 0,05.
2) Tidak terdeteksi outlier multivariat yang signifikan berdasarkan uji jarak Mahalanobis.

2. Membangkitkan Data Berdistribusi Normal Multivariat

Dalam analisis multivariat, sering kali diperlukan data simulasi yang mengikuti distribusi normal multivariat, baik untuk keperluan pembelajaran, simulasi metode statistik, maupun pengujian kinerja suatu prosedur analisis. Salah satu paket di R yang menyediakan fasilitas untuk membangkitkan data tersebut adalah `package MASS`. Paket ini menyediakan fungsi `mvrnorm`() yang digunakan untuk membangkitkan data acak dari distribusi normal multivariat.

Fungsi `rmvrnorm()` memerlukan tiga komponen utama, yaitu :

a. Ukuran sampel (n), yang menyatakan banyaknya pengamatan yang akan dibangkitkan,
b. Vektor rataan ($\boldsymbol{\mu}$), yang merepresentasikan nilai rata masing-masing variabel,
c. Matriks kovariansi ($\boldsymbol{\Sigma}$), yang menggambarkan variansi dan kovariansi antarvariabel.

Secara umum, data yang dibangkitkan dengan fungsi ini mengikuti model:

```
rmvnorm(n, mean = rep(0, nrow(sigma)), sigma =
       diag(length(mean)),  method=c("eigen",
       "svd", "chol")…)

Keterangan argumen :
n = banyaknya pengamatan
mean = vektor rata – rata
sigma = matriks kovariansi
method = metode untuk mencari matriks
dekomposisi, pilihannya menggunakan
eigenvalue ("eigen"), singular value
decomposition ("svd") atau Cholesky
decomposition ("chol").
```

Contoh

Misalkan ingin membangkitkan sampel berukuran 20 yang memiliki distribusi normal multivariat dengan rata-rata $\mu = \begin{bmatrix}1\\2\end{bmatrix}$ dan matriks kovariansi $\Sigma = \begin{bmatrix}4 & 2\\2 & 3\end{bmatrix}$, maka perintah dalam R adalah :

```
> library(MASS) #panggil paket MASS
> mu = c(1,2) #Membuat vektor rata – rata
> S = matrix(c(4,2,2,3),ncol=2) #Membuat matriks
kovariansi
> mu
[1] 1 2
> S
     [,1]  [,2]
[1,]   4     2
[2,]   2     3

> X = rmvnorm(n=20, mean=mu, sigma=S) #membangkitkan data
> X
             [,1]         [,2]
 [1,]  -2.5413701   -2.4302097
 [2,]   0.6238445    2.4159387
 [3,]  -0.2595549    1.6338186
 [4,]   1.0189431    4.0692178
 [5,]  -0.1435597    0.4150318
```

```
 [6,]   4.1725205    4.8089256
 [7,]  -0.6336919    1.9599231
 [8,]   2.8774370    2.4376712
 [9,]   2.0614234    1.8890704
[10,]   1.4194405   -0.3743147
[11,]   2.0250880    0.3201254
[12,]  -3.1328888    0.5186247
[13,]   0.9992352    1.7438064
[14,]   0.7917199    2.8112092
[15,]  -0.3654692    2.5875395
[16,]   2.8757324    1.5305870
[17,]   3.9755784    4.4089092
[18,]  -0.5880232    0.3131427
[19,]  -2.0185290    1.3426722
[20,]  -0.1987412   -0.8853036
```

Sekarang perhatikan kembali data sampel sebelumnya, kemudian dari sampel yang diperoleh dihitung vektor rata-rata dan matriks kovariansinya melalui cara berikut ini.

```
> colMeans(X)
[1] 0.6479567 1.5758193
> var(X)
          [,1]       [,2]
[1,]  4.007476   2.178010
[2,]  2.178010   3.188055
```

Perhatikan bahwa estimasi vektor rata-rata dan matriks kovariansinya tidak sama dengan parameternya tetapi hampir mendekati. Bagaimana jika ukuran sampelnya diperbesar menjadi 1000 dan 1.000.000?

```
> Y=mvrnorm(1000,mu,S) #membangkitkan data normal
multivariate dengan ukuran sampel 1000

> colMeans(Y)
[1] 1.109254 2.048369
> var(Y)
           [,1]     [,2]
[1,]   3.750351 1.798650
[2,]   1.798650 2.969581

> Z=mvrnorm(1000000,mu,S) #membangkitkan data normal
multivariate
Dengan ukuran sampel 1.000.000

> colMeans(Z)
[1] 1.000961 2.002571
> var(Z)
           [,1]       [,2]
[1,]   3.997107   1.999697
[2,]   1.999697   2.998300
```

Hasil simulasi sederhana di atas memperlihatkan bahwa semakin besar ukuran sampel maka estimasi vektor rata-rata dan matriks kovariansi mendekati nilai parameter yang sesungguhnya.

3. Memvisualisasikan Data Berdistribusi Normal Multivariat

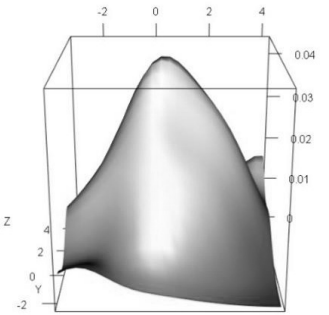
Pada subbagian sebelumnya telah dipelajari bagaimana membangkitkan data yang berdistribusi normal dengan parameter μ dan Σ tertentu. Selanjutnya akan dibahas mengenai bentuk distribusi normal multivariat (khususnya di sini *bivariate normal distribution*).

Gambar kurva fungsi normal multivariat akan dibentuk dalam tiga dimensi dengan menggunakan paket `rgl`. Gambar tiga dimensi dengan menggunakan perintah `pers3d()`. Pertama instal paket rgl.

```
> install.packages("rgl")
> library(rgl)
```

Setelah terinstal, proses membuat gambar tiga dimensi untuk fungsi distribusi normal multivariat dapat dilakukan dengan cara berikut ini.

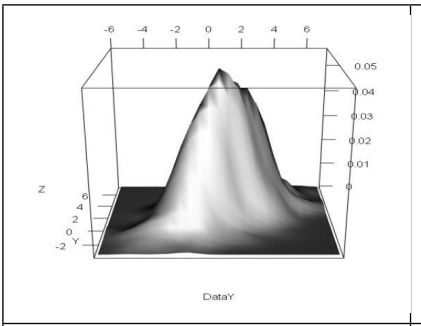
```
> DataX=kde2d(X[,1], X[,2]) #fungsi kernel dua dimensi
> col2 <- heat.colors(length(DataX$z))[rank(DataX$z)]
#pilihan warna
> persp3d(x=DataX, col = col2) #gambar 3 dimensi
```



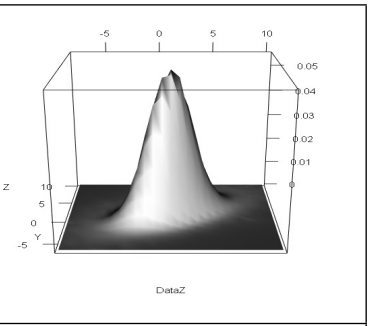
Bentuk gambar tiga dimensi untuk sampel yang berukuran 20 belum nampak seperti gambar lonceng atau bel. Sekarang dibandingkan bagaimana bentuk kurva untuk sampel berukuran n = 1000 dan n = 1.000.000. Berikut prosedurnya.

```
##Membuat gambar tiga dimensi n = 1000 (Y)
> DataY=kde2d(Y[,1], Y[,2])
> col2 <- heat.colors(length(DataY$z))[rank(DataY$z)]
> persp3d(x=DataY, col = col2)

##Membuat gambar tiga dimensi n = 1.000.000 (Z)

> DataZ=kde2d(Z[,1], Z[,2])
> col2 <- heat.colors(length(DataZ$z))[rank(DataZ$z)]
> persp3d(x=DataZ, col = col2)
```

Hasil outputnya dapat dilihat pada gambar di bawah ini



n = 1000



n = 1.000.000

Terlihat dari gambar di atas bahwa semakin besar ukuran sampel, bentuk kurva semakin membentuk gambar lonceng atau bel.

3.5 Latihan

1. Diberikan suatu variabel acak $\boldsymbol{X}$ berdistribusi bivariat dengan parameter sebagai berikut: $\mu_1 = 1; \mu_2 = 3; \sigma_{11} = 2; \sigma_{22} = 1; \rho = 12 = -0{,}80$
   a. Tuliskan fungsi normal bivariatnya!
   b. Tuliskan bentuk kuadrat jarak dari $(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ ke dalam bentuk fungsi kuadrat x_1 dan x_2!
2. Diberikan suatu variabel acak $\boldsymbol{X}$ berdistribusi bivariat dengan parameter sebagai berikut: $\mu_1 = 0; \mu_2 = 3; \sigma_{11} = 2; \sigma_{22} = 1; \rho = 12 = -0{,}50$
   a. Tuliskan fungsi normal bivariatnya!
   b. Tuliskan bentuk kuadrat jarak dari $(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (x - \mu)$ ke dalam bentuk fungsi kuadrat x_1 dan x_2!
3. Diberikan suatu variabel acak $\boldsymbol{X}$ berdistribusi bivariat dengan parameter sebagai berikut: $\mu_1 = 0; \mu_2 = 2; \sigma_{11} = 1; \sigma_{22} = 2; \rho = 12 = -0{,}40$
   a. Tuliskan fungsi normal bivariatnya!
   a. Tuliskan bentuk kuadrat jarak dari ke dalam bentuk fungsi kuadrat dan !

4. Berikut ini adalah data jumlah penduduk berdasarkan jenis kelamin.

Klasifikasi Jumlah Penduduk Di Jawa Tengah berdasarkan Umur		
Umur	Laki-Laki	Perempuan
0	281.795	266.667
1	278.832	263.491
2	277.803	261.509
3	281.321	265.525
4	275.477	258.862
5	272.990	258.363
6	281.602	268.261
7	290.830	274.502
8	290.119	275.740
9	316.362	300.601
10	330.575	312.715
11	283.874	267.390
12	294.229	279.220
13	308.395	291.985
14	312.337	294.314
15	306.405	284.670
16	289.910	269.781
17	285.370	266.641
18	267.851	254.492
19	247.411	240.281
20	246.523	250.734
21	223.106	230.943
22	223.914	231.082
23	223.939	234.061
24	235.694	245.818
25	249.717	259.097
26	242.368	255.184
27	264.101	275.901
28	255.161	265.137
29	258.043	263.702

Ujilah apakah data di atas berdistribusi normal multivariat!

5. Bangkitkan data berdistribusi normal multivariat dengan n = 100, 1000 dan 10.000 apabila parameter seperti di bawah ini.

$$\mu = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \text{ dan } \Sigma = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 3 & 0 \\ 1 & 0 & 3 \end{bmatrix}$$

6. Gambarkan bentuk tiga dimensi kurva pada soal no.5! Analisis!

BAB

IV

INFERENSI VEKTOR RATA-RATA

4.1 Pendahuluan

Statistika inferensi atau statistika induktif adalah cabang statistika yang bertujuan untuk membuat kesimpulan, prediksi, dan generalisasi suatu populasi berdasarkan sampel. Statistika inferensia melibatkan penggunaan teori probabilitas dan model statistik untuk menguji hipotesis dan memperkirakan karakteristik populasi yang tidak dapat diamati seluruhnya.

Dalam statistika inferensi ada dua pendekatan, yaitu pendugaan (estimasi) dan pengujian hipotesis. Pendugaan atau estimasi statistik adalah prosedur untuk memperkirakan nilai parameter populasi yang tidak diketahui berdasarkan data dari sampel. Tujuannya adalah untuk mendapatkan gambaran tentang populasi melalui data yang terbatas. Pendugaan ini dibagi menjadi dua jenis utama: 1). pendugaan titik yang menghasilkan satu nilai untuk parameter populasi dan 2). pendugaan interval yang memberikan rentang nilai (interval) di mana parameter populasi kemungkinan besar berada pada tingkat kepercayaan tertentu. Pengujian hipotesis statistik adalah sebuah metode inferensi statistik untuk menguji kebenaran sebuah pernyataan (hipotesis) dengan menganalisis data sampel yang bertujuan untuk memutuskan apakah data tersebut memberikan bukti yang cukup untuk menolak hipotesis nol (H_0) atau sebaliknya.

Pada bab ini akan membahas pengujian hipotesis untuk rata-rata dari beberapa variabel sekaligus atau uji hipotesis vektor rata-rata. Inferensi tentang vektor rata-rata, sebenarnya hampir sama pada uji hipotesis tentang rata-rata dengan menggunakan uji *t* atau uji *Z*. Misalnya suatu penelitian ingin mengetahui tingkat pendapatan, pengeluaran, pendidikan penduduk di provinsi DI Yogyakarta. Pada uji *t* atau uji *Z* dilakukan pengujian untuk setiap variabel-variabel tersebut. Sedangkan pada uji tentang vektor rata-rata, pengujian dilakukan secara serentak untuk beberapa variabel yang dikaji.

4.2 Uji Vektor Rata-rata untuk Satu Populasi

Untuk melakukan uji terhadap vektor rata-rata, perhatikan teorema berikut ini.

Teorema ***(Johnson dan Winchern, 2007)***

Misalkan, suatu sampel acak $X_1, X_2, ..., X_n$ diambil dari distribusi normal multivariat $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ maka :

$$\boldsymbol{T^2 = n(\bar{X} - \mu_0)' S^{-1}(\bar{X} - \mu_0)} \text{ akan berdistribusi } \frac{(n-1)p}{n-p} F_{p;(n-p)}$$

Keterangan:

1. $\boldsymbol{\mu_0} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix}; \boldsymbol{\bar{X}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$
2. $\boldsymbol{S}^{-1}$ adalah invers matriks variansi kovariansi
3. n = ukuran sampel
4. $F_{p;(n-p)}$ adalah distribusi F dengan derajat bebas p dan $n - p$.

Statistik $\boldsymbol{T^2}$ disebut dengan $\boldsymbol{T^2}$ **Hoteling** karena disesuaikan dengan penemu distribusi tersebut yaitu Harold Hoteling. Prosedur yang harus dilakukan dalam melakukan inferensi atau pengujian terhadap vektor rata-rata adalah :

1. Tentukan hipotesisnya

   $\boldsymbol{H_0 : \mu = \mu_0}$ versus $\boldsymbol{H_1 : \mu \neq \mu_0}$
2. Pilih tingkat signifikansinya α = 1%, 5% atau 10%
3. Statistik Uji

   $$\boldsymbol{T^2 = n(\bar{X} - \mu_0)' S^{-1}(\bar{X} - \mu_0)}$$
4. Daerah kritis

   $$\boldsymbol{T^2} > \frac{(n-1)p}{n-p} F_{p;(n-p)}$$
5. Hitung $\boldsymbol{T^2}$ sampel
6. Kesimpulan: Tolak H_0 apabila $\boldsymbol{T^2} > \frac{(n-1)p}{n-p} F_{p;(n-p)}$

Contoh

Untuk contoh perhitungan, digunakan data pendapatan di perusahaan Jaya. Data berikut ini adalah data pendapatan perusahaan Jaya dalam empat tahun.

	Tahun ke-1	Tahun ke-2	Tahun ke-3	Tahun ke-4
Pendapatan (x 100 juta)	45	50	43	42
Banyaknya karyawan	4	5	4	3

Lakukan uji T^2 **Hoteling,** $H_0 : \boldsymbol{\mu} = [40, 2]'$ versus $H_1 : \boldsymbol{\mu} \neq [40, 2]'$. Gunakan taraf signifikansi sebesar 1%!

Prosedur untuk melakukan uji **Hoteling** adalah :

1. Tentukan hipotesisnya
   $H_0 : \boldsymbol{\mu} = [40, 2]'$ versus $H_1 : \boldsymbol{\mu} \neq [40, 2]'$.
2. Pilih tingkat signifikansinya
3. Statistik Uji
$$\boldsymbol{T^2 = n(\bar{X} - \mu_0)' S^{-1}(\bar{X} - \mu_0)}$$
4. Daerah kritis
$$\boldsymbol{T^2} > \frac{(n-1)p}{n-p} F_{p;(n-p)}$$
   Berdasarkan tabel *F* dengan derajat bebas (2; 2) = 9, $n = 4$ dan $p = 2$, maka daerah penolakan H_0 adalah $\boldsymbol{T^2} > \frac{3(2)}{2} 9$ yaitu $\boldsymbol{T^2 > 27}$
5. Hitung $\boldsymbol{T^2}$ sampel
   Hasil perhitungan sebelumnya diperoleh: $\bar{\boldsymbol{x}} = \begin{bmatrix} 45 \\ 4 \end{bmatrix}$ dan $\boldsymbol{S} = \begin{bmatrix} 12{,}67 & 2{,}67 \\ 2{,}67 & 0{,}67 \end{bmatrix}$
   Cari invers matriks $\boldsymbol{S}$ yaitu
$$\boldsymbol{S^{-1}} = \frac{1}{12.67(0.67) - (2.67)^2} \begin{bmatrix} 0{,}67 & -2{,}67 \\ -2{,}67 & 0{,}67 \end{bmatrix} = \begin{bmatrix} 0.5 & -2 \\ -2 & 9.5 \end{bmatrix}$$
   Selanjutnya hitung
$$\boldsymbol{T^2} = 4 \left(\begin{bmatrix} 45 \\ 4 \end{bmatrix} - \begin{bmatrix} 40 \\ 2 \end{bmatrix} \right)' \begin{bmatrix} 0{,}67 & -2{,}67 \\ -2{,}67 & 0{,}67 \end{bmatrix} \left(\begin{bmatrix} 45 \\ 4 \end{bmatrix} - \begin{bmatrix} 40 \\ 2 \end{bmatrix} \right)$$
$$= 4 \, [5, \quad 2] \begin{bmatrix} 0{,}67 & -2{,}67 \\ -2{,}67 & 0{,}67 \end{bmatrix} \left(\begin{bmatrix} 5 \\ 2 \end{bmatrix} \right) = 10{,}5$$
6. Kesimpulan:
   Gagal tolak H_0 karena $\boldsymbol{T^2} > \frac{(n-1)p}{n-p} F_{p;(n-p)}$ artinya = $[40, 2]'$.

Dalam R perintah untuk menghitung ukuran data multivariat dapat menggunakan perintah berikut ini.

1. Menghitung vektor rata-rata sampel adalah: `colMeans()`

2. Menghitung matriks kovariansi sampel adalah: `var()`
3. Menghitung matriks korelasi sampel adalah: `cor()`

4.3 Uji T² Hotelling (*Hotelling T² Test*) untuk Satu Populasi

Inferensi tentang vektor rata-rata, sebenarnya hampir sama dengan melakukan uji hipotesis tentang rata-rata (univariat). Pada uji tentang vektor rata-rata, pengujian dilakukan secara serentak untuk semua variabel yang dikaji. Misalkan, penelitian mengenai kemampuan/prestasi siswa di sekolah. Di dalam pengujian mean vektor maka dilakukan pengujian terhadap variabel-variabel kemampuan siswa secara simultan, misalkan prestasi belajar, motivasi, kreativitas, kemandirian. Prosedur yang harus dilakukan dalam melakukan inferensi atau pengujian terhadap vektor rata-rata seperti di bawah ini.

1. Tentukan hipotesis
   $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ (tidak ada perbedaan)
   $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ (ada perbedaan)
2. Pilih tingkat signifikansinya

$$T^2 = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \boldsymbol{S}^{-1} (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim \frac{(n-1)p}{(n-p)} F_{p;(n-p)}$$

   Keterangan :
   - $\overline{\boldsymbol{x}} = \widehat{\boldsymbol{\mu}}$ yaitu vektor rata-rata yang dihitung dari sampel
   - $\boldsymbol{S} = \widehat{\boldsymbol{\Sigma}}$ yaitu matriks kovariansi yang dihitung dari sampel
   - adalah ukuran sampel
   - adalah hipotesis vektor rata-rata

   Apabila variabel acak $\boldsymbol{X}$ berdistribusi Normal Multivariat dengan parameter $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ atau $\boldsymbol{X} \sim NM(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ maka statistik $T^2 \sim \frac{p(n-1)}{(n-p)} F_{\alpha(p;n-p)}$
3. Keputusan adalah :
   Jika $\frac{(n-p)}{p(n-1)} T^2 > F_{\alpha(p;n-p)}$ maka tolak H_0 atau jika menggunakan nilai *p-value* adalah menolak H_0 apabila *p-value(sig)* < α.

Perintah R untuk melakukan uji Hotelling terdapat di *package* ICNSP. Pertama kali instal *package* tersebut melalui perintah berikut ini.

```
> install.packages("ICSNP")
> library("ICSNP")
```

Setelah menginstal paket ICSNP, selanjutnya lakukan uji Hotelling dengan menggunakan fungsi `HotellingsT2()`. Sintak perintahnya seperti di bawah ini.

```
HotellingsT2(X, Y = NULL, mu = NULL, test = "f", ...)

Arguments
X = matriks data
Y = matriks data kedua jika melakukan uji 2 sampel,
mu = parameter hipotesis,()
test = "f" di sini menggunakan distribusi F
```

Misalkan pada contoh di atas akan dilakukan uji Hotelling dengan hipotesis berikut ini.

Hipotesis adalah $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ [0,0,0] vs $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ [0,0,0]'

Uji Hotelling dilakukan seperti di bawah ini.

```
> HotellingsT2(X)
      Hotelling's one sample T2-test

data: X
T.2 = 7.619, df1 = 3, df2 = 2, p-value = 0.1182
alternative hypothesis: true location is not equal to
c(0,0,0)
```

Output di atas menjelaskan bahwa nilai T^2 = 7,619 dan *p_value* sebesar 0,1182. Berdasakan nilai *p_value* dapat diputuskan tidak menolak H0 karena nilai *p_value* lebih besar dibandingkan taraf signifikansi sebesar 0,05. Oleh karena itu dapat disimpulkan vektor rata-rata ketiga variabel tersebut sama dengan nol.

Apabila hipotesis yang diujikan sebagai berikut.

$H_0 : \boldsymbol{\mu}$ = [2,2,2] vs $H_1 : \boldsymbol{\mu} \neq$ [2,2,2]'

```
> HotellingsT2(X,mu=c(2,2,2))

      Hotelling's one sample T2-test

data: X
T.2 = 3.3595, df1 = 3, df2 = 2, p-value = 0.2378
alternative hypothesis: true location is not equal to
c(2,2,2)
```

Output di atas menjelaskan bahwa nilai T^2 = 3,3595 dan *p_value* sebesar 0,2378. Berdasakan nilai *p_value* dapat diputuskan tidak menolak H_0 karena nilai *p_value* lebih besar dibandingkan taraf signifikansi sebesar 0,05. Oleh karena itu dapat disimpulkan vektor rata-rata ketiga variabel tersebut sama dengan $\boldsymbol{\mu}_0 = (2,2,2)'$.

4.4 Uji T^2 Hotelling (*Hotelling T^2 Test*) untuk dua Populasi

Uji Hotelling T^2 untuk dua populasi adalah uji statistik multivariat yang merupakan perluasan uji-*t* dua sampel untuk membandingkan perbedaan antara dua kelompok sampel dengan dua atau lebih variabel dependen secara bersamaan. Uji ini menguji hipotesis nol bahwa rata-rata vektor (rata-rata multivariat) dari dua himpunan data berbeda.

Sama halnya dalam uji *t* dua populasi, di dalam uji Hotelling T^2 dibedakan sampel yang berpasangan (*dependent*) dan tidak berpasangan (*independent*). Tabulasi data untuk dua populasi disajikan pada Tabel di bawah ini.

Tabel 6. Data untuk Uji Dua Sampel Berpasangan

Observasi	Sesudah				Sebelum			
	Var1	Var2	…	Var-*p*	Var1	Var2	…	Var-*p*
1	X_{111}	X_{121}	…	X_{1p1}	X_{211}	X_{221}	…	X_{2p1}
2	X_{112}	X_{122}	…	X_{1p2}	X_{212}	X_{222}	…	X_{2p2}
n	X_{11n}	X_{12n}	…	X_{1pn}	X_{21n}	X_{22n}	…	X_{2pn}

Keterangan :

X_{ijk} = Pengamatan pada populasi ke-*i*, variabel ke-*j* dan observasi

ke-k, dengan :

i = 1, 2, (banyaknya populasi)

j = 1, 2, …, p (banyaknya variabel)

k = 1,2, ..., n (banyaknya pengamatan)

1. Uji Hotelling T^2 Dibedakan Sampel yang Berpasangan (*Dependent*)

Untuk melakukan uji T^2 Hotelling untuk data berpasangan maka dihitung dulu selisih setiap variabelnya.

Tabel 7. Tabulasi Data Selisih Setiap Variabel

Observasi	Selisih			
	Var1	**Var2**	**…**	**Var-p**
1	$D_{11} = X_{111} - X_{211}$	$D_{21} = X_{121} - X_{221}$	…	$D_{p1} = X_{1p1} - X_{2p1}$
2	$D_{12} = X_{112} - X_{212}$	$D_{22} = X_{122} - X_{222}$	…	$D_{p2} = X_{1p2} - X_{2p2}$
n	$D_{1n} = X_{11n} - X_{21n}$	$D_{2n} = X_{12n} - X_{22n}$	…	$D_{pn} = X_{1pn} - X_{22n}$

sehingga diperoleh vektor selisih/beda sebagai berikut :

$$\begin{aligned} D_{11} &= X_{111} - X_{211} \\ D_{22} &= X_{122} - X_{222} \\ &\vdots \\ D_{pn} &= X_{1pn} - X_{2pn} \end{aligned}$$

dan

$$D = \begin{bmatrix} D_{1k} \\ D_{2k} \\ \vdots \\ D_{pk} \end{bmatrix}; k = 1,2, \dots, n$$

Kemudian vektor selisih/beda rata-rata yaitu :

$$\overline{\boldsymbol{D}} = \begin{bmatrix} \overline{D}_1 \\ \overline{D}_2 \\ \vdots \\ \overline{D}_p \end{bmatrix}$$

di mana:

$$\overline{D}_j = \frac{1}{n}\sum_{k=1}^{n} D_{jk} \; untuk \; j \; = \; 1,2, \ldots , p.$$

$\overline{D}_j$ adalah rata-rata selisih dari variabel ke-j

Prosedur uji Hotelling T^2 sebagai berikut :

a. Tentukan hipotesisnya:
   Hipotesis
   $H_0 : \boldsymbol{\mu}_D = 0$ (tidak ada perbedaan)
   $H_1 : \boldsymbol{\mu}_D \neq 0$ (ada perbedaan)

b. Statistik Uji

$$T^2 = n(\overline{\boldsymbol{D}} - \boldsymbol{\mu}_D)' \boldsymbol{S}_D^{-1} (\overline{\boldsymbol{D}} - \boldsymbol{\mu}_D) \sim \frac{(n-1)p}{(n-p)} F_{p;(n-p)}$$

   Keterangan :
   $\overline{D}$ yaitu vektor selisih/beda rata-rata yang dihitung dari sampel
   S_D yaitu matriks kovariansi data selisih dihitung dari sampel
   n adalah ukuran sampel
   μ_D adalah hipotesis vektor selisih/beda rata-rata
   Apabila variabel acak berdistribusi Normal Multivariat dengan parameter atau maka statistik .

c. Keputusan adalah :
   Jika $\frac{(n-p)}{p(n-1)} T^2 > F_{\alpha(p;n-p)}$ maka tolak H_0 atau jika menggunakan nilai *p-value* adalah menolak H_0 apabila *p-value(sig)* < α.

Contoh

Buah bit adalah salah satu buang yang tinggi kandungan kalium, magnesium, dan nitrat dan konsumsi buat bit secara teratur membantu menurunkan tekanan darah dan kolesterol. Suatu penelitian ingin menguji kebenaran tersebut. Data sampel terdiri atas 10 pasien pria mengonsumsi buah bit secara teratur. Pasien diukur tekanan darah sistolik sebelum mengonsumi buah bit dan 60 menit sesudah pemberian buah bit. Peneliti ingin mengetahui apakah pemberian buah bit tersebut efektif untuk menurunkan tekanan

darah dan kadar kolesterol pasien tersebut dengan alpha 5%. Adapun data hasil pengukuran ditampilkan pada Tabel di bawah ini.

Tabel 8. Data Kesehatan Pasien

No	Sesudah		Sebelum	
	Tekanan Darah	Kolesterol	Tekanan Darah	Kolesterol
1	175	120	175	100
2	179	115	143	90
3	165	140	135	110
4	170	132	133	132
5	162	110	162	115
6	180	105	150	100
7	177	140	182	110
8	178	100	150	90
9	140	127	175	120
10	176	125	160	100

Misalkan data di atas disimpan dalam file `Kolesterol.csv`, langkah pertama memanggil data di R dengan perintah `read.csv`. Berikut outputnya di R.

```
> Kesehatan= read.csv("Kolesterol.csv", sep=";")
> Kesehatan
    TD1  KO1  TD2  KO2
1   175  120  140  100
2   179  115  143   90
3   165  140  135  110
4   170  132  133  132
5   162  110  162  115
6   180  105  150  100
7   177  140  182  110
8   178  100  150   90
9   140  127  175  120
10  176  125  160  100
```

Pada kasus ini, variabel TD1 = Tekanan Darah Sesudah dan TD2 = Tekanan Darah Sebelum. Sedangkan KO1 = Kolesterol Sesudah dan KO2

= Kolesterol Sebelum. Untuk melakukan uji Hotelling T^2 data perpasangan maka dicari selisih/beda setiap variabel sesudah dan sebelum dengan menggunakan perintah berikut ini.

```
> TD1=Kesehatan[,1]
> TD2=Kesehatan[,3]
> BedaTD=TD1-TD2
> BedaTD
 [1]  35  36  30  37  0  30  -5  28  -35  16
> TK1=Kesehatan[,2]
> TK2=Kesehatan[,4]
> BedaTK=TK1-TK2

> BedaTK
 [1]  20  25  30  0  -5  5  30  10  7  25
> DataBeda=cbind(BedaTD,BedaTK)
> DataBeda
       BedaTD   BedaTK
 [1,]      35       20
 [2,]      36       25
 [3,]      30       30
 [4,]      37        0
 [5,]       0       -5
 [6,]      30        5
 [7,]      -5       30
 [8,]      28       10
 [9,]     -35        7
[10,]      16       25
```

Selanjutnya lakukan uji Hotelling T^2 seperti biasa dengan menggunakan perintah `HotellingT2(X)` dengan hipotesis sebagai berikut :

```
#Uji Hotelling T2 untuk dua sampel dependent
(berpasangan)

> HotellingsT2(DataBeda)

 Hotelling's one sample T2-test

 data: DataBeda
 T.2 = 6.9033, df1 = 2, df2 = 8, p-value = 0.01811
 alternative hypothesis: true location is not equal to c(0,0)
```

Output di atas memberikan informasi nilai T^2 sebesar 6,9033 dengan *p_value* sebesar 0,01811 < 0,05, maka keputusannya tolak H_0. Dapat disimpulkan bahwa pemberian buah bit tersebut efektif untuk menurunkan tekanan darah dan kadar kolesterol pasien.

2. Uji Hotelling T^2 untuk Sampel Independent (Bebas/Tidak Berpasangan)

Kasus $\boldsymbol{\Sigma_1 = \Sigma_2 = \Sigma_3}$.

Untuk melakukan uji Hotelling T^2 untuk sampel independen dan matriks kovariansi kedua populasi sama perhatikan teorema berikut ini.

Teorema (Johnson dan Winchern, 2007)

Misalkan, suatu sampel acak $\boldsymbol{X_{11}}$, $\boldsymbol{X_{12}}$, ..., $\boldsymbol{X_{1n}}$ diambil dari distribusi normal multivariat N_p ($\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}$), dan $\boldsymbol{X_{21}}$, $\boldsymbol{X_{22}}$, ..., $\boldsymbol{X_{2n}}$ sampel acak diambil dari distribusi normal multivariat N_p ($\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$), maka :

$$\boldsymbol{T^2 = n(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))' \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{gab}^{-1} (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))}$$

akan berdistribusi

$$\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p;(n_1+n_2-p-1)}$$

Tabulasi data untuk dua sampel yang bebas/tidak berpasangan disajikan pada tabel di bawah ini.

Tabel 9. Data untuk Sampel Tidak Berpasangan/Bebas

Observasi	Populasi 1				Populasi 2			
	Var1	Var2	…	Var-*p*	Var1	Var2	…	Var-*p*
1	X_{111}	X_{121}	…	X_{1p1}	X_{211}	X_{221}	…	X_{2p1}
2	X_{112}	X_{122}	…	X_{1p2}	X_{212}	X_{222}	…	X_{2p2}
n	X_{11n}	X_{12n}	…	X_{1pn}	X_{21n}	X_{22n}	…	X_{2pn}

Prosedur uji Hotelling T^2 sebagai berikut.

a. Tentukan hipotesisnya:

   $H_0 : \boldsymbol{\mu_1 - \mu_2 = \delta_0}$ (tidak ada perbedaan rata-rata antara dua populasi)

   $H_1 : \boldsymbol{\mu_1 - \mu_2 \neq \delta_0}$ (ada perbedaan rata-rata antara dua populasi)

b. Statsistik Uji dengan

$$\boldsymbol{T^2 = n(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))' \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{gab}^{-1} (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))}$$

   Keterangan :

$$\boldsymbol{S_{gab}} = \frac{(n_1 - 1)\boldsymbol{S_1} + (n_2 - 1)\boldsymbol{S_2}}{n_1 + n_2 - 2}$$

c. Keputusan adalah Tolak H_0 jika

$$T^2 > \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p;(n_1+n_2-p-1)}$$

   jika menggunakan nilai *p-value* adalah menolak H_0 apabila *p-value(sig)* < 0,05.

Sedangkan jika $\boldsymbol{\Sigma_1 \neq \Sigma_1}$ kasus (matriks kovariansi kedua populasi tidak sama) kemudian diketahui n_1 - *p dan* n_2 - *p* berukuran besar maka statistik ujinya menjadi :

$$\boldsymbol{T^2 = n(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))' \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right)^{-1} (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))}$$

Keputusan adalah Tolak H_0 jika $T^2 > X_a^2$ jika menggunakan nilai *p-value* adalah menolak H_0 apabila *p-value(sig)* < 0,05.

Contoh

Pada kasus ini, masih menggunakan data Kesehatan yang disimpan dalam file Kolesterol.csv (lihat contoh 4.2), namun diasumsikan data sampel bebas artinya berasal dari dua populasi yang berbeda. Katakanlah data Kesehatan dari kelompok Laki-laki dan kelompok Perempuan.

Tabel 10. Data Kesehatan Kelompok Laki-laki dan Perempuan

No	Laki-laki		Perempuan	
	Tekanan Darah	Kolesterol	Tekanan Darah	Kolesterol
1	175	120	175	100
2	179	115	143	90
3	165	140	135	110
4	170	132	133	132
5	162	110	162	115
6	180	105	150	100
7	177	140	182	110
8	178	100	150	90
9	140	127	175	120
10	176	125	160	100

Berdasarkan data di atas, uji T^2 Hotelling independent dengan R sebagai berikut.

```
#PANGGIL DATA
> Kesehatan= read.csv("Kolesterol.csv", sep=";")

#Menghitung rata-rata setiap variabel
> colMeans(Kesehatan)
TD1 KO1 TD2 KO2
170.2 121.4 153.0 106.7

#uji kesamaan matriks kovariansi
> Laki=as.matrix(Kesehatan[,1 :2])
> Perempuan=as.matrix(Kesehatan[,3 :4])
> Gabung=rbind(Laki,Perempuan)
> Gabung
       TD1  KO1
```

```
 [1,]  175  120
 [2,]  179  115
 [3,]  165  140
 [4,]  170  132
 [5,]  162  110
 [6,]  180  105
 [7,]  177  140
 [8,]  178  100
 [9,]  140  127
[10,]  176  125
[11,]  140  100
[12,]  143   90
[13,]  135  110
[14,]  133  132
[15,]  162  115
[16,]  150  100
[17,]  182  110
[18,]  150   90
[19,]  175  120
[20,]  160  100

> Ind_Gab=factor(c(rep(1,10),rep(2,10)))
 Ind_Gab
[1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
Levels: 1 2

> Kullback(Gabung, Ind_Gab)
Kullback test for equal covariance matrices
        Chi*  df   P(Chi>Chi*)
 1 0.7197544   3      0.868548
```

Pada ouput di atas diperoleh informasi sebagai berikut.

a. Rata-rata tekanan darah laki laki (TD1) sebesar 170,2 dan rata-rata tekanan darah Perempuan (TD2) = 153,0
b. Rata-rata Tingkat kolesterol laki-laki (KO1) sebesar 121,4 dan rata-rata kolesterol Perempuan (KO2) sebesr 106,7
c. Uji kesamaan matriks kovariansi menggunakan uji Kullback diperoleh hasil *Chi-Square* = 0,71975 dengan p-*value* = 0,8685 (> 0,05) berarti gagal tolak H_0. Dapat disimpulkan matriks kovarinasi ke dua populasi sama ($\boldsymbol{\Sigma_1 = \Sigma_2}$)

Selanjutnya melakukan uji T^2 Hotelling untuk dua sampel yang berasal dari populasi yang independen. Berikut perintahnya.

```
#UJI T2 HOTELLING SAMPEL INDEPENDENT
> HotellingsT2Test(Gabung~Ind_Gab)
Hotelling's two sample T2-test

 data: Gabung by Ind_Gab
 T.2 = 5.438, df1 = 2, df2 = 17, p-value = 0.01494
 alternative hypothesis: true location difference is
not equal to c(0,0)
```

Output di atas memberikan informasi nilai T^2 sebesar 5,438 dengan *p_value* sebesar 0,01494 < 0,05, maka keputusannya tolak H_0. Dapat disimpulkan bahwa ada perbedaan tekanan darah dan kadar kolesterol pasien untuk kelompok Laki-laki dan Perempuan.

4.5 Latihan

1. Berikut ini adalah data pengunjung ke perpustakaan UIN Sunan Kalijaga Yogyakarta.

	Hari ke-				
	1	2	3	4	5
Jumlah pengunjung (X_1)	30	35	42	38	45
Banyaknya buku yang dipinjam (X_2)	10	15	20	12	28

   Lakukan uji Hoteling, versus . Gunakan taraf signifikansi sebesar 1%!

2. Andaikan suatu sampel acak diambil dari dua populasi yaitu dan . Asumsikan kedua populasi mempunyai matriks kovariansi yang sama. Vektor rata- rata adalah :

$$\bar{X}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} dan\ \bar{X}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

   Matriks kovariansi gabungan adalah :

$$S_{gab} = \begin{bmatrix} 7{,}3 & -1{,}1 \\ -1{,}1 & 4{,}8 \end{bmatrix}$$

Berdasarkan data di atas :

a. Uji perbedaan vektor rata-rata dengan menggunakan uji T^2 Hotellings, gunakan $\alpha = 5\%$!
b. Bentuklah fungsi diskriminan!
c. Tentukan keanggotaan $x_0 = (0\ 1)'$ apakah masuk pada kelompok/ populasi pertama atau kedua!

3. Data pengunjung perpustakaan di UIN Sunan Kalijaga berdasarkan fakultas diamati selama 1 minggu sebagai berikut.

Hari	Saintek	Ishum	Tarbiyah
Senin	20	22	33
Selasa	25	35	36
Rabu	30	36	42
Kamis	33	25	44
Jumat	40	25	45
Sabtu	15	10	30

Lakukan uji T^2 hotelling dengan menggunakan *soffware* R. Selidikilah apakah rata- rata pengunjung dari setiap fakultas sama, dengan hipotesis sebagai berikut.

$H_0 : \mu = [20, 20, 20]'$ *vs* $H_1 \neq [20, 20, 20]'$

4. Data penduduk berdasarkan Kabupaten/Kota di provinsi DI Yogyakarta.

	Yogyakarta	Sleman	Kulonprogo	Bantul	Gunungkidul
Jumlah penduduk (x 10.000)	40	120	40	95	73
Rata-rata pengeluaran per kapita (x 10.000)	280	245	190	240	160
Rata-rata lama sekolah (tahun)	12	11	8	9	7

Lakukan uji Hoteling, versus . Gunakan taraf signifikansi sebesar 1%!

5. Perkembangan dunia komputer dan teknologi informasi dewasa ini telah memengaruhi pola kerja dan aktivitas setiap individu. Tidak terkecuali dalam dunia pendidikan. Guru-guru di zaman komputerisasi seperti sekarang ini dituntut untuk bisa melengkapi kemampuan diri dengan keterampilan komputer dan penggunaan teknologi informasi dalam proses pembelajaran. Untuk menjawab permasalahan ini, Dinas Pendidikan memberikan pelatihan dan sosialisasi penggunaan komputer dan teknologi informasi bagi guru-guru. Pelatihan yang diberikan tentang bagaimana mengoperasikan dan menggunakan perangkat lunak Microsoft Excel dan Power Point (PPT). Sampel guru diambil sebanyak 11 orang dan nilai kemampuan Excell dan PPT sebelum dan sesudah mengikuti pelatihan disajikan pada tabel di bawah ini.

Sampel	Sebelum		Sesudah	
	Excell	PPT	Excell	PPT
1	56	55	85	77
2	56	53	88	73
3	48	72	66	84
4	48	79	55	84
5	51	51	65	70
6	54	64	64	75
7	48	60	62	86
8	41	64	74	84
9	43	76	64	79
10	53	60	79	70
11	50	71	89	74

Lakukan uji Hoteling, apakah terdapat perbedaan kemampuan Excell dan Power Point sebelum dengan sesudah mengikuti pelatihan? Gunakan taraf signifikansi sebesar 1%!

BAB

V

ANALISIS VARIANSI MULTIVARIAT/ *MULTIVARIATE ANALYSIS OF VARIANCE* (MANOVA)

5.1 Pendahuluan

MANOVA adalah singkatan dari *Multivariate Analysis of Variance* yang merupakan perluasan dari ANOVA (*Analysis of Variance*). Di dalam pengujian ANOVA hanya melibatkan satu variabel dependen (terikat/ respon) metrik dengan variabel independen non metrik. Contohnya suatu penelitian mengenai perbedaan tingkat pendapatan antara penduduk di kota/kabupaten Daerah Istimewa Yogyakarta. Pada kasus ini hanya ada satu variabel dependen yaitu tingkat pendapatan sedangkan variabel independennya adalah kota atau kabupaten di provinsi Daerah Istimewa Yogyakarta. Variabel independennya bersifat *non metric* (kategorikal) yaitu Kota Yogyakarta, Kabupaten Sleman, Kabupaten Bantul, Kabupaten Gunungkidul dan Kabupaten Kulonprogo.

MANOVA digunakan pada kondisi dimana variabel responnya lebih dari satu. Misalnya, pada penelitian mengenai perbedaan pendapatan, pengeluaran, pendidikan penduduk di kota/kabupatan prov DI Yogyakarta. Dalam kasus ini, variabel responnya ada tiga yaitu variabel pendapatan, pengeluaran, dan pendidikan.

Tujuan dilakukan analisis MANOVA yaitu untuk melihat pengaruh variabel bebas (independen) terhadap beberapa variabel dependen (terikat/ respon). MANOVA memiliki kemampuan tambahan yang tidak dimiliki metode ANOVA, yaitu dalam menangani beberapa variabel dependen. Jika ANOVA diterapkan pada sejumlah variabel dependen yang saling berkaitan, maka kemungkinan kesalahan dalam penarikan kesimpulan akan sangat besar. Metode MANOVA dapat mengatasi kelemahan tersebut dengan menguji secara simultan semua variabel sekaligus hubungan saling keterkaitannya.

MANOVA banyak digunakan dalam berbagai bidang, seperti psikologi, sosiologi, pendidikan, ilmu sosial, bisnis, dan bidang lainnya. Oleh karena itu MANOVA adalah alat analisis yang sangat berguna bagi peneliti yang ingin memahami hubungan antara beberapa variabel secara bersamaan.

5.2 Model MANOVA

MANOVA mirip dengan ANOVA hanya perbedaanya terletak pada

banyaknya variabel tak bebas/dependen Y. Di dalam ANOVA hanya ada satu variabel dependen, sedangkan di dalam MANOVA lebih dari satu, katakan ada *k* sehingga terdapat $Y_1, Y_2,...,Y_k$.

Model MANOVA :

$$Y_{ij} = \mu + \tau_i + e_{ij}; \qquad j = 1,2,..., n_i; \qquad i = 1,2, ..., k$$

dengan :

X_{ij} = pengamatan ke-j pada perlakuan ke-*i*

μ = vektor rata-rata

e_{ij} = *error* pada pengamatan ke-*j* dan perlakuan ke-*i*

τ_i = pengaruh perlakuan ke-*i*

Hipotesis :

H_0: $\tau_1 = \tau_2 = ... = \tau_k = 0$ (Tidak ada pengaruh perlakuan)

H_1: Minimal ada satu $\tau_i \neq 0$ (Terdapat pengaruh perlakuan)

- Statistik Uji dan distribusinya

$$\Lambda^* = \frac{|\boldsymbol{B}|}{|\boldsymbol{B} + \boldsymbol{W}|}$$

Asumsi-asumsi ini harus dipenuhi agar hasil analisis MANOVA yaitu :

1. Normalitas. Setiap variabel dependen harus mengikuti distribusi normal, dan kombinasi linear dari semua variabel dependen juga harus terdistribusi secara normal dalam setiap kelompok.
2. Homogenitas Matriks Kovarians. Matriks varians-kovarians (hubungan antara variabel dependen) di setiap kelompok harus sama.
3. Independensi Observasi. Setiap pengamatan atau data harus independen dari pengamatan lainnya, baik dalam kelompok yang sama maupun antar kelompok yang berbeda.
4. Linearitas. Variabel dependen yang digunakan dalam analisis diharapkan memiliki hubungan yang bersifat linear satu sama lain dan hubungan ini harus homogen di seluruh kelompok yang dibandingkan.

Pada umumnya variabel bebas di dalam MANOVA disebut perlakuan

atau kelompok. Biasanya pertanyaan penelitian adalah untuk melihat pengaruh perlakuan atau kelompok terhadap variabel *Y*.

Misalkan, terdapat *p* perlakuan dan *k* variabel dependen/terikat Y_1, Y_2, ..., Y_k,. Berdasarkan sampel acak berukuran *n* maka tabulasi data dapat ditabulasikan pada tabel di bawah ini.

Tabel 11. Data Pengamatan pada Model MANOVA

	Observasi	**Perlakuan/Kelompok**			
		1	2	…	*p*
Variabel Dependen ke-1	1	Y_{11}	Y_{12}	…	Y_{1k}
	2	Y_{21}	Y_{22}	…	Y_{2k}
	n	Y_{n1}	Y_{n2}	…	Y_{nk}
Variabel Dependen ke-2	1	Y_{11}	Y_{12}	…	Y_{1k}
	2	Y_{21}	Y_{22}	…	Y_{2k}
	n	Y_{n1}	Y_{n2}	…	Y_{nk}
⋮					
Variabel Dependen ke-*k*	1	Y_{11}	Y_{12}	…	Y_{1k}
	2	Y_{21}	Y_{22}	…	Y_{2k}
	n	Y_{n1}	Y_{n2}	…	Y_{nk}
Rata-rata					

Model MANOVA dirumuskan berikut ini.

$$\mathbf{Y}_{ij} = \mu + \tau_i + \boldsymbol{e}_{ij} \qquad (4.1)$$

dengan :

$\mathbf{Y}_{ij}$ = pengamatan ke-*j* pada perlakuan ke-*i*

μ = vektor rata-rata

$\mathbf{e}_{ij}$ = *error* pada pengamatan ke-*j* dan perlakuan ke-*i*

$\boldsymbol{\tau}_i$ = pengaruh perlakuan ke-i

j =1,2,..., n_i; i = 1,2, ..., k

Berdasarkan model di atas, maka setiap komponen pengamatan Y_{ij} dapat dipecahkan ke dalam bentuk :

$$\boldsymbol{y}_{ij} = \bar{\boldsymbol{y}} + (\bar{\boldsymbol{y}}_{\boldsymbol{i}} - \bar{\boldsymbol{y}}) + (\boldsymbol{y}_{\boldsymbol{ij}} - \bar{\boldsymbol{y}}_{\boldsymbol{i}})$$

(pengamatan) = (rata-rata seluruh sampel) + pengaruh perlakuan + residu

Catatan

Perlu diingat bahwa Hasil kali silang (*cross product*) suatu pengamatan adalah :

$$(\boldsymbol{y}_{\boldsymbol{ij}} - \bar{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})'$$

Bentuk di atas dapat ditulis kembali dalam bentuk :

$$\begin{aligned}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})' &= [(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)+(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}\ )\ ][(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)+(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}\ )\ ]' \\ &= (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)' + (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}\ )' \\ &\quad +(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}\ )(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)' + (\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}\ )(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}\ )'\end{aligned}$$

Dari persamaan di atas, terlihat bahwa ekpsresi kedua (di tengah) yaitu: $\sum_{j=1}^{n}(y_{ij} - \bar{y}_i) = 0$, atau menjadi matrik 0.

Selanjutnya, apabila ekspresi di atas dijumlahkan terhadap i dan j, menjadi:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})' = \sum_{i=1}^{k} n_i(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})' + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i)'$$

Jumlah kuadrat dalam (*Within sum of squares*) and *cross products* matriks dapat ditulis:

$$\boldsymbol{W} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'$$

$$= (\ n_1\text{-}1)\ S_1 + (n_2\text{-}1)S_2 + \ldots + (n_k - 1)S_k$$

Metode MANOVA menggunakan matriks *sums-of-square and*

cross-products (*SSCP*) untuk menguji perbedaan antarkelompok. Varian antarkelompok ditentukan dengan menetapkan sekat matriks SSCP total dan menguji signifikansiny Rasio F, diperoleh melalui rasio antara varians yang ada dalam satu kelompok dan matriks varians total kelompok, dan juga uji persamaan antarkelompok yang diteliti. Oleh karena itu tabel MANOVA ditampilkan di bawah ini.

Tabel 12. Matriks Sumber Variansi pada Model MANOVA

Source of Variation	Matriks SSCP (*Sum Squares and Cross Product*)	*Degree of Freedom* (*df*)
Treatment	$\boldsymbol{B} = \sum_{i=1}^{k} n_i(\boldsymbol{y_{ij}} - \bar{\boldsymbol{y}}_i)(\boldsymbol{y_{ij}} - \bar{\boldsymbol{y}}_i)'$	$k - 1$
Residual	$\boldsymbol{W} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{y_{ij}} - \bar{\boldsymbol{y}}_i)(\boldsymbol{y_{ij}} - \bar{\boldsymbol{y}}_i)'$	$\left(\sum n_i - k - 1\right)$
Total terkoreksi	$\mathbf{B+W} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{y_{ij}} - \bar{\boldsymbol{y}}_i)(\boldsymbol{y_{ij}} - \bar{\boldsymbol{y}}_i)'$	$\left(\sum n_i - 1\right)$

Tabel MANOVA sebenarnya hampir sama dengan Tabel ANOVA, perbedaanya adalah bentuk variansinya dalam vektor sedangkan ANOVA dalam bentuk skalar.

Hipotesis

H_0: $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots = \boldsymbol{\tau}_k = 0$ (Tidak ada pengaruh perlakuan)

H_1: Minimal ada satu $\boldsymbol{\tau}_1 \neq 0$ (Terdapat pengaruh perlakuan)

Statistik Uji yang dapat digunakan dalam MANOVA adalah :

Wilk's lamda = $\Lambda^* = \dfrac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}$

Lawley-Hotelling trace = $tr\,(\boldsymbol{BW})^{-1}$

Pillae Trace = $tr\,(\boldsymbol{W}(\boldsymbol{B} + \boldsymbol{W})^{-1})$

Roy's largest root = *Maximum eigen value of* $\boldsymbol{W}(\boldsymbol{B} + \boldsymbol{W})^{-1}$

Pada buku ini, statistik uji yang digunakan adalah *Wilk's lamda* yaitu :

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}$$

Bentuk distrubusi sampling dari dijelaskan pada tabel di bawah ini.

Distribusi sampling dari *wilk's lamda*

Banyaknya Variabel	Banyaknya perlakuan (kelompok)	Distribusi sampling
$p = 1$	$k \geq 2$	$\left(\frac{\sum n_i - k}{k-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{(k-1);(\sum n_i - k)}$
$p = 2$	$k \geq 2$	$\left(\frac{\sum n_i - k - 1}{k-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(k-1);2(\sum n_i - k - 1)}$
$p \geq 1$	$k = 2$	$\left(\frac{\sum n_i - p - 1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p;(\sum n_i - p - 1)}$
$p \geq 2$	$k = 3$	$\left(\frac{\sum n_i - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p;2(\sum n_i - p - 2)}$

H_0 akan ditolak apabila rasio:

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}$$

Sangat kecil dibandingkan dengan F_{tabel}.

Asumsi-asumsi dalam MANOVA adalah :

- Obeservasi harus independen,
- Matriks *variance-covariance* harus sama (atau dapat diperbandingkan) untuk setiap kelompok *treatment*.
- Variabel dependen harus memiliki distribusi normal multivariat.
- Normalitas multivariat dapat diasumsikan tetapi sulit dalam pengujian. Normalitas univariat tidak menjamin normalitas multivariat, namun jika seluruh variabel memenuhi normalitas univariat, maka kemencengan dari normalitas multivariat tidak konsekuensial,

- Criteria lainnya adalah Linearitas dan *multicollinearity* diantara variabel dependen serta sensitivitas terhadap *outliers*.

Contoh

Data berikut diambil dari buku Johnson dan Wincher (2007). Misalkan suatu penelitian ingin melihat pengaruh media tanam terhadap pertumbuhan tanaman Salak. Pertumbuhan tanaman terdiri dari tinggi tanaman dan jumlah helai daun. Media tanam ada tiga jenis, yaitu media tanam A = tanah; media tanam B = tanah + pupuk kandang dan media tanam C = pasir + pupuk kandang.

Setelah melalui percobaan, pada masing-masing jenis media tanam diambil sampel secara acak yaitu media tanam A diambil sampel sebanyak 3 pohon, media tanam B diambil sampel sebanyak 2 pohon dan media tanam C diambil sampel sebanyak 3 pohon. Diperoleh data seperti di bawah ini.

Tabel 13. Data Tanaman Salak

	Media Tanam A	Media Tanam B	Media Tanam C
Tinggi tanaman	9	0	3
	6	2	1
	9		2
Jumlah helai daun	3	4	8
	2	0	9
	7		7

Data di atas dibuat dalam bentuk matriks :

$$\begin{pmatrix} \begin{bmatrix} 9 \\ 3 \end{bmatrix} & \begin{bmatrix} 6 \\ 2 \end{bmatrix} & \begin{bmatrix} 9 \\ 7 \end{bmatrix} \\ \begin{bmatrix} 0 \\ 4 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \\ \begin{bmatrix} 3 \\ 8 \end{bmatrix} & \begin{bmatrix} 1 \\ 9 \end{bmatrix} & \begin{bmatrix} 2 \\ 7 \end{bmatrix} \end{pmatrix}$$

Diketahui: $k = 3$; $p = 2$; $n_1 = 3$; $n_2 = 2$; $n_3 = 3$;

Untuk melakukan Uji MANOVA maka langkah-langkahnya adalah :

1. Hitung rata-rata untuk setiap populasi/kelompok (media tanam)

$$\bar{y}_1 = \begin{bmatrix} 8 \\ 4 \end{bmatrix} \quad \bar{y}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \bar{y}_3 = \begin{bmatrix} 2 \\ 8 \end{bmatrix} \text{ dan } \bar{y} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

2. Menghitung jumlah kuadrat (*Sum of Square*/SS) untuk setiap variabel dependen.

   Jumlah kuadrat (*Sum of Square*/SS) untuk variabel dependen yang pertama yaitu variabel tinggi tanaman (Y_1) :

$$\begin{pmatrix} 9 & 6 & 9 \\ 0 & 2 & \\ 3 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 4 \\ 4 & 4 & \\ 4 & 4 & 4 \end{pmatrix} + \begin{pmatrix} 4 & 4 & 4 \\ -3 & -3 & \\ -2 & -2 & -2 \end{pmatrix} + \begin{pmatrix} 1 & -2 & 1 \\ -1 & 1 & \\ 1 & -1 & 0 \end{pmatrix}$$

   Pengamatan = mean + pengaruh perlakuan + residual dan

   $SS_{obs} = SS_{mean} + SS_{effect} + SS_{res}$

   $216 = 128 + 78 + 10$

   Total SS (*corrected*) = $SS_{obs} - SS_{mean}$

   $= 216 - 128 = 88$

   Operasi yang sama kemudian dilakukan terhadap variabel dependen lainnya (Y_2 = jumlah helai daun):

$$\begin{pmatrix} 3 & 2 & 7 \\ 4 & 0 & \\ 8 & 9 & 7 \end{pmatrix} = \begin{pmatrix} 5 & 5 & 5 \\ 5 & 5 & \\ 5 & 5 & 5 \end{pmatrix} + \begin{pmatrix} -1 & -1 & -1 \\ -3 & -3 & \\ 3 & 3 & 3 \end{pmatrix} + \begin{pmatrix} -1 & -2 & 3 \\ 2 & -2 & \\ 0 & 1 & -1 \end{pmatrix}$$

   Pengamatan = mean + pengaruh perlakuan + residual dan

   $SS_{obs} = SS_{mean} + SS_{effect} + SS_{res}$

   $272 = 200 + 48 + 24$

   Total SS (*corrected*) $= SS_{obs} - SS_{mean}$

   $= 272 - 200 = 72$

   Selanjutnya hitung hasil kali silang antara dua variabel tersebut yaitu:

   Mean $= 4(5) + 4(5) + \ldots + 4(5) = 8\,(4)(5) = 160$

   Treatmen $= 3(4)(-1) + (2\,(-3)(-3) + 3(-2)(3) = -12$

   Residual $= 1(1) + (-2)(-2) + 1(3) + \ldots + 0(-1) = 1$

   Total $= 9(3) + 6(2) + 9(7) + \ldots + 2(7) = 149$

   Total corrected cross product = Total cros product – means cross product $= 149 - 160 = -11$

3. Buat tabel MANOVA

Sumber Variansi	Matriks SSCP (*Sum Squares and Cross Product)*	Derajat Bebas
Treatmen	$\boldsymbol{B} = \begin{bmatrix} 78 & -12 \\ -12 & 48 \end{bmatrix}$	3 – 1 = 2
Residual	$\boldsymbol{W} = \begin{bmatrix} 10 & 1 \\ 1 & 24 \end{bmatrix}$	3 + 2 + 3 – 3 –1= 4
Total terkoreksi	$\boldsymbol{B} + \boldsymbol{W} = \begin{bmatrix} 88 & -11 \\ -11 & 72 \end{bmatrix}$	7

Statistik *wilk's lambda*

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{\begin{vmatrix} 10 & 1 \\ 1 & 24 \end{vmatrix}}{\begin{vmatrix} 88 & -11 \\ -11 & 72 \end{vmatrix}} = \frac{10(24)-(1)(1)}{88(72)-(-11)(-11)} = 0{,}0385$$

Selanjutnya lakukan prosedur uji hipotesis.

Nyatakan hipotesis penelitiannya yaitu:

H_0: $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \boldsymbol{\tau}_3 = 0$ (Tidak ada pengaruh perlakuan/media tanam)

H_1: Minimal ada satu $\boldsymbol{\tau}_1 \neq 0$ (Terdapat pengaruh perlakuan/media tanam)

Stat Uji :

$$F = \left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right)\frac{(\sum n_l - k - 1)}{(k-1)} = \left(\frac{1-\sqrt{0{,}0385}}{\sqrt{0{,}0385}}\right)\frac{(8-3-1)}{(3-1)} = 8{,}19$$

F tabel dengan $df\,\mathrm{v}_1 = 2(k\text{-}1) = 4$ dan $\mathrm{v}_2 = 2\,(\Sigma n_i - k - 1) = 8$, maka $F_{0{,}01(4;8)} = 7{,}01$.

Kesimpulan: Karena 8,91 > 7,01 maka keputusannya tolak $H_{0,}$ artinya ada perbedaan tinggi tanaman dan jumlah helai daun pada ketiga jenis media tanam tersebut.

5.3 Prosedur Uji MANOVA dengan R

Di dalam software R untuk melakukan analisis model MANOVA tersedia di dalam *package* mvnorm melalui perintah manova().

Contoh

Seorang peneliti ingin mengetahui apakah terdapat perbedaan kinerja

ekonomi dan fiskal daerah antara Kota dan Kabupaten. Kinerja daerah diukur menggunakan dua variabel dependen, yaitu :

1. Pendapatan Asli Daerah (PAD)
2. Produk Domestik Regional Bruto (PDRB)

Data diperoleh dari beberapa daerah Kota dan Kabupaten di suatu provinsi ditampilkan pada tabel 14.

Tabel 14. Data PAD dan PDRB berdasarkan jenis Kota dan Kabupaten

Kelompok	PAD	PDRB	Kelompok	PAD	PDRB
Kota	50	61	Kabupaten	33	66
Kota	45	63	Kabupaten	32	56
Kota	48	64	Kabupaten	37	55
Kota	36	65	Kabupaten	35	54
Kota	39	66	Kabupaten	42	48
Kota	41	60	Kabupaten	41	57
Kota	42	59	Kabupaten	43	55
Kota	35	59	Kabupaten	45	53
Kota	60	58	Kabupaten	41	52
Kota	55	57	Kabupaten	40	60
Kota	47	59	Kabupaten	31	50
Kota	36	57	Kabupaten	33	55
Kota	33	54	Kabupaten	36	54
Kota	38	52	Kabupaten	39	53
Kota	49	44	Kabupaten	38	52
Kota	51	46	Kabupaten	35	51
Kota	35	50	Kabupaten	32	49
Kota	42	47	Kabupaten	29	48
Kota	40	48	Kabupaten	40	47
Kota	39	45	Kabupaten	43	46

Hipotesis penelitian :

H_0 : $\tau_1 = \tau_2 = 0$ (Tidak ada pengaruh pekerjaan terhadap pembelian kendaraan).

H_1 : Minimal ada satu $\tau_1 \neq 0$; i = 1,2 (Ada pengaruh pekerjaan terhadap pembelian kendaraan).

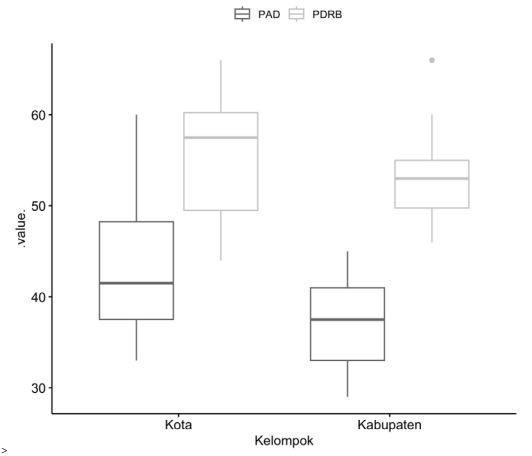
Misalkan data di atas disimpan dengan nama file Job.csv, maka pertama inputkan file tersebut melalui perintah berikut ini.

```
> setwd("/Users/ephadiana/Documents/Multivariat")
> Ekonomi= read.csv("PDRB.csv", sep=";")
> Ekonomi
    Kelompok  PAD  PDRB
1       Kota   50    61
2       Kota   45    63
3       Kota   48    64
4       Kota   36    65
5       Kota   39    66
6       Kota   41    60
7       Kota   42    59
8       Kota   35    59
9       Kota   60    58
10      Kota   55    57
11      Kota   47    59
12      Kota   36    57
13      Kota   33    54
14      Kota   38    52
15      Kota   49    44
16      Kota   51    46
17      Kota   35    50
18      Kota   42    47
19      Kota   40    48
20      Kota   39    45
21 Kabupaten   33    66
22 Kabupaten   32    56
23 Kabupaten   37    55
24 Kabupaten   35    54
25 Kabupaten   42    48
26 Kabupaten   41    57
27 Kabupaten   43    55
28 Kabupaten   45    53
29 Kabupaten   41    52
30 Kabupaten   40    60
31 Kabupaten   31    50
32 Kabupaten   33    55
33 Kabupaten   36    54
```

```
34 Kabupaten   39    53
35 Kabupaten   38    52
36 Kabupaten   35    51
37 Kabupaten   32    49
38 Kabupaten   29    48
39 Kabupaten   40    47
40 Kabupaten   43    46
```

Setelah data diinputkan maka langkah selanjutnya adalah melihat gambaran data (visualisasi data) dengan menggunakan boxplot.

```
> install.packages("ggplot2")
> install.packages("ggpubr")
> library(ggplot2)
> library(ggpubr)
> ggboxplotPDRB, x = "Kelompok", y = c("PAD", "PDRB"),
merge = TRUE, palette = "jco")
```



Boxplot di atas menggambarkan distribusi Pendapatan Asli Daerah (PAD) dan Produk Domestik Regional Bruto (PDRB) yang dikelompokkan berdasarkan Kota dan Kabupaten. Berdasarkan boxplot tersebut terlihat adanya perbedaan pola distribusi dan nilai median pada kedua variabel tersebut. Secara deskriptif, kelompok Kota cenderung memiliki nilai PAD dan PDRB yang lebih tinggi dibandingkan Kabupaten, dengan variasi data yang relatif lebih besar. Perbedaan ini mengindikasikan adanya ketidaksamaan

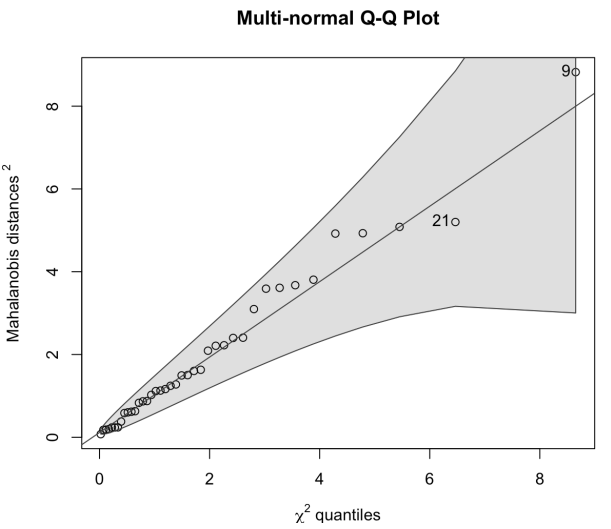
karakteristik ekonomi dan fiskal antarkelompok wilayah.

Namun demikian, analisis boxplot hanya bersifat eksploratif dan belum dapat digunakan untuk menarik kesimpulan inferensial. Oleh karena itu, untuk menguji secara simultan apakah terdapat perbedaan yang signifikan antarakelompok Kota dan Kabupaten berdasarkan kombinasi variabel PAD dan PDRB, digunakan uji *Multivariate Analysis of Variance* (MANOVA). Uji MANOVA digunakan untuk mengevaluasi perbedaan rata-rata vektor (mean vector) PAD dan PDRB antarkelompok dengan mempertimbangkan korelasi antarvariabel, sehingga memberikan hasil yang lebih komprehensif dibandingkan uji univariat terpisah.

1. Menguji Normalitas Data

Pada tahap ini akan diuji apakah variabel dependen (PAD dan PDRB) memiliki distribusi normal multivariat. Pertama dicek secara visual menggunakan plot-quatil-quantil. Grafik *qqplot* untuk data multivariat bisa menggunakan perintah `mqqnorm()` dimana perintah ini ada dalam *package* `RVAideMemoire`. Berikut cara membuat *qqplot* di R.

```
> install.packages("RVAideMemoire")
> library(RVAideMemoire)
> mqqnorm(Ekonomi[,2 :3], main = "Multi-normal Q-Q Plot")
[1] 9 21
```



Gambar di atas menunjukan hasil uji visual normalitas multivariat dengan menggunakan Q–Q plot multivariat terhadap variabel Pendapatan Asli Daerah (PAD) dan Produk Domestik Regional Bruto (PDRB). Metode ini membandingkan jarak Mahalanobis terurut dari data dengan kuantil distribusi chi-kuadrat, sehingga dapat digunakan untuk menilai kesesuaian data terhadap asumsi normalitas multivariat.

Hasil pemanggilan fungsi mqqnorm () menghasilkan indeks observasi 9 dan 21, yang mengindikasikan adanya pengamatan yang relatif paling jauh dari pola garis diagonal. Namun demikian, selama sebagian besar titik observasi berada di sekitar garis diagonal, maka secara umum data masih dapat dikatakan mendekati distribusi normal multivariat.

Untuk menguji data berdistribusi normal multivariat dapat dilakukan melalui uji Mardia. Uji normalitas multivariat Mardia dilakukan untuk mengevaluasi apakah variabel Pendapatan Asli Daerah (PAD) dan Produk Domestik Regional Bruto (PDRB) mengikuti distribusi normal multivariat. Uji Mardia dilakukan dengan menggunakan fungsi mvn() yang berada pada package MVN.

Hipotesisnya adalah :

H_0: Data berdistribusi normal multivariat

H_1: Data tidak berdistribusi normal multivariat

Berikut langkah-langkah menguji distribusi normal multivariat.

```
> # Uji Normalitas Mardia
> install.packages("MVN")#instal package MVN jika belum ada
> library(MVN) #panggil package MVN
> result = mvn(Ekonomi[,2 :3], mvn_test = "mardia")
> result$multivariate_normality
              Test  Statistic  p.value      Method       MVN
1  Mardia Skewness      5.367    0.252  asymptotic  ✓ Normal
2  Mardia Kurtosis     -0.285    0.776 asymptotic   ✓ Normal
```

Hasil pengujian menunjukkan bahwa nilai *p-value* untuk statistik Mardia Skewness sebesar 0,252, sedangkan nilai *p-value* untuk statistik Mardia Kurtosis sebesar 0,776. Kedua nilai *p-value* tersebut lebih besar dari taraf signifikansi α=0,05

Dengan demikian, tidak terdapat bukti yang cukup untuk menolak hipotesis nol yang menyatakan bahwa data PAD dan PDRB berdistribusi normal multivariat. Oleh karena itu, dapat disimpulkan bahwa data PAD dan PDRB memenuhi asumsi normalitas multivariat,

Hasil ini konsisten dengan analisis grafis melalui Q–Q plot sebelumnya, yang memperlihatkan bahwa distribusi variabel cenderung mengikuti pola normal.

2. Identifikasi Homoskedatisitas (variabel dependen memiliki variansi yang sama)

Salah satu asumsi dalam MANOVA yaitu variabel dependen harus memiliki variansi yang sama. Pengujian asumsi tersebut dapat dijawab dengan menggunakan uji Box's M. Uji Box's M digunakan untuk menguji asumsi kesamaan matriks kovariansi antar kelompok pada analisis MANOVA.

Berikut hipotesisnya :

H_0: Matriks kovariansi antar kelompok adalah sama (homogen)

H_1: Setidaknya ada satu kelompok dengan matriks kovariansi yang berbeda (heterogen)

Uji Box's M dapat dilakukan menggunakan perintah `box_m ()` yang terdapat pada paket biotools.

Berikut perintah uji homogenitas dengan menggunakan `box_m()` :

```
> install.packages("biotools")
> library(biotools)
> box_m(Ekonomi[, c("PAD", "PDRB")], Ekonomi$Kelompok)
# A tibble: 1 × 4
  statistic p.value parameter method
    <dbl>   <dbl>  <dbl>   <chr>
1    6.58  0.0864      3   Box's M-test for
Homogeneity of Covariance Matrices
```

Hasil uji Box's M menunjukkan nilai statistik sebesar 6,58 dengan p-value = 0,0864 dan derajat bebas 3. Pada taraf signifikansi α=0,05, nilai p-value lebih besar dari α, sehingga hipotesis nol tidak ditolak.

Hal ini mengindikasikan bahwa tidak terdapat perbedaan yang signifikan pada matriks kovariansi PAD dan PDRB antara kelompok Kota dan Kabupaten. Dengan demikian, asumsi kesamaan matriks kovariansi sebagai salah satu prasyarat utama dalam MANOVA dapat dianggap terpenuhi.

3. Identifikasi Multikolinieritas

Dalam analisis MANOVA, asumsi yang harus terpenuhi adalah variabel dependen tidak boleh ada korelasi. Uji untuk mengetahui adanya korelasi antara variabel dapat menggunakan uji korelasi Pearson karena datanya bersifat rasio. Hipotesis yang digunakan adalah :

H_0: Tidak ada korelasi antara variabel

H_1: Ada korelas antara variabel

Untuk melakuan uji korelasi dapat menggunakan perintah cor.test(), berikut perintahnya.

```
> cor.test(Ekonomi$PAD,Ekonomi$PDRB)

      Pearson's product-moment correlation

data: Ekonomi$PAD and Ekonomi$PDRB
t = 0.51351, df = 38, p-value = 0.6106
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval :
 -0.2345603 0.3845790
sample estimates :
       cor
0.08301479
```

Hasil uji korelasi Pearson antara Pendapatan Asli Daerah (PAD) dan Produk Domestik Regional Bruto (PDRB) menghasilkan koefisien korelasi sebesar r = 0,083 dengan *p-value* = 0,6106. Nilai p-value yang lebih besar dari taraf signifikansi α = 0,05 menunjukkan bahwa tidak terdapat hubungan linear yang signifikan antara PAD dan PDRB.

Selain itu, nilai koefisien korelasi yang sangat kecil mengindikasikan

bahwa hubungan antara kedua variabel bersifat sangat lemah. Interval kepercayaan 95% [−0,235;0,385] yang mencakup nilai nol semakin memperkuat bahwa korelasi antara PAD dan PDRB tidak signifikan secara statistik.

Dari perspektif asumsi MANOVA, hasil ini menunjukkan bahwa tidak terjadi multikolinieritas antar variabel dependen, karena nilai korelasi jauh di bawah batas kritis ($|r| \geq 0{,}80$). Dengan demikian, PAD dan PDRB layak dianalisis secara simultan dalam MANOVA tanpa risiko gangguan multikolinieritas.

4. Identifikasi *Outlier* (Pencilan)

Asumsi berikutnya dari MANOVA adalah tidak boleh ada *outlier* (pencilan) di dalam data. Outlier perlu diperiksa karena keberadaannya dapat memengaruhi :

- Normalitas multivariat,
- Matriks kovariansi,
- dan hasil uji MANOVA secara keseluruhan.

Salah satu teknik mendeteksi keberadaan *outlier* dalam data multivariat yaitu dengan menggunakan jarak Mahalanobis. Dalam R, perintah untuk menghitung jarak Mahalanobis dengan menggunakan fungsi `mahalanobis_distance()`. Perintah ini berada dalam `rstatix package`. Berikut cara untuk mengetahui keberadaan outlier dalam data.

```
> install.packages("rstatix")
> library(rstatix)
> mahalanobis_distance(Ekonomi[,2 :3])
   PAD  PDRB  mahal.dist  is.outlier
1   50    61       3.097       FALSE
2   45    63       2.400       FALSE
3   48    64       3.611       FALSE
4   36    65       3.674       FALSE
5   39    66       3.808       FALSE
6   41    60       0.868       FALSE
7   42    59       0.630       FALSE
8   35    59       1.277       FALSE
9   60    58       8.823       FALSE
10  55    57       4.930       FALSE
11  47    59       1.502       FALSE
12  36    57       0.617       FALSE
13  33    54       1.129       FALSE
14  38    52       0.237       FALSE
15  49    44       5.083       FALSE
16  51    46       4.922       FALSE
17  35    50       1.024       FALSE
18  42    47       1.631       FALSE
19  40    48       1.116       FALSE
20  39    45       2.407       FALSE
21  33    66       5.201       FALSE
22  32    56       1.603       FALSE
23  37    55       0.239       FALSE
24  35    54       0.585       FALSE
25  42    48       1.244       FALSE
26  41    57       0.197       FALSE
27  43    55       0.184       FALSE
28  45    53       0.602       FALSE
29  41    52       0.180       FALSE
30  40    60       0.876       FALSE
31  31    50       2.223       FALSE
32  33    55       1.165       FALSE
33  36    54       0.380       FALSE
34  39    53       0.075       FALSE
35  38    52       0.237       FALSE
36  35    51       0.832       FALSE
37  32    49       2.092       FALSE
38  29    48       3.591       FALSE
39  40    47       1.494       FALSE
40  43    46       2.211       FALSE
```

Hasil perhitungan jarak Mahalanobis untuk pasangan variabel Pendapatan Asli Daerah (PAD) dan Produk Domestik Regional Bruto (PDRB) menunjukkan bahwa seluruh observasi memiliki nilai jarak Mahalanobis yang lebih kecil dari nilai batas kritis distribusi Chi-square. Berdasarkan hasil yang ditampilkan, seluruh observasi diklasifikasikan sebagai bukan outlier multivariat, yang ditunjukkan oleh kolom `is.outlier = FALSE`.

Secara keseluruhan, hasil ini konsisten dengan terpenuhinya asumsi-asumsi MANOVA lainnya, seperti normalitas multivariat, homogenitas varians, kesamaan matriks kovariansi, dan tidak adanya multikolinieritas yang berlebihan antarvariabel dependen.

Setelah asumsi-asumsi tersebut terpenuhi, maka langkah terakhir melakukan analisis MANOVA.

Hipotesis :

H_0: Tidak ada perbedaan rata-rata PAD dan PDRB antara kelompok Kota dan Kabupaten

H_1: Ada perbedaan rata-rata PAD dan PDRB antara kelompok Kota dan Kabupaten

Terdapat beberapa kriteria untuk menguji hipotesis di atas. Berikut cara melakukan analisis MANOVA dalam R.

```
> Model=lm(cbind(PAD,PDRB)~Kelompok, Ekonomi)
> Model=lm(cbind(PAD,PDRB)~Kelompok, Ekonomi)
> M=manova(Model)
> summary(M, test = "Wilks")
          Df   Wilks approx F num Df den Df Pr(>F)
Kelompok   1 0.77489   5.3743      2     37 0.008932 **
Residuals  38
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

Berdasarkan output uji MANOVA dengan statistik Wilks' Lambda, diperoleh nilai Wilks sebesar 0,77489 dengan nilai approximate F = 5,3743, derajat bebas (2, 37), serta p-value = 0,008932. Pada taraf signifikansi α=0,05, nilai p-value lebih kecil dari α, sehingga hipotesis nol ditolak.

Hasil ini menunjukkan bahwa terdapat perbedaan yang signifikan secara simultan pada vektor rata-rata variabel dependen Pendapatan Asli Daerah (PAD) dan Produk Domestik Regional Bruto (PDRB) antara kelompok wilayah (Kota dan Kabupaten).

Meskipun MANOVA mampu mendeteksi adanya perbedaan secara bersama-sama, hasil tersebut belum memberikan informasi spesifik mengenai variabel dependen mana yang berkontribusi terhadap perbedaan tersebut. Oleh karena itu, diperlukan uji lanjutan (follow-up analysis) berupa ANOVA univariat untuk masing-masing variabel dependen. Uji ANOVA bertujuan untuk mengidentifikasi apakah perbedaan antarkelompok terjadi pada variabel PAD, PDRB, atau keduanya secara individual.

Untuk melakukan pengujian individual masing-masing variabel, pertama lakukan perintah berikut ini.

```
> grouped.data <- Ekonomi %>% gather(key = "variable",
value = "value", PAD, PDRB) %>% group_by(variable)
> grouped.data
# A tibble: 80 × 3
# Groups: variable [2]
   Kelompok  variable  value
   <chr>     <chr>     <int>
 1 Kota      PAD          50
 2 Kota      PAD          45
 3 Kota      PAD          48
 4 Kota      PAD          36
 5 Kota      PAD          39
 6 Kota      PAD          41
 7 Kota      PAD          42
 8 Kota      PAD          35
 9 Kota      PAD          60
10 Kota      PAD          55
#    70 more rows
#    Use `print(n = ...)` to see more rows
```

Objek grouped.data merupakan data yang telah ditransformasikan dari format wide menjadi long menggunakan fungsi gather (). Transformasi ini menggabungkan dua variabel dependen, yaitu PAD dan PDRB, ke dalam satu

kolom bernama value, dengan kolom variable sebagai penanda jenis variabel. Data kemudian dikelompokkan berdasarkan variable, sehingga setiap baris merepresentasikan satu pengamatan PAD atau PDRB pada masing-masing Kelompok (Kota/Kabupaten).

Struktur data long ini sangat berguna untuk analisis univariat lanjutan, khususnya uji ANOVA, karena memudahkan pemisahan analisis berdasarkan masing-masing variabel dependen. Setelah hasil MANOVA menunjukkan bahwa hipotesis nol ditolak (terdapat perbedaan multivariat antar kelompok), data ini memungkinkan peneliti untuk menguji secara terpisah apakah perbedaan tersebut signifikan pada variabel PAD, PDRB, atau keduanya.

Dengan demikian, pembentukan grouped.data merupakan langkah metodologis yang tepat sebagai jembatan antara MANOVA dan ANOVA, sehingga analisis dapat dilanjutkan secara sistematis dan interpretasi hasil menjadi lebih spesifik dan informatif.

Setelah data dikelompokkan, maka langkah selanjutnya melakukan ANOVA untuk masing-masing variabel dengan hipotesis sebagai berikut:

Untuk variabel PAD

Hipotesis :

H_0: Tidak ada perbedaan rata-rata PAD untuk kelompok Kota dan Kabupaten

H_1: Ada perbedaan rata-rata PAD untuk kelompok Kota dan Kabupaten.

Untuk variabel PDRB

Hipotesis :

H_0: Tidak ada perbedaan rata-rata PDRB untuk kelompok Kota dan Kabupaten

H_1: Ada perbedaan rata-rata PDRB untuk kelompok Kota dan Kabupaten.

Uji Anova di R dilakukan dengan menggunakan perintah anova_test()

```
> anova_test(grouped.data,value~Kelompok)
# A tibble: 2 × 8
   variable Effect      DFn   DFd     F      p  `p<.05` ges
*  <chr>    <chr>     <dbl> <dbl> <dbl>  <dbl>  <chr>   <dbl>
1  PAD      Kelompok      1    38  8.94  0.005  "*"      0.19
2  PDRB     Kelompok      1    38  1.97  0.169  ""       0.049
```

Hasil uji ANOVA univariat menunjukkan bahwa variabel PAD memiliki nilai statistik F=8,94 dengan p-value = 0,005, yang lebih kecil dari taraf signifikansi α=0,05. Dengan demikian, hipotesis nol ditolak, yang berarti terdapat perbedaan rata-rata PAD yang signifikan antara kelompok Kota dan Kabupaten. Nilai generalized eta squared (ges) sebesar 0,19 mengindikasikan pengaruh kelompok yang cukup besar, yaitu sekitar 19% variasi PAD dapat dijelaskan oleh perbedaan kelompok wilayah.

Sebaliknya, untuk variabel PDRB, diperoleh nilai F=1,97 dengan p-value = 0,169, yang lebih besar dari α=0,05. Hal ini menunjukkan bahwa tidak terdapat perbedaan rata-rata PDRB yang signifikan antara kelompok Kota dan Kabupaten. Nilai ges sebesar 0,049 menunjukkan besaran efek yang kecil, sehingga pengaruh kelompok terhadap PDRB relatif lemah.

Jika dikaitkan dengan hasil MANOVA yang signifikan, temuan ini mengindikasikan bahwa perbedaan multivariat antara kelompok terutama didorong oleh variabel PAD, sementara PDRB tidak memberikan kontribusi signifikan secara individual. Oleh karena itu, PAD dapat diidentifikasi sebagai variabel utama yang membedakan karakteristik ekonomi antara Kota dan Kabupaten.

5.4 Latihan

1. Suatu penelitian ingin mengetahui apakah ada perbedaan (pengaruh) 3 jenis pupuk terhadap produksi (ton) dan produktivitas (kw/ha) tanaman kentang. Hasil pengamatan disajikan pada tabel berikut ini.

	Pupuk A	Pupuk B	Pupuk C
Produksi	6	3	2
	5	1	5

	8	2	3
	4		2
	7		
Produktivitas	7	3	3
	9	6	1
	6	3	1
	9		3
	9		

Berdasarkan data di atas, apakah dapat disimpulkan bahwa ada pengaruh jenis pupuk terhadap produksi dan produktifitas tanaman kentang? Lakukan uji MANOVA dengan $\alpha = 5\%$!

2. Seorang guru ingin meneliti pengaruh metode pembelajaran A, B, dan C terhadap hasil belajar dan motivasi peserta didik dengan menggunakan metode MANOVA. Untuk kepentingan penelitian diambil sampel acak sebanyak 10 siswa. Berdasarkan hasil pengamatan diperoleh tabel MANOVA sebagai berikut.

Sumber Variansi	Matriks SSCP (*Sum Squares and Cross Product)*	Derajat Bebas
Perlakuan	$B = \begin{bmatrix} 178 & -52 \\ -52 & 148 \end{bmatrix}$	…
Residual/error	$W = \begin{bmatrix} \dots & \dots \\ \dots & \dots \end{bmatrix}$	…
Total terkoreksi	$B + W = \begin{bmatrix} 188 & -41 \\ -41 & 172 \end{bmatrix}$	...

   a. Lengkapilah tabel MANOVA di atas!
   b. Tuliskan hipotesis penelitiannya!

   Selidikilah apakah ada pengaruh metode pembelajaran terhadap motivasi dan hasil belajar peserta? Gunakan taraf nyata $\alpha = 0{,}01$!

3. Lakukanlah uji MANOVA pada data iris yang ada dalam *software* R. Data iris adalah data mengenai bunga iris yang terdiri dari 3 spesies Iris yaitu setosa, versicolor, and virginica. Sampel sebanyak 150 jenis bunga tersebut kemudian diukur 4 buah variabel yaitu Sepal.

Length, Sepal.Width, Petal.Length dan Petal.Width. Berikut data 5 pengamatan pertama dan lima pengamatan terakhir dari data iris.

```
> iris[1 :5,]
 Sepal.Length Sepal.Width Petal.Length Petal.Width  Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa

> iris[145 :150,]
 Sepal.Length Sepal.Width Petal.Length Petal.Width Species
145        6.7         3.3          5.7         2.5 virginica
146        6.7         3.0          5.2         2.3 virginica
147        6.3         2.5          5.0         1.9 virginica
148        6.5         3.0          5.2         2.0 virginica
149        6.2         3.4          5.4         2.3 virginica
150        5.9         3.0          5.1         1.8 virginica
```

Berdasarkan data tersebut :

a. Gambarkan (visualisasikan) dengan menggunakan boxplot! Analisis!
b. Lakukan pengujian terhadap asumsi-asumsi Manova!
c. Lakukan uji Manova, apakah ada perbedaan keempat variabel yaitu Sepal.Length, Sepal.Width, Petal.Length dan Petal.Width pada ketiga jenis bunga iris (setosa, vercolor dan virginica)?
d. Jika ada perbedaan, lakukan uji ANOVA untuk setiap variabel tersebut!

4. Suatu penelitian ingin mengelompokan pelanggan yaitu kelompok a (Royal) dan kelompok b (Tidak Royal) di Toko Z berdasarkan 3 variabel yaitu: Pendapatan, Pengeluaran, dan Konsumsi. Hasil survey terhadap 20 orang konsumen diperoleh data sbb :

No	Kelompok	Pendapatan	Pengeluaran	Konsumsi
1	a	191	131	53
2	a	185	134	50
3	a	200	137	52
4	a	173	127	50

No	Kelompok	Pendapatan	Pengeluaran	Konsumsi
5	a	171	128	49
6	a	160	118	47
7	a	188	134	54
8	a	186	129	51
9	a	174	131	52
10	a	163	115	47
11	b	186	107	49
12	b	211	122	49
13	b	201	144	47
14	b	242	131	54
15	b	184	108	43
16	b	211	118	51
17	b	217	122	49
18	b	223	127	51
19	b	208	125	50
20	b	199	124	46

Berdasarkan data konsumen tersebut :

a. Gambarkan (visualisasikan) dengan menggunakan boxplot! Analisis!
b. Lakukan pengujian terhadap asumsi-asumsi Manova!
c. Lakukan uji Manova, apakah ada perbedaan ketiga variabel yaitu Pendapatan, Pengeluaran, dan Konsumsi pada kelompok Royal (a) dan Tidak Royal (b)?
d. Jika ada perbedaan, lakukan uji ANOVA untuk setiap variabel tersebut!

BAB

VI

REGRESI BERGANDA MULTIVARIAT (*MULTIVARIATE MULTIPLE REGRESSION*)

6.1 Pendahuluan

Analisis regresi berganda (*multiple regression analysis*) adalah teknik stastitika yang mempelajari hubungan atau pengaruh variabel bebas terhadap variabel terikat (respon). Dikatakan berganda (*multiple*) karena variabel bebasnya lebih dari satu variabel.

Analisis regresi linear berganda adalah analisis yang mempelajari hubungan linear antara satu variabel terikat (Y) dan m variabel bebas (X_1, X_1, ..., X_m). Misalnya, pengaruh pendapatan, pendidikan dan usia terhadap pengeluaran. Dalam kasus ini, ada tiga variabel bebas yaitu pendapatan (X_1), pendidikan (X_2) dan usia (X_3) sedangkan variabel terikatnya yaitu tingkat pengeluaran (Y).

Dalam berbagai fenomena sering dihadapkan pada kondisi di mana beberapa variabel bebas memengaruhi lebih dari satu variabel terikat secara simultan. Sering ditemui di bidang ilmu, seperti ekonomi, sosial, pendidikan, kesehatan, dan teknik, suatu fenomena sering kali dipengaruhi oleh banyak faktor dan sekaligus menghasilkan lebih dari satu ukuran hasil (output). Pendekatan analisis regresi klasik yang hanya melibatkan satu variabel dependen sering kali belum mampu menangkap kompleksitas hubungan tersebut secara menyeluruh, maka diperlukan metode analisis yang lebih komprehensif, yaitu regresi berganda multivariat.

Regresi berganda multivariat (*multivariate multiple regression*) merupakan pengembangan dari regresi linear berganda dan regresi multivariat, di mana terdapat lebih dari satu variabel dependen yang dianalisis secara bersamaan dengan melibatkan satu atau lebih variabel independen. Metode ini memungkinkan untuk memodelkan hubungan antara sekumpulan variabel bebas dengan beberapa variabel terikat secara simultan, sekaligus mempertimbangkan adanya korelasi antarvariabel dependen.

Misalnya suatu penelitian ingin melihat pengaruh nilai ujian Metode Statistik, Kalkulus I dan Kalkulus II terhadap IPK dan Lama Waktu Studi. Dalam kasus ini banyaknya variabel bebas ada tiga dan variabel terikat ada dua.

Keunggulan utama regresi berganda multivariat terletak pada

kemampuannya untuk menangkap keterkaitan antarvariabel terikat yang sering kali tidak dapat diabaikan. Dalam banyak kasus, variabel terikat (dependen) memiliki hubungan yang erat satu sama lain, sehingga analisis regresi terpisah untuk setiap variabel terikat (dependen) dapat menghasilkan estimasi yang kurang efisien dan berpotensi mengabaikan informasi penting mengenai struktur hubungan data. Dengan menggunakan regresi berganda multivariat, informasi tersebut dapat dimanfaatkan secara optimal untuk memperoleh estimasi parameter yang lebih akurat dan interpretasi yang lebih holistik.

6.2 Model Analisis Regresi Linear Berganda

Misalkan terdapat p variabel bebas yaitu dengan satu variabel terikat maka model regresi linear berganda dirumuskan berikut ini.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \qquad (6.1)$$

Keterangan :

- Y = variabel terikat/dependen
- $X_1, X_1, ..., X_p$ = variabel bebas
- $\beta_0, \beta_1, ..., \beta_p$ = parameter
- ε = galat/*error*

Dalam memodelkan persamaan regresi linear memerlukan beberapa asumsi, yaitu :

1. Hubungan Linear atau adaptif anatara X dan Y
2. Galat (ε) berdistribusi normal
3. Tidak terjadi heteroskedastisitas
4. Tidak tejadi multikoliniertitas antarvariabel bebas
5. Tidak terjadi autokorelasi

Selanjutnya, berdasarkan sampel dari n pengamatan maka model regresi linear berganda menjadi :

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \varepsilon_2$$
$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$
$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \varepsilon_n$$

Persamaan di atas dapat diubah ke dalam bentuk matriks yaitu :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Atau

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

dengan :

$\boldsymbol{Y}$ = vektor dengan ukuran n x 1

$\boldsymbol{X}$ = matriks dengan ukuran n x $(p + 1)$

$\boldsymbol{\beta}$ = vektor dengan ukuran $(p + 1)$ x 1

$\boldsymbol{\varepsilon}$ = vektor dengan ukuran n x 1

Dengan menggunakan metode kuadrat terkecil (*Least Square Method*), maka estimasi parameter pada model regresi linear berganda dapat dihitung dengan menggunakan persamaan berikut ini.

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

sehingga estimasi model regresi linear berganda adalah :

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

Contoh

Data diambil dari buku Johnson dan Wichern (2007). Pada tabel di bawah ini, terdapat 5 pengamatan, 1 variabel Y dan 1 variabel X.

Y	1	4	3	8	9
X	0	1	2	3	4

Berdasarkan data di atas, tuliskan estimasi persamaan regresi linear!

Diketahui :

$$\boldsymbol{Y} = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{bmatrix}; \boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Kemudian dihitung :

$$\boldsymbol{X'X} = \begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix}; \ (\boldsymbol{X'X})^{-1} = \begin{bmatrix} 0.6 & -0.20 \\ -0.2 & 0.10 \end{bmatrix}; \ \boldsymbol{X'Y} = \begin{bmatrix} 25 \\ 70 \end{bmatrix}$$

Selanjutnya, dihitung :

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 0.6 & -0.20 \\ -0.2 & 0.10 \end{bmatrix} \begin{bmatrix} 25 \\ 70 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

sehingga, estimasi persamaan regresi adalah :

$$\hat{Y} = 1 + 2X$$

6.3 Model Analisis Regresi Linear Multivariat

Dalam analisis regresi multivariat, menganalisis pengaruh p variabel bebas yaitu X_1, X_2, ..., X_p terhadap m variabel terikat Y_1, Y_2, ..., Y_p. Misalkan diambil sampel acak berukuran n maka datanya dapat ditabulasikan pada tabel di bawah ini.

Tabel 15. Data Pengamatan pada Model Regresi Multivariat

	Variabel Terikat				Variabel Bebas			
Obs	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$	**…**	$\boldsymbol{Y_m}$	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$	**…**	$\boldsymbol{Y_p}$
1	Y_{11}	Y_{12}	…	Y_{1m}	X_{11}	X_{12}	…	X_{1p}
2	Y_{21}	Y_{22}	…	Y_{2m}	X_{21}	X_{22}	…	X_{2p}
n	Y_{n1}	Y_{n2}	…	Y_{nm}	X_{n1}	X_{n2}	…	X_{np}

Model regresi linear multivariat untuk setiap pengamatan dapat dirumuskan berikut ini.

$$\begin{aligned} Y_{i1} &= \beta_0 + \beta_{11}X_{i1} + \beta_{21}X_{i2} + \cdots + \beta_{p1}X_{ip} + \varepsilon_{i1} \\ Y_{i2} &= \beta_0 + \beta_{12}X_{i1} + \beta_{22}X_{i2} + \cdots + \beta_{p2}X_{ip} + \varepsilon_{i2} \\ &\vdots \\ Y_{im} &= \beta_0 + \beta_{1m}X_{i1} + \beta_{2m}X_{i2} + \cdots + \beta_{pm}X_{ip} + \varepsilon_{im} \end{aligned} \tag{6.2}$$

Untuk *n* pengamatan, persamaan di atas dapat diubah ke dalam bentuk matriks yaitu :

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_{(1)} \vdots & \boldsymbol{Y}_{(2)} \vdots & \cdots & \vdots \boldsymbol{Y}_{(m)} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_{(1)} \vdots & \boldsymbol{\beta}_{(2)} \vdots & \cdots & \vdots \boldsymbol{\beta}_{(p)} \end{bmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_{(1)} \vdots & \boldsymbol{\varepsilon}_{(2)} \vdots & \cdots & \vdots \boldsymbol{\varepsilon}_{(m)} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1m} \\ 1 & X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix}$$

Persamaan model regresi multivariat dapat dirumuskan berikut ini.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6.3}$$

dengan :

$\boldsymbol{Y}$ = matriks dengan ukuran

$\boldsymbol{X}$ = matriks dengan ukuran

$\boldsymbol{\beta}$ = vektor dengan ukuran

$\boldsymbol{\varepsilon}$ = vektor dengan ukuran

Asumsi-asumsi dalam model regresi linear berganda multivariat diantaranya adalah :

- Hubungan antara y_k dan x_j harus linear
- *Error* berdistribusi normal multivariat ($\varepsilon_{ik} \sim NM\ (\mathbf{0}, \boldsymbol{\Sigma})$)
- Variabel terikat memiliki variansi yang sama (*homoskedastisitas*)

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta};\ E(\boldsymbol{\varepsilon_{(i)}}) = \mathbf{0};\ Cov(\boldsymbol{\varepsilon_{(i)}}, \boldsymbol{\varepsilon_{(k)}}) = \sigma_{ik}\boldsymbol{I};\ Var(\boldsymbol{\varepsilon_{(ii)}}) = \sigma_{ii}\boldsymbol{I}$$

Estimasi parameter regresi multivariat adalah :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} \tag{6.3}$$

sehingga estimasi model regresi multivariate berganda adalah :

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \tag{6.3}$$

Contoh

Data diambil dari buku Johnson dan Wichern (2007). Pada tabel di bawah ini terdapat 5 pengamatan, 2 variabel Y dan 1 variabel X.

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

Berdasarkan data di atas, tuliskan estimasi persamaan regresi multivariat!

Jawab

Bentuk matriksnya :

$$\boldsymbol{Y} = \begin{bmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{bmatrix};\ \boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Berdasarkan matrik di atas, maka dapat dibuat matriks berikut ini.

$$\boldsymbol{X'X} = \begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix};\ (\boldsymbol{X'X})^{-1} = \begin{bmatrix} 0.6 & -0.20 \\ -0.2 & 0.10 \end{bmatrix};\ \boldsymbol{X'Y} = \begin{bmatrix} 25 & 5 \\ 70 & 20 \end{bmatrix}$$

sehingga, estimasi parameter regresi multivariat adalah :

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

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 0.6 & -0.20 \\ -0.2 & 0.10 \end{bmatrix}\begin{bmatrix} 25 & 5 \\ 70 & 20 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}$$

Estimasi persamaan regresi multivariatnya adalah :

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \end{bmatrix} = \begin{bmatrix} 1 & X \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}$$

Persamaan di atas dapat dituliskan menjadi berikut ini.

$$\hat{Y}_1 = 1 + 2X$$

$$\hat{Y}_2 = -1 + X$$

Prediksi Nilai Y

Setelah diperoleh nilai estimasi persamaan regresi multivariat, maka langkah berikutnya dalah menghitung berapa nilai prediksi dan residualnya.

Diketahui: $\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$

Prediksi nilai Y adalah :

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

$$= \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{bmatrix}$$

Selanjutnya, nilai vektor residual/*error* dapat dihitung dengan menggunakan persamaan berikut ini.

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

$$= \begin{bmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{bmatrix} - \begin{bmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 \\ -1 & 1 \\ 2 & -1 \\ -1 & -1 \\ 0 & 1 \end{bmatrix}$$

Selanjutnya dihitung matriks jumlah kuadrat silangnya.

$$\boldsymbol{Y'Y} = \begin{bmatrix} 171 & 43 \\ 43 & 19 \end{bmatrix}; \ \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} = \begin{bmatrix} 165 & 45 \\ 45 & 15 \end{bmatrix} \text{ dan } \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} 6 & -2 \\ -2 & 4 \end{bmatrix}$$

6.4 Uji Signifikansi

Dalam analisis regresi berganda, tabel ANOVA (*analysis of variance*) digunakan untuk mengevaluasi kebermaknaan model regresi. Pada analisis regresi multivariat analisis kebermaknaan atau signifikansi model dapat membuat tabel MANOVA (*multivariate analysis of variance*).

Ada beberapa uji yang dapat digunakan.

Wilk's lamda	$\Lambda^* = \dfrac{\lvert\mathbf{W}\rvert}{\lvert\mathbf{B}+\mathbf{W}\rvert}$
Lawley-Hotelling trace	$= tr\,(\boldsymbol{BW})^{-1}$
Pillae Trace	$= tr\,(\boldsymbol{W}\,(\boldsymbol{B}+\boldsymbol{W})^{-1})$
Roy's largest root	$=$ *Maximum eigen value of* $\boldsymbol{W}\,(\boldsymbol{B}+\boldsymbol{W})^{-1}$

6.5 Prosedur Analisis Model Regresi Berganda Multivariat dengan R

Dalam R, perintah membuat model regresi berganda multivariat menggunakan `lm()`. Berikut sintaknya.

```
lm(formula, data, ...)
Keterangan
formula adalah rumus atau persamaan yang digunakan
untuk memodelkan Y dan X.
data adalah matriks data yang diinputkan dalam model (
X dan Y)
```

Contoh

Data yang digunakan adalah data `mtcars` yang tersedia dalam R. Data ini berisi tentang jenis-jenis kendaraan (32 jenis kendaraan) dan 11 variabel. Berikut ini datanya.

Tabel 16. Data pada mtcars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472	205	2.93	5.250	17.98	0	0	3	4

Lincoln Continental	10.4	8	460	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121	109	4.11	2.780	18.60	1	1	4	2

Keterangan :

mpg	=	Miles/(US) *gallon* (konsumsi/pemakaian bahan bakar)
cyl	=	*Number of cylinders*
disp	=	*Displacement* (cu.in.) (*the combined volume of the engine's cylinders*)
hp	=	*Gross horsepower* (kekuatan mesin)
drat	=	*Rear axle ratio* (*this describes how a turn of the drive shaft corresponds to a turn of the wheels*
wt	=	*Weight* (1000 lbs)
qsec	=	1/4 *mile time* (kecepatan dan percepatan)
vs	=	*Engine* (0 = V-*shaped*, 1 = *straight*)
am	=	*Transmissio*n (0 = *automatic*, 1 = *manual*)

gear = *Number of forward gears*

carb = *Number of carburetors*

Misalkan ingin membuat model yang sederhana dulu yaitu melihat pengaruh "am" (jenis transmisi yaitu 0 = *matic* dan 1 = *manual*) dan "carb" (banyaknya karburator) terhadap "mpg=miles per galon" (pemakaian bahan bakar) dan "hp = horsepower" (kekuatan mesin kendaraan). Pada kasus ini terdiri dari 2 variabel terikat dan 2 variabel bebas. Berikut langkah-langkahnya.

```
> Y = as.matrix(mtcars[,c("mpg","hp")]) #Menentukan
variabel terikat
> anreg <- lm(Y ~ am + carb, data=mtcars) #membuat
model
> anreg #print hasilnya
Call :
lm(formula = Y1 ~ am + carb, data = mtcars)

Coefficients :
              mpg       hp
(Intercept)   23.146    71.233
am            7.653     -39.475
carb          -2.192    32.530
```

Berdasarkan output di atas, estimasi parameternya adalah :

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 23{,}146 & 71{,}233 \\ 7{,}653 & -39{,}475 \\ -2{,}192 & 32{,}530 \end{bmatrix}$$

Estimasi persamaan model regresi berganda multivariat adalah :

$$\hat{Y}_1 = 23{,}146 + 7{,}653X_1 - 2{,}192X_2$$

$$\hat{Y}_2 = 71{,}233 - 39{,}475X_1 + 32{,}530X_3$$

Keterangan :

X_1 =variabel "am" dan X_2 = variabel "carb"

Y_1 = variabel "mpg" dan Y_2 = variabel "hp"

Hasil estimasi menunjukkan pengaruh positif jenis transmisi ("am") terhadap pemakaian bahan bakar ("mpg"), sedangkan banyaknya carburator ("carb") berpengaruh negatif terhadap pemakaian bahan bakar kendaraan

(perhatikan tanda koefisien pada masing-masing variabel bebas).

Berbeda dengan pengaruh variabel "am" dan "carb" terhadap variabel mpg, pada variabel kekuatan mesin ("hp") hubungannya berlawanan arah. Pengaruh jenis tranmisi ("am") terhadap kekuatan mesin ("hp") adalah negatif sedangkan variabel banyaknya karburator berpengaruh positif terhadap kekuatan mesin kendaraan ("hp")

Selanjutnya, model di atas dianalisis diuji kecocokannya dengan melihat ringkasan (summary) melalui perintah berikut ini.

```
> Ringkas=summary(anreg)
```

Apabila ingin mengetahui pengaruh variabel bebas ("am" dan "carb") terhadap variabel "mpg", maka lakukan perintah berikut ini.

```
> Ringkas[[1]]

Call :
lm(formula = mpg ~ am + carb, data = mtcars)

Residuals :
 Min 1Q Median 3Q Max
-6.2320 -1.7415 -0.0706 2.3939 5.6377

Coefficients :
            Estimate  Std. error  t value  Pr(>|t|)
(Intercept)  23.1458      1.2941   17.885  < 2e-16 ***
am            7.6531      1.2230    6.258  7.87e-07 ***
carb         -2.1917      0.3778   -5.801  2.75e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 3.392 on 29 degrees of
freedom
Multiple R-squared: 0.7037,     Adjusted R-squared:
0.6832
F-statistic: 34.43 on 2 and 29 DF, p-value: 2.191e-08
```

Output di atas menjelaskan hasil analisis regresi untuk variabel terikat "mpg" dengan "am" dan "carb" .

Berdasarkan output di atas, diketahui bahwa koefisen determinasi (R^2) sebesar 0,7037. Nilai ini menunjukkan bahwa variabel "am" dan "carb" sudah menjelaskan keragaman (variansi) variabel mpg sebesar 70,37%.

Perhatikan bahwa nilai F_{hitung} = 34,43 dan *p_value* sebesar 1,191e-08 ≈ 0,00. Karena p-*value* lebih kecil dibandingkan taraf signifikansi sebesar 0,05, maka keputusannya tolak H_0 artinya model layak (sesuai) digunakan.

Karena model sudah sesuai (*fit*), maka langkah selanjutnya adala melakukan uji signifikansi setiap variabel bebasnya. Hipotesis untuk setiap variabel bebas dilakukan berikut ini.

Hipotesisnya adalah :

$H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$

Perhatikan bahwa nilai *p_value* untuk semua koefisien variabel bebas ("am" dan "carb") lebih kecil dari signifikansi sebesar 0.05 maka keputusannya tolak $H_{0,}$ artinya ada pengaruh variabel bebas ("am" dan "crab") terhadap variabel "mpg".

Selanjutnya, untuk melihat pengaruh variabel bebas ("am" dan "carb") terhadap variabel terikat "hp", maka ketikan perintah berikut ini.

```
> Ringkas[[2]]

Call :
lm(formula = hp ~ am + carb, data = mtcars)

Residuals :
    Min       1Q   Median       3Q      Max
-78.354  -15.396    8.706   19.298  102.121

Coefficients :
            Estimate Std. error  t value  Pr(>|t|)
(Intercept)   71.233     16.126    4.417  0.000128 ***
am           -39.475     15.239   -2.590  0.014843 *
carb          32.530      4.708    6.910  1.36e-07 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 42.27 on 29 degrees of freedom
Multiple R-squared: 0.6445, Adjusted R-squared: 0.62
F-statistic: 26.29 on 2 and 29 DF, p-value: 3.072e-07
```

Output di atas menjelaskan hasil analisis regresi untuk variabel terikat "`hp`" dengan "`am`" dan "`carb`".

Berdasarkan output di atas, diketahui bahwa koefisen determinasi sebesar 0,6445. Nilai ini menunjukkan bahwa variabel "`am`" dan "`carb`" sudah menjelaskan keragaman (variansi) variabel hp sebesar 64,45%.

Perhatikan bahwa nilai = 26,29 dan sebesar 3,072e-08 ≈ 0,00. Karena p-*value* lebih kecil dibandingkan taraf signifikansi sebesar 0,05 maka keputusannya tolak H_0, artinya model layak digunakan.

Kemudian perhatikan hasil uji signifikansi setiap variabel bebasnya. Nilai *p_value* untuk variabel am sebesar 0,01483 lebih kecil dari signifikansi sebesar 0,05 maka keputusannya tolak H_0, artinya ada pengaruh variabel am terhadap variabel hp.

Nilai *p_value* pada variabel "`carb`" sebesar 1,36e-07 ≈ 0,00. Nilai ini lebih kecil dari signifikansi sebesar 0,05, maka keputusannya tolak H_0 artinya ada pengaruh variabel "`carb`" terhadap variabel "`hp`".

Berdasarkan hasil uji kecocokan model dan uji signifikansi variabel bebas terhadap variabel terikat bahwa model sudah layak untuk digunakan. Oleh karena itu, model di atas dapat digunakan untuk memprediksi nilai variabel terikat ("mpg" dan "hp") apabila variabel terikat ("am" dan "carb") diketahui.

Misalkan diketahui "am" = 1 dan "carb" = 4, maka estimasi nilai "mpg" dan "hp" dapat diperoleh dengan cara sebagai berikut.

```
> x1 <- data.frame(am=1, carb=4)
> predict(anreg,x1, interval="confidence")
       mpg         hp
1 22.03196   161.8786
```

Hasil estimasi untuk variabel "mpg" = 22,03196 dan "hp"=161,8786. Untuk mengetahui seluruh estimasi dari model dapat dilakukan dengan mengetikkan perintah berikut ini.

```
> anreg$fitted.values
                          mpg          hp
Mazda RX4            22.03196   161.87855
Mazda RX4 Wag        22.03196   161.87855
Datsun 710           28.60721    64.28830
Hornet 4 Drive       20.95409   103.76354
Hornet Sportabout    18.76234   136.29362
Valiant              20.95409   103.76354
Duster 360           14.37884   201.35379
Merc 240D            18.76234   136.29362
Merc 230             18.76234   136.29362
Merc 280             14.37884   201.35379
Merc 280C            14.37884   201.35379
Merc 450SE           16.57059   168.82371
Merc 450SL           16.57059   168.82371
Merc 450SLC          16.57059   168.82371
Cadillac Fleetwood   14.37884   201.35379
Lincoln Continental  14.37884   201.35379
Chrysler Imperial    14.37884   201.35379
```

```
Fiat 128             28.60721     64.28830
Honda Civic          26.41546     96.81838
Toyota Corolla       28.60721     64.28830
Toyota Corona        20.95409    103.76354
Dodge Challenger     18.76234    136.29362
AMC Javelin          18.76234    136.29362
Camaro Z28           14.37884    201.35379
Pontiac Firebird     18.76234    136.29362
Fiat X1-9            28.60721     64.28830
Porsche 914-2        26.41546     96.81838
Lotus Europa         26.41546     96.81838
Ford Pantera L       22.03196    161.87855
Ferrari Dino         17.64847    226.93872
Maserati Bora        13.26497    291.99890
Volvo 142E           26.41546     96.81838
```

Kemudian, jika ingin mengetahui residual (*error*) dapat dilakukan dengan mengetikkan perintah berikut ini.

```
> anreg$residuals
                          mpg             hp
Mazda RX4          -1.03196384   -51.8785535
Mazda RX4 Wag      -1.03196384   -51.8785535
Datsun 710         -5.80720743    28.7117027
Hornet 4 Drive      0.44591160     6.2364641
Hornet Sportabout  -0.06234053    38.7063787
Valiant            -2.85408840     1.2364641
Duster 360         -0.07884480    43.6462079
Merc 240D           5.63765947   -74.2936213
Merc 230            4.03765947   -41.2936213
Merc 280            4.82115520   -78.3537921
Merc 280C           3.42115520   -78.3537921
Merc 450SE         -0.17059267    11.1762933
Merc 450SL          0.72940733    11.1762933
Merc 450SLC        -1.37059267    11.1762933
Cadillac Fleetwood -3.97884480     3.6462079
Lincoln Continental -3.97884480   13.6462079
```

```
Chrysler Imperial      0.32115520     28.6462079
Fiat 128               3.79279257      1.7117027
Honda Civic            3.98454043    -44.8183827
Toyota Corolla         5.29279257      0.7117027
Toyota Corona          0.54591160     -6.7635359
Dodge Challenger      -3.26234053     13.7063787
AMC Javelin           -3.56234053     13.7063787
Camaro Z28            -1.07884480     43.6462079
Pontiac Firebird       0.43765947     38.7063787
Fiat X1-9             -1.30720743      1.7117027
Porsche 914-2         -0.41545957     -5.8183827
Lotus Europa           3.98454043     16.1816173
Ford Pantera L        -6.23196384    102.1214465
Ferrari Dino           2.05153189    -51.9387243
Maserati Bora          1.73502762     43.0011050
```

6.6 Latihan

1. Pada tabel di bawah ini, terdapat 6 pengamatan dengan 2 variabel Y dan 1 variabel X.

Y_1	5	3	4	2	1
Y_2	-3	-1	-1	2	3
X	-2	-1	-1	2	3

   a. Cari estimasi persamaan regresi multivariat!
   b. Hitung matriks nilai prediksi dari dan matriks residual !

2. Suatu penelitian mengenai hubungan antara nilai Bahasa Inggris (X_1) dan nilai Matematika (X_2) dengan IPK (Y_1) dan Nilai TOEC (Y_2). Suatu sampel diambil dari 6 mahasiswa, hasilnya disajikan pada tabel di bawah ini.

Y_1	3,5	3,2	2,9	3,6	3,1
Y_2	3,5	4,5	3,5	4,8	3,0
X_1	7	8	6	8	7
X_2	8	7	6	9	6

   Berdasarkan data di atas :

a. Cari estimasi persamaan regresi multivariat!
b. Hitung matriks nilai prediksi dari dan matriks residual !

3. Lakukan analisis regresi berganda multivariat dengan *softwar*e R pada data mtcars (lihat data pada contoh di atas) yaitu untuk melihat hubungan antara "`am`", "`gear`" dan "`carb`" terhadap "`mpg`", "`hp`" dan "`wt`", sehingga pada kasus ini ada 3 variabel bebas dan 3 variabel terikat.
   a. Tentukan estimasi persamaan modelnya!
   b. Apakah model sudah cocok (layak)?
   c. Uji signifikansi semua variabel bebas terhadap variabel terikat! Analisis!
4. Pada tabel berikut adalah data Provinsi di Indonesia beserta variabel harapan hidup, harapan lama sekolah, rata-rata lama sekolah, pendapatan per kapita dan indeks pembanguna manusia. Berdasarkan data tersebut, lakukan analisis regresi multivariat untuk melihat apakah ada pengaruh variabel harapan hidup, harapan lama sekolah, rata-rata lama sekolah terhadap pendapatan per kapita dan indeks pembangunan manusia!

No	Provinsi	Harapan Hidup	Harapan Lama Sekolah	Lama Sekolah	Pendapatan Per Kapita	Indeks Pembangunan Manusia
1	Aceh	73.20	14.39	9.64	10.811	75.36
2	Sumatera Utara	73.90	13.49	9.93	11.46	75.76
3	Sumatera Barat	74.37	14.3	9.44	11.718	76.43
4	Riau	74.41	13.42	9.43	11.857	75.67
5	Jambi	74.06	13.14	8.9	11.621	74.36
6	Sumatera Selatan	74.26	12.64	8.57	12.015	73.84
7	Bengkulu	73.31	13.75	9.04	11.733	74.91
8	Lampung	74.39	12.78	8.36	11.258	73.13
9	Kep. Bangka Belitung	74.12	12.49	8.33	13.667	74.55
10	Jawa Barat	75.16	12.8	8.87	12.157	79.89
11	Jawa Tengah	74.91	12.86	8.02	12.276	84.15
12	Jawa Timur	75.07	13.43	8.28	12.852	74.92

No	Provinsi	Harapan Hidup	Harapan Lama Sekolah	Lama Sekolah	Pendapatan Per Kapita	Indeks Pembangunan Manusia
13	Banten	74.97	13.1	9.23	13.097	73.87
14	Nusa Tenggara Barat	72.25	13.98	7.87	11.606	81.62
15	Kalimantan Barat	73.94	12.68	7.78	10.321	75.35
16	Kalimantan Tengah	73.73	12.77	8.81	12.303	76.35
17	Kalimantan Selatan	74.18	12.87	8.62	13.399	78.63
18	Kalimantan Timur	74.94	14.03	10.02	13.793	73.1
19	Kalimantan Utara	73.57	13.21	9.35	10.197	69.14
20	Sulawesi Utara	74.08	12.97	9.84	11.998	71.19
21	Sulawesi Selatan	73.83	13.55	8.86	12.275	74.28
22	Kepulauan Riau	75.12	13.27	10.5	15.573	75.19
23	DKI Jakarta	75.99	13.51	11.49	19.953	78.79
24	D.I. Yogyakarta	75.36	15.7	9.92	15.361	73.41
25	Bali	75.10	13.62	9.54	14.92	75.68
26	Nusa Tenggara Timur	71.83	13.23	8.02	8.534	72.24
27	Sulawesi Tengah	70.84	13.34	9.04	10.536	75.18
28	Sulawesi Tenggara	71.88	13.71	9.42	10.606	73.62
29	Gorontalo	70.73	13.17	8.29	11.539	72.01
30	Sulawesi Barat	71.03	12.89	8.15	10.208	70.46
31	Maluku	70.68	14.09	10.26	9.684	73.4
32	Papua Barat	68.47	13.17	7.86	8.805	71.84
33	Papua Barat Daya	70.02	13.88	8.39	8.733	67.69
34	Papua	70.47	13.72	9.82	11.037	69.65
35	Papua Selatan	68.46	12.67	8.38	9.756	73.83
36	Maluku Utara	71.05	13.75	9.37	9.32	68.86
37	Papua Tengah	68.18	9.63	6.12	7.809	60.25
38	Papua Pegunungan	67.39	9.97	4.21	5.707	54.43

BAB

VIII

ANALISIS KORELASI KANONIK (CANONIC CORELATION ANALYSIS)

7.1 Pendahuluan

Analisis korelasi adalah analisis yang bertujuan untuk mengukur kekuatan hubungan antara dua variabel, misalnya hubungan antara promosi dengan penjualan, harga barang dengan penjualan atau pendapatan dengan tabungan (*saving*). Pada beberapa kasus, pola hubungan yang terjadi mungkin tidak hanya pada dua variabel tetapi pada sekumpulan variabel. Misalkan hubungan antara promosi dan harga terhadap penjualan dan pendapatan.

Pendekatan analisis bivariat atau regresi konvensional yang hanya melibatkan satu variabel dependen dan satu atau beberapa variabel independen sering kali tidak mampu menggambarkan hubungan kompleks antar dua kelompok variabel tersebut secara menyeluruh. Oleh karena itu, diperlukan suatu metode analisis multivariat yang mampu mengevaluasi hubungan simultan antara dua himpunan variabel, yaitu analisis kanonik.

Analisis kanonik (*Canonical Correlation Analysis/CCA*) merupakan salah satu teknik analisis multivariat yang bertujuan untuk mengukur dan menjelaskan keeratan hubungan antara dua set variabel secara bersamaan. Metode ini dilakukan dengan membentuk kombinasi linier dari masing-masing kelompok variabel sedemikian rupa sehingga korelasi antara kedua kombinasi linier tersebut menjadi maksimum.

Kelebihan analisis korelasi kanonik diantaranya bersifat multidimensional karena tidak hanya mengukur hubungan antara dua variabel saja namun antar 2 kelompok variabel secara keseluruhan.

Contoh

1. Hubungan antara variabel kesehatan (tekanan darah, kadar gula, kolesterol) dengan variabel gaya hidup (olahraga, pola makan);
2. Hubungan antara skor tes akademik (matematika, bahasa) dengan kemampuan *soft skill* (komunikasi, kepemimpinan, kerjasama, etos kerja).

Selain itu, korelasi kanonik dapat mengidentifikasi hubungan linear variabel di masing-masing kelompok yang memiliki korelasi paling tinggi. Artinya dapat mengetahui variabel yang paling berpengaruh dalam hubungan

antar kelompok tersebut.

7.2 Model Korelasi Kanonik

Analisis korelasi kanonik merupakan suatu teknik yang berguna untuk mengidentifikasi dan mengukur hubungan linier, yang melibatkan beberapa variabel dependen dan independen. Korelasi kanonik fokus pada korelasi antara kombinasi linier dari suatu himpunan variabel independen dengan kombinasi linier dari himpunan variabel dependen. Pasangan kombinasi liniernya disebut fungsi kanonik, dan korelasinya disebut koefisien korelasi kanonik.

Misalnya terdapat himpunan p variabel independen $X = (X_1, X_2, ..., X_p)$ dan himpunan q variabel dependen $Y = (Y_1, Y_2, ..., Y_q)$ dimana $p \leq q$.

dengan:

$$E(\boldsymbol{X}) = \boldsymbol{\mu_x} \text{ dan } E(\boldsymbol{Y}) = \boldsymbol{\mu_Y}$$

$$Var(\boldsymbol{X}) = \boldsymbol{\Sigma_{XX}} \; ; \; Var(\boldsymbol{Y}) = \boldsymbol{\Sigma_{YY}} \text{ dan } Cov(\boldsymbol{XY}) = \boldsymbol{\Sigma_{XY}}$$

Fungsi kanonik yang merupakan kombinasi linear dari kedua himpunan variabel tersebut dapat dituliskan berikut ini.

$$\boldsymbol{U} = \boldsymbol{a}'\boldsymbol{X} = a_1X_1 + a_2X_2 + \cdots + a_pX_p$$

$$\boldsymbol{V} = \boldsymbol{b}'\boldsymbol{Y} = b_1Y_1 + b_2Y_2 + \cdots + b_qY_q$$

sehingga

$$Var(\boldsymbol{U}) = \mathbf{a}'\boldsymbol{\Sigma_{XX}}\boldsymbol{a} \; ; \; Var(\boldsymbol{V}) = \mathbf{b}'\boldsymbol{\Sigma_{YY}}\boldsymbol{b} \text{ dan } Cov(\boldsymbol{UV}) = \mathbf{a}'\boldsymbol{\Sigma_{XY}}\boldsymbol{b}$$

Koefisien vektor $\boldsymbol{a}$ dan $\boldsymbol{b}$ dapat dicari dengan syarat memaksimumkan korelasi antara $\boldsymbol{U}$ dan $\boldsymbol{V}$ yaitu :

$$cor(\boldsymbol{UV}) = \frac{\mathbf{a}'\boldsymbol{\Sigma_{XY}}\boldsymbol{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma_{XX}}\boldsymbol{a}}\sqrt{\mathbf{b}'\boldsymbol{\Sigma_{YY}}\boldsymbol{b}}}$$

Banyaknya fungsi kanonik (variat) akan sama dengan banyaknya variabel dependen atau independen terkecil.

Pasangan pertama dari variabel kanonik adalah kombinasi linear antara $\boldsymbol{U}_1$ dan $\boldsymbol{V}_1$ yang memiliki variansi satu dan korelasinya terbesar. Sedangkan

pasangan kedua dari variabel kanonik adalah kombinasi linear $\boldsymbol{U}_2$ dan $\boldsymbol{V}_2$ yang memiliki variansi satu dan korelasi terbesar kedua serta tidak berkorelasi dengan variabel kanonik yang pertama. Pasangan ke-*k* dari variabel kanonik adalah kombinasi linear antara $\boldsymbol{U}_k$ dan $\boldsymbol{V}_k$ yang memiliki ragam satu dan korelasinya terbesar ke-*k* serta tidak berkorelasi dengan fungsi kanonik lainnya.

Fungsi kanonik pertama, korelasi terbesar pertama yaitu:

$$\boldsymbol{U_1} = \boldsymbol{a_1}'\boldsymbol{X};\ Var(\boldsymbol{U_1}) = 1$$

$$\boldsymbol{V_1} = \boldsymbol{b_1}'\boldsymbol{Y};Var(\boldsymbol{V_1}) = 1$$

maximum *corr* $(\boldsymbol{U}_1; \boldsymbol{V}_1) = \rho_1$

Fungsi kanonik kedua, korelasi terbesar kedua yaitu:

$$\boldsymbol{U_2} = \boldsymbol{a_2}'\boldsymbol{X};\ Var(\boldsymbol{U_2}) = 1$$

$$\boldsymbol{V_2} = \boldsymbol{b_2}'\boldsymbol{Y};Var(\boldsymbol{V_2}) = 1$$

maximum *corr* $(\boldsymbol{U}_2; \boldsymbol{V}_2) = \rho_2$

Fungsi kanonik ke-*k*, korelasi terbesar ke-*k* yaitu :

$$\boldsymbol{U_k} = \boldsymbol{a_k}'\boldsymbol{X};\ Var(\boldsymbol{U_k}) = 1$$

$$\boldsymbol{V_k} = \boldsymbol{b_k}'\boldsymbol{Y};Var(\boldsymbol{V_k}) = 1$$

maximum *corr* $(\boldsymbol{U}_k; \boldsymbol{V}_k) = \rho_k$

7.3 Estimasi Model Korelasi Kanonik

Suatu sampel acak berukuran yang terdiri dari himpunan p variabel independen $X = (X_1, X_2, ..., X_p,)$ dan himpunan q variabel dependen $Y = (Y_1, Y_2, ..., Y_q,)$ dimana $p \leq q$. Bentuk data dapat dituliskan berikut ini.

$$\boldsymbol{Z} = [\boldsymbol{X} \mid \boldsymbol{Y}]$$

$$= \left[\begin{array}{ccc|ccc} x_{11} & \cdots & x_{1p} & y_{11} & \cdots & y_{1q} \\ x_{21} & \cdots & x_{2p} & y_{21} & \cdots & y_{2q} \\ x_{31} & \cdots & x_{3p} & y_{31} & \cdots & y_{3q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} & y_{n1} & \cdots & y_{nq} \end{array}\right]$$

vektor rata-rata sampel adalah :

$$\overline{\boldsymbol{z}}_{(\boldsymbol{p+q},\boldsymbol{1})} = \begin{bmatrix} \overline{\boldsymbol{x}} \\ - \\ \overline{\boldsymbol{y}} \end{bmatrix}$$

dengan:

$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j$ dan $\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{y}_j$

Kemudian matriks kovariansi sampel adalah :

$$\boldsymbol{S}_{(p+q)\times(p+q)} = \begin{bmatrix} \boldsymbol{S_{xx}} & \boldsymbol{S_{xy}} \\ \boldsymbol{S_{yx}} & \boldsymbol{S_{yy}} \end{bmatrix}$$

Misalkan :

$$\boldsymbol{A} = \boldsymbol{S_{xx}^{-1/2} S_{xy} S_{yy}^{-1} S_{yx} S_{xx}^{-1/2}}$$

Selanjutnya dari matriks $\boldsymbol{A}$ dihitung akar ciri yaitu $\hat{\rho}_1 \geq \hat{\rho}_2 \geq \cdots \geq \hat{\rho}_p \geq 0$ yang berpasangan dengan vektor ciri $\hat{\boldsymbol{e}}_1 \geq \hat{\boldsymbol{e}}_2 \geq \cdots \geq \hat{\boldsymbol{e}}_p$. Kemudian :

$$\boldsymbol{B} = \boldsymbol{S_{yy}^{-1/2} S_{yx} S_{xx}^{-1} S_{xy} S_{yy}^{-1/2}}$$

Selanjutnya dari matriks $\boldsymbol{B}$ dihitung vektor ciri, yaitu $\hat{\boldsymbol{f}}_1 \geq \hat{\boldsymbol{f}}_2 \geq \cdots \geq \hat{\boldsymbol{f}}_p$ dimana vektor ciri pertama diperoleh dari $\hat{\boldsymbol{f}}_k = \frac{1}{\hat{\rho}_k} \boldsymbol{S_{yy}^{-1/2} S_{yx} S_{xx}^{-1} \hat{e}_k}$

Estimasi fungsi kanonik ke-k adalah :

$$\begin{aligned} \widehat{\boldsymbol{U}}_{\boldsymbol{k}} &= \boldsymbol{S_{xx}^{-1/2}} \hat{\boldsymbol{e}}_k \boldsymbol{X} = \widehat{\boldsymbol{a}}_k \boldsymbol{X} \\ \widehat{\boldsymbol{V}}_{\boldsymbol{k}} &= \boldsymbol{S_{yy}^{-1/2}} \hat{\boldsymbol{f}}_k \boldsymbol{Y} = \widehat{\boldsymbol{b}}_k \boldsymbol{Y} \end{aligned} \qquad (7.1)$$

dimana :

$$\widehat{\boldsymbol{a}}_k = \boldsymbol{S}_{\boldsymbol{xx}}^{-1/2}\widehat{\boldsymbol{e}}_k \text{ dan } \widehat{\boldsymbol{b}}_k = \boldsymbol{S}_{\boldsymbol{yy}}^{-1/2}\widehat{\boldsymbol{f}}_k$$

Kemudian, setiap korelasi kanonik ke-*k* adalah:

$$r_{\widehat{\boldsymbol{U}}_{\boldsymbol{k}};\widehat{\boldsymbol{V}}_{\boldsymbol{k}}} = \hat{\rho}_k; \quad k = 1,2,\dots,p \tag{7.2}$$

Contoh

Suatu penelitian ingin mengetahui hubungan hubungan antara Kemampuan Bahasa (X_1) dan Kemampuan Berhitung (X_2) terhadap Nilai IPK (Y_1) dan Masa Studi (Y_2). Berdasarkan hasil pengamatan diperolah matriks kovariansi sebagai berikut (data sudah distandardisasi) :

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{\boldsymbol{xx}} & \boldsymbol{S}_{\boldsymbol{xy}} \\ \boldsymbol{S}_{\boldsymbol{yx}} & \boldsymbol{S}_{\boldsymbol{yy}} \end{bmatrix} = \left[\begin{array}{cc|cc} 1{,}0 & 0{,}4 & 0{,}5 & 0{,}6 \\ 0{,}4 & 1{,}0 & 0{,}3 & 0{,}4 \\ \hline 0{,}5 & 0{,}3 & 1{,}0 & 0{,}2 \\ 0{,}6 & 0{,}4 & 0{,}2 & 1{,}0 \end{array}\right]$$

Buatlah fungsi kanonik dan hitunglah korelasi kanoniknya!

Jawab

Berdasarkan matriks variansi-kovariansi di atas diperoleh :

$$\boldsymbol{S}_{\boldsymbol{xx}} = \begin{bmatrix} 1{,}0 & 0{,}4 \\ 0{,}4 & 1{,}0 \end{bmatrix}; \boldsymbol{S}_{\boldsymbol{yy}} = \begin{bmatrix} 1{,}0 & 0{,}2 \\ 0{,}2 & 1{,}0 \end{bmatrix}$$

$$\boldsymbol{S}_{\boldsymbol{xy}} = \begin{bmatrix} 0{,}5 & 0{,}6 \\ 0{,}3 & 0{,}4 \end{bmatrix}; \boldsymbol{S}_{\boldsymbol{yx}} = \begin{bmatrix} 0{,}5 & 0{,}3 \\ 0{,}6 & 0{,}4 \end{bmatrix}$$

Kemudian dihitung :

$$\boldsymbol{S}_{\boldsymbol{xx}}^{-1} = \begin{bmatrix} 1{,}1905 & -0{,}4762 \\ -0{,}4762 & 1{,}1905 \end{bmatrix}; \boldsymbol{S}_{\boldsymbol{yy}}^{-1} = \begin{bmatrix} 1{,}0417 & -0{,}2083 \\ -0{,}2083 & 1{,}0417 \end{bmatrix}$$

$$\boldsymbol{S}_{\boldsymbol{xx}}^{-1/2} = \begin{bmatrix} 1{,}0681 & -0{,}2229 \\ -0{,}2229 & 1{,}0681 \end{bmatrix}; \boldsymbol{S}_{\boldsymbol{yy}}^{-1/2} = \begin{bmatrix} 1{,}0155 & -0{,}1026 \\ -0{,}1026 & 1{,}0155 \end{bmatrix}$$

Selanjutnya

$$\boldsymbol{A} = \boldsymbol{S}_{xx}^{-1/2}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}^{-1/2} = \begin{bmatrix} 0{,}4371 & 0{,}2178 \\ 0{,}2178 & 0{,}1096 \end{bmatrix}$$

Dari matrik *A* di atas dihitung akar ciri dan vektor ciri pasangannya yaitu :

$$\begin{aligned} |\boldsymbol{A} - \lambda \boldsymbol{I}| &= \boldsymbol{0} \\ \begin{vmatrix} 0{,}4371 - \lambda & 0{,}2178 \\ 0{,}2178 & 0{,}1096 - \lambda \end{vmatrix} &= (0{,}4371 - \lambda)(0{,}1096 - \lambda) - 0{,}2178^2 \\ &= \lambda^2 - 0{,}5467\lambda + 0{,}0005 = 0 \end{aligned}$$

Solusi dari persamaan di atas adalah :

$$\hat{\rho}_1 = \lambda_1 = 0{,}5458 \text{ dan } \hat{\rho}_2 = \lambda_2 = 0{,}0009$$

Kemudian dihitung vektor ciri yang berpasangan dengan akar ciri pertama $\hat{\rho}_1 = \lambda_1 = 0{,}5458$ adalah :

$$\boldsymbol{A}\boldsymbol{e_1} = \lambda_1 \boldsymbol{e_1}$$

$$\begin{bmatrix} 0{,}4371 & 0{,}2178 \\ 0{,}2178 & 0{,}1096 \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix} = 0{,}5458 \begin{bmatrix} e_{11} \\ e_{21} \end{bmatrix}$$

Diperoleh estimasi vektor ciri pertama adalah: $\hat{\boldsymbol{e}}_{\boldsymbol{1}}$ = [0,8947 0,4466]'. Selanjutnya dihitung koefisien fungsi kanonik yaitu :

$$\hat{\boldsymbol{a}}_1 = \hat{\boldsymbol{e}}_{\boldsymbol{1}}\boldsymbol{S}_{xx}^{-1/2}$$

$$\hat{\boldsymbol{a}}_1 = \begin{bmatrix} 0{,}8947 \\ 0{,}4466 \end{bmatrix} \begin{bmatrix} 1{,}0681 & -0{,}2229 \\ -0{,}2229 & 1{,}0682 \end{bmatrix} = \begin{bmatrix} 0{,}8561 \\ 0{,}2776 \end{bmatrix}$$

Kemudian, dicari vektor ciri dari matriks $\boldsymbol{B}$ dengan menggunakan rumus:

$$\hat{\boldsymbol{f}}_k = \frac{1}{\hat{\rho}_k}\boldsymbol{S}_{yy}^{-\frac{1}{2}}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}^{-1}\hat{\boldsymbol{e}}_k$$

Vektor ciri yang pertama adalah :

$$\hat{\boldsymbol{f}}_1 = \frac{1}{\hat{\rho}_1}\boldsymbol{S}_{yy}^{-\frac{1}{2}}\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}^{-1}\hat{\boldsymbol{e}}_1$$

$$\hat{\boldsymbol{f}}_1 = \frac{1}{0{,}5458}\begin{bmatrix} 1{,}0155 & -0{,}1026 \\ -0{,}1026 & 1{,}0155 \end{bmatrix}\begin{bmatrix} 0{,}5 & 0{,}3 \\ 0{,}6 & 0{,}4 \end{bmatrix}\begin{bmatrix} 1{,}1905 & -0{,}4762 \\ -0{,}4762 & 1{,}1905 \end{bmatrix}\begin{bmatrix} 0{,}8947 \\ 0{,}4466 \end{bmatrix}$$

$$\hat{\boldsymbol{f}}_1 = \begin{bmatrix} 0.7479 \\ 0.9442 \end{bmatrix}$$

Kemudian, dihitung koefisien *b* dengan rumus berikut ini.

$$\widehat{\boldsymbol{b}}_k = \boldsymbol{S}_{\boldsymbol{yy}}^{-1/2}\hat{\boldsymbol{f}}_k$$

$$\widehat{\boldsymbol{b}}_1 = \begin{bmatrix} 1{,}0155 & -0{,}1026 \\ -0{,}1026 & 1{,}0155 \end{bmatrix}\begin{bmatrix} 0.7479 \\ 0.9442 \end{bmatrix} = \begin{bmatrix} 0.6626 \\ 0.8821 \end{bmatrix}$$

Fungsi kanonik pertama dapat dituliskan dengan rumus berikut ini.

$$\widehat{\boldsymbol{U}}_{\boldsymbol{1}} = \widehat{\boldsymbol{a}}_1\boldsymbol{X}$$

$$\widehat{\boldsymbol{V}}_{\boldsymbol{1}} = \widehat{\boldsymbol{b}}_1\boldsymbol{Y}$$

Pada kasus ini, estimasi fungsi kanonik pertama adalah :

$$\widehat{U}_1 = \begin{bmatrix} 0{,}8561 \\ 0{,}2776 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = 0{,}8561X_1 + 0{,}2776X_2$$

Korelasi kanonik pertama adalah :

$$r_{\widehat{\boldsymbol{U}}_{\boldsymbol{1}};\widehat{\boldsymbol{V}}_{\boldsymbol{1}}} = \sqrt{\hat{\rho}_1} = \sqrt{0{,}5458} = 0{,}74$$

Korelasi antara himpunan variabel yaitu Kemampuan Bahasa (X_1) dan Kemampuan Berhitung (X_2) dan himpunan variabel yang terdiri dari Nilai IPK (Y_1) dan Masa Studi (Y_2) adalah positif dan hubungannya cukup kuat. Korelasi kanonik kedua yaitu $r_{\widehat{\boldsymbol{U}}_{\boldsymbol{2}};\widehat{\boldsymbol{V}}_{\boldsymbol{2}}} = \sqrt{\hat{\rho}_2} = \sqrt{0{,}0009} = 0{,}03$ sangat kecil, artinya tidak ada hubungan antara fungsi kanonik kedua ini (bisa diabaikan).

Uji Kesesuaian Model

Untuk mengetahui kesesuaian atau kecocokan model fungsi kanonik dapat dilakukan dengan dua acara yang diuraikan seperti di bawah ini.

1. Proporsi keragaman (variansi)

   Semakin besar nilai proporsi keragaman maka semakin baik

variabel-variabel kanonik yang dipilih dengan dapat menjelaskan keragaman asal. Sebagai acuan yang cukup baik yaitu proporsi kumulatifnya harus lebih besar dari 50%.

2. Uji hipotesis

Ada dua hipotesis yang akan diujikan dalam analisis korelasi kanonik yaitu uji hipotesis yang pertama untuk mengetahui apakah secara keseluruhan korelasi kanonik signifikan. Jika pada uji hipotesis yang pertama memperoleh kesimpulan bahwa paling tidak ada ada satu korelasi kanonik tidak bernilai nol, maka dilanjutkan dengan uji hipotesis kedua untuk mengetahui apakah ada sebagian korelasi kanonik signifikan.

Berikut hipotesis uji keseluruhan.

a. Uji Korelasi Kanonik secara serentak (bersama-sama) :

Hipotesisnya adalah :

$$H_0 : \rho_1 = \rho_2 = ... = \rho_k = 0 \;\; vs \;\; H_1 : Min\ ada\ 1\ \rho_1 \neq 0; \text{i} = 1,2,...,k$$

Statistik Uji adalah :

$$B = -\left[-n - 1 - \frac{1}{2}(p + q + 1)\right] \ln\left(\prod_{i=1}^{k}(1 - \rho_i^2)\right)$$

dimana n = jumlah pengamatan

Keputusan :

H_0 ditolak jika $B > X_a^2$ dengan derajat bebas p x q.

b. Uji Korelasi Kanonik secara individu

Hipotesisnya adalah :

$$H_0 : \rho_i = 0 \;\; vs \;\; H_1 : \rho_i \neq 0; \text{i} = 1, 2, ..., k$$

Statistik Uji adalah :

$$B = -\left[-n - 1 - \frac{1}{2}(p + q + 1)\right] \ln\left(\prod_{i=r}^{k}(1 - \rho_i^2)\right)$$

dimana n = jumlah pengamatan

Keputusan :

H_0 ditolak jika $B > X_a^2$ dengan derajat bebas $(p - q)$ x $(q - r)$.

7.4 Prosedur Korelasi Kanonik dengan R

Dalam software R, fungsi yang digunakan untuk melakukan teknik analisis korelasi kanonik adalah fungsi `cancor()` tersedia dalam stats *package* atau `cc()` dan yang tersedia dalam `CCA package`. Pada praktikum ini, analisis kanonik akan menggunakan perintah `cc()` karena hasilnya lebih lengkap dibandingkan perintah `cancor()`.

Sintak perintah cc() seperti di bawah ini.

```
cc(x, y)

Keterangan :
x = matriks data variabel X
y = matriks data variabel Y
```

Output yang dihasilkan dari perintah cc() terdiri dari :

`cor`	:	Korelasi kanonik
`names`	:	Daftar nama variabel dependen dan independen.
`xcoef`	:	Estimasi koefisien variabel *X*
`ycoef`	:	Estimasi koefisien variabel *Y*
`scores`	:	skor untuk setiap pengamatan serta korelasi untuk setiap fungsi kanonik (variat kanonik).

Contoh

Pada kasus ini akan menggunakan data yang tersedia di R yaitu `LifeCycleSavings`. Data ini terdiri dari 50 pengamatan (negara) dengan 5 variabel yaitu :

- sr = banyaknya alokasi tabungan
- pop15 = persentase penduduk dibawah usia 15 tahun
- pop75 = persentase penduduk di atas usia 75 tahun
- dpi= pendapatan yang siap digunakan (*disposable income*)
- ddpi = laju pertumbuhan dpi (pendapatan perkapita)

Pada kasus ini, misalkan variabel *X* adalah "sr", "dpi" dan "ddpi" dan variabel *Y* adalah "pop15" dan "pop75". Oleh karena itu, tujuannya adalah untuk mengetahui hubungan (keeratan) antara himpunan variabel X dan Y tersebut.

Detail data dapat diketahui dengan mengetikan perintah berikut ini :

```
> LifeCycleSavings #menampilkan data LifeCycleSavings

                sr  pop15  pop75     dpi   ddpi
Australia    11.43  29.35   2.87 2329.68   2.87
Austria      12.07  23.32   4.41 1507.99   3.93
Belgium      13.17  23.80   4.43 2108.47   3.82
Bolivia       5.75  41.89   1.67  189.13   0.22
Brazil       12.88  42.19   0.83  728.47   4.56
Canada        8.79  31.72   2.85 2982.88   2.43
Chile         0.60  39.74   1.34  662.86   2.67
China        11.90  44.75   0.67  289.52   6.51
Colombia      4.98  46.64   1.06  276.65   3.08
Costa Rica   10.78  47.64   1.14  471.24   2.80
Denmark      16.85  24.42   3.93 2496.53   3.99
Ecuador       3.59  46.31   1.19  287.77   2.19
Finland      11.24  27.84   2.37 1681.25   4.32
France       12.64  25.06   4.70 2213.82   4.52
Germany      12.55  23.31   3.35 2457.12   3.44
Greece       10.67  25.62   3.10  870.85   6.28
Guatamala     3.01  46.05   0.87  289.71   1.48
Honduras      7.70  47.32   0.58  232.44   3.19
Iceland       1.27  34.03   3.08 1900.10   1.12
India         9.00  41.31   0.96   88.94   1.54
Ireland      11.34  31.16   4.19 1139.95   2.99
Italy        14.28  24.52   3.48 1390.00   3.54
Japan        21.10  27.01   1.91 1257.28   8.21
Korea         3.98  41.74   0.91  207.68   5.81
Luxembourg   10.35  21.80   3.73 2449.39   1.57
Malta        15.48  32.54   2.47  601.05   8.12
Norway       10.25  25.95   3.67 2231.03   3.62
Netherlands  14.65  24.71   3.25 1740.70   7.66
New Zealand  10.67  32.61   3.17 1487.52   1.76
Nicaragua     7.30  45.04   1.21  325.54   2.48
Panama        4.44  43.56   1.20  568.56   3.61
Paraguay      2.02  41.18   1.05  220.56   1.03
Peru         12.70  44.19   1.28  400.06   0.67
```

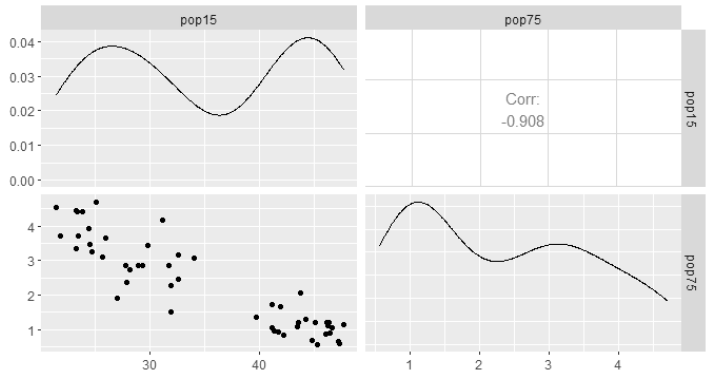
```
Philippines      12.78  46.26   1.12    152.01   2.00
Portugal         12.49  28.96   2.85    579.51   7.48
South Africa     11.14  31.94   2.28    651.11   2.19
South Rhodesia   13.30  31.92   1.52    250.96   2.00
Spain            11.77  27.74   2.87    768.79   4.35
Sweden            6.86  21.44   4.54   3299.49   3.01
Switzerland      14.13  23.49   3.73   2630.96   2.70
Turkey            5.13  43.42   1.08    389.66   2.96
Tunisia           2.81  46.12   1.21    249.87   1.13
United Kingdom    7.81  23.27   4.46   1813.93   2.01
United States     7.56  29.81   3.43   4001.89   2.45
Venezuela         9.22  46.40   0.90    813.39   0.53
Zambia           18.56  45.25   0.56    138.33   5.14
Jamaica           7.72  41.12   1.73    380.47  10.23
Uruguay           9.24  28.13   2.72    766.54   1.88
Libya             8.89  43.69   2.07    123.58  16.71
Malaysia          4.71  47.20   0.66    242.69   5.08
```

Sebelum melakukan analisis korelasi kanonik, pertama ditentukan dulu himpunan variabel *X* dan *Y*.

```
> Y=LifeCycleSavings[,2 :3]
> X=LifeCycleSavings[,c(1,4 :5)]
```

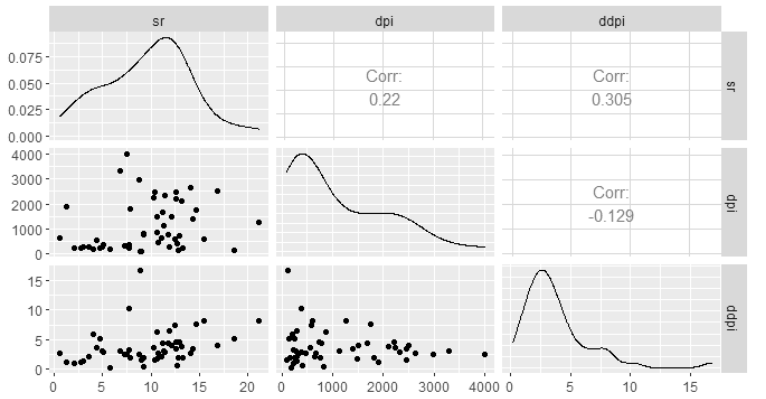
Selanjutnya, apabila ingin melihat hubungan antara variabel independen dan dependen dapat dilakukan dengan menggunakan perintah ggpairs yang tersedia dalam paket GGally.

```
> library(GGaly)
> ggpairs(Y)
```



Gambar di atas memperlihatkan pola sebaran antara variable "pop15" dan "pop75" serta korelasi diantara keduanya. Besarnya korelasi adalah -0,909 artinya ada hubungan negatif antara variabel pop15 dan pop75.

```
> ggpairs(X)
```



Pola hubungan dan korelasi diantara gugus (himpunan) variabel X digambarkan pada output di atas. Terlihat hubungan antara variabel dependen tidak terlalu besar, korelasi antara variabel "dpi" dengan "sr" sebesar 0,22; korelasi antara "ddpi" dan "sr" sebesar 0,305 sedangkan korelasi antara "dpi"

dengan “dppi” yaitu -0,129.

Selanjutnya, untuk mencari nilai korelasi kanonik dapat diperoleh melalui perintah berikut ini.

```
> Kanonik=cc(X,Y)
> Kanonik$cor #menampilkan korelasi kanonik
[1] 0.8247966 0.3652762
```

Berdasarkan output terlihat bahwa korelasi kanonik pertama sebesar 0,825 dan korelasi kanonik kedua sebesar 0,365.

Kemudian, fungsi kanonik dapat dicari melalui perintah berikut ini.

```
> Kanonik$xcoef
              [,1]           [,2]
sr   -0.0592971550   0.2336554912
dpi  -0.0009151786  -0.0005311762
ddpi -0.0291942000  -0.0858752749

> Kanonik$ycoef
              [,1]        [,2]
pop15   0.06377599  -0.2535544
pop75  -0.34053260  -1.8221811
```

Fungsi kanonik pertama yaitu:

$$U_1 = -0.0593\ sr - 0.0009\ dpi - 0.0292\ ddpi$$

$$V_1 = 0.0638\ pop15 - 0.3405\ pop75$$

Persamaan di atas dapat diinterpretasikan sebagai berikut: misalkan variabel “sr” pada fungsi kanonik yang pertama memiliki koefisien sebesar -0,0539 artinya setiap kenaikan 1 satuan variabel “sr” akan menurunkan 0,0539 terhadap fungsi kanonik pertama (U_1) dengan asumsi variabel “dpi” dan “ddpi” konstan. Begitu juga untuk variabel “pop15” pada fungsi kanonik yang pertama koefisien sebesar 0,0638 artinya setiap kenaikan 1% variabel “pop15” akan menaikan sebesar 0,0638 terhadap fungsi kanonik pertama (V_1) dengan asumsi variabel “pop75” konstan (tetap). Untuk koefisen yang lain dapat diinterpretasikan dengan cara yang sama.

Sedangkan, fungsi kanonik kedua adalah

$$U_2 = 0.2337\ sr - 0.0005\ dpi - 0.0859\ ddpi$$
$$V_2 = -0.2536\ pop15 - 1.8222\ pop17$$

Berdasarkan koefisien fungsi kanonik (*weight canonic*) dapat disimpulkan hasil sebagai berikut.

a. Pada fungsi kanonik pertama urutan kontribusi relatif variabel-variabel independen terhadap variabel kanonik adalah "sr", "ddpi" dan "dpi",
b. Pada fungsi kanonik pertama, urutan kontribusi relatif variabel-variabel dependen terhadap variabel kanonik adalah "pop75" dan "pop15",
c. Pada fungsi kanonik kedua, urutan kontribusi relatif variabel-variabel independen terhadap variabel kanonik adalah "sr", "ddpi" dan "dpi",
d. Sedangkan pada fungsi kedua, urutan kontribusi relatif variabel-variabel dependen terhadap variabel kanonik adalah "pop15" dan "pop75",

Analisis berikutnya adalah menghitung nilai loading. Loading kanonik dapat dihitung dari korelasi antara variabel asal dengan masing-masing variabel kanoniknya. Semakin besar nilai *loading* mencerminkan semakin dekat hubungan fungsi kanonik yang bersangkutan dengan variabel asal.

Perintah untuk menghitung loading kanonik dalam R menggunakan `comput()`. Berikut perintahnya :

```
> Kanonik2=comput(X,Y,Kanonik)
> Kanonik2$corr.X.xscores
           [,1]          [,2]
sr   -0.4910379     0.8557760
dpi  -0.9545172    -0.2637266
ddpi -0.0473377     0.1407737
> Kanonik2$corr.Y.xscores
            [,1]          [,2]
pop15  0.8107603   -0.06710179
pop75 -0.7998819   -0.08910177
> Kanonik2$corr.X.yscores
             [,1]         [,2]
sr    -0.40500636   0.31259455
dpi   -0.78728255  -0.09633306
ddpi  -0.03904398   0.05142128
> Kanonik2$corr.Y.yscores
             [,1]        [,2]
pop15   0.9829821  -0.1837015
pop75  -0.9697929  -0.2439299
```

Output di atas menggambarkan korelasi antara variabel dengan variatnya (fungsi kanonik). Perhatikan nilai loading -0,4910 menjelaskan korelasi dari variabel "sr" terhadap fungsi kanonik pertama artinya hubungan negatif dan cukup kuat, korelasi antara variabel "dpi" dengan fungsi kanonik pertama sebesar -0,9545 sedangkan korelasi antara variabel "ddpi" terhadap fungsi kanonik pertama sebesar -0,0473. Interpretasi untuk output lainnya dapat dilakukan dengan cara yang sama.

Berdasarkan *loading* kanonik dapat disimpulkan berikut ini.

a. Pada fungsi kanonik pertama, urutan variabel-variabel yang hubungannya paling erat dengan variabel kanonik dependen yaitu "pop15" kemudian "pop75" (lihat nilai pada `corr.Y.yscores`). Sedangkan urutan variabel-variabel yang hubungannya paling erat dengan variabel kanonik independen yaitu "dpi", "sr" terakhir "dppi" (lihat nilai pada `corr.X.xscores`).
b. Pada fungsi kanonik kedua, urutan variabel-variabel yang hubungannya paling erat dengan variabel kanonik dependen yaitu

"pop75" kemudian "pop15". Sedangkan urutan variabel-variabel yang hubungannya paling erat dengan variabel kanonik independen yaitu "sr", "dp" dan "dppi".

Apabila memperhatikan nilai *cross loading* dapat disimpulkan berikut ini.

a. Pada fungsi kanonik pertama, variabel dependen yang hubungannya paling erat dengan variabel kanonik independen yaitu "pop15" kemudian "pop75" (lihat nilai `corr.Y.xscore`). Sedangkan variabel independen yang hubungannya paling erat dengan variabel kanonik dependen yaitu "dpi", "sr" terakhir "dppi". (lihat nilai `corr.X.yscore`).
b. Pada fungsi kanonik kedua, variabel dependen yang hubungannya paling erat dengan variabel kanonik independen berturut-turut adalah "pop75" kemudian "pop15". Sedangkan variabel independen yang hubungannya paling erat dengan variabel kanonik dependen berturut-turut adalah "sr", "dp" dan "dppi".

Tahap berikutnya adalah menguji kesesuaian model kanonik. Perintah yang digunakan yaitu `p.asym()`. Perintah ini untuk menguji signifikansi dari fungsi kanonik dengan menggunakan pilihan statistik ujinya adalah uji Wilks, Hotelling, Pillai atau Roys. Sintak dari `p.asym()` adalah :

```
p.asym(rho, N, p, q, tstat = "Wilks")

Keterangan

rho = vektor koefisien korelasi kanonik
n = banyaknya pengamatan
p = banyaknya variabel independen
q = banyaknya variabel dependen
```

Untuk menggunakan perintah p.asym() harus menginstall paket "CCP". Berikut perintahnya :

```
> install.packages("CCP")
> library(CCP)
```

Setelah paket "CCP" terinstal maka langkah berikutnya adalah

melakukan uji serentak untuk fungsi kanonik.

```
> n=dim(LifeCycleSavings)[1]
> n
[1] 50
> p=length(X)
> p
[1] 3
> q=length(Y)
> q
[1] 2
> rho=Kanonik$cor
> rho
[1] 0.8247966 0.3652762
> p.asym(rho,n,p,q,tstat = "Wilks")
Wilks' Lambda, using F-approximation (Rao's F) :
              stat    approx  df1  df2        p.value
1 to 2:  0.2770526  13.49772    6   90   7.300349e-11
2 to 2:  0.8665733   3.54132    2   46   3.711268e-02
```

Pada output di atas menjelaskan bahwa fungsi kanonik 1 dan 2 signifikan secara statistik. Hal ini dibuktikan dengan *p_value* sebesar 7,3003e-11 ≈ 0,00 keputusannya H_0 ditolak artinya jadi fungsi kanonik pertama dan kedua dapat digunakan untuk diproses lebih lanjut.

Untuk uji individu, perhatikan statistic uji fungsi kanonik kedua (karena pada kasus ini hanya terbentuk 2 fungsi kanonik, fungsi kanonik yang pertama tidak keluar outputnya). Nilai statisitik uji sebesar 3,5413 dan *p_value* sebesar 0,03711 < . Keputusan H_0 ditolak artinya fungsi kanonik kedua signifikan dan dapat digunakan untuk proses lebih lanjut.

Berdasarkan hasil analisis korelasi kanonik maka dapat disimpulkan terdapat hubungan antara "sr", "dpi" dan "dppi" terhadap "pop15" dan "pop75".

7.5 Latihan

1. Suatu penelitian ingin mengetahui hubungan antara Motivasi Belajar (X_1) dan Minat (X_2) terhadap Nilai IPK (Y_1) dan Masa Studi (Y_2). Berdasarkan hasil pengamatan diperolah matriks kovariansi sebagai berikut:

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S_{xx}} & \boldsymbol{S_{xy}} \\ \boldsymbol{S_{yx}} & \boldsymbol{S_{yy}} \end{bmatrix} = \begin{bmatrix} 2{,}0 & 0{,}4 & 0{,}3 & 0{,}6 \\ 0{,}4 & 1{,}0 & 0{,}8 & 0{,}4 \\ 0{,}3 & 0{,}8 & 2{,}0 & 0{,}6 \\ 0{,}6 & 0{,}4 & 0{,}6 & 3{,}0 \end{bmatrix}$$

Berdasarkan data di atas, buatlah fungsi kanonik dan hitunglah korelasi kanoniknya!

2. Suatu penelitian ingin mengukur hubungan antara tingkat kepuasan nasabah dengan kualitas layanan. Variabel kualitas layanan terdiri dari fasilitas (X_1) dan kepedulian (X_2) sedangkan tingkat kepuasan diukur kepuasan produk (Y_1) dan jasa (Y_2). Berdasarkan hasil pengamatan diperolah matriks kovariansi sebagai berikut.

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S_{xx}} & \boldsymbol{S_{xy}} \\ \boldsymbol{S_{yx}} & \boldsymbol{S_{yy}} \end{bmatrix} = \begin{bmatrix} 8 & 2 & 3 & 1 \\ 2 & 5 & -1 & 3 \\ 3 & -1 & 6 & -2 \\ 1 & 3 & -2 & 7 \end{bmatrix}$$

Berdasarkan data di atas, buatlah fungsi kanonik dan hitunglah korelasi kanoniknya!

3. Lakukan analisis korelasi kanonik pada data mtcars (lihat data pada contoh di atas) yaitu untuk melihat hubungan antara variabel independen yaitu "am", "gear" dan "carb" terhadap variabel Y yaitu "mpg", "hp" dan "wt". Dengan demikian pada kasus ini ada 3 variabel independen dan 3 variabel dependen.
   a. Tentukan fungsi kanoniknya! Interpretasikan!
   b. Hitung loading kanonik dan interpretasikan?
   c. Uji signifikansi secara serentak dan individu untuk fungsi kanonik yang terbentuk tersebut! Analisis!

4. Lakukan analisis korelasi kanonik untuk melihat apakah ada hubungan serentak variabel harapan hidup, harapan lama sekolah, rata-rata lama sekolah dengan variabel pendapatan per kapita dan indeks pembangunan manusia!

No	Provinsi	Harapan Hidup	Harapan Lama Sekolah	Lama Sekolah	Pendapatan Per Kapita	Indeks Pembangunan Manusia
1	Aceh	73.20	14.39	9.64	10.811	75.36

No	Provinsi	Harapan Hidup	Harapan Lama Sekolah	Lama Sekolah	Pendapatan Per Kapita	Indeks Pembangunan Manusia
2	Sumatera Utara	73.90	13.49	9.93	11.46	75.76
3	Sumatera Barat	74.37	14.3	9.44	11.718	76.43
4	Riau	74.41	13.42	9.43	11.857	75.67
5	Jambi	74.06	13.14	8.9	11.621	74.36
6	Sumatera Selatan	74.26	12.64	8.57	12.015	73.84
7	Bengkulu	73.31	13.75	9.04	11.733	74.91
8	Lampung	74.39	12.78	8.36	11.258	73.13
9	Kep. Bangka Belitung	74.12	12.49	8.33	13.667	74.55
10	Jawa Barat	75.16	12.8	8.87	12.157	79.89
11	Jawa Tengah	74.91	12.86	8.02	12.276	84.15
12	Jawa Timur	75.07	13.43	8.28	12.852	74.92
13	Banten	74.97	13.1	9.23	13.097	73.87
14	Nusa Tenggara Barat	72.25	13.98	7.87	11.606	81.62
15	Kalimantan Barat	73.94	12.68	7.78	10.321	75.35
16	Kalimantan Tengah	73.73	12.77	8.81	12.303	76.35
17	Kalimantan Selatan	74.18	12.87	8.62	13.399	78.63
18	Kalimantan Timur	74.94	14.03	10.02	13.793	73.1
19	Kalimantan Utara	73.57	13.21	9.35	10.197	69.14
20	Sulawesi Utara	74.08	12.97	9.84	11.998	71.19
21	Sulawesi Selatan	73.83	13.55	8.86	12.275	74.28
22	Kepulauan Riau	75.12	13.27	10.5	15.573	75.19
23	DKI Jakarta	75.99	13.51	11.49	19.953	78.79
24	D.I. Yogyakarta	75.36	15.7	9.92	15.361	73.41
25	Bali	75.10	13.62	9.54	14.92	75.68

No	Provinsi	Harapan Hidup	Harapan Lama Sekolah	Lama Sekolah	Pendapatan Per Kapita	Indeks Pembangunan Manusia
26	Nusa Tenggara Timur	71.83	13.23	8.02	8.534	72.24
27	Sulawesi Tengah	70.84	13.34	9.04	10.536	75.18
28	Sulawesi Tenggara	71.88	13.71	9.42	10.606	73.62
29	Gorontalo	70.73	13.17	8.29	11.539	72.01
30	Sulawesi Barat	71.03	12.89	8.15	10.208	70.46
31	Maluku	70.68	14.09	10.26	9.684	73.4
32	Papua Barat	68.47	13.17	7.86	8.805	71.84
33	Papua Barat Daya	70.02	13.88	8.39	8.733	67.69
34	Papua	70.47	13.72	9.82	11.037	69.65
35	Papua Selatan	68.46	12.67	8.38	9.756	73.83
36	Maluku Utara	71.05	13.75	9.37	9.32	68.86
37	Papua Tengah	68.18	9.63	6.12	7.809	60.25
38	Papua Pegunungan	67.39	9.97	4.21	5.707	54.43

BAB

VIII

ANALISIS DISKRIMINAN (DISCRIMINAT ANALYSIS)

8.1 Pendahuluan

Salah satu metode dalam analisis multivariat adalah analisis diskriminan. Analisis diskriminan bertujuan untuk mengklasifikasikan suatu individu atau observasi ke dalam kelompok yang saling bebas (*mutually exclusive/disjoint*) dan menyeluruh (*exhaustive*) dengan berdasarkan sejumlah variabel penjelas.

Analisis diskriminan termasuk dalam kategori teknik dependensi, artinya hubungan antara variabel dapat dibedakan menjadi variabel terikat (respon) dan variabel bebas (penjelas). Konsep analisis diskriminan hampir sama dengan analisis regresi, perbedaanya adalah di dalam analisis diskriminan variabel terikatnya (respon) berupa data kualitatif (nominal atau ordinal) dan variabel bebas berupa data kuantitatif (interval atau rasio).

Analisis diskriminan digunakan untuk membuat satu model prediksi keanggotaan kelompok didasarkan pada karakteristik-karakteristik yang diobservasi untuk masing-masing kasus. Prosedur ini akan menghasilkan fungsi diskriminan yang didasarkan pada kombinasi-kombinasi linier yang berasal dari variabel-variabel prediktor atau bebas yang dapat menghasilkan perbedaan paling baik antara kelompok-kelompok yang dianalisis.

Misalnya divisi Marketing suatu perusahaan ingin mengklasifikasikan tipe-tipe konsumen berdasarkan kriteria/variabel tertentu. Dalam kasus ini tipe-tipe konsumen dibedakan menjadi dua yaitu konsumen loyal dan tidak loyal, sedangkan variabel penjelasnya terdiri dari faktor usia, pendapatan, dan tingkat kepuasan konsumen.

Ada dua tujuan utama dalam analisis diskriminan, yaitu:

1. Fungsi diskriminan yang terbentuk digunakan untuk menggambarkan atau menjelaskan perbedaan antara kelompok-kelompok tersebut. Tujuannya adalah dapat mengidentifikasi kontribusi variabel bebas dalam memisahkan kelompok dan menjelaskan gambaran terbaik setiap kelompok.
2. Fungsi diskriminan dapat digunakan untuk memprediksikan suatu observasi menjadi anggota salah satu kelompok.

Untuk menggunakan teknik analisis ini syarat-syarat yang harus dipenuhi diantaranya :

1. Variabel tergantung hanya satu dan bersifat non-metrik, artinya data harus kategorikal dan berskala nominal.
2. Variabel bebas terdiri lebih dari dua variabel dan berskala interval.
3. Semua kasus harus independen.
4. Semua variabel prediktor sebaiknya mempunyai distribusi normal multivariat, dan matrik *variance-covariance* dalam kelompok harus sama untuk semua kelompok.
5. Keanggotaan kelompok diasumsikan ekseklusif, maksudnya tidak satupun kasus yang termasuk dalam kelompok lebih dari satu. Dan *exhaustive* secara kolektif, maksudnya semua kasus merupakan anggota satu kelompok

Ada beberapa metode untuk melakukan analisis diskriminan, diantaranya:

a. Analisis Diskriminan Linear Fisher.
b. Analisis Diskriminan Kuadratik.
c. Analisis Diskriminan Non Parametrik.

8.2 Model Analisis Diskriminan Linear

Pada buku ini, analisis diskriminan linear Fisher (analisis diskriminan) akan dijelaskan karena model ini paling popular digunakan. Pada prinsipnya analisis diskriminan bertujuan untuk mengklasifikasikan setiap obyek (individu atau observasi) ke dalam dua atau lebih kelompok berdasarkan pada sejumlah variabel bebas tertentu. Pengklasifikasian ini bersifat *mutually exclusive* artinya jika objek A sudah masuk ke dalam kelompok pertama, maka objek tersebut tidak mungkin menjadi anggota kelompok kedua, ketiga dan seterusnya.

Menurut Supranto (2004), teknik analisis diskriminan dibedakan menjadi dua, yaitu analisis diskriminan dua kelompok dan analisis diskriminan berganda. Analisis diskriminan dua kelompok terbentuk apabila variabel terikat (Y) terdiri dari dua kategori atau dua kelompok ($k = 2$). Pada kasus ini akan terbentuk satu fungsi diskriminan. Sedangkan, analisis diskriminan berganda terjadi jika variabel dependen (Y) dikelompokkan menjadi lebih dari dua kategori/kelompok ($k > 2$) maka diperlukan fungsi diskriminan sebanyak

= (k - 1).

Misalkan ada dua kelompok atau populasi yaitu π_1 dan π_2 dengan p variabel bebas yaitu $\boldsymbol{X} = [X_1, X_2, \ldots, X_p,]'$ Pada masing-masing populasi diperoleh rata-rata:

$$\boldsymbol{\mu_1} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \text{ dan } \boldsymbol{\mu_2} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

dan matriks variansi kovariansi yaitu :

$$\boldsymbol{\Sigma_1} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} \cdots & \sigma_{2p} \\ \vdots & \vdots \ddots & \cdots \\ \sigma_{p1} & \sigma_{p2} \cdots & \sigma_{pp} \end{bmatrix} \text{ dan } \boldsymbol{\Sigma_2} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} \cdots & \sigma_{2p} \\ \vdots & \vdots \ddots & \cdots \\ \sigma_{p1} & \sigma_{p2} \cdots & \sigma_{pp} \end{bmatrix}$$

Model fungsi diskriminan untuk dua kelompok adalah :

$$\boldsymbol{Y} = \lambda \boldsymbol{X} = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\boldsymbol{\Sigma^{-1}}\boldsymbol{X} \qquad (8.1)$$

Pada prakteknya parameter populasi $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ dan $\boldsymbol{\Sigma}_2$ tidak diketahui oleh karena itu fungsi diskriminan diestimasi menggunakan data sampel. Misalkan sampel acak diambil dari populasi pertama (π_1) sebanyak n_1 dan sampel acak diambil dari populasi kedua kedua (π_1) sebanyak n_1, sehingga diperoleh vektor rata-rata sampel adalah :

$$\bar{\boldsymbol{x}}_{\boldsymbol{1}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \text{ dan } \bar{\boldsymbol{x}}_{\boldsymbol{2}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

dan matriks variansi kovariansi sampel yaitu :

$$\mathbf{S_1} = \begin{bmatrix} s_{11} & s_{12} \cdots & s_{1p} \\ s_{21} & s_{22} \cdots & s_{2p} \\ \vdots & \vdots \ddots & \cdots \\ s_{p1} & s_{p2} \cdots & s_{pp} \end{bmatrix} \text{ dan } \mathbf{S_2} = \begin{bmatrix} s_{11} & s_{12} \cdots & s_{1p} \\ s_{21} & s_{22} \cdots & s_{2p} \\ \vdots & \vdots \ddots & \cdots \\ s_{p1} & s_{p2} \cdots & s_{pp} \end{bmatrix}$$

Diasumsikan bahwa kedua matriks variansi kovariansi sama maka matriks kovariansi gabungan adalah :

$$\mathbf{S}_{gab} = \frac{(n_1 - 1)}{(n_1 - 1) + (n_2 - 1)}\mathbf{S}_1 + \frac{(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}\mathbf{S}_2 \qquad (8.2)$$

Estimasi fungsi diskriminan adalah :

$$\hat{Y} = (\bar{\boldsymbol{X}}_{\mathbf{1}} - \bar{\boldsymbol{X}}_{\mathbf{2}})' \boldsymbol{S}_{\boldsymbol{gab}}^{-\mathbf{1}} \boldsymbol{X} = \hat{\boldsymbol{\lambda}} \boldsymbol{X} \qquad (8.2)$$

Keterangan :

$\hat{\boldsymbol{\lambda}}$ adalah koefisien fungsi diskriminan, dengan $\hat{\boldsymbol{\lambda}} = (\bar{\boldsymbol{X}}_{\mathbf{1}} - \bar{\boldsymbol{X}}_{\mathbf{2}})' \boldsymbol{S}_{\boldsymbol{gab}}^{-\mathbf{1}}$.

$\hat{Y}$ adalah skor diskriminan

$\boldsymbol{S}_{\boldsymbol{gab}}^{-\mathbf{1}}$ adalah invers matriks kovariansi gabungan

Contoh

Suatu penelitian ingin mendeteksi penyakit kanker berdasarkan gaya hidup dan pola makan. Sampel pasien diambil dari dua populasi/kelompok yaitu Pasien Kanker dan Bukan Pasien Kanker. Hasil pengamatan disajikan pada matrik dibawah ini.

$$\boldsymbol{X}_{\mathbf{1}} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 3 \\ 4 & 5 \\ 5 & 5 \end{bmatrix} \text{ dan } \boldsymbol{X}_{\mathbf{2}} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 1 \\ 3 & 2 \\ 5 & 3 \\ 6 & 5 \end{bmatrix}$$

Berdasarkan data di atas buatlah fungsi diskriminan yang dapat membedakan/ mengklasifikasikan pasien kanker dan bukan pasien kanker!

Jawab

Berdasarkan data di atas, diketahui $n_1 = 5$ dan $n_2 = 6$

Kemudian dihitung vektor rata-rata sampel:

$$\bar{\boldsymbol{X}}_1 = \begin{bmatrix} 3{,}0 \\ 3{,}6 \end{bmatrix} \text{ dan } \bar{\boldsymbol{X}}_2 = \begin{bmatrix} 3{,}3 \\ 2{,}0 \end{bmatrix}$$

Dihitung matriks variansi kovariansi adalah :

$$\boldsymbol{S}_1 = \begin{bmatrix} 2{,}5 & 2{,}0 \\ 2{,}0 & 1{,}8 \end{bmatrix} \text{ dan } \boldsymbol{S}_2 = \begin{bmatrix} 3{,}5 & 3{,}2 \\ 3{,}2 & 3{,}2 \end{bmatrix}$$

Selanjutnya dihitung matriks matriks kovariansi gabungannya yaitu:

$$\boldsymbol{S_{gab}} = \frac{(5-1)}{(5-1)+(6-1)}\boldsymbol{S_1} + \frac{(6-1)}{(5-1)+(6-1)}\boldsymbol{S_2}$$

$$= \begin{bmatrix} 1{,}111 & 0{,}889 \\ 0{,}889 & 0{,}800 \end{bmatrix} + \begin{bmatrix} 1{,}945 & 1{,}778 \\ 1{,}778 & 1{,}778 \end{bmatrix} = \begin{bmatrix} 3{,}056 & 2{,}667 \\ 2{,}667 & 2{,}578 \end{bmatrix}$$

Selanjutnya cari matriks inversnya :

$$\boldsymbol{S_{gab}^{-1}} = \begin{bmatrix} 3{,}368 & -3{,}484 \\ -3{,}484 & 3{,}992 \end{bmatrix}$$

Estimasi koefisien fungsi disrkiminan adalah :

$$\hat{\lambda} = (\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}})' \boldsymbol{S_{gab}^{-1}}$$

$$= [0{,}3 \quad 1{,}2] \begin{bmatrix} 3{,}368 & -3{,}484 \\ -3{,}484 & 3{,}992 \end{bmatrix} = \begin{bmatrix} -3{,}1704 \\ 3{,}7452 \end{bmatrix}$$

sehingga, estimasi persamaan diskriminan :

$$\hat{Y} = -3{,}1704X_1 + 3{,}7452X_2$$

Asumsi dalam model diskriminan adalah:

- Variabel bebas harus terdistribusi normal.
- Matriks kovariansi variabel bebas harus sama (*equal*).
- Tidak terjadi multikolinearitas (tidak berkorelasi) antarvariabel bebas.
- Tidak terdapat data yang ekstrim (*outlier*).

Langkah-langkah membuat model diskriminan yaitu:

- Memilah variabel-variabel menjadi variabel terikat (*dependent*) dan variabel bebas (*independent*).
- Estimasi fungsi diskriminan.
- Menguji signifikansi Fungsi Diskriminan yang terbentuk, dengan menggunakan `Wilk's Lambda`, `Pilai`, `F test`, dan lainnya.
- Menguji ketepatan klasifikasi dari fungsi diskriminan.
- Melakukan interpretasi fungsi diskriminan.
- Melakukan uji validasi fungsi diskriminan.

Salah satu hal yang penting dalam melakukan analisis diskriminan adalah mengevaluasi ketepatan metode klasifikasi. Kriteria yang digunakan adalah ukuran *apparent error rate* (APER) dimana metode ini bertujuan untuk menghitung persentase kesalahan pengklasifikasian (*misclassified*) dengan menggunakan fungsi klasifikasi.

Perhitungan APER dapat dikerjakan dengan mudah dengan menggunakan matriks konfusi. Matriks konfusi menunjukkan keanggotaan kelompok aktual terhadap keanggotaan kelompok yang diprediksi. Misalkan diambil sampel acak dari populasi pertama sebanyak dan sampel acak diambil dari populasi kedua kedua sebanyak , maka matriks konfusinya adalah :

		Prediksi Keanggotaan		Total
		π_1	π_2	
Keanggotaan Aktual	π_1	π_{1C}	$n_{1M} = n_1 - \pi_1$	n_1
	π_1	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

Keterangan :

- n_{1C} adalah banyaknya sampel dari populasi/kelompok 1 yang sesuai klasifikasi,
- n_{1M} adalah banyaknya sampel dari populasi/kelompok 1 yang sesuai misklasifikasi,
- n_{2C} adalah banyaknya sampel dari populasi/kelompok 2 yang sesuai klasifikasi,
- n_{1M} adalah banyaknya sampel dari populasi/kelompok 2 yang sesuai misklasifikasi.

Rumus APER adalah :

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \times 100\% \qquad (8.2)$$

Apabila suatu fungsi diskriminan memiliki nilai APER kecil maka model diskriminan sudah baik. Selanjutnya dapat dihitung ketepatan klasfikasi melalui rumus:

Ketepatan Klasifikasi = 1 - *APER*

Rumus di atas menunjukkan angka ketepatan prediksi dari model diskriminan. Pada umumnya ketepatan di atas 50% dianggap memadai atau valid.

Untuk mengevaluasi fungsi diskriminan fungsi diskriminan pada contoh pengklasifikasian pasien kanker maka diambil sampel sebanyak 12 orang dari masing- masing kelompok. Hasil pengujian keanggotaan ditampilkan dalam matriks konfusi berikut ini.

Tabel 17. Matriks Konfusi

		Prediksi Keanggotaan		Total
		π_1	π_1	
Keanggotaan Aktual	π_1	10		
	π_2	1	11	12

Jawab

Berdasarkan matriks konfusi di atas, nilai APER adalah :

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \times 100\%$$

Persentase ketepatan pengklasifikasian sebesar 87,5%. Angka ketepatan prediksi dari model diskriminan lebih dari 50% sehingga dapat disimpulkan model disrkiminan sudah memadai atau valid.

8.3 Prosedur Analisis Diskriminan dengan R

Software R sudah menyediakan fasilitas analisis diskriminan pada *package* MASS dengan perintah lda(). Sintak perintahnya seperti di bawah ini.

```
lda(formula, data, ..., subset, na.action)
Keterangan :
Formula: group ~ x1 + x2 + …
Data: matriks data
```

Contoh

Pada kasus ini diambil data dari software yaitu data iris. Data iris adalah data mengenai bunga iris yang terdiri dari 3 spesies Iris yaitu setosa, versicolor, and virginica. Sampel sebanyak 150 jenis bunga tersebut kemudian diukur 4 buah variabel yaitu Sepal.Length, Sepal.Width, Petal.Length dan Petal.Width. Berikut tampilan data iris untuk 10 pengamatan pertama dan 1 teakhir.

```
> iris[1 :10,]
    Sepal.Length Sepal.Width Petal.Length Petal.Width  Species
1            5.1         3.5          1.4         0.2   setosa
2            4.9         3.0          1.4         0.2   setosa
3            4.7         3.2          1.3         0.2   setosa
4            4.6         3.1          1.5         0.2   setosa
5            5.0         3.6          1.4         0.2   setosa
6            5.4         3.9          1.7         0.4   setosa
7            4.6         3.4          1.4         0.3   setosa
8            5.0         3.4          1.5         0.2   setosa
9            4.4         2.9          1.4         0.2   setosa
10           4.9         3.1          1.5         0.1   setosa
> iris[140 :150,]
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
140          6.9         3.1          5.4         2.1 virginica
141          6.7         3.1          5.6         2.4 virginica
142          6.9         3.1          5.1         2.3 virginica
143          5.8         2.7          5.1         1.9 virginica
144          6.8         3.2          5.9         2.3 virginica
145          6.7         3.3          5.7         2.5 virginica
146          6.7         3.0          5.2         2.3 virginica
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

Berdasarkan data di atas dibuat fungsi diskriminan yang membedakan ketiga jenis spesies bunga iris tersebut. Dalam kasus ini ada 3 kelompok yaitu *setosa, versicolor,* dan *virginica*. Apabila ada 3 kelompok (grup) maka akan terbentuk 2 fungsi diskriminan. Berikut pengolahan data dengan menggunakan R melalui perintah `lda()`.

```
> D=lda(Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.
Width,data=iris)
> D
Call :
lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length +
Petal.Width,
 data = iris)

Prior probabilities of groups :
    setosa  versicolor  virginica
 0.3333333   0.3333333  0.3333333

Group means :
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants :
                     LD1          LD2
Sepal.Length   0.8293776   0.02410215
Sepal.Width    1.5344731   2.16452123
Petal.Length  -2.2012117  -0.93192121
Petal.Width   -2.8104603   2.83918785

Proportion of trace :
   LD1     LD2
0.9912  0.0088
```

Informasi yang ada dalam output di atas terdiri dari:

- `Prior probabilities of groups` yaitu proporsi (persentase) pengamatan yang berada pada setiap kelompok (group). Dalam hal ini setiap kelompok terdiri dari 50 pengamatan, sehingga peluang sampel terambil untuk setiap kelas adalah 50/150 = 0,3333.
- `Group Means` adalah rata-rata variabel di setiap kelompok (group). Kelompok sentosa rata-rata Sepal.Length= 5,006; Sepal.Width= 3,428; Petal.Length = 1,462 dan rata - rata Petal.Width = 0.246 demikian untuk kelompok yang lainnya.
- `Coefficients of linear discriminants` adalah fungsi diskriminan linear yang terbentuk. Perhatikan dari output di atas, fungsi diskriminan yang terbentuk adalah :

LD1 = 0,8294 Sepal.Lenght + 1,5345 Sepal.Widht – 2,2012 Petal.Length – 2,8105 Petal.Length

LD2 = 0,0241 Sepal.Lenght + 2,1645 Sepal.Widht – 0,9329 Petal.Length + 2,8392 Petal.Length

Keterangan :

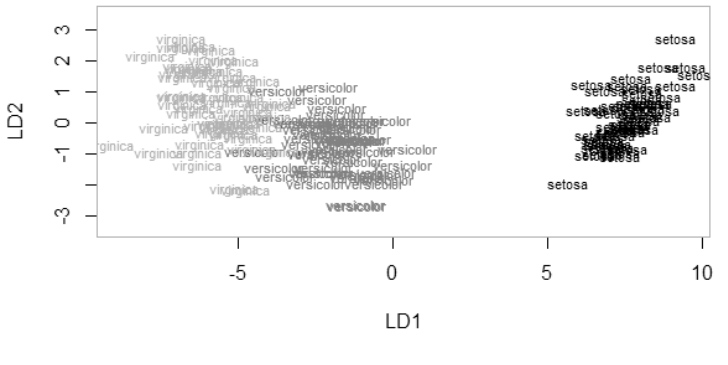
LD1= *Linear Discriminant* 1

LD2=*Linear Discriminan*t 2

- `Proportion of trace` adalah proporsi variansi yang dijelaskan oleh fungsi diskriminan. Dalam hal ini proporsi variansi yang dijelaskan oleh LD1 sebesar 99,12% dan LD2 sebesar 0,88%.

Apabila digambarkan kedua fungsi diskriman akan nampak seperti gambar di bawah ini

```
> plot(D, col="blue" col=as.integer(iris$Species)
```



Terlihat dari gambar di atas, sebaran data membentuk tiga kelompok yang berbeda. Spesies sentosa berada dalam satu kelompok (group) pada posisi di sebelah kanan (warna hitam), spesies versicolor berada dalam kelompok (group) di tengah (warna merah) sedangkan kelompok spesies virginica berada disebelah kiri (warna hijau). Ada sebagian pengamatan kelompok virginica dan versicolor yang beririsan tetapi hanya sebagian kecil.

Langkah selanjutnya adalah melakukan uji signifikansi terhadap fungsi diskriminan yaitu dengan menggunakan uji Wilks. Caranya seperti berikut ini.

```
> DataIris=as.matrix(iris[,-5])
> DataManova=manova(DataIris~iris$Species)
> DataManova
Call :
  manova(DataIris ~ iris$Species)

Terms :
                 iris$Species   Residuals
Sepal.Length          63.2121     38.9562
Sepal.Width           11.3449     16.9620
Petal.Length         437.1028     27.2226
Petal.Width           80.4133      6.1566
Deg. of Freedom             2         147

Residual standard errors: 0.5147894 0.3396877 0.4303345 0.20465
Estimated effects may be unbalanced

> UjiWilks=summary(DataManova, test="Wilks")
> UjiWilks
              Df    Wilks approx F num Df  den Df Pr(>F)
iris$Species   2 0.023439   199.15      8     288   < 2.2e-16 ***
Residuals    147
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Berdasarkan hasil uji Wilks diperoleh nilai p_value sebenar 2,2e-16 ≈ 0,000 menunjukkan kalo H_0 ditolak. Hal ini berarti variabel-variabel pada ketiga kelompok (group) tersebut berbeda sehingga model diskriminan yang dibentuk tidak bisa ditolak.

Langkah berikutnya dalam analisis diskriminan adalah menguji ketepatan fungsi diskriminan dalam memprediksi pengelompokan (mengklasifikasikan). Hal ini bisa dilakukan melalui cara di bawah ini.

```
> prediksi=predict(D)
> prediksi
$class
  [1] setosa     setosa    setosa    setosa    setosa    setosa    setosa
  [8] setosa     setosa    setosa    setosa    setosa    setosa    setosa
 [15] setosa     setosa    setosa    setosa    setosa    setosa    setosa
…
[134] versicolor virginica virginica virginica virginica virginica virginica
[141] virginica  virginica virginica virginica virginica virginica virginica
[148] virginica  virginica virginica
Levels: setosa versicolor virginica
```

Output di atas menjelaskan fungsi diskriman memprediksikan setiap pengamatan ke dalam setiap kelompok. Pengamatan 1 masuk kedalam kelompok sentosa, pengamatan ke-2 masuk kelompok sentosa, sampai pengamatan ke-148 masuk kelompok virginica.

Selanjutnya dibuatkan klasifikasi ketepatan pengelompokan dengan data asal, perintahnya seperti di bawah ini.

```
> Y=prediksi$class
> X=iris$Species
> table(Y,X)
                        X
Y             setosa  versicolor  virginica
  setosa          50           0          0
  versicolor       0          48          1
  virginica        0           2         49
```

Hasil ketepatan prediksi dapat dilihat dari jumlah pengamatan yang berada di diagonal matriks. Jumlah yang tepat sebanyak 147 pengamatan dan yang tidak tepat sebanyak 3 pengamatan.

Berdasarkan plot sebaran data, ada sebagaian pengamatan kelompok versicolor dan virginica yang beririsan, ada kemungkinan misklasifikasi. Pada kasus ini persentase ketepatan klasifikasi sebesar 98 sudah sangat baik.

8.4 Latihan

1. Suatu penelitian ingin membuat klasifikasi dua jenis tanaman yaitu tanaman pertama A dan tanaman kedua B, berdasarkan dua variabel bebas yaitu lebar daun dan bentuk daun. Sampel acak diambil dari kedua tanaman tersebut dimana $n_1 = n_1 = 3$. Hasil pengamatan kedua tanaman tersebut disajikan pada matriks di bawah ini.

$$X_1 = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix} dan\, X_2 = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}$$

   Asumsikan kedua matriks kovariansi sama. Berdasarkan data di atas,

   a. Hitunglah vektor rata-rata dan matriks kovariansi untuk data di atas!
   b. Estimasi fungsi diskriminan!
   c. Apabila suatu pengamatan $\boldsymbol{x}_0 = (3\ \ 6)'$, tentukan termasuk jenis tanaman mana? Asumsikan bahwa peluang prior dan biaya/ risiko kedua kelompok tersebut sama.

2. Pada contoh di atas, misalkan *sample training* diambil sebanyak 12 pada masing- masing jenis tanaman. Hasilnya disajikan pada tabel di bawah ini.

		Prediksi Keanggotaan		Total
		π_1 (A)	π_2 (B)	
Keanggotaan Aktual	π_1 (A)	10		
	π_2 (B)	…	9	12

   a. Lengkapilah tabel di atas!
   b. Hitunglah nilai APER! Apakah fungsi diskriminan sudah valid?

3. Andaikan suatu sampel acak diambil dari dua populasi yaitu dan . Asumsikan kedua populasi mempunyai matriks kovariansi yang sama. Vektor rata-rata adalah :

$$\bar{X}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} dan\, \bar{X}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Matriks kovariansi gabungan adalah :

$$S_{gab} = \begin{bmatrix} 7{,}3 & -1{,}1 \\ -1{,}1 & 4{,}8 \end{bmatrix}$$

Berdasarkan data di atas :

a. Uji perbedaan vektor rata-rata dengan menggunakan uji T^2 Hotellings, gunakan $\alpha = 5\%$!
b. Bentuklah fungsi diskriminan!
c. Tentukan keanggotaan $\boldsymbol{x}_0$ = (3 1)' apakah masuk pada kelompok/populasi pertama atau kedua!

4. Tabel di bawah ini merupakan hasil pengamatan dari 20 konsumen beserta Pendapatan, Pengeluaran dan Konsumsi. Konsumen dikelompokan menjadi kelompok a (Royal) dan kelompok b (Tidak Royal).

No	Kelompok	Pendapatan	Pengeluaran	Konsumsi
1	a	191	131	53
2	a	185	134	50
3	a	200	137	52
4	a	173	127	50
5	a	171	128	49
6	a	160	118	47
7	a	188	134	54
8	a	186	129	51
9	a	174	131	52
10	a	163	115	47
11	b	186	107	49
12	b	211	122	49
13	b	201	144	47
14	b	242	131	54
15	b	184	108	43
16	b	211	118	51
17	b	217	122	49

18	b	223	127	51
19	b	208	125	50
20	b	199	124	46

Berdasarkan data di atas, lakukan analisis diskriminan yang membedakan kedua kelompok konsumen berdasarkan 3 variabel Pendapatan, Pengeluaran, dan Konsumsi!

BAB

IX

ANALISIS KLASTER (*CLUSTER ANALYSIS*)

9.1 Pendahuluan

Analisis klaster (*cluster analysis*) adalah salah satu teknik interdependensi dalam analisis multivariat karena dalam analisis ini tidak ada variabel bebas dan variabel terikatnya. Analisis klaster bertujuan untuk mengidentifikasi sekelompok objek yang mempunyai kemiripan karakteristik tertentu yang dapat dipisahkan dengan kelompok objek lainnya. Pada prinsipnya, objek yang berada dalam satu kelompok akan mempunyai karakteristik yang relatif sama (homogen) sedangkan objek antarkelompok memiliki karakteristik yang berbeda.

Objek dalam analisis klaster dapat berupa barang, jasa, hewan, manusia (responden, konsumen, atau yang lain). Misalnya suatu penelitian ingin mengelompokkan kota-kota berdasarkan tingkat polutannya. Variabel-variabel (karakteristik) adalah debu (TSP = *total suspended particulate*), Carbon Monoksida (CO), Amonia(NH_3), Nitrogen Dioksida (NO_2), dan Sulfur Dioksida (SO_2).

Contoh kasus berikutnya adalah mengelompokkan provinsi di Indonesia berdasarkan indikator kesejahteraan rakyat yang terdiri dari PDRB per kapita (X_1), kepadatan penduduk (X_2), penduduk miskin (X_3), jumlah angkatan kerja (X_4), pengeluaran riil perkapita (X_5), dan angka harapan hidup (X_6).

Pengelompokan objek ke dalam suatu klaster atau kelompok menggunakan suatu prosedur. Prosedur analisis klaster ini digunakan untuk mengidentifikasi suatu objek ke dalam kelompok tertentu yang karakteristiknya relatif sama menggunakan algoritma tertentu. Algoritma yang digunakan mengharuskan kita membuat spesifikasi jumlah kluster-kluster yang akan dibuat. Metode yang digunakan untuk membuat klasifikasi dapat dipilih satu dari dua metode, yaitu memperbarui kelompok-kelompok kluster secara iteraktif atau hanya melakukan klasifikasi.

Prosedur dalam analisis klaster, secara garis besarnya dapat dijelakan sebagai berikut.

1. Pengelompokan data pada analisis klaster dapat didasarkan pada kesamaan (*similarity*) atau jarak (*distance*) dari masing-masing objek;
2. Proses dimulai dengan pengambilan p pengukuran variabel dan n objek pengamatan (matriks n x p). Matriks tersebut

ditransformasikan ke dalam bentuk matriks similaritas berukuran *n* x *n* yang dihitung berdasarkan pasangan-pasangan objek *p* variabel;
3. Kemudian suatu algoritma pengelompokan dipilih;
4. Terdapat dua pendekatan dalam algoritma pengelompokan, yaitu pendekatan hierarkis (*hierarchical clustering*) dan pendekatan non-hierarkis (*nonhierarchical clustering*).

9.2 Konsep Jarak (*Distance*)

Analisis klaster adalah suatu alat untuk mengelompokkan sejumlah objek berdasarkan variabel yang secara relatif mempunyai kesamaan karakteristik diantara objek-objek tersebut, sehingga keragaman dalam suatu kelompok tersebut lebih kecil dibandingkan dengan keragaman antarkelompok.

Konsep dasar pengukuran analisis klaster adalah konsep pengukuran jarak (*distance*) dan kesamaan (*similarity*). *Distance* adalah ukuran tentang jarak pisah antar objek sedangkan *similarity* adalah ukuran kedekatan. Konsep ini penting karena pengelompokan pada analisis klaster didasarkan pada kedekatan. Pengukuran jarak (*distance type measure*) digunakan untuk data-data yang bersifat metrik, sedangkan pengukuran kesesuaian (*matching tipe measure*) digunakan untuk data-data yang bersifat kualitatif.

Jika nilai ukuran kemiripan atau jarak antardua buah objek semakin kecil maka antara dua objek tersebut cenderung berada dalam suatu kelompok. Sebaliknya, semakin besar nilai ukuran kemiripan atau jarak antardua buah objek artinya semakin besar pula perbedaan antara dua objek tersebut, sehingga makin cenderung untuk tidak menganggapnya ke dalam kelompok yang sama.

Terdapat beberapa cara dalam mengukur jarak, diantaranya adalah :

1. Jarak *Euclidean*, yaitu jarak berupa akar jumlah kuadrat perbedaan nilai untuk tiap variabel.

   Misalkan ada dua pengamatan A dan B, masing-masing terdiri dari *p* variabel yaitu $A = (x_1, x_1, ..., x_p)$ dan $B = (y_1, y_1, ..., y_p)$ maka jarak *euclidien* dihitung dengan rumus berikut ini.

$$d(A,B) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + \left(x_p - y_p\right)^2}$$

2. Jarak City Block atau Manhatan jumlah nilai perbedaan mutlak untuk tiap variabel. Rumusnya adalah :

$$d(A,B) = |x_1 - y_1| + |x_2 - y_2| + \cdots + \left|x_p - y_p\right|$$

Secara garis besarnya, prosedur analisis cluster dapat dijelaskan sebagai berikut.

1. Pengelompokan data pada analisis klaster dapat didasarkan pada kesamaan (*similarity*) atau jarak (*distance*) dari masing-masing objek.
2. Proses dimulai dengan pengambilan *p* pengukuran variabel dan *n* objek pengamatan (matriks *n* x *p*). Matriks tersebut ditransformasikan ke dalam bentuk matriks similaritas berukuran *n* x *n* yang dihitung berdasarkan pasangan-pasangan objek *p* variabel.
3. Kemudian suatu algoritma pengelompokan dipilih.
4. Terdapat dua pendekatan dalam algoritma pengelompokan, yaitu pendekatan hierarkis (*hierarchical clustering*) dan pendekatan non-hierarkis (*nonhierarchical clustering*).

Misalkan suatu pengamatan dengan dan yang dapat disajikan pada tabel di bawah ini :

Tabel 18. Data Pengamatan dalam Analisis Klaster

Objek	Variabel			
	1	2	...	*p*
1	x_{11}	x_{12}		x_{1p}
2	x_{11}	x_{12}		x_{2p}
⋮				
n	x_{n1}	x_{n2}		x_{np}

Berdasarkan matriks data di atas dibuat matriks jarak untuk setiap pasangan pengamatan yaitu :

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n1} & \cdots & d_{nn} \end{bmatrix}$$

Contoh

Misakan ada 5 objek yaitu A, B, C, D dan E dengan dua variabel yaitu X_1 dan X_2, data ditampillkan pada tabel di bawah ini.

Objek	Variabel X_1	Variabel X_2
A	1	1
B	4	1
C	1	2
D	3	4
E	5	4

Buatlah matriks jarak untuk kelima pengamatan tersebut dengan menggunakan konsep jarak Manhatan!

Jawab

Jarak antara objek dengan dirinya sendiri adalah 0, berikutnya perhitungannya.

- $d_{AA} = |1 - 1| + |1 - 1| = 0$
- $d_{BB} = |4 - 4| + |1 - 1| = 0$
- $d_{CC} = |1 - 1| + |2 - 2| = 0$
- $d_{DD} = |3 - 3| + |4 - 4| = 0$
- $d_{EE} = |5 - 5| + |4 - 4| = 0$

Jarak antara objek lainnya adalah

- d_{AB} = d_BA = $|1 - 4| + |1 - 1| = 3$
- d_{AC} = d_CA = $|1 - 1| + |1 - 2| = 1$
- d_{AD} = d_DA = $|1 - 3| + |1 - 4| = 5$
- d_{AE} = d_EA = $|1 - 5| + |1 - 4| = 7$
- d_{BC} = d_CB = $|4 - 1| + |1 - 2| = 4$
- d_{BD} = d_DB = $|4 - 3| + |4 - 1| = 4$
- d_{BE} = d_EB = $|4 - 5| + |1 - 4| = 4$

- $d_{CD} = d\_DC = |1 - 3| + |2 - 4| = 4$
- $d_{CE} = d\_EC = |1 - 5| + |2 - 4| = 6$
- $d_{ED} = d\_DE = |3 - 5| + |4 - 4| = 2$

Matriks jarak adalah :

$$\boldsymbol{D} = \begin{bmatrix} 0{,}0 & 3{,}0 & 1{,}0 & 5{,}0 & 7{,}0 \\ 3{,}0 & 0{,}0 & 4{,}0 & 4{,}0 & 4{,}0 \\ 1{,}0 & 4{,}0 & 0{,}0 & 4{,}0 & 6{,}0 \\ 5{,}0 & 4{,}0 & 4{,}0 & 0{,}0 & 2{,}0 \\ 7{,}0 & 4{,}0 & 6{,}0 & 2{,}0 & 0{,}0 \end{bmatrix}$$

Matrik di atas dapat dibuat dalam bentuk tabel berikut ini.

	A	B	C	D	E
A	0,0	3,0	1,0	5,0	7,0
B	3,0	0,0	4,0	4,0	4,0
C	1,0	4,0	0,0	4,0	6,0
D	5,0	4,0	4,0	0,0	2,0
E	7,0	4,0	6,0	2,0	0,0

Ada dua teknik yang digunakan dalam analisis klaster yaitu teknik hierarki dan non-hierarki. Pada metode hierarki, jumlah klaster belum atau tidak diketahui sebelumnya, sementara pada metode non-hierarki jumlah klaster ditetapkan terlebih dahulu, sebelum melakukan pengklasteran objek.

Ada dua teknik di dalam analisis klaster yaitu :

1. Teknik Hierarki

   Teknik hierarki adalah teknik pengelompokan (klaster) dengan cara membentuk konstruksi hierarki atau berdasarkan tingkatan tertentu seperti struktur pohon (struktur pertandingan). Dengan demikian proses pengelompokannya dilakukan secara bertingkat atau bertahap.

   Di dalam teknik hierarki ada dua pendekatan yang digunakan yaitu :

   a. Metode *Agglomerative*

      Metode ini dimulai dengan kenyataan bahwa setiap objek membentuk clusternya masing-masing. Kemudian dua objek dengan jarak terdekat bergabung. Selanjutnya objek ketiga akan

bergabung dengan cluster yang ada atau bersama objek yang lain dan membentuk klaster baru. Hal ini tetap memperhitungkan jarak kedekatan antarobjek. Proses akan berlanjut hingga akhirnya terbentuk satu klaster yang terdiri dari keseluruhan objek.

Di dalam metode *agglomerative* ada beberapa cara yaitu :

- single linkage (nearest neighbor methods)
- complete linkage (furthest neighbor methods)
- average linkage methods (between groups methods)
- within groups methods
- median methods
- centroid methods
- ward's error sum of squares methods

b. Metode Divisive

Metode divisive berlawanan dengan metode *agglomerative*. Metode ini pertama-tama dimulai dengan satu klaster besar yang mencakup semua observasi (objek). Selanjutnya objek yang mempunyai ketidakmiripan yang cukup besar akan dipisahkan sehingga membentuk cluster yang lebih kecil. Pemisahan ini dilanjutkan sehingga mencapai sejumlah klaster yang diinginkan. Salah satu teknik yang termasuk metode divisive yaitu *Splinter average distance methods.*

2. Teknik Non Hierarki

Berbeda dengan metode hierarkikal, prosedur nonhierarkikal dimulai dengan memilih sejumlah nilai klaster awal sesuai dengan jumlah yang diinginkan dan kemudian objek digabungkan ke dalam klaster-klaster tersebut.

a. Metode *sequential threshold procedure*

Metode ini melakukan pengelompokan dengan terlebih dahulu memilih satu objek dasar yang akan dijadikan nilai awal klaster, kemudian semua objek yang ada di dalam jarak terdekat dengan cluster ini akan bergabung lalu dipilih klaster kedua dan semua objek yang mempunyai kemiripan dimasukkan dalam klaster ini. Demikian seterusnya hingga terbentuk beberapa klaster

dengan keseluruhan objek di dalamnya.

b. *Parallel threshold procedure*

Secara prinsip sama dengan prosedur *sequential threshold*, hanya saja dilakukan pemilihan terhadap beberapa objek awal klaster sekaligus dan kemudian melakukan penggabungan objek ke dalamnya secara bersamaan.

c. *Optimizing*

Merupakan pengembangan dari kedua metode di atas dengan melakukan optimasi pada penempatan objek yang ditukar untuk klaster lainnya dengan pertimbangan kriteria optimasi.

Prosedur analisis klaster K-means digunakan untuk mengelompokkan sejumlah kasus besar yang lebih dari 200 dengan lebih efisien. Metode ini berdasarkan *nearest centroid sorting*, yaitu pengelompokan berdasarkan jarak terkecil antara kasus dengan pusat dari klaster. Teknik ini membutuhkan jumlah klaster yang ditentukan terlebih dahulu oleh pemakai. Untuk tujuan tersebut dapat menggunakan analisis hierarkikal dalam menentukan jumlah klaster. Teknik ini juga dapat digunakan untuk menempatkan data baru untuk dikelompokkan ke dalam klaster terdekat.

Namun dalam perkembangan saat ini, metode analisisi klaster semakin berkembang diantaranya *Fuzzy Clustering, Density Based Clustering* dan lain-lain.

9.3 Metode Hierarki

Di dalam analisis klaster ada beberapa teknik yang termasuk metode hierarki. Pada buku ini, akan dibahas tiga teknik yang popular dan banyak digunakan yaitu metode *Single linkage*, *Complete linkage* dan *Average Linkage*.

1. Metode Pautan Tunggal (*Single Lingkage*)

Pada metode *single linkage* jarak antar *cluster* dapat dilakukan dengan melihat kesamaan jarak antar dua pasang objek yang ada, dan memilih jarak jarak paling dekat atau aturan tetangga dekat (*nearest neighbor rule*).

Berikut adalah gambaran secara visual dari metode *single linkage*, pada gambar di bawah ini.

Gambar 6. Pengelompokan Objek dengan Metode Single Linkage

Adapun langkah-langkah dalam metode *single linkage* sebagai berikut (Johnson & Wichern, 1992) :

a. Mencari atau menentukan jarak terkecil di dalam D = $\{d_{ik}\}$
b. Menggabungkan *cluster U* dan *V* untuk menjadikan *cluster* baru (*UV*); dimana

$$d_{(UV)} = \min \{d_{UV}\} ; d_{UV} \in D$$

c. Setelah menerapkan algoritma di atas, maka jarak antara dan stiap *cluster* lainnya dihitung dengan cara :

$$d_{(UV)W} = \min \{d_{UW}, d_{VW}\} ;$$

Di sini d_{UW} dan d_{VW} masing-masing merupakan jarak terpendek (*nearest neighbours*) antara *cluster-cluster U* dan *W* dan juga *cluster-cluster V* dan *W*.

Proses di atas akan berhenti jika semua objek bergabung kedalam suatu kelompok.

Contoh

Lakukan analisis klaster dengan metode *Single Lingkage* untuk matriks jarak di bawah ini.

	A	B	C	D	E
A	0,0	3,0	1,0	5,0	7,0
B	3,0	0,0	4,0	4,0	4,0
C	1,0	4,0	0,0	4,0	6,0
D	5,0	4,0	4,0	0,0	2,0

E	7,0	4,0	6,0	2,0	0,0

Jawab

1. Mencari objek dengan jarak minimum
   A dan C mempunyai jarak terdekat, yaitu 1,0 maka objek A dan C bergabung menjadi satu klaster

2. Menghitung jarak antara klaster AC dengan objek lainnya

   $d_{(AC)B} = \min \{d_{AB}, d_{CB}\} = \min \{3, 4\} = 3,0$
   $d_{(AC)D} = \min \{d_{AD}, d_{CD}\} = \min \{5, 4\} = 4,0$
   $d_{(AC)E} = \min \{d_{AE}, d_{CE}\} = \min \{7, 6\} = 6,0$

   Dengan demikian terbentuk matriks jarak yang baru

	AC	B	D	E
AC	0,0	3,0	4,0	6,0
C	3,0	0,0	4,0	6,0
D	4,0	4,0	0,0	2,0
E	6,0	6,0	2,0	0,0

3. Selanjutnya, dari matriks di atas mencari objek dengan jarak terdekat D dan E mempunyai jarak terdekat, yaitu 2,0 maka objek D dan E bergabung menjadi satu klaster
4. Kemudian, menghitung jarak antara klaster dengan objek lainnya yaitu :

   $d_{(AC)(DE)} = \min \{d_{AD}, d_{AE}, d_{CD}, d_{CE}\} = \min \{5, 7, 4, 6\} = 4,0$
   $d_{B,(DE)} = \min \{d_{BD}, d_{BE}\} = \min \{4, 6\} = 4,0$

   maka terbentuklah matriks jarak yang baru, yaitu:

	AC	B	DE
AC	0,0	3,0	4,0
B	3,0	0,0	4,0
DE	4,0	4,0	0,0

5. Dari matriks di atas mencari objek dengan jarak terdekat klaster AC dan B mempunyai jarak terdekat, yaitu 3,0 maka klaster AC dan B bergabung menjadi klaster baru yaitu ABC

6. Kemudian, menghitung jarak antara klaster dengan objek lainnya yaitu :

$d_{(ABC)(DE)}$ = min {d_{AD}, d_{AE}, $d_{BD,}$ $d_{BE,}$ d_{CD}, d_{CE} } = min {5,7,4,6,4,4 } = 4,0

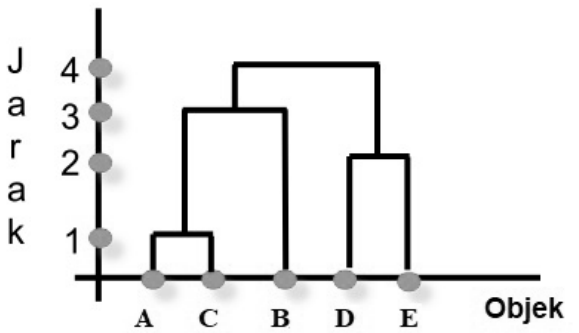
maka terbentuk matriks jarak yang baru, yaitu :

	ABC	DE
ABC	0,0	4,0
DE	4,0	0,0

Pada langkah yang terakhir, pada jarak sebesar 4, klaster ABC bergabung dengan DE sehingga terbentuk klaster tunggal yaitu ABCDE.

Pada umumnya, hasil dari pengelompokan/pengklasteran dapat ditampilkan dalam bentuk *dendogram*, atau diagram pohon. Cabang-cabang pohon tersebut merupakan hasil dari pengelompokan. Cabang-cabang yang bergabung pada diagram pohon menunjukkan banyaknya klaster yang terjadi.

Pada kasus di atas, bentuk dendogrammnya adalah :



Gambar 7. Bentuk Dendogram dengan Metode Single Lingkage

2. Metode Pautan Lengkap (*Complete Lingkage*)

Pada metode *complete linkage* jarak antar-*cluster* dapat dilakukan dengan melihat kesamaan jarak antardua pasang objek yang ada, dan memilih jarak yang terjauh (jarak yang terbesar). Di bawah ini adalah gambaran secara visual dari metode *complete linkage*

Gambar 8. Pengelompokan Objek dengan Metode Complete Lingkage

Adapun langkah-langkah dalam metode *complete linkage* sebagai berikut (Johnson & Wichern, 1992) :

a. Mencari atau menentukan jarak terkecil di dalam $D = \{d_{ik}\}$
b. Menggabungkan *cluster U* dan *V* untuk menjadikan *cluster* baru (*UV*); dimana

$$d_{(UV)} = \min \{d_{UV}\} ; d_{UV} \in D$$

c. Setelah menerapkan algoritma di atas, maka jarak antara dan setiap *cluster* lainnya dihitung dengan cara :

$$d_{(UV)W} = \min \{d_{UW} , d_{VW}\} ;$$

di sini dan masing-masing merupakan jarak terpendek (*nearest neighbours*) antara *cluster-cluster U* dan *W* dan juga *cluster-cluster V* dan *W*. Proses di atas akan berhenti jika semua objek bergabung ke dalam suatu kelompok.

Contoh

Lakukan analisis klaster dengan metode *Complete Lingkage* untuk matriks jarak di bawah ini.

	A	B	C	D	E
A	0,0	3,0	1,0	5,0	7,0
B	3,0	0,0	4,0	4,0	4,0
C	1,0	4,0	0,0	4,0	6,0
D	5,0	4,0	4,0	0,0	2,0
E	7,0	4,0	6,0	2,0	0,0

Jawab

1. Mencari objek dengan jarak minimum

   A dan C mempunyai jarak terdekat, yaitu 1,0 maka objek A dan C bergabung menjadi satu klaster

2. Menghitung jarak antara klaster AC dengan objek lainnya

   $d_{(AC)B} = \max \{d_{AB}, d_{CB}\} = \max \{3, 4\} = 4,0$

   $d_{(AC)D} = \max \{d_{AD}, d_{CD}\} = \max \{5, 4\} = 5,0$

   $d_{(AC)E} = \max \{d_{AE}, d_{CE}\} = \max \{7, 6\} = 7,0$

   Dengan demikian terbentuk matriks jarak yang baru.

	AC	B	D	E
AC	0,0	4,0	5,0	7,0
C	4,0	0,0	4,0	6,0
D	5,0	4,0	0,0	2,0
E	7,0	6,0	2,0	0,0

3. Selanjutnya, dari matriks di atas mencari objek dengan jarak terdekat D dan E mempunyai jarak terdekat, yaitu 2,0 maka objek D dan E bergabung menjadi satu klaster.

4. Kemudian, menghitung jarak antara klaster dengan objek lainnya yaitu :

   $d_{(AC)(DE)} = \max \{d_{AD}, d_{AE}, d_{CD}, d_{CE}\} = \max \{5, 7, 4, 6\} = 7,0$

   $d_{B,(DE)} = \max \{d_{BD}, d_{BE}\} = \min \{4, 4\} = 4,0$

   maka terbentuklah matriks jarak yang baru, yaitu:

	AC	B	DE
AC	0,0	4,0	7,0

B	4,0	0,0	4,0
DE	7,0	4,0	0,0

5. Dari matriks di atas objek B memiliki jarak sebesar 4 baik ke klaster AC maupun ke klaster DE. Misalkan pada kasus ini, dipilih klaster AC bergabung dengan B sehingga terbentuk klaster baru yaitu ABC. Kemudian, menghitung jarak antara klaster ABC dengan objek lainnya yaitu :

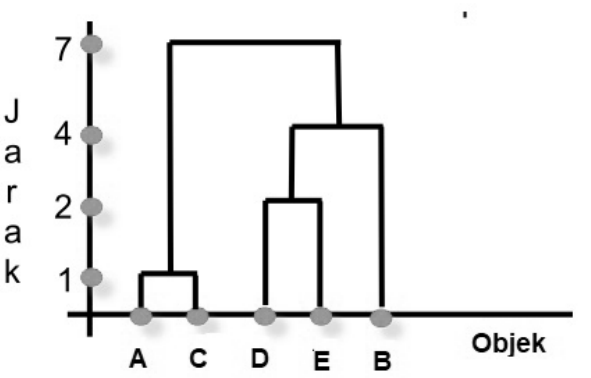
$d_{(ABC)(DE)}$ $= \max \{d_{AD}, d_{AE}, d_{BD,} d_{BE,} d_{CD}, d_{CE}\} = \max \{5,7,4,6,4,4\} = 7,0$

maka terbentuk matriks jarak yang baru, yaitu :

	ABC	DE
ABC	0,0	7,0
DE	7,0	0,0

Pada langkah yang terakhir, pada jarak sebesar 7, klaster ABC bergabung dengan DE sehingga terbentuk klaster tunggal yaitu ABCDE.

Bentuk dendogram dengan metode *Complete Linkage* adalah :

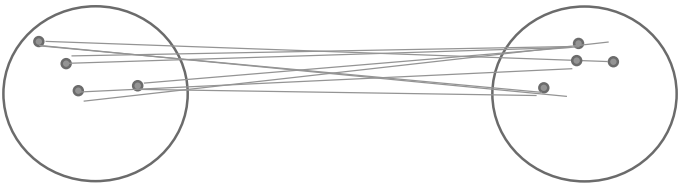


Gambar 9. Bentuk Dendogram dengan Metode Complete Lingkage

3. Metode Pautan Rata-rata (*Average Lingkage*)

Pada metode ini, prosedurnya sama dengan metode *single linkage* dan

complete linkage tetapi ukuran jarak yang digunakan adalah jarak rata-rata (*average*) antar tiap pasangan objek. Di bawah ini adalah gambaran secara visual dari metode *average linkage*



Gambar 10. Pengelompokan Objek dengan Metode Average Lingkage

Adapun langkah-langkah dalam metode *single linkage* sebagai berikut (Johnson & Wichern, 1992) :

a. Mencari atau menentukan jarak terkecil di dalam $D = \{d_{ik}\}$
b. Menggabungkan *cluster* U dan V untuk menjadikan *cluster* baru ; dimana $d_{(UV)}$ = average $\{d_{(UV)}\}$; $d_{UV} \in D$
c. Setelah menerapkan algoritma di atas maka jarak antara (UV) dan stiap *cluster* lainnya dihitung dengan cara :

$$d_{(UV)W} = \text{average } \{d_{UV}, d_{VW}\} ; d_{UV; VW} \in D$$

Proses di atas akan berhenti jika semua objek bergabung kedalam suatu kelompok.

Contoh

Lakukan analisis klaster dengan metode *Complete Lingkage* untuk matriks jarak di bawah ini.

	A	B	C	D	E
A	0,0	3,0	1,0	5,0	7,0
B	3,0	0,0	4,0	4,0	4,0
C	1,0	4,0	0,0	4,0	6,0
D	5,0	4,0	4,0	0,0	2,0
E	7,0	4,0	6,0	2,0	0,0

Jawab

1. Mencari objek dengan jarak minimum

   A dan C mempunyai jarak terdekat, yaitu 1,0 maka objek A dan C bergabung menjadi satu klaster

2. Menghitung jarak antara klaster AC dengan objek lainnya

   $d_{(AC)B}$ = average $\{d_{AB}, d_{CB}\}$ = average {3, 4} = 3,5

   $d_{(AC)D}$ = average $\{d_{AD}, d_{CD}\}$ = average {5, 4} = 4,5

   $d_{(AC)E}$ = average $\{d_{AE}, d_{CE}\}$ = average {7, 6} = 6,5

   Dengan demikian terbentuk matriks jarak yang baru

	AC	B	D	E
AC	0.0	3,5	4,5	6,5
B	3,5	0,0	4,0	4,0
D	4,5	4,0	0,0	2,0
E	6,5	4,0	2,0	0,0

3. Selanjutnya, dari matriks di atas mencari objek dengan jarak terdekat D dan E mempunyai jarak terdekat, yaitu 2,0 maka objek D dan E bergabung menjadi satu klaster.

4. Kemudian, menghitung jarak antara klaster dengan objek lainnya yaitu :

   $d_{(AC)(DE)}$ = average $\{d_{AD}, d_{AE}, d_{CD}, d_{CE}\}$ = average {5, 7, 4, 6 }= 5,25

   $d_{B,(DE)}$ = average $\{d_{BD}, d_{BE}\}$ = average {4, 4 } = 4,0

   maka terbentuklah matriks jarak yang baru, yaitu:

	AC	B	DE
AC	0,00	4,0	5,25
B	4,00	0,0	4,00
DE	5,25	4,0	0,00

5. Dari matriks di atas objek B memiliki jarak sebesar 4 baik ke klaster AC maupun ke klaster DE. Misalkan pada kasus ini, dipilih klaster B bergabung dengan DE sehingga terbentuk klaster baru yaitu BDE.
6. Kemudian, menghitung jarak antara klaster BDE dengan objek lainnya yaitu :

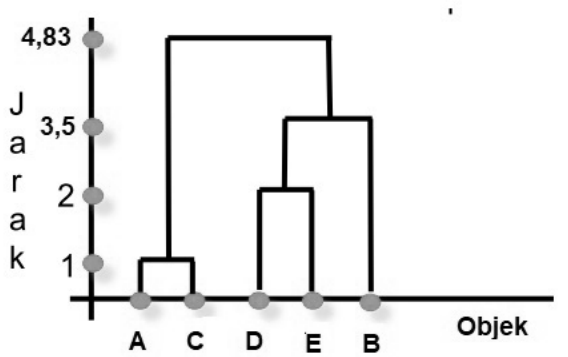
$d_{(AC)(BDE)}$=average {$d_{AB,}$ $d_{AD,}$ d_{AE}, $d_{CB,}$ $d_{CD,}$ d_{CE} } = average {3,5,7,4,4,6 } = 4,83

maka terbentuk matriks jarak yang baru, yaitu :

	ABC	DE
ABC	0,00	4,83
DE	4,83	0,00

Pada langkah yang terakhir, pada jarak sebesar 4,83, klaster AC bergabung dengan BDE sehingga terbentuk klaster tunggal yaitu ABCDE.

Bentuk dendogram dengan metode *Complete Linkage* adalah :



Gambar 11. Pengelompokan Objek dengan Metode Average Lingkage

9.4 Metode Non-Hierarki

Prosedur analisis klaster dengan menggunakan teknik non-hierarki berbeda dengan teknik hierarki, pada metode non-hierarki banyaknya klaster ditentukan terlebih dahulu. Informasi banyaknya klaster yang terbentuk diperoleh dari penelitian sebelumnya atau dari teori yang mendasarinya. Misalnya suatu penelitian untuk mengelompokkan kabupaten berdasarkan indikator Indeks Pembangunan Manusia (IPM). Berdasarkan penelitian sebelumnya dan teori bahwa kelompok

kabupaten berdasarkan IPM dapat dibagi menjadi tiga kelompok yaitu kelompok IPM kecil, sedang, dan tinggi. Prosedur dengan metode non-hierarki terdiri dari *Sequential threshold procedure, Parallel threshold procedure* dan *Optimizing*. Salah satu teknik yang popular digunakan adalah prosedur K-Means.

Ada beberapa kelebihan metode K-Means, yaitu:

1. Metode ini mudah untuk diimplementasikan dan dijalankan;
2. Waktu yang dibutuhkan untuk menjalankan prosedur K-Means relatif cepat;
3. Metode K-Means mudah untuk diadaptasi dan umum digunakan.

Prosedur pengelompokan dengan metode K-Means adalah :

1. Menentukan banyaknya klaster, misalkan ada kelompok/klaster;
2. Pilih setiap objek ke dalam setiap klaster secara acak;
3. Hitung titik pusat (*centroid*) dari setiap kelompok;
4. Hitung jarak setiap objek terhadap titik pusat (*centroid*) setiap kelompok (pada umunya menggunakan jarak *Euclidian*);
5. Tempatkan kembali setiap objek pada klaster dengan jarak yang minimum;
6. Lakukan langkah 3-5 sampai semua objek berada pada kelompok dengan jarak ke *centroid* paling kecil.

Contoh

Berikut ini akan diberikan ilustrasi metode K-Means. Misalkan pengelompokan objek A, B, C, dan D dengan menggunakan dua variabel yaitu (data dari Johnson dan Winchern, 2007), seperti tersaji di bawah ini.

Objek	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	3	-2

Berdasarkan data di atas, lakukan pengelompokan dengan metode K-Means!

Jawab

Langkah-langkahnya :

1. Tentukan banyaknya kelompok, misalkan pada kasus ini
2. Pilih secara acak objek yang ditempatkan di setiap kelompok, misalkan anggota kelompok pertama terdiri dari objek A dan B sedangkan kelompok kedua terdiri dari objek C dan D.
3. Menghitung *centroid* setiap kelompok yaitu :

Klaster	Koordinat *centroid*	
	$\bar{X}_1$	$\bar{X}_2$
(AB)	$\frac{5+(-1)}{2} = 2$	$\frac{3+1}{2} = 2$
(CD)	$\frac{1+(-3)}{2} = -1$	$\frac{-2+(-2)}{2} = -2$

4. Menghitung jarak setiap objek terhadap setiap *centroid* dengan menggunakan rumus jarak *Euclidian* yaitu :

$$d_{A(AB)} = (5-2)^2 + (3-2)^2 = 10$$
$$d_{A(CD)} = (5-(-1))^2 + (3-(-2))^2 = 61$$
$$d_{B(AB)} = (-1-2)^2 + (1-2)^2 = 10$$
$$d_{B(CD)} = (-1-(-1))^2 + (1-(-2))^2 = 9$$
$$d_{C(AB)} = (1-2)^2 + (-2-2)^2 = 17$$
$$d_{C(CD)} = (1-(-1))^2 + (-2-(-2))^2 = 4$$
$$d_{D(AB)} = (-3-2)^2 + (-2-2)^2 = 41$$
$$d_{D(CD)} = (-3-(-1))^2 + (-2-(-2))^2 = 4$$

5. Menempatkan setiap objek ke kelompok yang jarak ke *centroid* paling kecil yaitu :

Objek	Jarak ke klaster pertama	Jarak ke klaster kedua	Keputusan
A	10	61	Tetap di klaster pertama
B	10	9	Pindah ke klaster kedua
C	17	4	Tetap di klaster kedua
D	41	4	Tetap di klaster kedua

6. Hitung kembali centroid akibat perpindahan objek B ke klaster kedua yaitu :

Klaster	Koordinat *centroid*	
	$\bar{X}_1$	$\bar{X}_2$
(AB)	5	3
(CD)	$\frac{-1+1+(-3)}{3} = -1$	$\frac{1+-2+(-2)}{3} = -1$

7. Menghitung jarak setiap objek terhadap setiap *centroid* dengan menggunakan rumus jarak *Euclidian* yaitu :

$$d_{A(A)} = (5-5)^2 + (3-3)^2 = 0$$
$$d_{A(BCD)} = (5-(-1))^2 + (3-(-1))^2 = 52$$
$$d_{B(A)} = (-1-5)^2 + (1-3)^2 = 40$$
$$d_{B(BCD)} = (-1-(-1))^2 + (1-(-1))^2 = 4$$
$$d_{C(A)} = (1-5)^2 + (-2-3)^2 = 41$$
$$d_{C(BCD)} = (1-(-1))^2 + (-2-(-1))^2 = 5$$
$$d_{D(A)} = (-3-5)^2 + (-2-3)^2 = 89$$
$$d_{D(BCD)} = (-3-(-1))^2 + (-2-(-1))^2 = 5$$

8. Menempatkan setiap objek ke kelompok yang jarak ke *centroid* paling kecil yaitu :

Objek	Jarak ke klaster pertama	Jarak ke klaster kedua	Keputusan
A	0	52	Tetap di klaster pertama
B	40	4	Tetap di klaster kedua
C	41	5	Tetap di klaster kedua
D	89	5	Tetap di klaster kedua

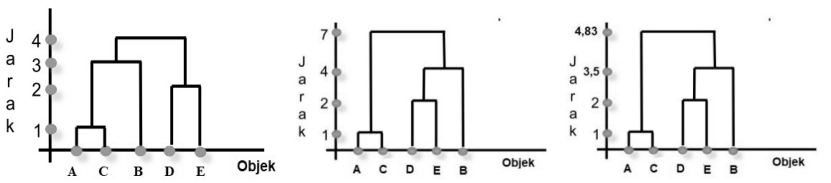
9. Semua objek sudah menempati klaster dengan jarak ke *centroid* terkecil, artinya tidak ada yang pindah klaster. Oleh karena itu prosedur K-Means berhenti dengan menghasilkan dua kelompok yaitu kelompok pertama beranggotakan objek A dan kelompok kedua terdiri dari objek C, D, dan E.

9.5 Cara Menentukan Banyaknya Klaster

Isu utama di dalam analisis klaster adalah menentukan berapa banyak klaster yang terbentuk. Tidak ada aturan baku untuk menentukan berapa banyaknya klaster tetapi ada beberapa kriteria yang bisa dipergunakan, yaitu (Supranto, 2004) :

1. Pertimbangan *teoritis*, *koseptual*, *praktis*, mungkin bisa diusulkan/ disarankan untuk menentukan berapa banyaknya *cluster* yang sebenarnya.
2. Di dalam pengelompokan hierarki, jarak dimana *cluster* digabung bisa dipergunakan sebagai kriteria. Informasi ini bisa diperoleh dari skedul *aglomerasi*.
3. Di dalam pengelompokan non-hierarki, rasio jumlah varian dalam *cluster* dengan jumlah varian antar-*cluster* dapat diplotkan melawan banyaknya *cluster*. Titik dimana suatu siku (*an elbow*) atau lekukan tajam (*a sharp bend*) terjadi, menunjukkan banyaknya *cluster*, di luar titik ini biasanya tidak berguna atau tidak perlu.
4. Besarnya relatif *cluster* seharusnya berguna/bermanfaat.

Perhatikan data pada contoh sebelumnya mengenai pengelompokan dengan menggunakan ketiga metode yaitu *Single Linkage*, *Complete Linkage* dan *Average Linkage*.



a. Single Lingkage **b. Complete Lingkage** **c. Average Lingkage**

Gambar 12. Dendogram untuk Metode Single, Complete dan Average Linkage

Pada gambar dendogram di atas, nampak bahwa dua kemungkinan pengelompokan yaitu :

Opsi pertama terbentuk dua kelompok:

1. Kelompok pertama terdiri dari objek A dan C,

2. Kelompok kedua terdiri dari objek B, D dan E.

Opsi kedua terbentuk tiga kelompok:

1. Kelompok pertama terdiri dari objek A dan C,
2. Kelompok kedua terdiri dari objek B,
3. Kelompok ketiga terdiri dari objek D dan E.

Langkah pengelompokan dalam analisis klaster mencakup 3 hal berikut :

1. Mengukur kesamaan jarak,
2. Membentuk klaster secara hirarki /non-hierarki,
3. Menentukan jumlah klaster.

9.6 Prosedure Analisis Klaster Menggunakan Software R

Ada beberapa teknik hierarki yang tersedia dalam *software* yaitu teknik *single linkage, complete lingkage, average lingkage* dan *ward.* R menyediakan fasilitas untuk melakukan analisis klaster dengan teknik hierarki yaitu fungsi hclust() dengan sintaknya seperti di bawah ini.

```
hclust(d, method = "complete", members = NULL)

Keterangan :

d       = matriks jarak (distance)
method  = metode hirarki yang digunakan, pilihannya
:"single", "average", "complete", "ward.D2",
"ward.D",…
members = anggota klaster
```

Dalam menjalankan fungsi `hclust()` harus membuat matriks jarak sebagai inputnya. Perintah untuk membuat matriks jarak tersedia dengan menjalankan fungsi `dist()`, sintaknya seperti di bawah ini.

```
dist(x, method = "euclidean",…)

Keterangan:
X = data
method = pilihan metode yang tersedia: "Euclidian",
"maximum", "manhattan", "Canberra", "binary" atau
"minkowski".
```

Contoh

Seorang peneliti ingin membuat klaster kabupaten berdasarkan indikator pertumbuhan ekonomi. Ada enam variabel yang dipakai yaitu: X_1 = *Produk Domestic Bruto* (PDB) (Milyar); X_2 = Tingkat Pengangguran (%); X_3 = Inflasi (%); X_4 = Konsumsi (Milyar); X_5 = Investasi (Milyar) dan X_6 = Pendapatan Per Kapita (juta).

Tabel 19. Data Kabupaten berdasarkan Indikator Ekonomi

Kabupaten	X_1	X_2	X_3	X_4	X_5	X_6
A	6	5	6	3	3	4
B	2	3	2	4	5	4
C	6	3	6	4	2	3
D	3	7	4	5	2	7
E	2	4	2	2	7	4
F	6	4	6	3	3	4
G	5	3	6	3	3	4
H	7	2	7	4	2	4
I	2	7	2	3	7	3
J	3	5	3	6	4	6
K	2	5	2	3	5	3
L	5	4	5	4	2	4
M	2	3	2	5	4	4
N	4	6	4	6	3	6
O	6	5	4	2	1	4
P	3	5	4	6	5	7
Q	4	4	7	2	2	5
R	3	7	2	6	4	3

S	4	6	3	6	3	6
T	3	4	3	4	7	3

Misalkan data di atas disimpan dalam folder "`C :/R_Kerja/Multivariat`" dengan nama file "`DataEkonomi.csv`".

Langkah pertama, inputkan data tersebut ke dalam R dengan seperti di bawah ini.

```
> X = read.csv2("DataEkonomi.csv")
> D=dist(X[,2 :7],method="euclidean")
> D
          1        2        3        4        5        6        7         8
1  0.000000
2  6.403124 0.000000
3  2.645751 6.480741 0.000000
4  5.567764 6.324555 6.782330 0.000000
5  7.071068 3.000000 7.937254 7.549834 0.000000
6  1.000000 6.164414 2.000000 6.000000 7.000000 0.000000
7  2.236068 5.477226 2.000000 6.164414 6.557439 1.414214 0.000000
8  3.605551 7.745967 2.000000 7.745967 9.110434 2.828427 2.828427  0.000000
9  7.280110 4.690416 8.602325 7.071068 3.316625 7.615773 7.615773 10.099505
10 5.656854 3.872983 6.244998 3.316625 5.656854 5.744563 5.567764  7.280110
11 6.082763 2.449490 6.782330 6.164414 2.645751 6.164414 5.830952  8.366600
12 2.236068 5.291503 2.000000 4.898979 6.855655 2.000000 2.000000  3.464102
13 6.403124 1.414214 6.164414 5.830952 4.358899 6.164414 5.477226  7.483315
14 4.690416 5.385165 5.567764 2.236068 6.928203 5.000000 5.196152  6.557439
15 3.000000 6.633250 3.741657 5.656854 7.549834 3.162278 3.741657  4.898979
16 5.916080 4.690416 6.782330 3.741657 5.916080 6.000000 5.830952  7.483315
17 3.000000 6.633250 3.741657 5.656854 7.416198 2.828427 2.449490  4.242641
18 6.324555 4.795832 7.000000 5.000000 6.000000 6.708204 6.855655  8.660254
19 5.196152 5.099020 6.000000 2.449490 6.708204 5.477226 5.656854  7.071068
20 6.082763 2.828427 6.633250 7.211103 2.645751 6.000000 5.656854  7.874008
          9       10       11       12       13       14       15        16
1
2
3
4
5
6
7
8
9  0.000000
```

```
10 5.744563 0.000000
11 2.828427 4.582576 0.000000
12 7.348469 4.582576 5.477226 0.000000
13 5.477226 3.316625 3.162278 4.898979 0.000000
14 6.557439 2.000000 5.567764 3.872983 4.795832 0.000000
15 7.874008 6.244998 6.164414 2.828427 6.480741 5.385165 0.000000
16 6.164414 1.732051 5.477226 5.291503 4.472136 2.645751 7.071068 0.000000
17 8.246211 6.244998 6.633250 3.162278 6.633250 5.567764 4.000000 6.324555
18 4.358899 3.741657 3.872983 5.567764 4.358899 4.000000 6.557439 5.000000
19 6.324555 1.732051 5.291503 4.242641 4.472136 1.000000 5.477226 2.828427
20 3.464102 4.795832 2.828427 5.830952 3.741657 5.916080 7.211103 5.099020
         17       18       19       20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17 0.000000
18 7.681146 0.000000
19 6.164414 3.605551 0.000000
20 7.071068 4.795832 5.830952 0.000000
```

Output di atas menghasilkan matriks jarak dengan menggunakan jarak *Euclidian*. Jarak dari data pertama dengan data pertama adalah 0 karena jarak dengan variabelnya itu sendiri adalah 0 atau tidak ada. Kemudian, jarak dari data ke-1 (kabupaten pertama) dengan data ke-2 (kabupaten kedua) sebesar 6,403124, sedangkan jarak data ke-1 (kabupaten pertama) dengan data ke-3 (kabupaten ketiga) sebesar 2,645751. Demikian seterusnya sampai data terakhir.

Perhatikan bahwa, matriks jarak yang dibuat adalah data pada kolom ke 2-7. Hal ini disebabkan kolom pertama berisi objek (kabupaten) yang akan dikelompokkan.

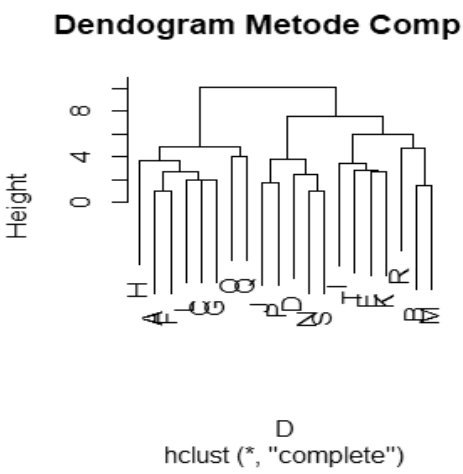
Setelah matriks jarak diketahui, maka langkah selanjutnya adalah membuat klaster dengan menggunakan perintah `hclust()`. Caranya seperti di bawah ini.

```
> y=hclust(D,method="complete")
> y

Call :
hclust(d = D, method = "complete")

Cluster method    : complete
Distance          : euclidean
Number of objects : 20
```

Informasi yang diperoleh dari output di atas adalah analisis klaster yang digunakan adalah metode *Complete* dengan menggunakan konsep jarak *Euclidian* dimana objek yang akan dikelompokkan sebanyak 20 pengamatan.

Untuk melihat hasil dendogram analisis klaster, gunakan perintah plot () dan `rect.hclust()` seperti di bawah ini.

```
> plot(y,X$Kabupaten,labels=X$Kabupaten,
main="Dendogram Metode Complete")
```



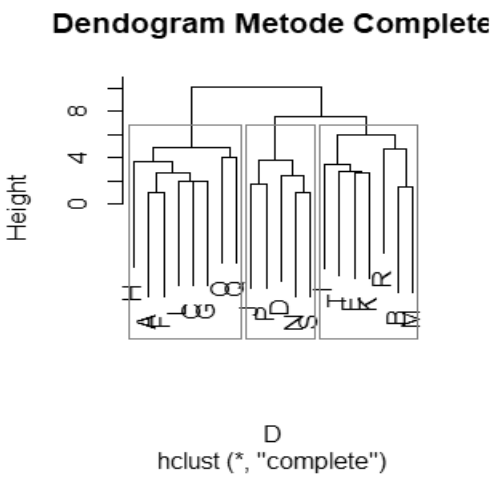
Hasil dendogram menunjukkan ada beberapa pengelompokan yang

mungkin terbentuk. Untuk menentukan berapa banyak kelompok yang terbentuk dapat menggunakan perintah rect.hclust().

```
rect.hclust(tree, k = NULL, which = NULL, x = NULL, h =
          NULL, border = 2, cluster = NULL)
Keterangan :
tree = objek
k = banyaknya kelompok yang terbentuk
h = besarnya jarak sebagai patokan penentuan kelompok
(cut poin)
border = warna
```

Misalkan pada data contoh kasus di atas, kelompok yang ingin dibentuk sebanyak 3 kelompok maka perintahnya seperti di bawah ini.

```
> rect.hclust(y,k=3, border="red")
```



Gambar di atas menunjukkan dendrogram hasil analisis klaster hierarki dengan menggunakan metode *Complete Linkage*. Metode *complete linkage* mengelompokkan objek berdasarkan jarak maksimum antaranggota klaster, sehingga cenderung menghasilkan klaster yang relatif kompak dan homogen.

Berdasarkan dendrogram, dengan memotong pohon klaster pada ketinggian (*height*) tertentu sebagaimana ditandai oleh kotak merah, dapat

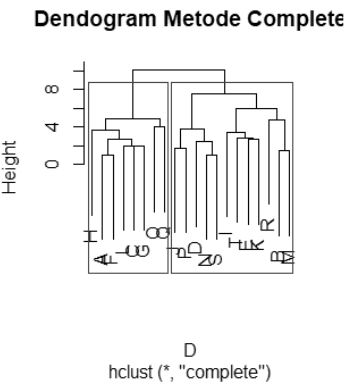
diidentifikasi tiga klaster utama. Setiap klaster berisi kabupaten yang memiliki kemiripan variabel/karakteritik ekonomi yang mirip/sama.

```
> anggota=data.frame(id=X$Kabupaten, cutree(y,k=3))
> anggota
    id   cutree.y..k...3.
1    A                  1
2    B                  2
3    C                  1
4    D                  3
5    E                  2
6    F                  1
7    G                  1
8    H                  1
9    I                  2
10   J                  3
11   K                  2
12   L                  1
13   M                  2
14   N                  3
15   O                  1
16   P                  3
17   Q                  1
18   R                  2
19   S                  3
20   T                  2
```

Berdasarkan output di atas terlihat bahwa anggota pada klaster 1 terdiri dari kabupaten A, C, F, G,H, L, O dan Q. Anggota pada klaster 2 terdiri dari kabupaten B, E, I, K, R, T dan sedangkan anggota pada klaster ke-3 terdiri dari kabupaten D, J, N, P dan S.

Misalkan kelompok yang ingin dibentuk adalah dengan menggunakan kriteria jarak terjauh `(h = 10)` maka gunakan perintah seperti di bawah ini.

```
> rect.hclust(y,h=10, border="blue")
```



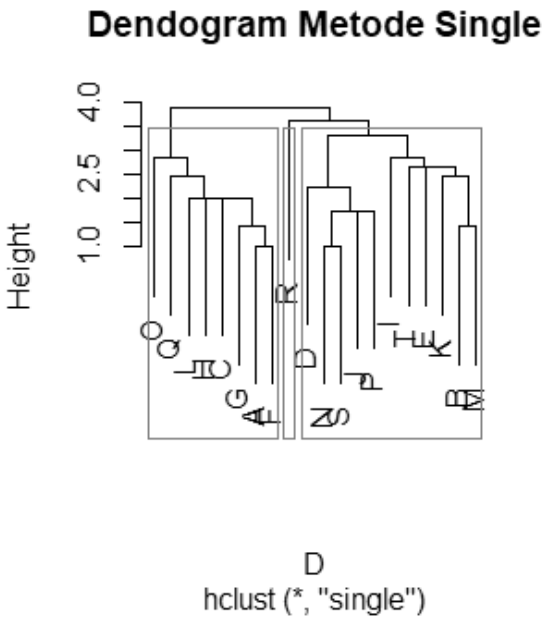
Nampak pada gambar *dendogram* di atas, klaster yang terbentuk terdiri dari 2 klaster (kelompok). Selanjutnya, keanggotaan setiap kelompok dapat dilihat melalui cara seperti di bawah ini.

```
> anggota=data.frame(id=X$Kabupaten, cutree(y,h=10))
> anggota
 id cutree.y..h...10.
1 A 1
2 B 2
3 C 1
4 D 2
5 E 2
6 F 1
7 G 1
8 H 1
9 I 2
10 J 2
11 K 2
12 L 1
13 M 2
14 N 2
15 O 1
16 P 2
17 Q 1
18 R 2
19 S 2
20 T 2
```

Berdasarkan output di atas, pada klaster 1 terdiri dari 8 anggota yaitu kabupaten A, C, F, G, H, L, O dan Q sedangkan klaster 2 terdiri dari kabupaten B, D, E, I, J, K, M, N, P, R, S dan T.

Pembentukan klaster dengan menggunakan metode lainnya misalnya *Single, Average* atau *Ward* pada dasarnya sama dengan metode *Complete.* Di sini akan diberikan contoh dengan menggunakan metode *Single* dengan banyaknya klaser ada 3.

```
> Z = hclust(D, method =”single”)
> plot(Z,X$Kabupaten a,labels=X$Kabupaten,
main=”Dendogram Metode Single”)
> rect.hclust(Z,k=3)
```




```
> anggota=data.frame(id=X$Mahasiswa, cutree(Z,k=3))
> anggota
 id cutree.Z..k...3.
1 A 1
2 B 2
3 C 1
4 D 2
5 E 2
6 F 1
7 G 1
8 H 1
9 I 2
10 J 2
```

```
11 K 2
12 L 1
13 M 2
14 N 2
15 O 1
16 P 2
17 Q 1
18 R 3
19 S 2
20 T 2
```

Berdasarkan output di atas, terbentuk tiga klaster dengan masing-masing anggota setiap klaster sebagai berikut :

- Klaster 1 anggotanya terdiri dari kabupaten A, C, F, G, H, L, O dan Q
- Klaster 2 anggotanya terdiri dari kabupaten B, D, E, I, J, K, M, N, P, S dan T
- Klater 3 hanya terdiri dari 1 anggota yaitu kabupaten R.

9.7 Teknik K-Means dengan Menggunakan Software R

K-Means adalah salah satu metode klaster non-hierarki. Teknik ini berusaha untuk membuat kelompok (klaster) dengan cara mempartisi data yang memiliki karakteristik yang sama dimasukan ke dalam satu kelompok.

Berbeda dengan teknik hierarki, pada teknik K-*Means* ini banyaknya klaster ditentukan terlebih dahulu misalnya dua, tiga atau empat klaster. *K-mean* sangat efektif dan efisien jika digunakan untuk mengelompokkan objek yang berjumlah besar.

Perintah untuk melakukan teknik K-Means dalam R adalah kmeans(), dengan sintak seperti di bawah ini.

```
kmeans(x, centers, iter.max = 10, nstart = 1,
 algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
 "MacQueen"), trace=FALSE)

Keterangan :

x          = data matriks
centers    = banyaknya klaster yang ingin dibentuk
algorithm  = pilihan metode algoritmanya "Hartigan-
Wong", "Lloyd", "Forgy", atau "MacQueen".
```

Sebelum melakukan klaster dengan teknik K-Means perlu diinstal *package* ggplot2 dan factoextra.

```
> install.packages("ggplot2")
> library(ggplot2)
> install.packages("factoextra")
> library(factoextra)
```

Contoh

Pada kasus ini, data yang digunakan adakah data Kabupaten (lihat contoh di atas). Pertama dipanggil datanya dengan menggunakan perintah seperti di bawah ini.

```
> X
   Kabupaten  X1  X2  X3  X4  X5  X6
1          A   6   5   6   3   3   4
2          B   2   3   2   4   5   4
3          C   6   3   6   4   2   3
4          D   3   7   4   5   2   7
5          E   2   4   2   2   7   4
6          F   6   4   6   3   3   4
7          G   5   3   6   3   3   4
8          H   7   2   7   4   2   4
9          I   2   7   2   3   7   3
10         J   3   5   3   6   4   6
11         K   2   5   2   3   5   3
12         L   5   4   5   4   2   4
13         M   2   3   2   5   4   4
14         N   4   6   4   6   3   6
15         O   6   5   4   2   1   4
16         P   3   5   4   6   5   7
17         Q   4   4   7   2   2   5
18         R   3   7   2   6   4   3
19         S   4   6   3   6   3   6
20         T   3   4   3   4   7   3
```

Langkah pertama dalam teknik K-*Means* adalah menentukan berapa banyak klaster yang akan dibentuk. Di dalam R, fungsi yang dapat digunakan untuk menentukan banyaknya klaster adalah fviz_nbclust(). Sintak perintah seperti di bawah ini.

```
fviz_nbclust( x, FUNcluster, method = c("silhouette",
"wss", "gap_stat"), …)

Keterangan :

x = matriks data
method = pilihan metode within cluster sums of squares
(wss), average silhouette (silhouette) atau gap
statistics (gap_stat).
```

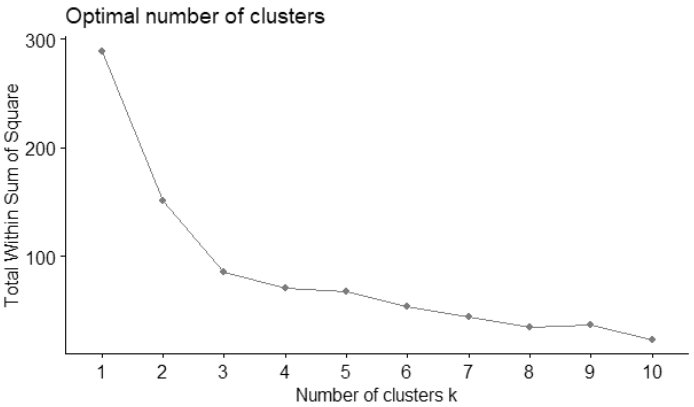
Ada tiga metode yang dapat digunakan yaitu metode `within cluster sums of squares (wss)`, `average silhouette (silhouette)` dan `gap statistics (gap_stat)`.

Pada contoh kasus ini, banyaknya klaster yang terbentuk akan dicari

dengan menggunakan 2 metode (wss dan silhouette). Caranya melalui perintah berikut ini :

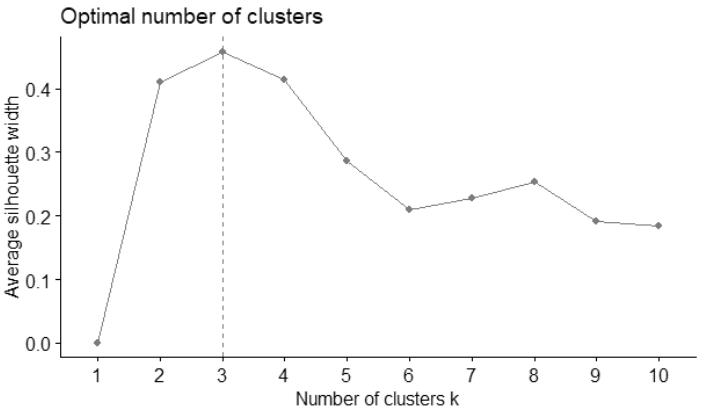
```
> fviz_nbclust(X[,2 :7], FUNcluster = kmeans, method="wss")
> fviz_nbclust(X[,2 :7], FUNcluster = kmeans, method="silhouette")
```

Perhatikan bahwa, data yang diolah hanya data numerik. Dengan demikian, data yang digunakan hanya data yang terdapat dalam kolom ke 2 sampai dengan kolom ke-7. Hasilnya dapat dilihat pada gambar di bawah ini.



Gambar 13. Plot Banyaknya Cluster dengan Metode WSS

Metode wss disebut juga dengan metode Elbow karena kriteria penentuan banyaknya klaster dilihat dari grafik yang turun secara drastis sehingga membentuk siku (*elbow*). Perhatikan pada gambar di atas, pada $k = 3$ terlihat grafik menurun drastis dan setelah itu pergerakannya tidak terlalu banyak (landai).



Gambar 14. Plot Banyaknya Cluster dengan Metode Silhoutte

Perhatikan gambar di atas, dengan metode *silhouette* menunjukkan banyaknya klaster yang terbentuk pada kasus ini adalah 3 kelompok (klaster).

Berdasarkan analisis kedua metode tersebut dapat disimpulkan bahwa banyaknya klaster yang terbentuk pada kasus ini adalah 3 klaster (kelompok).

Tahap selanjutnya dalam analisis klaster adalah menentukan keanggotaan setiap klaster. Caranya dengan melakukan perintah seperti di bawah ini.

```
> klaster=kmeans(X[,2 :7],centers=3)
> klaster


K-means clustering with 3 clusters of sizes 6, 8, 6

Cluster means :
        X1        X2        X3        X4        X5        X6
1 2.166667  4.333333  2.166667  3.500000  5.833333  3.500000
2 5.625000  3.750000  5.875000  3.125000  2.250000  4.000000
3 3.333333  6.000000  3.333333  5.833333  3.500000  5.833333

Clustering vektor :
 [1] 2 1 2 3 1 2 2 2 1 3 1 2 1 3 2 3 2 3 3 1

Within cluster sum of squares by cluster :
[1] 28.83333 30.62500 25.83333
 (between_SS / total_SS = 70.4 %)

Available components :
[1] "cluster" "centers" "totss" "withinss" "tot.withinss"
[6] "betweenss" "size" "iter" "ifault"
```

Hasil analisis k-Means dengan tiga klaster menunjukkan bahwa 20 kabupaten dapat dikelompokkan ke dalam tiga kelompok ekonomi yang berbeda, masing-masing beranggotakan 6, 8, dan 6 kabupaten. Profil klaster dibedakan berdasarkan nilai rata-rata keenam indikator ekonomi.

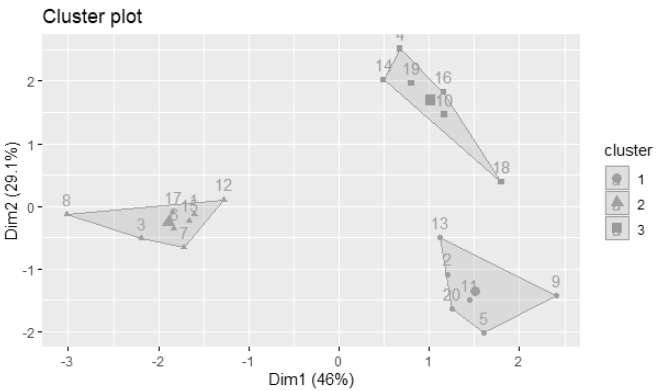
Pada klaster pertama rata-rata tertinggi berada pada variabel X_5 yaitu Investasi, klater kedua mempunyai rata-rata tertinggi pada variabel X_1 (PDB) dan X_3 (Inflasi) sedangkan klaster ketiga mempunyai rata-rata tertinggi pada variabel X_2 (Tingkat Pengangguran), X_4 (Konsumsi) dan X_6 (Pendapatan per Kapita).

Selain itu, dari output di atas diperoleh informasi mengenai jumlah kuadrat dalam setiap klaster yaitu berturut-turut untuk setiap klaster adalah 28,833; 30,625 dan 25,833. Sehingga ketiga klaster tersebut menjelaskan variansi sebesar 70,4%. Menunjukkan bahwa pengelompokan yang dihasilkan cukup baik, karena sebagian besar variasi data dapat dijelaskan oleh perbedaan antarklaster. Dengan demikian, hasil analisis ini berhasil

mengungkap pola heterogenitas ekonomi antar kabupaten.

Apabila ingin melihat visualisasi pengelompokan dapat menggunakan fungsi seperti di bawah ini.

```
> fviz_cluster(klaster, data = X[,2 :7])
```



Gambar di atas menunjukkan visualisasi dari 3 klaster yang terbentuk. Klaster 1 berwarna merah, klaster 2 berwarna hijau dan klaster tiga berwarna biru. Terlihat tiga klaster yang terbentuk sudah optimal. Untuk melihat anggota setiap klaster dilakukan dengan menggunakan perintah seperti di bawah ini.

```
> final=data.frame(X, klaster$cluster)
> final
   Kabupaten  X1  X2  X3  X4  X5  X6  klaster.cluster
1          A   6   5   6   3   3   4                2
2          B   2   3   2   4   5   4                1
3          C   6   3   6   4   2   3                2
4          D   3   7   4   5   2   7                3
5          E   2   4   2   2   7   4                1
6          F   6   4   6   3   3   4                2
7          G   5   3   6   3   3   4                2
8          H   7   2   7   4   2   4                2
9          I   2   7   2   3   7   3                1
10         J   3   5   3   6   4   6                3
11         K   2   5   2   3   5   3                1
12         L   5   4   5   4   2   4                2
13         M   2   3   2   5   4   4                1
14         N   4   6   4   6   3   6                3
15         O   6   5   4   2   1   4                2
16         P   3   5   4   6   5   7                3
17         Q   4   4   7   2   2   5                2
18         R   3   7   2   6   4   3                3
19         S   4   6   3   6   3   6                3
20         T   3   4   3   4   7   3                1
```

Berdasarkan output di atas diperoleh anggota masing-masing kelompok (klaster) yaitu :

- Klaster pertama anggotanya adalah kabupaten B, E, I, K,M dan T
- Klaster kedua anggotanya adalah kabupaten A,C,F,G,H,L,O dan Q
- Klaser ketiga anggotanya adalah kabupaten D, J. N, P, R dan S

9.8 Latihan

1. Pada Contoh di atas, data pengelompokan Kabupaten, lakukan pengelompokan dengan menggunakan metode Average dan Ward!
2. Pada Contoh di atas, data pengelompokan Kabupaten, lakukan pengelompokan dengan menggunakan metode *Average* tetapi konsep jarak yang digunakan adalah jarak maksimum!
3. Pada Contoh di atas, data pengelompokan Kabupaten tentukan

keanggotaan setiap kelompok apabila banyaknya kelompok adalah 4.

4. Pada tabel berikut adalah data Kabupaten/Kota di Provinsi Jawa Tengah berdasarkan tenaga kesehatan tahun 2015 seperti tenaga medis, tenaga keperawatan, tenaga kebidanan, tenaga kefarmasian dan tenaga kesehatan lainnya. Berdasarkan data tersebut lakukan analisis klaster dengan menggunakan metode K-Means. Analisis!

No	Kabupaten/ kota	Tenaga medis	Tenaga keperawatan	Tenaga kebidanan	Tenaga kefarmasian	Tenaga Kesehatan lainnya
1	Cilacap	87	432	630	30	129
2	Banyumas	286	934	730	60	312
3	Purbalingga	78	317	317	35	150
4	Banjarnegara	92	389	487	45	165
5	Kebumen	215	1004	726	90	298
6	Purworejo	133	550	415	63	154
7	Wonosobo	109	394	359	59	113
8	Magelang	169	488	474	60	223
9	Boyolali	204	691	487	187	247
10	Klaten	247	1100	557	102	393
11	Sukoharjo	454	1320	542	141	469
12	Wonogiri	157	529	277	72	247
13	Karanganyar	271	576	544	57	204
14	Sragen	247	841	649	95	314
15	Grobogan	215	1091	736	151	300
16	Blora	112	579	418	11	155
17	Rembung	81	426	372	31	149
18	Pati	183	995	649	95	266
19	Kudus	227	1044	450	98	246
20	Jepara	189	703	442	208	221
21	Demak	112	461	399	55	137
22	Semarang	271	579	297	86	218
23	Temanggung	217	623	410	154	188
24	Kendal	166	628	464	68	221
25	Batang	108	553	474	86	141
26	Pekalongan	215	773	574	105	232
27	Pemalang	137	612	463	50	173
28	Tegal	169	735	683	53	204

No	Kabupaten/ kota	Tenaga medis	Tenaga keperawatan	Tenaga kebidanan	Tenaga kefarmasian	Tenaga Kesehatan lainnya
29	Brebes	168	742	876	79	206
30	Kota Magelang	269	850	129	167	165
31	Kota Surakarta	666	2334	292	349	601
32	Kota Salatiga	185	604	158	76	260
33	Kota Semarang	1024	3371	443	565	889
34	Kota Pekalongan	204	586	230	131	193
35	Kota Tegal	176	629	131	156	240

5. Pada tabel berikut adalah data Provinsi di Indonesia berdasarkan indikator Indeks Pembanguan Manusia (IPM) yang meliputi Harapan hidup, harapan lama sekolah dan pendapatan per kapita. Berdasarkan data tersebut, lakukan analisis klaster untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan indikator IPM!

No	Provinsi	Harapan Hidup	Harapan Lama Sekolah	Lama Sekolah	Pendapatan Per Kapita
1	Aceh	73.20	14.39	9.64	10.811
2	Sumatera Utara	73.90	13.49	9.93	11.46
3	Sumatera Barat	74.37	14.3	9.44	11.718
4	Riau	74.41	13.42	9.43	11.857
5	Jambi	74.06	13.14	8.9	11.621
6	Sumatera Selatan	74.26	12.64	8.57	12.015
7	Bengkulu	73.31	13.75	9.04	11.733
8	Lampung	74.39	12.78	8.36	11.258
9	Kep. Bangka Belitung	74.12	12.49	8.33	13.667
10	Jawa Barat	75.16	12.8	8.87	12.157
11	Jawa Tengah	74.91	12.86	8.02	12.276
12	Jawa Timur	75.07	13.43	8.28	12.852

No	Provinsi	Harapan Hidup	Harapan Lama Sekolah	Lama Sekolah	Pendapatan Per Kapita
13	Banten	74.97	13.1	9.23	13.097
14	Nusa Tenggara Barat	72.25	13.98	7.87	11.606
15	Kalimantan Barat	73.94	12.68	7.78	10.321
16	Kalimantan Tengah	73.73	12.77	8.81	12.303
17	Kalimantan Selatan	74.18	12.87	8.62	13.399
18	Kalimantan Timur	74.94	14.03	10.02	13.793
19	Kalimantan Utara	73.57	13.21	9.35	10.197
20	Sulawesi Utara	74.08	12.97	9.84	11.998
21	Sulawesi Selatan	73.83	13.55	8.86	12.275
22	Kepulauan Riau	75.12	13.27	10.5	15.573
23	DKI Jakarta	75.99	13.51	11.49	19.953
24	D.I. Yogyakarta	75.36	15.7	9.92	15.361
25	Bali	75.10	13.62	9.54	14.92
26	Nusa Tenggara Timur	71.83	13.23	8.02	8.534
27	Sulawesi Tengah	70.84	13.34	9.04	10.536
28	Sulawesi Tenggara	71.88	13.71	9.42	10.606
29	Gorontalo	70.73	13.17	8.29	11.539
30	Sulawesi Barat	71.03	12.89	8.15	10.208
31	Maluku	70.68	14.09	10.26	9.684
32	Papua Barat	68.47	13.17	7.86	8.805

No	Provinsi	Harapan Hidup	Harapan Lama Sekolah	Lama Sekolah	Pendapatan Per Kapita
33	Papua Barat Daya	70.02	13.88	8.39	8.733
34	Papua	70.47	13.72	9.82	11.037
35	Papua Selatan	68.46	12.67	8.38	9.756
36	Maluku Utara	71.05	13.75	9.37	9.32
37	Papua Tengah	68.18	9.63	6.12	7.809
38	Papua Pegunungan	67.39	9.97	4.21	5.707

BAB

X

PENSKALAAN MULTIDIMENSIONAL (*MULTIDIMENSIONAL SCALLING*)

10.1 Pendahuluan

Dalam berbagai bidang sering dihadapkan pada kebutuhan untuk memahami tingkat kemiripan atau ketidakmiripan (*similarity/dissimilarity*) antarobjek yang diteliti. Informasi mengenai kemiripan tersebut tidak selalu dapat dijelaskan secara langsung melalui variabel-variabel terukur, tetapi sering kali diperoleh dalam bentuk jarak, peringkat, atau penilaian subjektif. Oleh karena itu, diperlukan suatu metode analisis yang mampu merepresentasikan hubungan antarobjek tersebut secara visual dan intuitif, yaitu penskalaan multidimensional (*Multidimensional Scalling)*.

Multidimensional Scalling (MDS) adalah suatu analisis yang dapat digunakan untuk memetakan atau mencari konfigurasi sejumlah objek dalam ruang dimensi rendah, misalnya dua atau tiga, yang diharapkan dapat merefleksikan sebaik mungkin ukuran ketakmiripan yang diketahui antarobjek tersebut. Dengan demikian, MDS bertujuan untuk mengidentifikasi pola, struktur, dan pengelompokan objek berdasarkan hubungan kedekatannya secara relatif. Keunggulan utama penskalaan multidimensional terletak pada kemampuannya untuk menyederhanakan data yang kompleks menjadi representasi visual yang mudah dipahami.

Penskalaan multidimensional banyak digunakan dalam bidang pemasaran, psikologi, pendidikan, ilmu sosial, dan kesehatan, misalnya untuk memetakan persepsi konsumen terhadap merek, menganalisis persepsi mahasiswa terhadap kualitas layanan pendidikan, atau mempelajari pola kemiripan antargejala penyakit. Peta yang dihasilkan oleh MDS membantu peneliti dalam menafsirkan posisi relatif objek dan memberikan dasar yang kuat untuk pengambilan keputusan.

10.2 *Multidimensional Scalling* (MDS)

Analisis MDS ialah suatu kelas prosedur untuk menyajikan persepsi dan preferensi pelanggan secara spatial dengan menggunakan tayangan yang bisa dilihat (*a visual display*).

Persepsi atau hubungan antara stimulus secara psikologis ditunjukkan sebagai hubungan geografis antar titik-titik di dalam suatu ruang

multidimensional (peta spatial).

Peta spatial ialah hubungan antara merek atau stimulus lain yang dipersepsikan, dinyatakan sebagai hubungan geometris antara titik-titik di dalam ruang yang multidimensional koordinat, menunjukkan posisi (letak) suatu merek atau stimulus dalam peta spatial. Berdasarkan tipe datanya, MDS dibagi menjadi 2 yaitu :

1. MDS Metrik (Klasik) digunakan apabila datanya bersifat kuantitatif dan memiliki skala interval dan rasio;
2. MDS Non Metrik digunakan apabila data kualitatif dengan skala nominal atau ordinal.

Manfaat penskalaan multidimesional (*multidimensional scalling*) yaitu :

1. Ukuran citra (*image measurement*). Membandingkan persepsi pelanggan dan bukan pelanggan dari perusahaan dengan persepsi perusahaan.
2. Segmentasi pasar (*market segmentation*).
3. Pengembangan produk baru (*new product development*). Melihat adanya celah (*gap*) dalam peta spatial, yang menunjukkan adanya peluang untuk menempatkan produk baru.

Prosedur analisis penskalaan multidimesional (*multidimensional scalling*) sebagai berikut :

1. Merumuskan Masalah
   Tentukkan tujuan dari hasil penelitian
   Memilih merek/stimulus yang akan digunakan. Biasanya banyaknya stimulus harus lebih besar dari 8 tetapi kurang dari 25.
2. Memperoleh Input Data
   a. Data Persepsi → Responden diminta memberikan penilaian terhadap seluruh kemungkinan pasangan merek mengenai kemiripan atau ketidakmiripan. Kalau ada *n* merek, maka akan terdapat *n(n-1)/2* pasangan.
   b. Data preferensi → Responden diminta untuk membuat peringkat dari yang paling disukai sampai paling tidak disukai
   Memilih Prosedure Penskalaan Multidimensional
3. Memilih prosedur MDS
   a. Tergantung data input: data ordinal berupa matriks atau non-matriks.

b. Tingkat responden/individu atau tingkat agregrat/kelompok

4. Menentukan banyaknya Dimensi
   a. Pengetahuan sebelumnya (teori atau riset)
   b. Kemudahan untuk menginterpretasikan dimensi, biasanya dipilih 2 or 3 dimensi.
   c. Kriteria siku (*elbow criterion*). Titik pada saat suatu siku atau bengkokan tajam terjadi, menunjukkan banyaknya dimensi yang tepat.
   d. Mudah dipergunakan
   e. Pendekatan statistik
5. Membentuk *perceptual map*
6. Berikan label nama dimensi dan interpretasikan dari *perceptual map*
   a. Pemberian label pada dimensi memerlukan pertimbangan subjektif dari pihak peneliti.
   b. Konfigurasi dalam peta spatial dapat diinterpretasikan dengan mengkaji koordinat dan posisi relatif dari merek. Merek yang berdekatan akan bersaing keras (kemiripan yang besar) dan merek yang terisolas menunjukkan suatu citra yang unik. Celah (*Gap*) dalam peta spatial mungkin menunjukkan adanya peluang potensal untuk mengenalkan produk baru.
   c. Semakin dekat posisi sebuah objek dengan objek lainnya maka dikatakan objek tersebut semakin mirip, sedangkan semakin jauh posisi sebuah objek dengan objek lainnya maka dikatakan objek tersebut berbeda dengan yang lainnya.
7. Evaluasi Keandalan dan Kesahihan

   Stress adalah ukuran ketidakcocokkan (*a lak of fit measure*), semakin tinggi nilai *stress* semakin tidak cocok. Nilai stres menunjukkan indikasi mutu pemecahan MDS. Adapun kriterianya seperti di bawah ini.

Tabel 20. Kriteria Kebaikan MDS

Stress	*Goodness of Fit* (GOF)

20,00%	Jelek
10,01% - 20,00 %	Cukup
5,01% - 10,00%	Bagus
2,51% - 5,00%	Sangat Bagus
< 2,51%	Sempurna

dimana nilai Stress adalah :

$$Stress = \sqrt{\frac{\sum_{ij}^{n}(d_{ij}-\hat{d}_{ij})^2}{\sum_{ij}^{n} d_{ij}^2}}$$

Nilai *Goodness of fit* adalah:

$$R^2 = 1 - Stress = 1 - \sqrt{\frac{\sum_{ij}^{n}(d_{ij}-\hat{d}_{ij})^2}{\sum_{ij}^{n} d_{ij}^2}}$$

Adapun algortima MDS yaitu :

a. Untuk *n* merk/stimulus, hitung jarak/kemiripan setiap pasangan stimulus, Urutkan dari yang terkecil ke terbesar,
b. Lakukan regresi (kuadrat terkecil) monotonik d_{ij} terhadap σ_{ij}. Hasil dugaan,
c. $\sigma_{ij} = a + b\, d_{ij,}$
d. Hitung stres.

Prosedur MDS dengan Menggunakan R

Dalam software R fungsi yang digunakan untuk *multidimensoinal scalling* dibedakan menjadi dua yaitu *classical multidimensional scalling* (*Classical* MDS) dan *nonmetric multidimensional scaling* (Nonmetric MDS).

Untuk *classical* MDS dapat digunakan fungsi cmdscale() dengan sintak seperti di bawah ini.

```
cmdscale(d, k = 2, eig = FALSE, …)

Keterangan :
d = matriks jarak
k = dimensi peta spatial
eig = eigenvalue (akar ciri)
```

Contoh Kasus MDS Klasik

Suatu penelitian ingin mengetahui kemiripan kota-kota besar di Indonesia berdasarkan keadaan geografisnya yaitu X_1 = Ketinggian; X_2 = Suhu Minimum; X_3 = Kecepatan Angin; X_4 = Kelembapan Udara dan X_5 = Curah Hujan.

Tabel 21. Data Provinsi Berdasarkan Keadaan Geografis

Kota	Ketinggian	Suhu Minimum	Kecepatan Angin	Kelembapan Udara	Curah Hujan
Aceh	21	24,9	11,8	69,7	5,5
Medan	25	25,8	9,8	70,3	80,3
Padang	3	25,8	11,8	68,9	178,5
Jambi	25	26,7	3,0	69,5	147,1
Bengkulu	16	26,8	12,6	66,8	400,6
Jakarta	2	25,3	15,3	77,3	404,5
Bandung	740	27,1	18,5	73,6	168,0
Semarang	3	25,3	14,2	73,8	301,5
Yogyakarta	107	25,7	7,9	73,6	364,4
Surabaya	3	28,4	17,6	69,7	216,8
Denpasar	1	25,6	11,1	74,2	294,2
Kupang	108	25,5	7,3	75,2	136,3
Samarinda	230	20,3	5,2	76,3	145,8
Manado	80	25,7	8,9	74,8	158,0
Ambin	12	25,9	7,1	72,2	60,4
Jayapura	99	27,0	11,9	74,6	115,8

Berdasarkan data di atas, lakukan analisis *multidimensional scalling* (MDS) untuk mengetahui kedekatan kota-kota tersebut berdasarkan keadaan geografisnya!

Langkah pertama, inputkan data tersebut ke dalam R. Misalnya data disimpan dalam Excel dengan nama Provinsi.csv yaitu :

```
> Prov=read.csv2("Provinsi.csv")
```

Setelah data berhasil diinputkan maka langkah berikutnya adalah membentuk matriks jarak terlebih dahulu. Untuk menghitung jarak hanya kolom-kolom yang berisi angka numerik saja yaitu data pada kolom ke 2 sampai kolom ke-6. Berikut membuat matrik jarak untuk data Prov :

```
> Jarak=dist(Prov[,2 :6]) #Membuat matriks jarak
> round(Jarak,digits=2) # Membulatkan matrik jarak dengan digit = 2

        1      2      3      4      5      6      7      8      9
2   74.94
3  173.94 100.66
4  141.94  67.16  39.35
5  395.15 320.46 222.49 253.86
6  399.54 325.14 226.19 258.84  18.19
7  737.18 720.42 737.12 715.49 760.50 774.99
8  296.59 222.36 123.12 156.43 100.22 103.07 749.01
9  369.10 295.72 213.10 232.35  98.29 112.70 662.85 121.71
10 212.17 138.51  38.83  74.55 184.36 187.90 738.63  84.92 180.88
11 289.43 215.28 115.84 149.34 107.72 110.43 749.74   8.19 127.18
12 157.25 100.28 113.43  84.01 280.03 288.51 632.90 195.87 228.11
13 251.94 215.41 229.62 205.23 333.03 345.02 510.71 275.47 250.92
14 163.62  95.31  79.95  56.64 251.06 258.64 660.15 162.94 208.16
15  55.90  24.00 118.58  87.81 340.31 344.38 736.00 241.38 318.50
16 135.20  82.22 114.81  81.00 296.75 304.60 643.16 209.07 248.77
       10     11     12     13     14     15
2
3
4
5
6
7
8
9
10
```

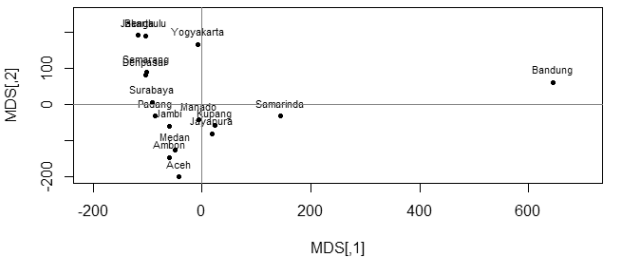
```
11  77.88
12 132.85 190.78
13 238.40 273.00 122.50
14  97.44 157.47  35.46 150.65
15 157.05 234.10 122.42 234.24 118.99
16 139.55 203.55  22.91 134.74  46.40 103.29
```

Selanjutnya, setelah membuat matriks jarak maka analisis MDS. Misalkan dalam kasus ini MDS akan dibentuk dengan dua dimensi ($k = 2$). Berikut perintahnya.

```
> MDS=cmdscale(Jarak, k=2)
> MDS
               [,1]          [,2]
 [1,]   -43.514473   -201.135556
 [2,]   -50.013737   -126.539616
 [3,]   -85.510528    -32.362058
 [4,]   -59.390003    -60.492928
 [5,]  -103.671578    189.358368
 [6,]  -118.020417    191.345827
 [7,]   645.767409     60.241305
 [8,]  -102.660752     89.474135
 [9,]    -8.509285    166.186692
[10,]   -90.824668      5.643597
[11,]  -103.638067     81.928236
[12,]    24.351285    -59.518291
[13,]   143.818041    -33.111587
[14,]    -6.397249    -41.922093
[15,]   -60.114143   -148.084791
[16,]    18.328165    -81.011240
```

Output di atas adalah koordinat yang diperlukan untuk menggambarkan ke-16 kota dalam 2 dimensi. Untuk menampilkan gambar peta spatial dapat dilanjutkan dengan perintah-perintah berikut ini :

```
> X=MDS[,1]
> Y=MDS[,2]
>Kota=c("Aceh","Medan","Padang","Jambi","Bengkulu",
"Jakarta","Bandung","Semarang","Yogyakarta",
"Surabaya","Denpasar","Kupang","Samarinda","Manado",
"Ambon","Jayapura")

> plot(MDS, type="p", pch=20, ylim=c(-200,250), xlim=c(-200,700))
> text(X, Y, labels = Kota, cex=0.7, pos=3)
> abline (v = 0, col="red") #Membuat garis vertical di koordinat 0
> abline (h = 0, col="red") #Membuat garis horizontal di koordinat 0
```



Gambar di atas menunjukkan setiap titik-titik yang berdekatan menunjukkan kemiripan yang tinggi begitu pula sebaliknya. Dalam kasus ini terlihat Bengkulu dan Jakarta posisi sangat dekat sehingga disimpulkan mempunyai kemiripan yang besar, begitu pula dengan Semarang dan Denpasar. Sebaliknya kota Bandung berada pada posisi yang berjauhan dengan kota-kota yang lain, sehingga kota Bandung tidak memiliki kemiripan dekat kota-kota lainnya.

Apabila dilihat berdasarkan posisi di setiap kuadran, maka kota-kota tersebut dapat dikelompokkan sebagai berikut :

Kelompok I : Jakarta, Bengkulu, Yogyakarta, Semarang, Denpasar, Surabaya.
Kelompok II : Bandung.
Kelompok III : Samarinda, Kupang, Jayapura.
Kelompok IV : Padang, Manado, Jambi, Medan, Ambon, Aceh

Untuk melihat kecocokan model analisis *multidimensional scalling*

(MDS) dapat dilihat dari nilai *Goodness of Fit* (GOF). Apabila nilai GOF semakin mendekati 1, maka modelnya semakin baik. Perintah untuk menghitung GOF dalam R dapat diperoleh melalui cara seperti di bawah ini.

```
> MDS2=cmdscale(Jarak, eig=TRUE)
> MDS2$GOF
[1] 0.9994687 0.9994687
```

Terlihat dari output di atas nilai GOF pada kasus ini sebesar 99,95% sehingga dapat disimpulkan model MDS sudah bagus.

10.3 Latihan

1. Lakukan analisis MDS pada data mtcars dengan dimensi k = 2, karena variabel vs dan am bersifat kategori maka tidak digunakan dalam kasus ini.
   a. Buat matriks jarak!
   b. Buat peta spatial dua dimensinya
   c. Berdasarkan letak/posisi pada peta spatial, jenis mobil mana yang mempunyai kemiripan yang tinggi?
   d. Kelompokkan jenis kendaraan tersebut apabila dilihat dari posisi di kuadrannya!

2. Indeks Pembangunan Manusia (IPM) adalah ukuran yang menunjukkan tingkat pencapaian pembangunan manusia di suatu wilayah, dan mengukur kualitas hidup masyarakat berdasarkan tiga dimensi utama: umur panjang dan sehat, pengetahuan, serta standar hidup layak. Dimensi ini diukur melalui harapan hidup saat lahir, lama bersekolah, dan pendapatan per kapita. IPM digunakan untuk mengklasifikasikan status pembangunan suatu negara atau wilayah, mulai dari maju, berkembang, hingga terbelakang. Berdasarkan data dari Badan Pusat Statsisk (BPS), IPM Menurut Komponen Pembentuk dan Provinsi, 2022–2024 disajikan pada tabel seperti di bawah ini.

Provinsi	Umur Harapan Hidup Saat Lahir (tahun)			Harapan lama sekolah (tahun)			Rata-rata Lama Sekolah (tahun)		
	2022	2023	2024	2022	2023	2024	2022	2023	2024
Aceh	72,92	73,06	73,2	14,37	14,38	14,39	9,44	9,55	9,64
Sumatera Utara	73,39	73,67	73,9	13,31	13,48	13,49	9,71	9,82	9,93
Sumatera Barat	73,88	74,14	74,37	14,1	14,11	14,3	9,18	9,28	9,44
Riau	73,95	74,18	74,41	13,29	13,3	13,42	9,22	9,32	9,43
Jambi	73,61	73,84	74,06	13,05	13,13	13,14	8,68	8,81	8,9
Sumatera Selatan	73,76	74,04	74,26	12,55	12,63	12,64	8,37	8,5	8,57
Bengkulu	72,9	73,11	73,31	13,68	13,74	13,75	8,91	9,03	9,04
Lampung	73,95	74,17	74,39	12,74	12,77	12,78	8,18	8,29	8,36
Kep Bangka Belitung	73,68	73,9	74,12	12,18	12,31	12,49	8,11	8,25	8,33
Kepulauan Riau	74,62	74,9	75,12	12,99	13,05	13,27	10,37	10,41	10,5
DKI Jakarta	75,54	75,81	75,99	13,08	13,33	13,51	11,31	11,45	11,49
Jawa Barat	74,65	74,91	75,16	12,62	12,68	12,8	8,78	8,83	8,87
Jawa Tengah	74,58	74,69	74,91	12,81	12,85	12,86	7,93	8,01	8,02
D.I Yogyakarta	75,11	75,18	75,36	15,65	15,66	15,7	9,75	9,83	9,92
Jawa Timur	74,57	74,87	75,07	13,37	13,38	13,43	8,03	8,11	8,28
Banten	74,46	74,77	74,97	13,05	13,09	13,1	9,13	9,15	9,23
Bali	74,6	74,88	75,1	13,48	13,58	13,62	9,39	9,45	9,54
Nusa Tenggara Barat	71,66	72,02	72,25	13,96	13,97	13,98	7,61	7,74	7,87
Nusa Tenggara Timur	71,3	71,57	71,83	13,21	13,22	13,23	7,7	7,82	8,02
Kalimantan Barat	73,47	73,71	73,94	12,66	12,67	12,68	7,59	7,71	7,78
Kalimantan Tengah	73,34	73,54	73,73	12,75	12,76	12,77	8,65	8,73	8,81
Kalimantan Selatan	73,7	73,97	74,18	12,82	12,86	12,87	8,46	8,55	8,62
Kalimantan Timur	74,45	74,72	74,94	13,84	14,02	14,03	9,92	9,99	10,02
Kalimantan Utara	73,51	73,54	73,57	13,06	13,2	13,21	9,27	9,34	9,35
Sulawesi Utara	73,59	73,85	74,08	12,95	12,96	12,97	9,68	9,77	9,84
Sulawesi Tengah	70,49	70,66	70,84	13,32	13,33	13,34	8,89	8,96	9,04
Sulawesi Selatan	73,4	73,63	73,83	13,53	13,54	13,55	8,63	8,76	8,86
Sulawesi Tenggara	71,7	71,79	71,88	13,69	13,7	13,71	9,25	9,31	9,42

Provinsi	Umur Harapan Hidup Saat Lahir (tahun)			Harapan lama sekolah (tahun)			Rata-rata Lama Sekolah (tahun)		
	2022	2023	2024	2022	2023	2024	2022	2023	2024
Gorontalo	70,22	70,5	70,73	13,12	13,16	13,17	8,02	8,1	8,29
Sulawesi Barat	70,42	70,76	71,03	12,87	12,88	12,89	8,08	8,13	8,15
Maluku	70,16	70,45	70,68	14	14,08	14,09	10,19	10,2	10,26
Maluku Utara	70,47	70,76	71,05	13,73	13,74	13,75	9,24	9,26	9,37
Papua Barat	68,06	68,28	68,47	13,14	13,16	13,17	7,43	7,66	7,86
Papua Barat Daya	69,58	69,83	70,02	13,86	13,87	13,88	8,22	8,23	8,39
Papua	70,03	70,26	70,47	13,6	13,71	13,72	9,63	9,64	9,82
Papua Selatan	68,06	68,27	68,46	12,44	12,56	12,67	8,04	8,24	8,38
Papua Tengah	67,86	68,04	68,18	9,61	9,62	9,63	5,91	6,02	6,12
Papua Pegunungan	67,11	67,27	67,39	9,95	9,96	9,97	3,89	4,01	4,21
Indonesia	73,7	73,93	74,15	13,1	13,15	13,21	8,69	8,77	8,85

Berdasarkan data di atas, lakukan teknik MDS dengan mengambil data pada tahun 2024 kemudian analisis hasilnya.

BAB

XI

ANALISIS KOMPONEN UTAMA (*PRINCIPAL COMPONENT ANALYSIS*)

11.1 Pendahuluan

Dalam penelitian yang melibatkan banyak variabel, sering dihadapi pada permasalahan kompleksitas data dan tingginya keterkaitan antarvariabel. Banyaknya variabel yang saling berkorelasi dapat menyulitkan proses analisis dan interpretasi hasil, serta berpotensi menimbulkan masalah seperti multikolinearitas. Misalkan, suatu penelitian mengenai faktor-faktor yang memengaruhi tingkat inflasi. Tingkat inflasi dipengaruhi oleh beberapa variabel diantaranya, Bahan makanan (X_1), Makanan jadi, minuman, tembakau, rokok (X_2), Perumahan, air, listrik, gas, bahan bakar (X_3), Sandang (X_4), Kesehatan (X_5), Pendidikan, rekreasi, olahraga (X_6), Transportasi, komunikasi, dan jasa keuangan (X_7), Nilai Tukar (X_8), Ekspor (X_9) dan Impor (X_{10}). Penggunaan seluruh variabel tersebut secara bersamaan dalam analisis lanjutan, seperti regresi, dapat menimbulkan multikolinearitas dan menyulitkan interpretasi pengaruh masing-masing variabel. Oleh karena itu, diperlukan suatu teknik analisis statistik yang mampu menyederhanakan struktur data tanpa menghilangkan informasi penting yang terkandung di dalamnya, yaitu analisis komponen utama.

Analisis komponen utama (AKU) merupakan suatu teknik analisis statistik untuk mentransformasi variabel-variabel asli yang masih saling berkorelasi satu dengan yang lain menjadi satu set variabel baru yang tidak berkorelasi lagi. Peubah-peubah baru itu disebut sebagai komponen utama (Johnson dan Wichern, 1982).

Keunggulan utama PCA terletak pada kemampuannya dalam menyederhanakan data berdimensi tinggi menjadi struktur yang lebih ringkas dan mudah diinterpretasikan, tanpa kehilangan informasi yang signifikan. Selain itu, PCA termasuk dalam metode saling ketergantungan (*interdependence methods*), karena tidak membedakan variabel dependen dan independen. Metode ini banyak digunakan sebagai tahap awal eksplorasi data, serta sebagai alat bantu dalam analisis lanjutan seperti analisis klaster, regresi, dan pemodelan struktural.

11.2 Model Analisis Komponen Utama (AKU)

Analisis Komponen Utama (*Principal Component Analysis/PCA*)

merupakan salah satu metode analisis multivariat yang bertujuan untuk mereduksi dimensi data dengan cara mentransformasikan sekumpulan variabel asal yang saling berkorelasi menjadi sejumlah variabel baru yang saling bebas (tidak berkorelasi), yang disebut sebagai komponen utama. Komponen utama tersebut dibentuk sebagai kombinasi linier dari variabel-variabel asal dan disusun sedemikian rupa sehingga komponen pertama mampu menjelaskan variasi data terbesar, diikuti oleh komponen-komponen berikutnya dengan kontribusi variasi yang semakin kecil.

Misalkan terdapat p variabel acak $\boldsymbol{X} = (X_1, X_2, ..., X_p)$, maka model AKU adalah kombinasi linear dari p variabel acak tersebut, dirumuskan seperti di bawah ini.

$$KU_1 = \boldsymbol{a_1'X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$
$$KU_2 = \boldsymbol{a_2'X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$
$$\vdots$$
$$KU_p = \boldsymbol{a_p'X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

Keterangan :

α_i' adalah vektor koefisien persamaan komponen utama ke-i, untuk i = 1, 2, … , p dimana keragaman (variansi) masing-masing Komponen Utama (KU) adalah :

$$Var(KU_i) = \boldsymbol{a_i'\Sigma a_i} = \lambda_i; \qquad \text{untuk } i = 1,2,\cdots,p$$

Komponen utama memiliki sifat varians yang semakin mengecil yaitu

$$Var(KU_1) \geq Var(KU_2) \geq \cdots \geq Var\left(KU_p\right) \geq 0$$

Atau

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

Oleh karena itu, sebagian besar variasi (keragaman atau informasi) cenderung berkumpul pada beberapa KU pertama, dan semakin sedikit informasi dari variabel asal yang terkumpul pada KU terakhir. Hal ini berarti bahwa KU pada urutan terakhir dapat diabaikan tanpa kehilangan banyak informasi.

Seperti disebutkan sebelumnya bahwa persamaan Komponen Utama yang terbentuk tidak saling berkorelasi. Oleh karena itu :

Komponen Utama ke-i = kombinasi linear dari $\boldsymbol{\alpha}_i'\mathbf{X}$

yang memaksimumkan $\boldsymbol{\alpha}_i' \boldsymbol{\Sigma}\boldsymbol{\alpha}_i$

dengan kendala $\boldsymbol{\alpha}_i' \boldsymbol{\alpha}_i = \mathbf{1}$

$$Cov\,(\boldsymbol{\alpha}_i'\mathbf{X}, \boldsymbol{\alpha}_k'\mathbf{X}) = 0;\ k < i$$

Teorema

Diberikan matriks variansi-kovariansi $\boldsymbol{\Sigma}$ dari variabel acak $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$. Matriks variansi-kovariansi $\boldsymbol{\Sigma}$ mempunyai akar ciri dan vektor ciri berpasangan yaitu $(\lambda_1\ \boldsymbol{e}_1)$, $(\lambda_2\ \boldsymbol{e}_2)$,…, $(\lambda_p\ \boldsymbol{e}_p)$ dimana $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, maka persamaan komponen utama ke-i adalah :

$$Y_i = \boldsymbol{e_1'X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p; \qquad i = 1,2, \ldots, p$$

dengan $Var(Y_i) = \boldsymbol{e_i'\Sigma e_i} = \lambda_i; \quad i = 1, 2, \ldots, p$

$$Y_i = \boldsymbol{e_1'X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p; \qquad i = 1,2, \ldots, p$$

Keterangan :

$\boldsymbol{e}_i' = (e_{11}, e_{12}, \ldots, e_{1k}, \ldots, e_{1p})'$adalah vektor koefisien pada komponen utama ke-*i*. Setiap koefisien mengukur bobot variabel ke-*k* terhadap komponen utama ke-*i*.

Contoh

Misalkan terdapat variabel acak X_1 dan X_2 dengan matriks variansi-kovarinasi seperti di bawah ini.

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

Tentukan persamaan komponen utamanya!

Jawab

Persamaan komponen utama adalah :

$$Y_1 = \boldsymbol{e_1'X} = e_{11}X_1 + e_{12}X_2;$$
$$Y_2 = \boldsymbol{e_1'X} = e_{11}X_1 + e_{12}X_2;$$

Langkah pertama adalah menghitung akar ciri dan vektor ciri dari matriks $\boldsymbol{\Sigma}$, dengan menggunakan persamaan seperti di bawah ini.

$$\begin{aligned}
|\boldsymbol{\Sigma} - \boldsymbol{\lambda I}| &= \left|\begin{bmatrix}5 & 2\\ 2 & 2\end{bmatrix} - \lambda\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right| \\
&= \left|\begin{bmatrix}5-\lambda & 2\\ 2 & 2-\lambda\end{bmatrix}\right| \\
&= (5-\lambda)(2-\lambda) - 4 \\
&= \lambda^2 - 7\lambda + 6 \\
&= (\lambda - 6)(\lambda - 1) = 0
\end{aligned}$$

Akar ciri dari matriks $\boldsymbol{\Sigma}$ adalah $\lambda_1 = 6$ dan $\lambda_2 = 1$ (perhatikan bahwa $\lambda_1 \geq \lambda_2$)

Selanjutnya, mencari vektor ciri yang berpasangan dengan setiap akar ciri. Vektor ciri untuk akar ciri pertama diperoleh dengan menyelesaikan persamaan seperti di bawah ini.

$$\boldsymbol{\Sigma x} = \lambda \boldsymbol{x}$$
$$\begin{bmatrix}5 & 2\\ 2 & 2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} = 6\begin{bmatrix}x_1\\ x_2\end{bmatrix}$$
$$\begin{bmatrix}5x_1 & 2x_2\\ 2x_1 & 2x_2\end{bmatrix} = \begin{bmatrix}6x_1\\ 6x_2\end{bmatrix}$$

Bentuk matriks di atas dapat ditulis kembali menjadi

$$5x_1 + 2x_2 = 6x_1$$
$$2x_1 + 2x_2 = 6x_2$$

Berdasarkan persamaan di atas, solusi penyelesaian ada nilai yang memenuhi persamaan seperti di bawah ini.

$$x_1 = 2x_2$$

Misalkan di pilih secara acak maka Oleh karena itu, vektor ciri yang

berpasangan dengan akar ciri pertama adalah

$$\boldsymbol{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Untuk vektor ciri yang berpasangan dengan akar ciri kedua dapat dihitung dengan cara yang sama, yaitu :

$$\boldsymbol{\Sigma x} = \lambda \boldsymbol{x}$$
$$\begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$
$$\begin{bmatrix} 5x_1 & 2x_2 \\ 2x_1 & 2x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Bentuk matriks di atas dapat ditulis kembali menjadi

$$5x_1 + 2x_2 = x_1$$
$$2x_1 + 2x_2 = x_2$$

Berdasarkan persamaan di atas, diperoleh persamaan berikut ini.

$$x_2 = -2x_1$$

Misalkan di pilih secara acak maka Oleh karena itu, vektor ciri yang berpasangan dengan akar ciri kedua adalah:

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

Ingat di dalam analisis komponen utama ada syaratnya yaitu $\boldsymbol{\alpha}_i' \, \boldsymbol{\alpha}_i = \mathbf{1}$ maka vektor ciri harus memenuhi syarat $\boldsymbol{e}_i' \, \boldsymbol{e}_i = \mathbf{1}$. Oleh karena itu, vektor ciri dicari dengan menggunakan persamaan seperti di bawah ini.

$$\boldsymbol{e} = \frac{\boldsymbol{x}}{\sqrt{\boldsymbol{x}'\boldsymbol{x}}}$$

sehingga, vektor ciri pertama adalah :

$$\boldsymbol{e_1} = \frac{\begin{bmatrix} 2 \\ 1 \end{bmatrix}}{\sqrt{\begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}}} = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$$

Vektor ciri kedua adalah :

$$\boldsymbol{e_2} = \frac{\begin{bmatrix} 1 \\ -2 \end{bmatrix}}{\sqrt{\begin{bmatrix} 1 & -2 \end{bmatrix}\begin{bmatrix} 1 \\ -2 \end{bmatrix}}} = \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix}$$

Persamaan komponen utama dapat dituliskan seperti di bawah ini.

$$Y_1 = \frac{2}{\sqrt{5}}X_1 + \frac{1}{\sqrt{5}}X_2;$$
$$Y_2 = \frac{1}{\sqrt{5}}X_1 - \frac{2}{\sqrt{5}}X_2;$$

11.3 Proporsi Keragaman dan Korelasi

Analisis komponen utama bertujuan untuk mereduksi variabel tanpa mengurangi informasi yang terkandung didalam data asal. Salah satu kriteria untuk menentukan banyaknya persamaan komponen utama adalah dengan menggunakan proporsi keragaman (variansi) yang dijelaskan oleh komponen utama.

Pada analisis komponen utama, total variansi dari data asal akan sama dengan total variansi komponen utamanya, artinya:

$$\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} Var(Y_i) \qquad (11.1)$$
$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

Poporsi keragaman (variansi) setiap komponen utama dapat dihitung menggunakan rumus seperti di bawah ini.

$$Proporsi\ keragaman\ Y_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad (11.2)$$

Persamaan di atas, menjelaskan persentase atau proporsi setiap komponen utama terhadap variansi data asal. Kriteria menentukan banyaknya komponen utama adalah proporsi kumulatif yang dijelaskan oleh komponen utama. Pada umumnya batas minimum proporsi kumulatif variansi sebesar 60% - 80%.

Selanjutnya, korelasi antara persamaan komponen utama ke-*i* dengan

variabel ke-k dapat dihitung dengan menggunakan rumus iseperti di bawah ini.

$$\rho_{Y_i;X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}; \qquad i,k = 1,2,\dots,p \qquad (11.2)$$

Contoh

Pada kasus sebelumnya, diketahui :

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

Hitunglah proorsi keragaman setiap komponen utama!

Jawab

Berdasarkan matriks $\boldsymbol{\Sigma}$ diketahui bahwa $\sigma_{11} = 5$; $\sigma_{22} = 2$ dan $\sigma_{12} = \sigma_{21} = 2$;

Kemudian diketahui juga bahwa $\lambda_1 = 6$; $\lambda_2 = 1$

Proporsi keragaman pada komponen utama pertama adalah :

$$Proporsi\ keragaman\ Y_1 = \frac{6}{6+1} = 0{,}857$$

Proporsi keragaman pada komponen utama kedua adalah :

$$Proporsi\ keragaman\ Y_2 = \frac{1}{6+1} = 0{,}143$$

Komponen utama pertama sudah menjelaskan keragaman data sebesar 85,7% dan komponen utama kedua menjelaskan sebesar 14,3%, sehingga pada kasus ini, komponen utama pertama sudah cukup baik menjelaskan variansi (keragaman) data.

Contoh

Pada kasus sebelumnya, hitunglah korelasi antara persamaan komponen utama pertama variabel ke-1 dan ke-2.

Jawab

Hasil perhitungan sebelumnya diketahui :

$$Y_1 = \frac{2}{\sqrt{5}}X_1 + \frac{1}{\sqrt{5}}X_2;$$

$$Y_2 = \frac{1}{\sqrt{5}}X_1 - \frac{2}{\sqrt{5}}X_2;$$

Diketahui bahwa $\sigma_{11} = 5; \sigma_{22} = 2$ dan $\sigma_{12} = \sigma_{21} = 2; \lambda_1 = 6 ; \lambda_2 = 1$

Oleh karena itu dengan menggunakan persamaan di atas, korelasi antara persamaan komponen utama pertama dengan variabel ke-1 adalah:

$$\rho_{Y_1;X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}}$$

$$\rho_{Y_1;X_1} = \frac{2/\sqrt{5}(\sqrt{6})}{\sqrt{5}} = 0{,}98$$

Hubungan antara komponen utama pertama dengan variabel pertama sebesar 0,98 artinya hubungannya positif dan kuat.

Kemudian, korelasi antara persamaan komponen utama pertama dengan variabel ke-2 adalah

$$\rho_{Y_1;X_2} = \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}}$$

$$\rho_{Y_1;X_2} = \frac{1/\sqrt{5}(\sqrt{6})}{\sqrt{2}} = 0{,}775$$

Hubungan antara komponen utama pertama dengan variabel kedua sebesar 0,775, artinya hubungannya positif dan cukup kuat.

Selanjutnya, korelasi antara persamaan komponen utama kedua dengan variabel ke-1 adalah

$$\rho_{Y_2;X_1} = \frac{e_{21}\sqrt{\lambda_2}}{\sqrt{\sigma_{11}}}$$

$$\rho_{Y_2;X_1} = \frac{1/\sqrt{5}(1)}{\sqrt{5}} = 0{,}20$$

Hubungan antara komponen utama kedua dengan variabel pertama sebesar 0,20, artinya hubungannya positif tetapi lemah.

Kemudian, korelasi antara persamaan komponen utama kedua dengan variabel ke-2 adalah:

$$\rho_{Y_2;X_2} = \frac{e_{22}\sqrt{\lambda_2}}{\sqrt{\sigma_{22}}}$$

$$\rho_{Y_1;X_2} = \frac{-2/\sqrt{5}(\sqrt{1})}{\sqrt{2}} = -0{,}32$$

Hubungan antara komponen utama kedua dengan variabel kedua sebesar -0,32, artinya hubungannya negatif dan lemah.

11.4 Prosedur Analisis Komponen Utama (AKU) dengan R

Perintah yang dapat digunakan untuk melalukan AKU dalam software R adalah prcomp(). Berikut sintaknya.

```
prcomp(formula, data = NULL, subset, na.action, ...)
```

Keterangan :
`formula:` adalah formula untuk variabel numerik
`data:` matriks data
`subset:` vektor yang digunakan untuk memilih baris (pengamatan) dari data
`na.action:` fungsi yang mengindikasikan jika ada data hilang

Contoh

Pada praktikum ini, contoh kasus akan menggunakan data *mtcars* yang sudah tersedia dalam software R (lihat lagi contoh kasus pada praktikum anareg berganda multivariat). Dikarenakan data pada variabel "`vs`" dan "`am`" bersifat kategorikal, maka kedua variabel ini tidak digunakan. Berikut cara menginput datanya.

```
> Mobil=mtcars[,c(1 :7,10,11)]
> Mobil
                   mpg cyl  disp  hp  drat    wt  qsec gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90    4    2
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30    4    4
Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90    4    4
Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40    3    3
Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60    3    3
Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00    3    3
Cadillac Fleetwood 10.4  8 472.0 205 2.93 5.250 17.98    3    4
Lincoln Continental 10.4 8 460.0 215 3.00 5.424 17.82    3    4
Chrysler Imperial 14.7   8 440.0 230 3.23 5.345 17.42    3    4
Fiat 128          32.4    4 78.7  66 4.08 2.200 19.47    4    1
Honda Civic       30.4    4 75.7  52 4.93 1.615 18.52    4    2
Toyota Corolla    33.9    4 71.1  65 4.22 1.835 19.90    4    1
Toyota Corona     21.5   4 120.1  97 3.70 2.465 20.01    3    1
Dodge Challenger  15.5   8 318.0 150 2.76 3.520 16.87    3    2
AMC Javelin       15.2   8 304.0 150 3.15 3.435 17.30    3    2
Camaro Z28        13.3   8 350.0 245 3.73 3.840 15.41    3    4
Pontiac Firebird  19.2   8 400.0 175 3.08 3.845 17.05    3    2
Fiat X1-9         27.3    4 79.0  66 4.08 1.935 18.90    4    1
Porsche 914-2     26.0   4 120.3  91 4.43 2.140 16.70    5    2


Lotus Europa      30.4    4 95.1 113 3.77 1.513 16.90    5    2
Ford Pantera L    15.8   8 351.0 264 4.22 3.170 14.50    5    4
Ferrari Dino      19.7   6 145.0 175 3.62 2.770 15.50    5    6
Maserati Bora     15.0   8 301.0 335 3.54 3.570 14.60    5    8
Volvo 142E        21.4   4 121.0 109 4.11 2.780 18.60    4    2
```

Setelah data dimasukan, syarat melakukan analisis komponen utama adalah ada korelasi antarvariabel-variabel. Berikut cara melakuan identifikasi keberadaannya korelasi antarvariabel.

```
> korelasi=cor(Mobil) #menghitung korelasi
> round(korelasi,digits=2) #melakukan pembulatan 2 digit
       mpg   cyl  disp    hp  drat    wt  qsec  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00 -0.21 -0.66
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66  0.27  1.00
```

Di atas adalah matriks korelasi untuk ke-9 variabel. Korelasi tertinggi terjadi pada varibel "`disp`" dan "`cyl`" yaitu 0,90, sedangkan korelasi terendah terjadi antara variabel "`drat`" dan "`carb`". Secara keseluruh ada identifikasi korelasi antara variabel-variabel tersebut, sehingga analisis komponen utama dapat dilakukan.

Perintah untuk melakukan analisis KU adalah :

```
> PCA=prcomp(Mobil, center=TRUE, scale. = TRUE)
> summary(PCA)
Importance of components :
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     2.3782 1.4429 0.71008 0.51481 0.42797 0.35184
Proportion of Variance 0.6284 0.2313 0.05602 0.02945 0.02035 0.01375
Cumulative Proportion  0.6284 0.8598 0.91581 0.94525 0.96560 0.97936
                           PC7    PC8     PC9
Standard deviation     0.32413 0.2419 0.14896
Proportion of Variance 0.01167 0.0065 0.00247
Cumulative Proportion  0.99103 0.9975 1.00000
```

Output di atas menampilan nilai deviasi standar, proporsi variansi dan proporsi kumulatif masing-masing Komponen Utama (*Principal Component*). Standar deviasi diperoleh dari nilai eigen matriks kovarianasi variabel X.

Standar deviasi untuk KU pertama sebesar 2,3781, KU kedua sebear 1,4429 demikan seterusnya sampai standar deviasi terakhir (KU-9) sebesar

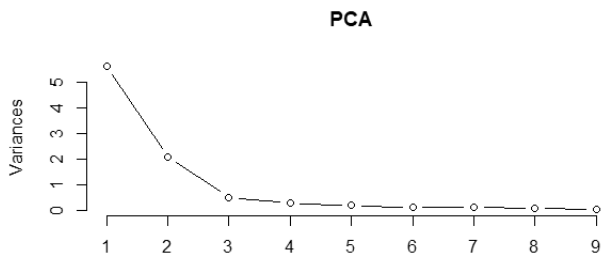
0,14896.

Terlihat dari output di atas, terbentuk 9 persamaan Komponen Utama (KU) karena ada 9 variabel. KU pertama menjelaskan (variansi) data sebesar 62,84%, KU kedua menjelaskan keragaman data sebesar 23,13% demikian seterusnya sampai KU kesembilan hanya menjelaskan sebesar 0,247% saja dari total variansi data.

Proporsi kumulatif KU-1 dan KU-2 sudah menjelaskan sebesar 85,98%, sehingga hanya dengan menggunakan dua Komponen Utama saja sudah mencukupi.

Untuk membuat dan menampilkan *scree plot* dilakukan dengan cara sebagai berikut.

```
> plot(PCA, type="l")
```



Dari *scree plot* tersebut terlihat bahwa kurva mulai landai pada titik ketiga, artinya bahwa dengan dua komponen saja sudah mencukupi untuk mewakili kesembilan variabel.

Selanjutnya, nilai loading persamaan KU dapat diperoleh melalui perintah berikut ini.

```
 > round(PCA$rotation, digits=3)
        PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9
mpg  -0.393  0.028 -0.221 -0.006 -0.321  0.720 -0.381 -0.125  0.115
cyl   0.403  0.016 -0.252  0.041  0.117  0.224 -0.159  0.810  0.163
disp  0.397 -0.089 -0.078  0.339 -0.487 -0.020 -0.182 -0.064 -0.662
hp    0.367  0.269 -0.017  0.068 -0.295  0.354  0.696 -0.166  0.252
drat -0.312  0.342  0.150  0.846  0.162 -0.015  0.048  0.135  0.038
wt    0.373 -0.172  0.454  0.191 -0.187 -0.084 -0.428 -0.198  0.569
qsec -0.224 -0.484  0.628 -0.030 -0.148  0.258  0.276  0.356 -0.169
gear -0.209  0.551  0.207 -0.282 -0.562 -0.323 -0.086  0.316  0.047
carb  0.245  0.484  0.464 -0.214  0.400  0.357 -0.206 -0.108 -0.320
```

Persamaan komponen utama 1 dan 2 dapat dituliskan seperti di bawah ini.

$$KU1 = -0{,}393\text{mpg} + 0{,}403\text{cyl} + 0{,}397\text{disp} + 0{,}367\text{hp} - 0{,}312\text{drat} + 0{,}373\text{wt} - 0{,}224\text{qsec} - 0{,}209\text{gear} + 0{,}245\text{car}$$

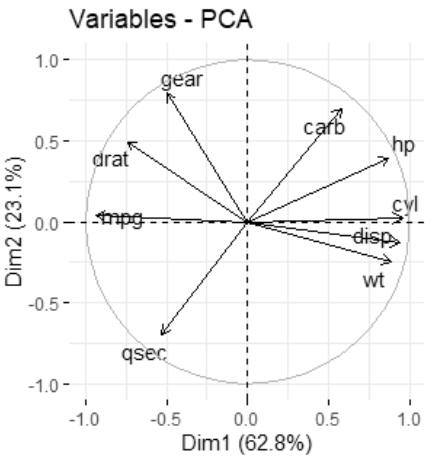
$$KU2 = 0{,}028\text{mpg} + 0{,}116\text{cyl} - 0{,}089\text{disp} + 0{,}269\text{hp} + 0{,}342\text{drat} - 0{,}172\text{wt} - 0{,}484\text{qsec} + 0{,}551\text{gear} + 0{,}484\text{car}$$

Variabel yang memberikan bobot tertinggi pada KU1 adalah variabel "cyl", "mpg" dan "disp" sedangkan variabel yang memberikan bobot tertinggi pada KU2 adalah variabel "qsec", "gear" dan "carb".

Analisis komponen utama telah mereduksi 9 variabel menjadi hanya 2 variabel (2 komponen utama) dimana proporsi variansi (keragaman) dari kedua KU telah menjelaskan sebesar 85,98% dan sebesar 14,02% dijelaskan oleh KU lainnya.

Apabila ingin melihat posisi setiap variabel terhadap KU1 dan KU2 dapat dilakukan dengan menggunakan perintah seperti di bawah ini.

```
> install.packages("factoextra")
> library(factoextra)
> fviz_pca_var(PCA)
```



Gambar di atas memperlihatkan biplot variabel-variabel dimana sumbu-x menunjukkan KU1 dan sumbu-y menunjukkan KU2. Terlihat garis panah menunjukkan arah masing-masing variabel. Koordinat masing-masing variabel sama dengan bobot/loading (koefisien) dari persamaan KU.

Kelebihan dari KU ini telah membuat suatu variabel baru yang tidak berkorelasi (lihat KU1 dan KU2 saling tegak lurus), sehingga hasil dari analisis KU dapat digunakan untuk mengatasi masalah kolinieritas pada analisis regresi.

Selanjutnya, apabila ingin menampilkan *score* untuk KU1 dan KU2 dapat diperoleh dengan mengetikkan perintah seperti di bawah ini.

```
> round(PCA$x[,1 :2], digits=3)
 PC1 PC2
Mazda RX4           -0.664    1.173
Mazda RX4 Wag       -0.637    0.977
Datsun 710          -2.300   -0.327
Hornet 4 Drive      -0.215   -1.977
Hornet Sportabout    1.587   -0.829
Valiant              0.050   -2.447
Duster 360           2.714    0.361
Merc 240D           -2.044   -0.801
Merc 230            -2.295   -1.306
Merc 280            -0.383    0.581
Merc 280C           -0.367    0.412
Merc 450SE           1.885   -0.724
Merc 450SL           1.671   -0.714
Merc 450SLC          1.777   -0.841
Cadillac Fleetwood   3.650   -0.948
Lincoln Continental  3.710   -0.843
Chrysler Imperial    3.332   -0.481
Fiat 128            -3.452   -0.433
Honda Civic         -3.855    0.708
Toyota Corolla      -3.855   -0.387
Toyota Corona       -1.904   -1.573
Dodge Challenger     1.804   -1.134
AMC Javelin          1.465   -0.978
Camaro Z28           2.601    0.765
Pontiac Firebird     1.874   -0.979
Fiat X1-9           -3.148   -0.255
Porsche 914-2       -2.779    1.637
Lotus Europa        -2.909    1.396
Ford Pantera L       1.548    3.021
Ferrari Dino         0.080    2.835
Maserati Bora        2.963    3.999
Volvo 142E          -1.904    0.108
```

11.4 Latihan

1. Suatu variabel acak dan mempunyai matriks variansi

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}$$

a. Cari vektor ciri dan akar ciri dari matriks!
b. Tentukan persamaan komponen utamanya!
c. Hitung proporsi keragaman masing-masing komponen utama yang terbentuk! Analisis hasilnya!
d. Hitung korelasi setiap komponen utama dengan setiap variabel, interpretasikan!

2. Suatu variabel acak X_1 dan X_2 mempunyai matriks variansi

$$\boldsymbol{\Sigma} = \begin{bmatrix} 9 & -2 \\ -2 & 6 \end{bmatrix}$$

a. Cari vektor ciri dan akar ciri dari matriks!
b. Tentukan persamaan komponen utamanya!
c. Hitung proporsi keragaman masing-masing komponen utama yang terbentuk! Analisis hasilnya!
d. Hitung korelasi setiap komponen utama dengan setiap variabel, interpretasikan!

3. Suatu variabel acak X_1 dan X_2 mempunyai matriks variansi

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}$$

a. Cari vektor ciri dan akar ciri dari matriks!
b. Tentukan persamaan komponen utamanya!
c. Hitung proporsi keragaman masing-masing komponen utama yang terbentuk! Analisis hasilnya!
d. Hitung korelasi setiap komponen utama dengan setiap variabel, interpretasikan!

4. Suatu penelitian ingin mengetahui tingkat kepuasan mahasiswa terhadap layanan perpustakaan. Adapun variabel penjelasnya yaitu:
X_1 = keramahan pustakawan,
X_2 = professional pustakawan,
X_3 = kecepatan layanan,
X_4 = jaringan internet/Wi-Fi yang stabil,

X_5 = kenyamanan ruang baca,
X_6 = kemudahan akses katalog online (OPAC);
X_7 = kelengkapan koleksi buku.

Hasil pengumpulan data terhadap 25 orang mahasiswa, disajikan pada tabel di bawah.

Responden	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	3,0	3,0	5,0	5,0	6,0	2,0	3,0
2	4,0	3,0	3,0	4,0	5,0	3,0	3,0
3	4,0	3,0	2,0	2,0	3,0	3,0	4,0
4	2,0	2,0	3,0	3,0	4,0	2,0	2,0
5	6,0	5,0	5,0	6,0	5,0	6,0	5,0
6	5,0	5,0	4,0	3,0	4,0	5,0	4,0
7	4,0	4,0	5,0	4,0	5,0	5,0	4,0
8	4,0	5,0	7,0	5,0	6,0	6,0	6,0
9	5,0	5,0	5,0	4,0	5,0	6,0	5,0
10	4,0	4,0	4,0	4,0	4,0	5,0	4,0
11	5,0	5,0	3,0	2,0	4,0	5,0	5,0
12	6,0	5,0	4,0	3,0	3,0	6,0	5,0
13	5,0	4,0	5,0	5,0	6,0	4,0	4,0
14	3,0	4,0	4,0	5,0	5,0	3,0	2,0
15	4,0	2,0	5,0	6,0	6,0	4,0	3,0
16	6,0	6,0	3,0	4,0	3,0	7,0	5,0
17	4,0	4,0	5,0	6,0	5,0	5,0	5,0
18	3,0	4,0	3,0	4,0	5,0	5,0	6,0
19	7,0	5,0	3,0	4,0	5,0	4,0	5,0
20	5,0	4,0	5,0	6,0	6,0	4,0	3,0
21	5,0	5,0	6,0	7,0	7,0	5,0	6,0
22	4,0	5,0	4,0	5,0	4,0	4,0	5,0
23	5,0	6,0	2,0	2,0	2,0	5,0	6,0
24	6,0	5,0	5,0	5,0	4,0	5,0	6,0
25	4,0	6,0	2,0	2,0	4,0	5,0	6,0

Berdasarkan data di atas, lakukanlah analisis komponen utama menggunakan *software* R kemudian jawablah pertanyaan seperti di bawah ini.

a. Berapa komponen utama yang terbentuk?
b. Tuliskan model komponen utama tersebut!
c. Berapa proporsi variansi yang dijelaskan oleh komponen utama tersebut?
d. Gambarkan biplot komponen utama tersebut!
e. Hitung *score* komponen utama yang terbentuk tersebut!

BAB

XII

ANALISIS FAKTOR (FACTOR ANALYSIS)

12.1 Pendahuluan

Dalam penelitian empiris yang melibatkan banyak variabel, sering dihadapkan pada permasalahan kompleksitas data dan keterkaitan yang tinggi antarvariabel. Banyak variabel yang diukur secara bersamaan sering kali mencerminkan konsep atau konstruk yang sama, sehingga informasi yang diperoleh menjadi tumpang tindih dan sulit diinterpretasikan secara langsung. Kondisi ini menuntut adanya suatu teknik analisis statistik yang mampu mengidentifikasi struktur laten yang mendasari hubungan antarvariabel tersebut, yaitu analisis faktor.

Analisis faktor diperkenalkan pertama kali oleh Spearman sekitar tahun 1927-1930 dan dikembangkan oleh Thurstone pada tahun 1935. Analisis faktor dan analisis komponen utama memiliki kemiripan yaitu kedua metode sama-sama mempelajari hubungan antarvariabel. Selain itu, tujuan kedua analisis yaitu membuat variabel baru yang dimensinya lebih rendah/kecil dibandingkan dengan variabel asal.

Perbedaan antara analisis faktor dan analisis komponen utama adalah :

1. Analisis komponen utama: membuat variabel baru (komponen utama) yang tidak berkorelasi dimana variabel tersebut dapat menjelaskan sebanyak mungkin keragaman (variansi) data asal.
2. Analisis faktor: mengidentifikasi suatu variabel baru yang tidak teramati (*unobservable*) dari hubungan antarvariabel dimana variabel baru tersebut dinamakan "*factor*".

Analisis faktor merupakan salah satu metode analisis multivariat yang bertujuan untuk menjelaskan hubungan antarvariabel teramati melalui sejumlah faktor laten yang jumlahnya lebih sedikit. Faktor-faktor laten ini tidak dapat diukur secara langsung, tetapi direpresentasikan oleh pola korelasi antarvariabel yang diamati. Dengan demikian, analisis faktor membantu peneliti dalam memahami dimensi dasar atau konstruk utama yang membentuk suatu fenomena kompleks. Contoh suatu penelitian ingin mengetahui faktor-faktor yang melandasi calon mahasiswa memilih perguruan tinggi. Variabel-variabel yang disertakan terdiri dari enam variabel yaitu X_1 = Kualitas PT; X_2 = Nasihat teman; X_3 = Informasi tentang PT; X_4 = Memperluas pergaulan; X_5 = Biaya kuliah dan X_6 = Waktu kuliah.

Hasil analisis menunjukkan bahwa faktor-faktor yang mempengaruhi calon mahasiswa memilih PT adalah *faktor akademik dan biaya kuliah* **(F_1)** dan *faktor non akademik* **(F_2).** Dengan menggunakan analisis faktor dapat mereduksi jumlah variabel menjadi lebih sedikit tanpa mengurangi informasi yang dikandung di dalam data.

12.2 Model Analisis Faktor

Analisis faktor merupakan salah satu teknik analisis multivariat yang termasuk kelompok independensi. Analisis ini bertujuan untuk mereduksi banyak variabel yang saling berkorelasi menjadi sedikit variabel baru yang tidak saling berkorelasi yang disebut faktor. Terdapat dua tujuan utama analisis faktor yaitu :

1. *Data Summarization*, yakni mengenali atau mengidentifikasi adanya hubungan antarvariabel dengan melakukan uji korelasi;
2. *Data Reduction*, yakni meringkas informasi yang terkandung dalam sejumlah variabel awal menjadi sebuah set faktor yang hanya terdiri dari beberapa faktor saja.

Misalkan $\boldsymbol{X} = (X_1, X_2, ..., X_p)'$ adalah variabel yang teramati (*observable variables*) dimana vektor rata-rata $\boldsymbol{\mu}$ dan matriks variansi-kovariansi $\boldsymbol{\Sigma}$. Dalam analisis faktor dinyatakan bahwa setiap variabel teramati dapat dijelaskan oleh beberapa variabel yang tidak teramati (*unobservable variables*) yaitu F_1, F_2, ..., F_m, yang disebut *common factors*. Selain itu, setiap variabel mengandung faktor *error*/spesifik $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$, yang disebut juga *specific factors*. Oleh karena itu, model analisis faktor dirumuskan seperti di bawah ini.

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_i - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p \end{aligned} \qquad (12.1)$$

dimana :

- l_{ij} adalah loading faktor (bobot faktor) ke-*j* pada pengamatan ke -*i*
- F_j adalah variabel yang tidak teramati atau disebut juga *common*

factor ke-*j* untuk $j = 1, 2, \ldots, m$; $m \leq p$.

- ε_j adalah variabel *error* atau disebut juga *spesific factor* ke-*j* untuk $j = 1, 2, \ldots, m$; $m \leq p$.

Nilai *loading* menunjukkan korelasi antara faktor umum yang terbentuk dengan variabel asal, semakin besar nilai loading maka semakin erat hubungan diantara keduanya.

Persamaan (12.1) dapat dituliskan dalam bentuk matriks bentuknya menjadi :

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_p \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & l_{13} & \cdots & l_{1m} \\ l_{21} & l_{22} & l_{23} & \cdots & l_{2m} \\ l_{31} & l_{32} & l_{33} & \cdots & l_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & l_{p3} & \cdots & l_{pm} \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ F_3 \\ \vdots \\ F_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_p \end{pmatrix} \qquad (12.2)$$

$$\boldsymbol{X}_{(p\times 1)} - \boldsymbol{\mu}_{(p\times 1)} = \boldsymbol{L}_{(p\times m)}\boldsymbol{F}_{(m\times 1)} + \boldsymbol{\varepsilon}_{(p\times 1)}$$

Keterangan :

$\boldsymbol{X}$ = vektor variabel asal

$\boldsymbol{\mu}$ = vektor rata-rata variabel asal

$\boldsymbol{L}$ = matriks *loading factor*

$\boldsymbol{F}$ = vektor faktor bersama (*common factor*)

ε = vektor faktor spesifik (*specific factor*)

Asumsi-asumsi yang harus dipenuhi dalam analisis faktor adalah :

1. Asumsi yang harus dipenuhi oleh variabel random $\boldsymbol{F}$ adalah
   $E(\boldsymbol{F}) = \boldsymbol{0}$; $Cov(\boldsymbol{F}) = \boldsymbol{I}$;
2. Asumsi yang harus dipenuhi oleh variabel random adalah
   $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$; $Cov(\boldsymbol{\varepsilon}) = \boldsymbol{\psi}$;
   dimana

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

3. Variabel $\boldsymbol{F}$ dan $\boldsymbol{\varepsilon}$ saling bebas, sehingga $Cov(\boldsymbol{\varepsilon}, \boldsymbol{F}) = \boldsymbol{0}$;

Dengan menggunakan asumi-asumsi tersebut, maka dapat diuraikan

persamaan seperti di bawah ini.

$$\begin{aligned}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})' &= (\boldsymbol{LF}+\boldsymbol{\varepsilon})(\boldsymbol{LF}+\boldsymbol{\varepsilon})' \\ &= \boldsymbol{LF}(\boldsymbol{LF})' + \boldsymbol{LF\varepsilon}' + \boldsymbol{\varepsilon}(\boldsymbol{LF})' + \boldsymbol{\varepsilon\varepsilon}'\end{aligned} \tag{12.3}$$

Berdasarkan asumsi di atas, diperoleh bahwa $(\boldsymbol{X}) = \boldsymbol{\Sigma}$; E($\boldsymbol{FF}$) = $\boldsymbol{I}$ dan $Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = \mathrm{E}(\boldsymbol{\varepsilon}\mathrm{F}) = \mathbf{0}$ maka:

$$\begin{aligned}\boldsymbol{\Sigma} = Cov(\boldsymbol{X}) &= E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})' \\ &= E(\boldsymbol{LF}(\boldsymbol{LF})' + \boldsymbol{LF\varepsilon}' + \boldsymbol{\varepsilon}(\boldsymbol{LF})' + \boldsymbol{\varepsilon\varepsilon}') \\ &= \boldsymbol{L}E(\boldsymbol{FF}')\boldsymbol{L}' + \boldsymbol{L}E(\boldsymbol{F\varepsilon}') + E(\boldsymbol{\varepsilon F}')\boldsymbol{L}' + E(\boldsymbol{\varepsilon\varepsilon}') \\ \boldsymbol{\Sigma} &= \boldsymbol{LL}' + \boldsymbol{\psi}\end{aligned} \tag{12.4}$$

Persamaan (12.4) dapat diuraikan menjadi :

$$\begin{aligned}\sigma_{ii} &= l_{i1} + l_{i2} + \cdots + l_{im} + \psi_i \quad ; i = 1,2,\ldots,p \\ Var(X_i) &= komunalitas + Spesifik\ variansi\end{aligned}$$

Persamaan di atas menjelaskan mengenai keragaman setiap variabel $Var(X_i)$ dapat diterangkan oleh sejumlah *m common factor* disebut communality dan faktor spesifik disebut uniqueness atau keragaman spesifik (spesific varians).

Selanjutnya, misalkan komunalitas ke-*i* dirumuskan seperti di bawah ini.

$$h_i = l_{i1} + l_{i2} + \ldots + l_{im}$$

sehingga :

$$\sigma_{ii} = h_i + \psi_i$$

12.3 Metode Estimasi Model Analisis Faktor

Ada beberapa metode estimasi dalam membuat analisis faktor yaitu metode *principal component* dan *maximum likelihood*. Pada buku ini, dibahas estimasi model analisis faktor dengan menggunakan komponen utama (*component principal*).

Metode komponen utama bertujuan untuk mengestimasi parameter

dalam model analisis faktor yaitu matriks loading, komunalitas $\boldsymbol{\psi}$ dan faktor spesifik $\boldsymbol{\varepsilon}$. Data yang digunakan dalam estimasi model analisis faktor bisa matriks variansi-kovariansi atau matriks korelasi.

Diberikan matriks sampel variansi-kovariansi $\boldsymbol{S}$ dari sampel acak teramati dari p variabel yaitu $\boldsymbol{X} = (X_1, X_2, ..., X_p)'$. Matriks sampel variansi-kovariansi $\boldsymbol{S}$ mempunyai akar ciri dan vektor ciri berpasangan yaitu $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), (\hat{\lambda}_2, \hat{\boldsymbol{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ dimana $\hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$, maka estimasi *matriks loading* adalah:

$$\hat{\boldsymbol{L}} = \left[\sqrt{\hat{\lambda}_1}\ \hat{\boldsymbol{e}}_1 \vdots \ \sqrt{\hat{\lambda}_2}\ \hat{\boldsymbol{e}}_2 \vdots \ \cdots \vdots \ \sqrt{\hat{\lambda}_m}\ \hat{\boldsymbol{e}}_m\right] \tag{12.5}$$

Estimasi komunalitas untuk variabel ke-i adalah :

$$\hat{h}_i = \hat{l}_{i1} + \hat{l}_{i2} + \cdots + \hat{l}_{im} \tag{12.6}$$

Kemudian, estimasi dari faktor spesifik untuk variabel ke-i adalah :

$$\hat{\psi}_i = s_{ii} - \hat{h}_i \tag{12.7}$$

sehingga diperoleh, matriks faktor spesifik berikut ini.

$$\hat{\boldsymbol{\psi}} = \begin{bmatrix} \hat{\psi}_1 & 0 & \cdots & 0 \\ 0 & \hat{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\psi}_p \end{bmatrix}$$

Contoh

Pada soal ini menggunakan data pada bab analisis komponen utama, misalkan terdapat variabel acak X_1 dan X_2 dengan matriks variansi-kovariansi sampelnya adalah:

$$\boldsymbol{S} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

Lakukanlah analisis faktor untuk data di atas!

Jawab

Dari matriks $\boldsymbol{S}$ diketahui $s_{11} = 5$; $s_{22} = 2$ dan $s_{11} = s_{11} = 2$;

Kemudian diketahui juga bahwa akar ciri dan vektor ciri pasangannya berturut – turut adalah $\hat{\lambda}_1 = 6; \hat{\boldsymbol{e}}_1 = (2/\sqrt{5} \quad 1/\sqrt{5})'$; dan $\hat{\lambda}_2 = 1; \hat{\boldsymbol{e}}_2 = (1/\sqrt{5} \quad -2/\sqrt{5})$;

Misalkan, jumlah faktor yang digunakan adalah 2 ($m = 2$), maka estimasi faktor loadingnya adalah :

$$\begin{aligned}\hat{\boldsymbol{L}} &= \left[\sqrt{\hat{\lambda}_1}\ \hat{\boldsymbol{e}}_1 \vdots \quad \sqrt{\hat{\lambda}_2}\ \hat{\boldsymbol{e}}_2\right] \\ &= \left[\sqrt{6}\begin{pmatrix}2/\sqrt{5}\\ 1/\sqrt{5}\end{pmatrix} \vdots \quad \sqrt{1}\begin{pmatrix}1/\sqrt{5}\\ -2/\sqrt{5}\end{pmatrix}\right] \\ &= \left[\begin{pmatrix}2{,}191\\ 1{,}095\end{pmatrix} \quad \begin{pmatrix}0{,}449\\ -0{,}894\end{pmatrix}\right]\end{aligned}$$

Estimasi komunalitas (faktor umum) setiap variabel adalah :

$$\hat{h}_1 = (2{,}191)^2 + (0{,}449)^2 = 5$$
$$\hat{h}_2 = (1{,}095)^2 + (-0{,}894)^2 = 2$$

Kemudian, estimasi dari faktor spesifik untuk variabel ke-i adalah :

$$\hat{\psi}_1 = s_{11} - \hat{h}_1 = 5 - 5 = 0$$
$$\hat{\psi}_2 = s_{22} - \hat{h}_2 = 2 - 2 = 0$$

Hasil estimasi di atas dapat disajikan dalam bentuk tabel seperti di bawah ini.

Tabel 22. Hasil Analisis Faktor dengan m = 2

Variabel	*Loading*		Komunalitas	Spesifik *Variance*
	F_1	F_2		
	2,191	0,449	5	0
	1,095	-0,894	5	0
Akar Ciri	6	1		

Nilai loading menyatakan proporsi setiap variabel acak ke-*i* terhadap faktor ke-*j*. Pada tabel di atas, terlihat proporsi variabel sebesar 2,191 terhadap , sedangkan proporsi variabel terhadap sebesar 0,449. Variabel memberikan kontribusi yang positif terhadap dan .

Selanjutnya, proporsi variabel X_2 sebesar 1,095 terhadap F_1, sedangkan proporsi variabel X_2 terhadap F_2 sebesar -0,894. Variabel X_2 memberikan kontribusi yang positif terhadap F_1 tetapi memberikan kontribusi negatif terhadap F_2.

Misalkan, jumlah faktor yang digunakan adalah 1 ($m = 1$), maka estimasi faktor loadingnya adalah :

$$\hat{\boldsymbol{L}} = \left[\sqrt{\hat{\lambda}_1}\ \hat{\boldsymbol{e}}_1\right] = \left[\sqrt{6}\begin{pmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix}\right]$$
$$= \left[\begin{pmatrix} 2{,}191 \\ 1{,}095 \end{pmatrix}\right]$$

Estimasi komunalitas (faktor umum) setiap variabel adalah :

$$\hat{h}_1 = (2{,}191)^2 = 4{,}80$$
$$\hat{h}_2 = (1{,}095)^2 = 1{,}20$$

Kemudian, estimasi dari faktor spesifik untuk variabel ke-i adalah :

$$\hat{\psi}_1 = s_{11} - \hat{h}_1 = 5 - 4{,}80 = 0{,}20$$
$$\hat{\psi}_2 = s_{22} - \hat{h}_2 = 2 - 1{,}20 = 0{,}80$$

Hasil estimasi di atas dapat disajikan dalam bentuk tabel seperti di bawah ini.

Tabel 23. Hasil analisis Faktor dengan m = 1

Variabel	*Loading*	Komunalitas	Spesifik Variansi
	2,191	4,80	0,20
	1,095	1,20	0,80
Akar Ciri	6		

12.4 Penentuan Jumlah Faktor Umum (*Common Factors*)

Analisis faktor digunakan untuk mencari faktor-faktor yang mampu menjelaskan hubungan atau korelasi antara berbagai variabel yang teramati. Variabel-variabel dalam satu faktor mempunyai korelasi yang tinggi sedangkan korelasi dengan faktor lain relative lebih rendah.

Dalam analisis faktor ada dua pendekatan yang digunakan yaitu analisis faktor eksplorasi (*Exploratory Factor Analysis/EFA*) dan analisis faktor konfirmatori (*Confirmatory Factor Analysis/CFA*).

Analisis faktor eksplorasi adalah teknik analisis faktor dimana jumlah faktor yang terbentuk belum dapat ditentukan sebelum analisisnya selesai dikerjakan. Teknik ini bisa digunakan, apabila seorang peneliti belum mempunyai pengetahuan/teori/hipotesa mengenai struktur faktor yang akan terbentuk. Oleh karena itu, analisis faktor eksploratori digunakan untuk membangun suatu teori baru.

Analisis faktor konfirmatori adalah teknik analisis faktor dimana jumlah faktor yang terbentuk sudah ditentukan sebelumnya. Penentuan jumlah faktor ini berdasarkan teori atau konsep yang sudah ada sebelumnya, sehingga tujuan analisis faktor konfirmatori adalah menguji sejauh mana data (variabel teramati) konsisten berada dalam konstruk (faktor) tersebut atau tidak.

Apabila jumlah faktor yang terbentuk belum ditentukan maka ada beberapa kriteria untuk menentukan sejumlah *m* faktor yaitu :

1. Kriteria nilai eigen (*eigen value*)
   Suatu eigen value menunjukkan besar sumbangan variansi dari faktor terhadap data. Suatu faktor dengan nilai eigen lebih besar dari 1 dimasukkan ke dalam model, sedangkan jika nilai eigen kurang dari 1 tidak dimasukan ke dalam model.
2. Kriteria proporsi keragaman (variansi) yang dijelaskan oleh faktor
   Penentuan banyak faktor ditentukan oleh persentase kumulatif variansi yang dijelaskan oleh sejumlah faktor. Jumlah faktor ditentukan jika kumulatif presentase variansi sudah mencapai paling sedikit 60% dari seluruh keragaman (variansi) variabel asli.
   Rumusnya proporsi keragaman faktor ke-*j* adalah

$$Proporsi\ keragaman\ F_j = \frac{\hat{\lambda}_j}{s_{11} + s_{22} + \cdots + s_{pp}} \qquad (12.8)$$

3. Kriteria *Scree plot*
   Penentuan jumlah faktor dapat dilihat dari grafik *scree plot*. *Scree*

plot adalah plot nilai *eigen* dari matriks . Apabila grafik *scree plot* terjadi penurunan yang sangat tajam kemudian diikuti oleh grafik yang mulai mendatar, maka pada titik tersebut menunjukkan banyaknya faktor yang dapt diekstraksi.

Contoh

Pada kasus sebelumnya, diperoleh persamaan faktor seperti di bawah ini.

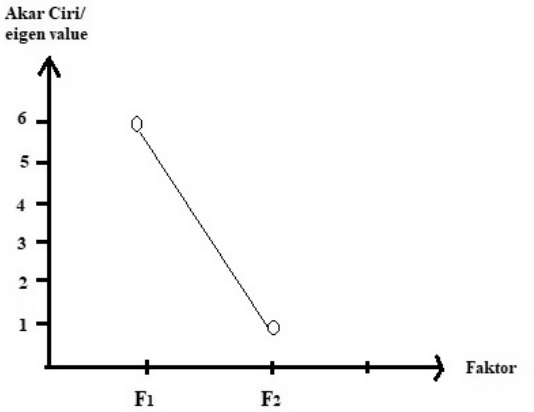
$$X_1 = 2{,}191F_1 + 0{,}449F_2$$
$$X_2 = 1{,}095F_1 - 0{,}894F_2$$

Dan $\hat{\lambda}_1 = 6$ dan $\hat{\lambda}_2 = 1$

1. Kriteria akar ciri (*eigen value*)
   Faktor pertama mempunyai nilai eigen lebih dari 1 sehingga dimasukan ke dalam model.
2. Kriteria proporsi keragaman
   Proporsi keraman setiap faktor adalah :

$$proporsi\ keragaman\ F_1 = \frac{6}{5+2} = 0{,}857$$
$$proporsi\ keragaman\ F_2 = \frac{1}{5+2} = 0{,}143$$

   Proporsi keragaman yang dijelaskan oleh faktor pertama = 0,857, sehingga menggunakan 1 faktor sudah dapat menjelaskan keragaman data sebesar 85,7%.
3. Kriteria *scree plot*
   Grafik *scree plot* diperlihatkan pada gambar di bawah ini.



Gambar 15. Gambar Scree Plot

Pada gambar di atas, terlihat pada titik grafik *scree plot* terjadi penurunan yang tajam, sehingga jumlah faktor yang dapat diekstraksi hanya 1 yaitu

Pada kasus ini, berdasarkan analisis dari ketiga kriteria maka dapat disimpulkan bahwa jumlah faktor adalah satu.

12.5 Prosedur Analisis Faktor dengan R

Software R menyediakan fasilitas untuk melakukan analisis faktor dengan perintah `factanal()` yang tersedia dalam *package* `stats`. Paket stats sudah tersedia di R sehingga tidak perlu diinstal. Untuk mengaktifkannya cukup diketik perintah berikut ini: `library(stat)`. Selanjutnya perintah `factanal()` dapat dilakukan, dimana sintaknya seperti di bawah ini.

```
factanal(x, factors, data = NULL, covmat = NULL, n.obs
= NA,
 subset, na.action, start = NULL,
 scores = c("none", "regression", "Bartlett"),
 rotation = "varimax", control = NULL, ...)
Keterangan :
X = matriks data
Factors = banyaknya faktor
Data = optional, diisi jika x adalah rumus/formula/
model
Covmat = matriks kovariansi
n.obs= banyaknya pengamatan
scores = score dari FA, pilihannya "none",
"regression" atau "Bartlett".
Rotation = metode rotasi matriks, defaultnya metode
"varimax".
```

Contoh

Pada kasus ini, menggunakan data pada contoh di atas yaitu data kabupaten beserta indikator pertumbuhan ekonomi. Ada enam variabel yang dipakai yaitu: X_1 = *Produk Domestic Bruto* (PDB) (Milyar); X_2 = Tingkat Pengangguran (%); X_3 = Inflasi (%); X_4 = Konsumsi (Milyar); X_5 = Investasi (Milyar) dan X_6 = Pendapatan Per Kapita (juta).

Panggil datanya dengan mengetikkan perintah seperti di bawah ini.

```
> X                                          #Memanggil data
    Kabupaten  X1  X2  X3  X4  X5  X6
1           A   6   5   6   3   3   4
2           B   2   3   2   4   5   4
3           C   6   3   6   4   2   3
4           D   3   7   4   5   2   7
5           E   2   4   2   2   7   4
6           F   6   4   6   3   3   4
7           G   5   3   6   3   3   4
8           H   7   2   7   4   2   4
9           I   2   7   2   3   7   3
10          J   3   5   3   6   4   6
11          K   2   5   2   3   5   3
12          L   5   4   5   4   2   4
13          M   2   3   2   5   4   4
14          N   4   6   4   6   3   6
15          O   6   5   4   2   1   4
16          P   3   5   4   6   5   7
17          Q   4   4   7   2   2   5
18          R   3   7   2   6   4   3
19          S   4   6   3   6   3   6
20          T   3   4   3   4   7   3
```

Pada kasus ini permasalahannya adalah apakah kabupaten berdasarkan 6 variabel bisa direduksi menjadi sejumlah faktor?

Analisis faktor didasari oleh adanya korelasi antara variabel-variabel yang digunakan. Oleh karena itu variabel yang digunakan merupakan variabel yang saling berkorelasi satu dengan yang lainnya. Setelah dilakukan analisis faktor maka akan terbentuk set variabel baru yang lebih sedikit dan tidak berkorelasi yang disebut dengan faktor.

Kriteria untuk mengukur interkorelasi antarvariabel adalah KMO (*Kaiser Meyer Olkin*). Berikut kriteria KMO.

Tabel 24. Kriteria KMO (Kaiser Meyer Olkin)

Ukuran KMO	Rekomendasi
≥ 0,90	Sangat baik (*marvelous*)
0,80 – 0,89	Berguna (*meritorious*)

0,70 – 0,79	Biasa (*middling*)
0,60 – 0,69	Cukup (*mediocre*)
0,50 – 0,59	Buruk (*miserable*)
≤ 0,49	Tidak terima (*unacceptable*)

Sedangkan untuk menilai kelayakan setiap variabel untuk dianalisis faktor digunakan kriteria *Measure of Sampling Adequacy* (MSA). Nilai MSA berkisar antara 0 sampai 1, dan berdasarakan nilai MSA yang didapat diambil kesimpulan sebagai berikut :

- Jika MSA ≥ 0,5 artinya variabel masih bisa diprediksi dan dianalisis lebih lanjut
- Jika MSA < 0,5 artinya variabel tidak bisa diprediksi dan harus dikeluarkan dari analisis.

Dalam R tersedia fasilitas untuk menghitung KMO dan MSA yaitu dengan perintah `KMO()` yang tersedia dalam `package psych`. Berikut langkah-langkahnya :

```
> install.packages("psych")
> library(psych)
```

Setelah diinstal paket, maka langkah selanjutnya menghitung matriks korelasi dari data yaitu :

```
> M=cor(X[,2 :7]) #membuat matriks korelasi
> round(M, digits=3) #matriks dibulatkan dengan 3
digit
> round(M, digits=3)
        X1      X2      X3      X4      X5      X6
X1   1.000  -0.344   0.847  -0.220  -0.733  -0.078
X2  -0.344   1.000  -0.418   0.336   0.131   0.306
X3   0.847  -0.418   1.000  -0.285  -0.676   0.089
X4  -0.220   0.336  -0.285   1.000  -0.014   0.520
X5  -0.733   0.131  -0.676  -0.014   1.000  -0.257
X6  -0.078   0.306   0.089   0.520  -0.257   1.000
```

Matriks korelasi di atas menjelaskan hubungan antara variabel. Korelasi antara X_1 dan X_2 sebesar -0,344 artinya hubungan negatif dan cukup kuat, korelasi antara X_1 dan X_3 sebesar 0,847 artinya hubungan positif dan sangat kuat demikian seterusnya.

Selanjutnya mencari ukuran KMO dan MSA melalui perintah seperti di bawah ini.

```
> KMO(M)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = M)
Overall MSA = 0.58
MSA for each item =
  X1    X2    X3    X4    X5    X6
0.58  0.71  0.59  0.52  0.75  0.33
```

Hasil ouput memperlihatkan nilai KMO sebesar 0,58 artinya termasuk dalam kategori kurang bagus (buruk). Hal ini menunjukkan bahwa ukuran sampel yang digunakan termasuk dalam kategori buruk dan data tidak dapat dianalisis lebih lanjut menggunakan analisis faktor.

Berdasarkan nilai MSA terlihat bahwa hanya variabel X1-X5 yang layak digunakan dalam analisis faktor sedangkan variabel X6 tidak layak dan harus dikeluarkan dari model. Oleh karena itu, variabel X6 dikeluarkan dari model kemudian dilakukan uji KMO dan MSA. Hasilnya seperti di bawah ini.

```
> MO=cor(X[,2 :6])
> round(MO, digits=3)
        X1      X2      X3      X4      X5
X1   1.000  -0.344   0.847  -0.220  -0.733
X2  -0.344   1.000  -0.418   0.336   0.131
X3   0.847  -0.418   1.000  -0.285  -0.676
X4  -0.220   0.336  -0.285   1.000  -0.014
X5  -0.733   0.131  -0.676  -0.014   1.000
> KMO(MO)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = MO)
Overall MSA = 0.73
MSA for each item =
  X1    X2    X3    X4    X5
0.71  0.76  0.73  0.63  0.75
```

Berdasarkan hasil di atas nilai KMO sebesar 0,738, artinya termasuk dalam kategori bagus, oleh karena itu data dapat dianalisis lebih lanjut menggunakan analisis faktor. Berdasarkan nilai MSA terlihat bahwa semua variabel X_1-X_5 di atas 0,60, sehingga semua variabel layak digunakan dalam analisis faktor.

Langkah selanjutnya adalah membuat analisis faktor. Untuk menentukan berapa banyak faktor yang harus dibentuk dapat dilihat dari kriteria nilai eigen. Nilai eigen > 1 menjadi patokan jumlah faktor.

R menyediakan fasilitas menghitung jumlah faktor melalui perintah eigen() yang ada dalam `package nFactors`. Sehingga perlu diinstal terlebih dahulu. Berikut mencari nilai eigen dari suatu matriks.

```
> install.packages("nFactors")
> library(nFactors)
> E
eigen() decomposition
$values
[1] 2.7561285 1.1953201 0.6452967 0.2574285 0.1458262

$vektors
           [,1]        [,2]        [,3]      [,4]        [,5]
[1,]  0.5580114 -0.15456552 -0.07427840 0.3485457  0.73330173
[2,] -0.3199204 -0.53214377 -0.75559460 0.2043394 -0.04238068
[3,]  0.5622611 -0.04930268 -0.03579794 0.4805123 -0.67026720
[4,] -0.2186644 -0.70186827  0.64448119 0.2096603 -0.01591786
[5,] -0.4715109  0.44484007  0.08319871 0.7496077  0.10469455
```

Perhatikan bahwa hanya ada dua nilai eigen yang nilainya lebih besar dari 1 yaitu 2,756 dan 1,195. Oleh karena itu, pada kasus ini banyaknya faktor adalah 2. Analisis faktor dapat dilakukan melalui perintah seperti di bawah ini.

```
> AF=factanal(X[,2 :6], factors=2, rotation
="varimax")
> AF

Call :
factanal(x = X[, 2 :6], factors = 2)

Uniquenesses :
   X1     X2     X3     X4     X5
0.149  0.640  0.137  0.664  0.231
```

```
Loadings :
   Factor1  Factor2
X1   0.863   -0.325
X2  -0.191    0.569
X3   0.810   -0.456
X4            0.578
X5  -0.874

                 Factor1  Factor2
SS loadings        2.202    0.976
Proportion Var     0.440    0.195
Cumulative Var     0.440    0.636

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.02 on 1 degree of freedom.
The p-value is 0.897
```

Perhatikan output di atas, nilai *uniqueness* setiap variabel berturut-turut adalah X_1 = 0,149; X_2 = 0,640, X_3=0,137, X_4 = 0,664 dan X_5=0,231. Nilai ini merupakan proporsi variansi (keragaman) yang tidak dapat dijelaskan oleh model analisis faktor. Semakin kecil nilainya semakin baik.

Model analisis faktor yang terbentuk adalah

X_1 = 0,683F1 – 0,325 F2

X_2 = -0,191F1+ 0,569F2

X_3 = 0,810F1 – 0,456F2

X_4 = 0,578 F2

X_5 = -0,874F1

Nilai loading mengidentifikasikan korelasi antara variabel dengan faktor yang terbentuk. Semakin tinggi nilai loading maka semakin erat hubungan variabel terhadap faktor. Hair (1998) dalam tulisannya menyatakan bahwa nilai minimal loading yang digunakan adalah lebih besar dari ± 0,30; loading ± 0,40 dianggap penting; dan loading ± 0,50 atau lebih besar dinyatakan signifikan.

Berdasarkan *output* terlihat bahwa F1 memiliki bobot *loading* yang besar dari variabel X_1 (*Produk Domestic Bruto*), X_3 (Inflasi) dan X_5 (Investasi).

Sedangkan F2 mempunyai bobot *loading* yang tinggi pada variabel X_2 (Tingkat Pengangguran) dan X_4 (Pendapatan Per Kapita).

Selanjutnya adalah menilai kelayakan model analisis faktor yang terbentuk. Salah satu kriteria adalah dengan melihat persentase variansi total yang dapat dijelaskan oleh banyaknya faktor yang terbentuk.

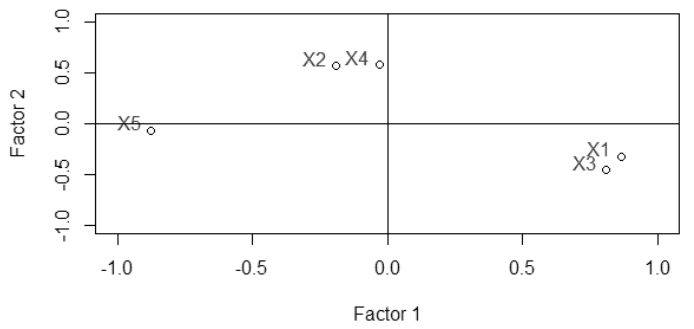
Penentuan jumlah faktor dilihat dari persentase kumulatif keragaman (variansi) yang bisa dijelaskan oleh faktor yang terbentuk. Dalam penelitian ilmu sosial batas total keragaman (variansi) yang digunakan minimal 60%.

Berdasarkan output, diketahui faktor pertama (F1) mampu menjelaskan 44% dari total variansi, faktor kedua (F2) mampu menjelaskan 19,5% dari total variansi, sehingga kedua faktor tersebut mampu menjelaskan sebesar 63,5% dari total variansi. Oleh karena itu, pembentukan dua faktor sudah cukup baik.

Langkah terakhir dari analisis faktor adalah menginterpretasikan faktor yang terbentuk. Pada kasus ini pemilihan PT berdasarkan lima variabel (X_1, X_2, X_3, X_4 dan X_5) dapat direduksi menjadi 2 faktor saja yaitu, **F1** adalah faktor ekonomi dan **F2** yaitu faktor pengangguran. Hal ini disebabkan F1 memiliki hubungan yang kuat dengan variabel X_1 (*Produk Domestic Bruto*), X_3 (Inflasi) dan X_5 (Investasi). Sedangkan F2 mempunyai hubungan yang kuat dengan variabel X_2 (Tingkat Pengangguran) dan X_4 (Pendapatan Per Kapita).

Penggambaran secara visual, bagaimana *scatter plot* setiap variabel terhadap faktor yang terbentuk.

```
plot(AF$loadings[,1], AF$loadings[,2], xlab = "Factor
1",
 ylab = "Factor 2", ylim = c(-1,1),xlim = c(-1,1),)

text(AF$loadings[,1]-0.08, AF$loadings[,2]+0.08,
 colnames(X[,2 :6]), col="blue")
abline(h = 0, v = 0)
```



Scatter plot di atas memperlihatkan kedudukan (posisi) setiap variabel pada masing-masing faktor. Gambar tersebut memperkuat interpretasi bahwa variabel X_1, X_3 dan X_5 berkontribusi besar dalam menyusun faktor pertama (F1) sedangkan variabel X_2 dan X_4 memberikan kontribusi yang besar terhadap faktor kedua (F2).

12.6 Latihan

1. Suatu variabel acak dan mempunyai matriks variansi:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 9 & -2 \\ -2 & 6 \end{bmatrix}$$

   a. Tentukan model analisis faktor!
   b. Hitung loading, komunalitas dan faktor spesifik!
   c. Interpetasikan nilai loadingnya!
   d. Berapa banyak faktor yang dapat diekstraksi? Gunakan kriteria

akar ciri, proporsi keragaman dan *scree plot*!

2. Suatu variabel acak dan mempunyai matriks variansi:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}$$

Berdasarkan matriks di atas, jawablah pertanyaa di bawah ini.

a. Tentukan model analisis faktor!
b. Hitung loading, komunalitas dan faktor spesifik!
c. Interpetasikan nilai loadingnya!
d. Berapa banyak faktor yang dapat diekstraksi? Gunakan kriteria akar ciri, proporsi keragaman dan *scree plot*!

3. Suatu penelitian ingin melakukan analisis faktor dari tiga variabel yaitu X_1, X_2 dan X_3 Hasil perhitungan akar ciri dan vektor ciri diperoleh nilai berikut ini

$$\hat{\lambda}_1 = 1{,}96 \text{ dan } \hat{\boldsymbol{e}}_1 = [0{,}625 \quad 0{,}593 \quad 0{,}507]'$$
$$\hat{\lambda}_2 = 0{,}68 \text{ dan } \hat{\boldsymbol{e}}_2 = [-0{,}219 \quad -0{,}491 \quad 0{,}843]'$$
$$\hat{\lambda}_3 = 0{,}36 \text{ dan } \hat{\boldsymbol{e}}_3 = [0{,}749 \quad -0{,}638 \quad -0{,}177]'$$

Kemudian, diketahui $s_{11} = s_{22} = s_{33} = 1$
Berdasarkan data di atas,

a. Buatlah analisis faktor dengan $m = 1$! Analisis dan interpretasikan!
b. Buatlah analisis faktor dengan $m = 2$! Analisis dan interpretasikan!
c. Bandingkan hasil analisis pada soal 3a dan 3b!

4. Seorang peneliti ingin melihat faktor-faktor yang mempengaruhi tingkat kepuasan masiswa terhadap pembelajaran e-learning. Tingkat kepuasan diukur oleh 6 variabel yaitu X_1 = pelayanan dosen, X_2 = pengetahuan pemakai, X_3 = produk system, X_4 = kemudahan penggunaan, X_5 = aksesibilitas aplikasi dan X_6 = karakteristik aplikasi. Berdasarkan teori/hipotesa, jumlah faktor sebanyak 2. Hasil pengolahan data dengan $m = 2$ disajikan pada tabel seperti di bawah ini.

Variabel	Loading	
	F_1	F_2

X_1	0,602	0,200
X_2	0,467	0,154
X_3	0,926	0,143
X_4	1,000	0,000
X_5	0,874	0,476
X_6	0,894	0,327

Berdasarkan tabel di atas :

a. Tuliskan estimasi model analisis faktor!
b. Hitung komunalitas (*common factors*) masing-masing variabel!
c. Hitung spesifik variansi (*spesific factors*) masing-masing variabel! Jika diasumsikan

5. Berikut ini adalah matriks korelasi data tentang Olympic Decathlon dari tahun 1960 sampai 2004 (Johnson and Wichern, p. 499) :

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1,000	0,639	0,475	0,323	0,552	0,326	0,351	0,401	0,182	-0,035
X_2	0,639	1,000	0,495	0,567	0,471	0,352	0,400	0,517	0,310	0,101
X_3	0,475	0,495	1,000	0,436	0,254	0,281	0,793	0,473	0,468	-0,012
X_4	0,323	0,567	0,436	1,000	0,345	0,350	0,366	0,604	0,234	0,238
X_5	0,552	0,471	0,254	0,345	1,000	0,155	0,210	0,421	0,212	0,413
X_6	0,326	0,352	0,281	0,350	0,155	1,000	0,255	0,416	0,171	0,000
X_7	0,351	0,400	0,793	0,366	0,210	0,255	1,000	0,404	0,418	0,011
X_8	0,401	0,517	0,473	0,604	0,421	0,416	0,404	1,000	0,315	0,240
X_9	0,182	0,310	0,468	0,234	0,212	0,171	0,418	0,315	1,000	0,098
X_{10}	-0,035	0,101	-0,012	0,238	0,413	0,000	0,011	0,240	0,098	1,000

Decathlon adalah jenis permainan yang terdiri dari 10 cabang olaharaga yaitu lari 100m (X_1), lompat jauh (X_2), tembakan (X_3), lompat tinggi (X_4), lari 400m (X_5), rintangan 100m (X_6), lempar cakram (X_7), lompat galah (X_8), lempar lembing (X_9), dan lari 1500m (X_{10}).

Berdasarkan matriks korelasi di atas, lakukan analisisi faktor dengan software R. Jawablah pertanyaan berikut ini.

a. Hitung KMO dan MSA, analisis!

b. Lakukan analisis faktor dengan banyaknya faktor 2!
c. Berapa total variansi yang disumbangkan oleh dua faktor tersebut?
d. Tuliskan hasil model analisis faktornya!
e. Interpretasikan kedua faktor tersebut!

6. Sebuah perguruan tinggi ingin mengukur kepuasan mahasiswa terhadap layanan akademik dan fasilitas kampus. Terdapat 10 indikator kepuasan (X1–X10) yang diduga membentuk beberapa faktor laten. Untuk tujuan tersebut, dilakukan survei terhadap 50 mahasiswa dengan menggunakan kuesioner skala Likert (1 = sangat tidak puas sampai 5 = sangat puas). Hasil survey disajikan pada tabel dibawah ini :

Responden	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	1	5	1	3	5	5	4	5	2	5
2	1	5	1	1	3	1	1	3	3	2
3	1	1	3	3	2	4	4	4	1	5
4	1	1	5	3	3	4	1	2	2	2
5	1	4	4	2	4	3	5	1	4	5
6	1	1	1	3	2	3	1	5	5	3
7	1	3	3	3	1	1	2	5	1	5
8	1	2	2	2	2	2	5	2	5	3
9	1	2	1	5	2	1	1	4	1	4
10	1	5	4	4	2	5	5	1	5	5
11	1	1	3	2	5	2	4	2	4	3
12	1	1	5	4	1	5	1	4	4	5
13	1	3	5	4	2	1	2	5	5	4
14	1	4	3	1	4	1	5	3	4	3
15	2	1	4	2	3	5	4	1	3	1
16	2	4	5	1	4	5	2	2	5	1
17	2	4	3	3	3	3	4	1	2	3
18	2	4	3	4	5	2	4	3	3	5
19	2	1	4	2	3	3	4	1	3	5
20	2	2	3	2	5	2	5	5	4	4

Responden	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
21	2	2	3	1	5	1	3	4	2	4
22	2	5	3	2	1	5	2	4	3	5
23	2	4	2	3	1	4	4	3	5	2
24	2	5	5	3	1	3	1	1	4	2
25	2	5	1	1	1	2	1	1	1	1
26	3	2	3	1	3	5	1	4	1	3
27	3	3	5	3	3	5	3	5	4	3
28	3	4	3	4	1	4	3	5	3	4
29	3	3	3	4	1	3	2	5	5	1
30	3	5	3	1	2	3	2	5	2	4
31	3	3	2	3	4	3	1	1	2	3
32	3	4	3	5	2	3	3	2	1	4
33	3	5	5	2	3	4	3	3	1	1
34	4	2	3	3	4	3	3	1	5	2
35	4	1	1	2	2	3	3	5	5	3
36	4	5	5	5	2	2	3	2	5	3
37	4	2	5	2	4	3	4	1	4	5
38	4	5	4	5	1	1	5	5	5	2
39	4	4	1	5	5	1	3	1	2	2
40	4	5	5	2	5	2	5	4	3	4
41	4	5	1	1	4	4	3	2	1	2
42	5	1	3	5	5	3	5	2	1	3
43	5	3	2	1	2	2	1	5	1	4
44	5	5	5	2	3	5	5	3	5	3
45	5	5	2	2	5	5	1	4	1	3
46	5	3	2	5	1	1	4	2	1	2
47	5	4	5	1	2	2	3	2	3	2
48	5	4	2	5	5	4	4	2	1	2
49	5	2	2	1	5	2	2	3	4	5
50	5	3	2	3	1	5	5	5	5	3

Keterangan :

X_1: Kejelasan informasi akademik

X_2: Kemudahan akses layanan akademik

X_3: Kecepatan pelayanan administrasi

X_4: Keramahan staf administrasi
X_5: Ketersediaan fasilitas ruang kuliah
X_6: Kenyamanan ruang kuliah
X_7: Kualitas fasilitas teknologi informasi
X_8: Akses internet kampus
X_9: Kualitas layanan perpustakaan
X_{10}: Ketersediaan bahan pustaka

Berdasarkan data di atas, lakukan analisis faktor dengan software R!

BAB

XIII

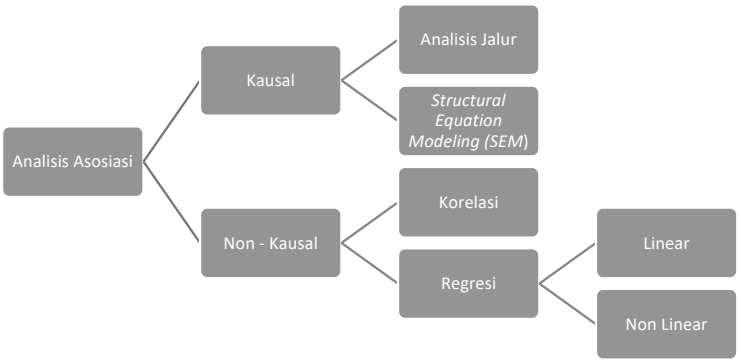
ANALISIS JALUR (*PATH ANALYSIS*)

13.1 Pendahuluan

Analisis jalur (*path analysis*) pertama kali dikembangkan oleh Sewall Wright, seorang ahli genetika dan ahli statistik asal Amerika Serikat, pada awal abad ke-2. Wright mengembangkan metode ini untuk menjelaskan hubungan sebab-akibat antarvariabel biologis yang saling berkaitan secara kompleks. Konsep utama yang diperkenalkan adalah penggunaan diagram jalur (*path diagram*) dan koefisien jalur (*path coefficient*) untuk menggambarkan serta mengukur pengaruh langsung dan tidak langsung antarvariabel dalam suatu sistem kausal.

Analisis jalur adalah salah satu teknik dalam analisis asosiasi (hubungan). Analisis ini merupakan perluasan dari model regresi linier ganda yaitu menganalisis hubungan sebab akibat yang terjadi pada regresi berganda jika variabel independen (bebas) mempengaruhi variabel dependen secara langsung maupun tidak langsung.

Gambar di bawah ini menjelaskan peta kedudukan analisis jalur dalam analisis asosiasi.



Gambar 16. Peta Alur dalam Analisis Asosiasi

Seiring perkembangannya, analisis jalur mulai diadopsi secara luas di luar bidang genetika, khususnya dalam ilmu sosial, ekonomi, psikologi, pendidikan, dan manajemen. Pada dekade 1950-an hingga 1970-an, metode ini berkembang pesat seiring dengan kemajuan analisis regresi dan ketersediaan perangkat komputasi statistik. Para peneliti memanfaatkan

analisis jalur untuk menguji model hubungan kausal yang didasarkan pada teori, dengan asumsi bahwa arah hubungan antarvariabel telah ditentukan sebelumnya.

Selanjutnya, analisis jalur menjadi dasar bagi pengembangan metode yang lebih kompleks, yaitu *Structural Equation Modeling* (SEM). SEM memperluas analisis jalur dengan mengakomodasi variabel laten dan kesalahan pengukuran secara eksplisit. Meskipun demikian, analisis jalur tetap relevan dan banyak digunakan dalam penelitian empiris yang melibatkan variabel terukur secara langsung. Hingga saat ini, analisis jalur masih menjadi metode analisis yang penting dalam penelitian kuantitatif untuk memahami struktur hubungan kausal dan mekanisme pengaruh antarvariabel secara sistematis.

13.2 Model Analisis Jalur

Analisis jalur merupakan sebuah metode statistik yang dapat digunakan untuk mengidentifikasi dan menguji hubungan kausal antara berbagai variabel dalam suatu model. Metode ini bertujuan untuk memahami sejauh mana variabel-variabel tersebut saling memengaruhi dan bagaimana pengaruh tersebut terjadi melalui jalur atau lintasan yang saling terhubung.

Analisis jalur mempunyai kedekatan dengan analisis regresi berganda atau bisa dikatakan regresi berganda adalah bentuk khusus dari analisis jalur. Teknik analisis jalur disebut juga sebagai model sebab akibat (*causal model*). Hal ini didasarkan bahwa analisis jalur memungkinkan pengguna melakukan pengujian teoritis terhadap hubungan sebab dan akibat terhadap variabel-variabelnya.

Teknik analisis jalur digunakan untuk menguji besarnya sumbangan (kontribusi) yang ditunjukkan oleh koefisien jalur pada setiap diagram jalur dari hubungan kausal antarvariabel X_1, X_2, dan X_3 terhadap *Y* terhadap *Z*.

Oleh sebab itu, rumusan masalah penelitian dalam kerangka *path analysis* berkisar pada: a). Apakah variabel eksogen (X_1, X_2, …, X_k) berpengaruh terhadap variabel endogen Y? dan b). Berapa besar pengaruh kausal langsung, kausal tidak langsung, kausal total, maupun simultan seperangkat variabel eksogen (X_1, X_2, …, X_k) terhadap variabel endogen Y?

Tujuan analisis jalur adalah :

1. Untuk melihat hubungan antara variabel-variabel yang didasarkan pada suatu teori;
2. Untuk menjelaskan hubungan (korelasi) variabel dengan menggunakan suatu model
3. Menguji suatu model matematis dengan menggunakan persamaan yang mendasarinya;
4. Mengidentifikasi jalur penyebab suatu variabel tertentu terhadap variabel lain yang diperngaruhinya;
5. Menghitung besarnya pengaruh variabel independen exogenous terhadap variabel dependen *endogenousous*.

Berikut perbedaan antara analisis regresi dengan analisis jalur seperti di bawah ini.

Tabel 25. Perbedaan Analisis Jalur dan Analisis Regresi

	Analisis Regresi	Analisis Jalur
Tujuan	Memprediksi nilai Y jika X_1, X_2, …, X_k diketahui	- Menganalisis pola hubungan kausal antara variabel Y dan X - Melihat pengaruh langsung ataupun tidak langsung
Jenis variabel	- Variabel Independen/ Bebas (X) - Variabel Dependen/ Terikat (Y)	- Variabel Penyebab (eksogenous) - Variabel Akibat (endogenous)
Rumusan Masalah	- Apakah variabel X_1, X_2,…, X_k berpengaruh terhadap Y? - Berapa besar variasi perubahan Y yang dapat dijelaskan oleh variabel bebas (X)?	- Apakah variabel X_1, X_2,…, X_k berpengaruh langsung atau tidak langsung terhadap Y? - Berapa besar pengaruh langsung dan tidak langsung variabel bebas tersebut?
Input Data	Matriks data dalam skala interval dan dalam bentuk data mentah.	Matriks dalam skala interval dan dalam bentuk skor baku.

Asumsi-asumsi dari analisis jalur hampir sama denga analisis regresi berganda yaitu :

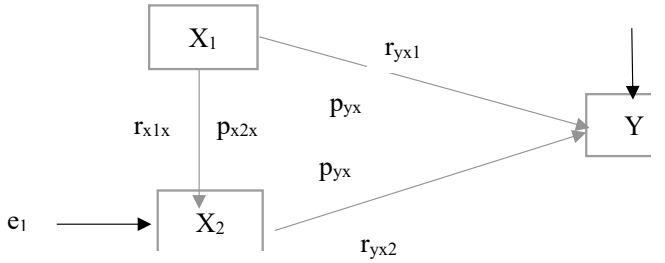
1. Linearitas artinya hubungan antara variabel bersifat linear.
2. *Error* berdistribusi normal dan homogen.

3. Tidak terjadi multikolinieritas antara variabel *X*.
4. Tidak ada autokorelasi (residual bersifat independen).
5. Tidak ada arah yang bersifat rekursif (satu arah).
6. Model yang diuji memiliki kerangka teoritis yang kuat.

Persyaratan tambahan yang harus dipenuhi saat kita akan menggunakan *path analysis* yaitu :

1. Data metrik berskala interval,
2. Terdapat variabel independen exogenous dan dependen endogenousous untuk model regresi berganda dan variabel perantara untuk model mediasi dan model gabungan mediasi dan regresi berganda serta model kompleks,
3. Ukuran sampel yang memadai, sebaiknya di atas 100 dan idealnya 400-1000
4. Pola hubungan antarvariabel: pola hubungan antarvariabel hanya satu arah tidak boleh ada hubungan timbal balik (r*eciprocal*),
5. Hubungan sebab akibat didasarkan pada teori yang sudah ada dengan asumsi sebelumnya menyatakan bahwa memang terdapat hubungan sebab akibat dalam variabel-variabel yang sedang kita teliti,
6. Pertimbangkan hal-hal yang sudah dibahas dalam asumsi dan prinsip-prinsip dasar di bab sebelumnya.

Pemodelan analisis jalur biasanya ditampilkan dalam bentuk diagram jalur (*path diagram*), yang menggambarkan hubungan antara variabel-variabel dengan menggunakan panah atau jalur yang menghubungkan satu variabel ke variabel lainnya. Panah menunjukkan arah pengaruh kausal yang menggambarkan hubungan kausalitas antara variabel *eksogenous*, *endogenous,* dan *intervening*.



Keterangan :

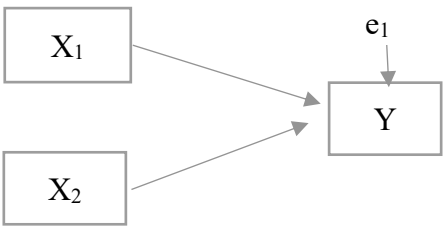
- X_1 dan X_2 adalah variabel eksogenous
- Y adalah variabel endogenous
- p_{ij} adalah koefisien jalur antaravariabel endogenous ke-i dengan variabel eksogenous ke-j. Koefisien jalur menunjukkan besarnya pengaruh langsung variabel eksogenous terhadap variabel endogenous.
- r_{ij} adalah korelasi antara variabel ke-*i* dengan variabel ke- *j*.
- e_i adalah faktor *error* (galat) ke-*i*

Menurut Sarwono (2010), model-model yang ada dalam analisis jalur terdiri dari :

1. Model Regresi Linear Berganda

Model regresi linear berganda merupakan model yang menganalisis hubungan variabel bebas/independen (lebih dari 1) terhadap variabel dependen/terikat.

Misalnya, suatu penelitian ingin mengetahui pengaruh Remunerasi (X_1) dan Motivasi Kerja (X_2) terhadap Kinerja Pegawai (Y). Diagram jalurnya seperti di bawah ini.



Persamaan: $Y = \beta_1 X_1 + \beta_2 X_2 + e_1$

Keterangan :

X_1 dan X_2 Adalah variabel eksogenous

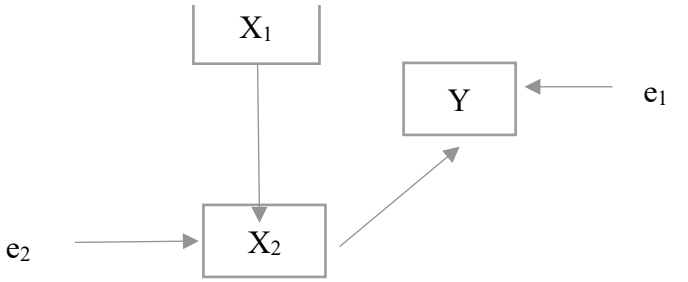
Y = variabel endogenous

2. Model Mediasi

Model kedua dari analisis jalur adalah model mediasi. Model ini terjadi apabila variabel eksogenous tidak langsung mempengaruhi terhadap variabel endogenous tetapi melalui variabel perantara atau *intervening variable*.

Contoh

Suatu penelitian ingin mengetahui pengaruh Remunerasi (X_1) dan Motivasi Kerja (X_2) terhadap Kinerja Pegawai (Y). Pada kasus ini, pengaruh Remunerasi tidak langsung mempengaruhi variabel Kinerja Pegawai tetapi berpengaruh langsung terhadap Motivasi Kerja. Pada kasus ini variabel Motivasi Kerja (X_2) sebagai variabel perantara. Diagram path dapat digambarkan seperti di bawah ini.



Persamaan struktural adalah :

$$Y = p_{YX_2}X_2 + e_1$$

$$X_2 = p_{X_2X_1}X_1 + e_2$$

Keterangan :

X_1 = variabel eksogenous

X_2 = variabel perantara sekaligus variabel endogenous bagi X_1

Y = variabel endogenous

p_{YX_2} = koefisien jalur menunjukkan pengaruh X_2 terhadap Y

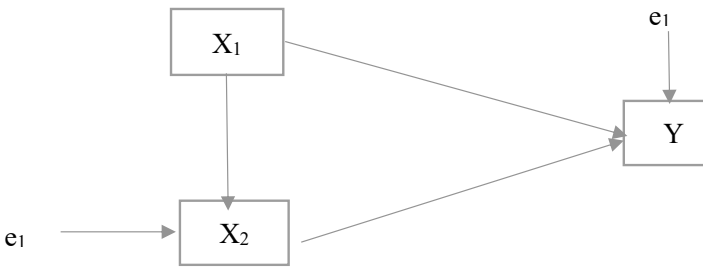
$p_{X_2X_1}$= koefisien jalur menunjukkan pengaruh dari X_1terhadap X_2

3. Model Gabungan antara regresi berganda dengan mediasi

Model gabungan antara regresi berganda dengan model mediasi terjadi apabila ada variabel eksogenousous yang berpengaruh langsung dan tidak langsung terhadap variabel endogenousous.

Contohnya: Penelitian mengenai pengaruh Remunerasi (X_1) dan Motivasi Kerja (X_2) terhadap Kinerja Pegawai (Y). Dalam hal ini, secara teori managemen bahwa pemberian Remunerasi berpengaruh langsung terhadapa kinerja pegawai serta berpengaruh terhadap motivasi kerja.

Bentuk diagram jalurnya adalah



Persamaan strukturalnya adalah :

$$Y = p_{YX_1}X_1 + p_{YX_2}X_2 + e_1$$
$$X_2 = p_{X_2X_1}X_1 + e_2$$

Keterangan :

X_1 = variabel eksogenous
X_2 = variabel perantara sekaligus variabel endogenous bagi X_1
Y = variabel endogenous
p_{YX_1} = koefisien jalur menunjukkan pengaruh X_1 terhadap Y
p_{YX_2} = koefisien jalur menunjukkan pengaruh X_2 terhadap Y
$p_{X_2X_1}$ = koefisien jalur menunjukkan pengaruh dari X_1terhadap X_2

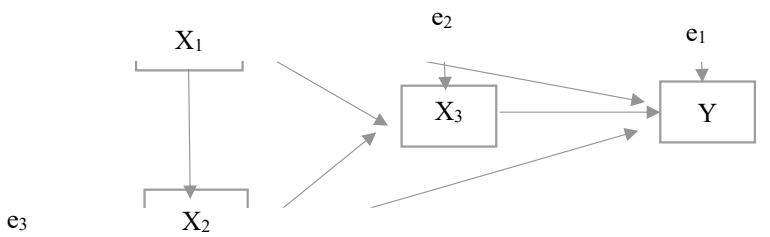
4. Model Kompleks

Model keempat dari model analisis jalur adalah model kompleks. Model

ini pengembangan dari model ketiga dimana pola hubungan antara variabel lebih rumit (kompleks).

Contoh

Pada penelitian mengenai pengaruh Remunerasi (X_1) dan Motivasi Kerja (X_2) terhadap Kinerja Pegawai (Y). Dalam kasus ini ternyata ada variabel lain yang terlibat yaitu Suasana Kerja (X_3). Pola hubungannya digambarkan seperti di bawah ini.



Persamaan struktural adalah :

$$Y = p_{YX_1}X_1 + p_{YX_2}X_2 + p_{YX_3}X_3 + e_1$$

$$X_3 = p_{X_3X_1}X_1 + p_{X_3X_2}X_2 + e_2$$

$$X_2 = p_{X_2X_1}X_1 + e_3$$

Keterangan :

X_1 = variabel eksogenous

X_2 = variabel perantara sekaligus variabel endogenous untuk X1)

X_3 = variabel perantara sekaligus variabel endogenous untuk X1 dan X_2)

Y = variabel endogenous

p_{YX_1} = koefisien jalur menunjukkan pengaruh X_1 terhadap Y

p_{YX_2} = koefisien jalur menunjukkan pengaruh X_2 terhadap Y

p_{YX_3} = koefisien jalur menunjukkan pengaruh X_3 terhadap Y

$p_{X_3X_1}$ = koefisien jalur menunjukkan pengaruh dari X_1 terhadap X_3

$p_{X_3X_2}$ = koefisien jalur menunjukkan pengaruh dari X_2 terhadap X_3

$p_{X_2X_1}$ = koefisien jalur menunjukkan pengaruh dari X_1terhadap X_2

Berdasarkan gambar di atas dapat dijelaskan bahwa variabel Remunerasi berpengaruh langsung terhadap Kinerja Pegawai serta pengaruh tidak langsung melalui variabel Motivasi Kerja dan Suasana Kerja. Kemudian, variabel Motivasi Kerja berpengaruh langsung terhadap Kinerja Pegawai dan berpengaruh tidak langsun melalui variabel Suasana Kerja. Selain itu, Motivasi Kerja berpengaruh lanngsung terhadap Kinerja Pegawai. Pada kasus ini, ada dua variabel perantara yaitu Suasana Kerja dan Motivasi Kerja.

13.3 Tahapan Analisis jalur

Berikut prosedur dalam melakukan analisis jalur adalah :

1. Merancang model sesuai dengan teori.
   Langkah pertama adalah menyusun model teoretis yang jelas, berdasarkan literatur atau kerangka pemikiran yang ada. Model ini harus menggambarkan hipotesis mengenai hubungan antarvariabel, termasuk variabel mediasi jika ada.
2. Menyusun hipotesis penelitian
3. Membuat diagram jalur
   Pemodelan analisis jalur biasanya ditampilkan dalam bentuk diagram jalur (*path diagram*) yang menggambarkan hubungan antara variabel-variabel dengan menggunakan panah atau jalur yang menghubungkan satu variabel ke variabel lainnya.
4. Melakukan estimasi parameter model
   Mengestimasi koefisien jalur yang menghubungkan variabel-variabel dalam model. Proses ini melibatkan perhitungan regresi berganda dan analisis hubungan kausal berdasarkan data yang telah dikumpulkan.
   Berikut Langkah-langkah mengestimasi koefisien jalur
   a. Menyusun matriks korelasi antara variabel eksogenousous

$$R_{\boldsymbol{X}} = \begin{bmatrix} 1 & r_{x_1x_2} & \cdots & r_{x_1x_k} \\ r_{x_2x_1} & 1 & \cdots & r_{x_2x_k} \\ \cdots & \cdots & \ddots & \cdots \\ r_{x_kx_1} & r_{x_kx_2} & \cdots & 1 \end{bmatrix}$$

   Vektor korelasi antara variabel eksogenouseous (X) dengan

variabel endogenousous (Y) yaitu :

$$R_{YX} = \begin{bmatrix} r_{yx_1} \\ r_{yx_2} \\ \vdots \\ r_{yx_k} \end{bmatrix}$$

b. Mencari invers matriks korelasi eksogenousousnya yaitu :

$$R_{XX}^{-1} = \begin{bmatrix} 1 & r_{x_1x_2} & \cdots & r_{x_1y} \\ r_{x_2x_1} & 1 & \cdots & r_{x_2y} \\ \cdots & \cdots & \ddots & \cdots \\ r_{yx_1} & r_{yx_2} & \cdots & 1 \end{bmatrix}^{-1}$$

notasi invers matriks korelasi adalah :

$$C = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1k} \\ C_{21} & C_{22} & \cdots & C_{2k} \\ \cdots & \cdots & \ddots & \cdots \\ C_{k1} & C_{k2} & \cdots & C_{kk} \end{bmatrix}$$

c. Menentukan koefisien jalur dengan menggunakan rumus seperti di bawah ini.

$$P_{YX} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1k} \\ C_{21} & C_{22} & \cdots & C_{2k} \\ \cdots & \cdots & \ddots & \cdots \\ C_{k1} & C_{k2} & \cdots & C_{kk} \end{bmatrix} \begin{bmatrix} r_{yx_1} \\ r_{yx_2} \\ \vdots \\ r_{yx_k} \end{bmatrix}$$

Jika sudah diperoleh estimasi koefisien jalur, maka lakukan uji parsial dengan hipotesis seperti di bawah ini.

$H_0: P_{ij} = 0$ (Tidak ada pengaruh variabel eksogenous ke-*j* terhadap endogenous ke-*i*)

$H_1: P_{ij} > 0$ (Ada pengaruh variabel eksogenous ke-*j* terhadap endogenous ke-*i*)

Statistik uji :

$$t = \frac{P_{YX_i}}{\sqrt{\frac{(1 - R^2_{(Y,X1,X2,\ldots,Xk)})C_{ii}}{n-k-1}}}$$

Keterangan :

k = banyaknya variabel eksogenous

Keputusan: H_0 ditolak apabila $t_{hit} > t_{\alpha;(n-k-1)}$

5. Melakukan uji kesesuaian model (*goodness of fit*)

Menguji sejauh mana model tersebut cocok dengan data. Beberapa indeks kecocokan model seperti *Chi-square, CFI* (*Comparative Fit Index*), *RMSEA* (*Root Mean Square error of Approximation*) digunakan untuk menilai kesesuaian model.

a. *Chi-square*

Chi-square digunakan untuk menguji seberapa dekat kecocokan antara matriks kovarian sampel S dengan matriks kovarian model Uji statistik adalah:

$$\chi^2 = (n-1)F(S,\Sigma)$$

dimana :

$$F(S,\Sigma) = \text{tr}(S\Sigma^{-1}) - (p+q) + \ln\Sigma - S$$

Nilai X^2 yang kecil menunjukkan *good fit* (kecocokan yang baik). Kriteria yang dapat digunakan adalah adalah *p-value*. Apabila nilai p_value 0,05 menunjukkan bahwa data empiris yang diperoleh memiliki perbedaan dengan teori yang telah dibangun. Sedangkan nilai *p-value* tidak signifikan (> 0,05) menunjukkan bahwa data empiris sesuai dengan model. Oleh karena itu diperoleh kesimpulan hipotesis diterima jika nilai *p-value* yang diharapkan lebih besar daripada 0,05.

b. *Goodness of Fit Index* (GFI)

GFI dapat dijadikan sebagai ukuran kecocokan mutlak karena pada dasarnya GFI membandingkan model yang dihipotesiskan dengan tidak ada model sama sekali Rumus dari GFI adalah

seperti di bawah ini.

$$GFI = 1 - \frac{\hat{F}}{F_0}$$

Keterangan :

$\hat{F}$ = Nilai minimum dari F untuk model yang dihipotesiskan

F_0 = Nilai minimum dari F, ketika tidak ada model yang dihipotesiskan

Nilai GFI berkisar antara 0 sampai 1. Apabila nilai GFI 0,90 maka model sudah cocok atau fit (*good fit*) sedangkan jika nilanya diantara 0,80 GFI 0,90 sering disebut marginal fit.

c. *Root Mean Square error of Approximation* (*RMSEA*)

Kriteria kecocokan model yang ketiga adalah RMSEA (*Root Mean Square error of Approximation*). Rumus perhitungan RMSEA adalah seperti di bawah ini.

$$RMSEA = \sqrt{\frac{\hat{F}_0}{df}}$$

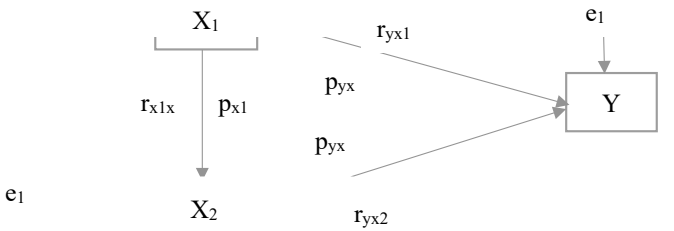
Semakin kecil nilai RMSEA maka semakin baik (cocok) modelnya. Menurut Brown dan Cudeck (1993) sebagaimana dikutip oleh Wijanto (2008), nilai 0,05 RMSEA 0,08 menunjukkan *good fit*.

6. Interpretasi Hasil

Tahap ini melakukan evaluasi signifikansi koefisien jalur, mengidentifikasi jalur mediasi, dan memahami pengaruh langsung serta tidak langsung antarvariabel. Hasil ini kemudian dapat digunakan untuk mengonfirmasi atau merevisi hipotesis awal.

Contoh

Penelitian mengenai pengaruh Remunerasi (X_1) dan Motivasi Kerja (X_2) terhadap Kinerja Pegawai (Y).



Hasil pengamatan terhadap responden sebanyak n = 20 diperoleh korelasi setiap variabel sebagai berikut :

Korelasi	X_1	X_2	Y
X_1	1,000	0,951	0,969
X_2	0,951	1,000	0,970
Y	0,969	0,970	1,000

Jawab

Hipotesis penelitian

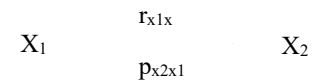
Ada pengaruh langsung Remunerasi (X_1) dan Motivasi Kerja (X_2) terhadap Kinerja Pegawai (Y) baik secara parsial maupun gabungan.

Persamaan strukturalnya :

$$X_2 = p_{X_2X_1}X_1 + e_2$$
$$Y = p_{YX_1}X_1 + p_{YX_2}X_2 + e_1$$

Estimasi Koefisien Jalur :

1. Persamaan struktural: $X_2 = p_{X_2X_1}X_1 + e_2$



Pada persamaan ini, variabel X_2 adalah variabel endogenousous karena dipengaruhi oleh variabel X_1. Koefisien jalurnya adalah

$$r_{x_1x_2} = p_{x_2x_1} = 0{,}951$$

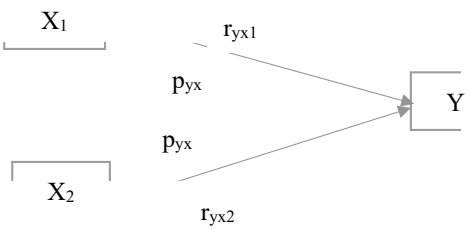
maka persamaan struktural adalah:

$$X_2 = 0{,}951X_1 + e_1$$

Koefisien determinasi dapat diperoleh dengan cara: $R^2_{x_2x_1} = 0{,}951^2 = 0{,}9044$. Kemudian pengaruh lain yang tidak dimasukan ke dalam model (*error*) terhadap variabel X_2 adalah:

$$e_1 = 1 - R^2_{x_2x_1} = 1 - 0{,}9044 = 0{,}0956$$

2. Selanjutnya pada persamaan struktural kedua yaitu



Mencari koefisien jalur p_{yx_1} dan p_{yx_1} dapat dicari melalui langkah-langkah berikut ini.

a. Tentukan matriks korelasi variabel eksogenousous :

$$R_{XX} = \begin{bmatrix} 1 & 0{,}951 \\ 0{,}951 & 1 \end{bmatrix}$$

Korelasi antara variabel eksogenous dan endogenous yaitu :

$$R_{YX} = \begin{bmatrix} 0{,}969 \\ 0{,}970 \end{bmatrix}$$

b. Cari invers matrik korelasi dari yaitu :

$$C = \begin{bmatrix} 10{,}4603 & -9{,}9478 \\ -9{,}9478 & 10{,}4603 \end{bmatrix}$$

c. Menghitung koefisien jalur adalah :

$$\begin{bmatrix} p_{yx_1} \\ p_{yx_2} \end{bmatrix} = \begin{bmatrix} 10{,}4603 & -9{,}9478 \\ -9{,}9478 & 10{,}4603 \end{bmatrix} \begin{bmatrix} 0{,}969 \\ 0{,}970 \end{bmatrix} = \begin{bmatrix} 0{,}4867 \\ 0{,}5071 \end{bmatrix}$$

sehingga diperoleh p_{yx_1} = 0,4867 dan p_{yx_1} = 0,5071. Persamaan struktruralnya dapat dituliskan seperti di bawah ini.

$$Y = 0{,}4867X_1 + 0{,}5071X_2 + e_2$$

Koefisien determinasi adalah :

$$R^2_{y.x_2x_1} = r_{yx_1}p_{yx_1} + r_{yx_2}p_{yx_2}$$
$$= 0{,}969(0{,}4867) + 0{,}970(0{,}5071) = 0{,}9635$$

Kemudian faktor *error*-nya adalah :

$$e_2 = 1 - R^2_{y.x_2x_1} = 1 - 0{,}9635 = 0{,}0365$$

Hasil estimasi model jalur diperoleh :

$$X_2 = 0{,}951X_1 + e_1; \text{ dengan } R^2_{x_2x_1} = 0{,}9044$$
$$Y = 0{,}4867X_1 + 0{,}5071X_2 + e_2; \text{ dengan } R^2_{y.x_2x_1} = 0{,}9635$$

Langkah berikutnya dari analisis jalur adalah melakukan uji signifikansi model yaitu melalui uji *t*.

a. Uji koefisien jalur X_1 ke X_2 ($p_{x_2x_1}$)

   Hipotesisnya adalah :

   H_0: $P_{X_2X_1}$ = (Tidak ada pengaruh X_2 terhadap X_1)

   H_1: $P_{X_2X_1}$ > (Ada pengaruh X_2 terhadap X_1)

   Statistik uji :

$$t_{hit} = \frac{0{,}951}{\sqrt{\frac{0{,}0956}{20-1-1}}} = 13{,}049$$

   Sedangkan: $t_{\alpha;(n-k-1)} = t_{0{,}05;(18)=1{,}73}$

   Keputusan: H_0 ditolak karena $t_{hit} > t_{\alpha;(n-k-1)}$, artinya ada pengaruh variabel X_1 terhadap X_2.

b. Uji koefisien jalur X_1 dan X_2 terhadap Y yaitu (p_{yx_1} *dan* p_{yx_2})

   Hipotesis untuk pengaruh X_1 terhadap Y yaitu

   H_0: p_{yx_1} = (Tidak ada pengaruh X_1 terhadap Y)

   H_1: p_{yx_1} > (Ada pengaruh X_1 terhadap Y)

Statistik uji :

Diketahui: $k = 2$, $C_{11} = 10{,}4603$, $p_{yx_1} = 0{,}4867$, $R^2_{y.x_2x_1} = 0{,}9635$

$$t_{hit} = \frac{0{,}4867}{\sqrt{\frac{(1-0{,}9635)10{,}4603}{20-2-1}}} = 3{,}2476$$

Sedangkan: $t_{\alpha;(n-k-1)} = t_{0{,}05;(17)=1{,}74}$

Keputusan: H_0 ditolak karena $t_{hit} > t_{\alpha;(n-k-1)}$, artinya ada pengaruh variabel X_1 terhadap Y.

Untuk menguji pengaruh X_2 terhadap Y dilakukan prosedur berikut ini.

H_0: $P_{YX_2} = 0$ (Tidak ada pengaruh X_2 terhadap Y)

H_1: $P_{YX_2} > 0$ (Ada pengaruh X_2 terhadap Y)

Statistik uji :

Diketahui: $k = 2$, $C_{22} = 10{,}4603$, $p_{yx_2} = 0{,}5071$; $R^2_{y.x_2x_1} = 0{,}9635$

$$t_{hit} = \frac{0{,}5071}{\sqrt{\frac{(1-0{,}9635)10{,}4603}{20-2-1}}} = 3{,}3838$$

Sedangkan: $t_{\alpha;(n-k-1)} = t_{0{,}05;(17)=1{,}74}$

Keputusan: H_0 ditolak karena $t_{hit} > t_{\alpha;(n-k-1)}$, artinya ada pengaruh variabel X_2 terhadap Y.

13.4 Prosedur Analisis Jalur dengan R

Dalam perangkat lunak R, analisis jalur (*path analysis*) dapat dilakukan dengan menggunakan *package lavaan*, yang merupakan salah satu paket statistik yang banyak digunakan untuk analisis *Structural Equation Modeling* (SEM), termasuk analisis jalur. Package ini menyediakan fungsi utama `sem()` yang digunakan untuk mengestimasi model hubungan kausal antarvariabel yang telah dirumuskan berdasarkan kerangka teori.

Secara umum, model analisis jalur dalam *lavaan* dituliskan dalam bentuk persamaan struktural menggunakan sintaks khusus, kemudian diestimasi dengan memanggil fungsi `sem()`. Adapun struktur umum penggunaan fungsi `sem()` dalam `package lavaan` adalah sebagai berikut :

```
sem(model = NULL, data = NULL, ordered = NULL, sampling.
weights = NULL,
  sample.cov = NULL, sample.mean = NULL, sample.th = NULL,
  sample.nobs = NULL,r group = NULL, cluster = NULL,
  constraints = "", WLS.V = NULL, NACOV = NULL, ov.order = "model",

Keterangan
- model = A description of the user-specified model
- data = An optional data frame containing the observed
  variables used in the model
- ordered = Character vektor. Only used if the data is in a
  data.frame
```

```
- sampling.weights = A variable name in the data frame
  containing sampling weight information.
- sample.cov = A sample variance-covariance matrix
- sample.mean sample.mean = A sample mean vektor.
- sample.th vektor of sample-based thresholds.
- sample.nobs = Number of observations if the full data frame
  is missing and only sample moments are given.
- Group = A variable name in the data frame defining the groups
  in a multiple group analysis.
- Cluster = A (single) variable name in the data frame defining
  the clusters in a two-level dataset
- Constraints = Additional (in)equality constraints not yet
  included in the model syntax
- WLS.V = A user provided weight matrix to be used by
  estimator "WLS"; if the estimator is "DWLS", only the
  diagonal of this matrix will be used.
- NACOV = A user provided matrix containing the elements of
  (N times) the asymptotic variance-covariance matrix of the
  sample statistics
- ov.order = If "model" (the default), the order of the
  observed variable names (as reflected for example in the
  output of lavNames()) is determined by the model syntax.
```

Contoh

Seorang peneliti ingin menganalisis faktor-faktor yang memengaruhi kinerja pegawai. Variabel yang digunakan dalam penelitian ini adalah Remuneras (X_1), Motivasi (X_2), dan Suasana Kerja (X_3) sebagai variabel eksogen, kepuasan kerja (Z) sebagai variabel intervening, serta kinerja pegawai (Y) sebagai variabel endogen. Untuk menjawab penelitian ini, maka diambil sampel karyawan sebanyak 199 orang.

Data diambil melalui survei yang dibagikan pada karyawan di suatu Perusahaan. Responden memberikan penilaian untuk variabel Remunerasi, Motivasi, Suasana Kerja, Kepuasan dan Kinerja Karyawan dengan skala likert 1-5 dengan ketentuan yaitu 1 = sangat tidak setuju (*strongly disagree*) s/d 5 = sangat setuju (*strongly agree*). Sedangkan variabel *behavior*, responden memberikan nilai 1 = tidak pernah (*never*) s/d 5 = sering (*all of the time*).

- X_1 = Variabel remunerasi mencerminkan penilaian/persepsi karyawan terhadap remun yang diterima
- X_2 = Variabel motivasi kerja mencerminkan perasaan karyawan terhadap motivasi kerja
- X_3 = Variabel suasana kerja mencerminkan penilaian/persepsi karyawan terhadap suasana kerja
- Z = Variabel kepuasan mencerminkan penilaian/persepsi karyawan terhadap Tingkat kepuasaan (variabel perantara/*intervening variables*).
- *Y* = Variabel kinerja karyawan mencerminkan persepsi karyawan tentang kinerja

Datanya dapat diakses pada link berikut ini: (https://drive.google.com/file/d/1zkIKUaGWBZAz0LmHZTR7lB6DRpId7wfk/view?usp=sharing)

Tabel di bawa ini merupakan ringkasan saja :

Tabel 26. Data Kinerja Karyawan

No	Remunerasi	Motivasi	Suasana	Kepuasan	Kinerja
1	2.31	2.31	2.03	2.5	2.62
2	4.66	4.01	3.63	3.99	3.64
3	3.85	3.56	4.2	4.35	3.83
…					
…					
196	2.29	1.43	1.93	2.01	1.31
197	4.13	3.34	3.45	3.21	2.68
198	1.39	2.64	2.9	1.81	4.21
199	4.28	4.29	4.16	4.62	3.41

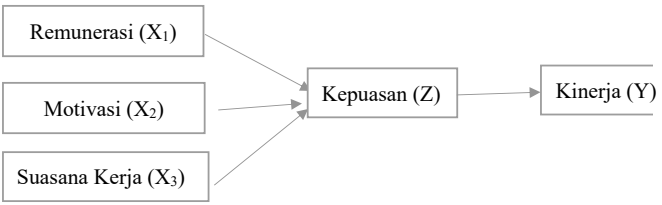
Pada penelitian ini hipotesisi adalah sebagai berikut

Berdasarkan informasi tersebut :

1. Buatlah diagram jalur (*path diagram*) yang menggambarkan hubungan antarvariabel.
2. Tuliskan persamaan struktural yang sesuai dengan diagram jalur tersebut.
3. Jelaskan pengaruh langsung dan tidak langsung masing-masing variabel eksogen terhadap kinerja pegawai

Jawab:

1. Pada kasus ini, diagram jalur sebagai berikut :



Gambar 17. Peta Jalur antara Remunerasi, Motivasi, Suasana Kerja, Kepuasan dan Kinerja Karyawan

2. Pada kasus ini model analisis jalur yang diestimasi yaitu :

$$Z = p_{ZX_1}X_1 + p_{ZX_2}X_2 + p_{ZX_3}X_3 + e$$
$$Y = p_iZ + e$$

Untuk mengestimasi model analisis jalur (*path analysis*) digunakan fungsi sem () yang tersedia dalam *package* `lavaan` pada perangkat lunak R. Prosedur estimasi dilakukan melalui beberapa tahapan, yang meliputi pendefinisian model struktural, pemanggilan fungsi `sem()`, serta evaluasi hasil estimasi parameter. Langkah-langkah pelaksanaan analisis jalur menggunakan *package* `lavaan` dijelaskan secara sistematis sebagai berikut.

```
>install.packages("lavaan")
>library(lavaan)

> # Membuat model analisi jalur
> Manajemen=read.csv2("Kinerja.csv")
> View(Manajemen)
> specmod = "
+ Kepuasan ~ Remunerasi + Motivasi + Suasana
+ Kinerja ~ Kepuasan
+ "
>
> # Mengestimate model analisis jalur
> fitmod <- sem(specmod, # name of specified model object
+ data=Manajemen) # name of data frame object
>
> # Menampilkan hasil ringkasan model analiss jalur
> summary(fitmod,          # name of fitted model object
+     fit.measures=TRUE,   # request model fit indices
+     rsquare=TRUE)        # request R-squared
estimates
lavaan 0.6-20 ended normally after 1 iteration

 Estimator                                         ML
 Optimization method                           NLMINB
 Number of model parameters                         6

 Number of observations                           199
Model Test User Model :

 Test statistic                                 2.482
 Degrees of freedom                                 3
 P-value (Chi-square)                           0.479
```

```
Parameter Estimates :

 Standard errors Standard
 Information Expected
 Information saturated (h1) model Structured

Regressions :
 Estimate Std.Err z-value P(>|z|)
 Kepuasan ~
 Remunerasi 0.335 0.060 5.594 0.000
 Motivasi 0.129 0.061 2.105 0.035
 Suasana 0.287 0.059 4.882 0.000
 Kinerja ~
 Kepuasan 0.474 0.070 6.739 0.000

Model Test Baseline Model :

 Test statistic                                129.338
 Degrees of freedom                                  7
 P-value                                         0.000

User Model versus Baseline Model :

 Comparative Fit Index (CFI)                     1.000
 Tucker-Lewis Index (TLI)                        1.010

Loglikelihood and Information Criteria :

 Loglikelihood user model (H0)              -1993.280
 Loglikelihood unrestricted model (H1)             NA

 Akaike (AIC)                                3998.560
 Bayesian (BIC)                              4018.320
 Sample-size adjusted Bayesian (SABIC)       3999.312

Root Mean Square Error of Approximation :

 RMSEA                                           0.000
 90 Percent confidence interval - lower          0.000
 90 Percent confidence interval - upper          0.111
 P-value H_0: RMSEA <= 0.050                     0.665
 P-value H_0: RMSEA >= 0.080                     0.162
```

```
Standardized Root Mean Square Residual :

 SRMR                                                0.021

Variances :
 Estimate Std.Err z-value P(>|z|)
 .Kepuasan 1065.259 106.793 9.975 0.000
 .Kinerja 1613.626 161.768 9.975 0.000

R-Square :
 Estimate
 Kepuasan 0.351
Kinerja 0.186
```

Pada output di atas, dapat dijelaskan seperti di bawah ini.

1. Uji kebaikan model (*goodness of fit*) menggunakan kriteria seperti di bawah ini.

Tabel 26. Kriteria Goodness of FIt

No	Kriteria	Cutoff kesesuian model
1	Chi-Square ()	P-value ≥ 0,05
2	CFI	≥ 0,90
3	TLI	≥ 0,95
4	RMSEA	≤ 0,08
5	SMSR	≤ 0,08

   a. Chi-Square
      Nilai = 2,203 dengan *p_value* sebesar 0,568. Nilai *chi-square* yang kecil dan nilai *p_value* ≤ 0,05 menunjukkan bahwa data empiris yang diperoleh memiliki perbedaan dengan teori yang telah dibangun. Hal ini menunjukkan *good fit* (kecocokan yang baik).
   b. *The comparative fit index* (CFI)
      Nilai CFI sebesar 1,000 yang menunjukkan nilai lebih besar dari *cutoff* (0,90). Hal ini menunjukkan *good fit* (kecocokan yang baik).
   c. *The Tucker-Lewis index* (TLI)
      Nilai LTI sebesar 1,019 yang menunjukkan nilai lebih besar dari *cutoff* (0,90). Hal ini menunjukkan *good fit* (kecocokan yang baik).

d. *The root mean square error of approximation* (RMSEA)
Nilai RMSEA sebesar 0,000 < *cutoff* 0,08. Hal ini memberikan bukti yang signifikan bahwa model sudah cocok.

e. *The standardized root mean square residual* (SRMR)
Nilai SRMS sebesar 0,019 < *cutoff* 0,08. Hal ini memberikan bukti yang signifikan bahwa model sudah cocok.

Hasil uji kesesuaian model menunjukkan bahwa model yang dibangun memiliki kecocokan yang baik dengan data. Hal ini ditunjukkan oleh nilai statistik Chi-square sebesar 2,482 dengan derajat bebas 3 dan nilai p-value sebesar 0,479 (> 0,05), sehingga tidak terdapat perbedaan yang signifikan antara matriks kovariansi sampel dan matriks kovariansi model.

Indeks goodness-of-fit lainnya juga mendukung kesesuaian model. Nilai CFI = 1,000 dan TLI = 1,010 menunjukkan tingkat kecocokan yang sangat baik. Selain itu, nilai RMSEA = 0,000 dengan p-value untuk hipotesis RMSEA ≤ 0,05 sebesar 0,665 serta SRMR = 0,021 mengindikasikan bahwa model dapat diterima secara empiris.

2. Uji Parsial dengan hipotesis :

H_0: $P_{ij} = 0$ (Tidak ada pengaruh variabel eksogenous ke-*j* terhadap endogenous ke-*i*)

H_1: $P_{ij} > 0$ (Ada pengaruh variabel eksogenous ke-*j* terhadap var endogenous ke-*i*)

a. Pengaruh Kepuasan (Z) terhadap Kinerja (Y).
Hasil estimasi koefisien jalur antara Z dan Y (p_{YZ} = 0,474) dengan *p_value* sebesar 0,00, artinya ada pengaruh positif yang signifikan antara Kepuasan dengan Kinerja.

b. Pengaruh Remunerasi (X_1) terhadap Kepuasaan (Z)
Hasil estimasi koefisien jalur antara X_1 dan Z (p_{ZX_1} = 0,335) dengan *p_value* sebesar 0,000, artinya ada pengaruh positif yang signifikan Remunerasi terhadap Kepuasan.

c. Pengaruh Motivasi (X_2) terhadap Kepuasan (Z)
Hasil estimasi koefisien jalur antara X_1 dan Z (p_{ZX_2} = 0,129) dengan *p_value* sebesar 0,035, artinya ada pengaruh positif

yang signifikan Motivasi terhadapa Kepuasan.

d. Pengaruh Suasana Kerja (X_3) terhadap Kepuasan (Z)
Hasil estimasi koefisien jalur antara X_1 dan Z (p_{ZX_3} = 0,287) dengan *p_value* sebesar 0,000, artinya ada pengaruh positif yang signifikan Suasana Kerja terhadap Kepuasan.

Analisis Hubungan Antarvariabel

Hasil estimasi parameter struktural menunjukkan bahwa remunerasi, motivasi, dan suasana kerja berpengaruh signifikan terhadap kepuasan kerja. Remunerasi memiliki pengaruh positif paling besar terhadap kepuasan kerja (β = 0,335; p < 0,001), diikuti oleh suasana kerja (β = 0,287; p < 0,001), dan motivasi kerja (β = 0,129; p = 0,035). Hal ini mengindikasikan bahwa peningkatan pada ketiga faktor tersebut akan meningkatkan tingkat kepuasan kerja pegawai.

Selanjutnya, kepuasan kerja terbukti berpengaruh positif dan signifikan terhadap kinerja pegawai dengan koefisien sebesar 0,474 dan p-value < 0,001. Temuan ini menunjukkan bahwa kepuasan kerja berperan penting sebagai variabel intervening yang menjembatani pengaruh faktor-faktor manajerial terhadap kinerja. Hasil estimasi model Analisis Jalur untuk kasus ini dapat dirumuskan sebagai berikut:

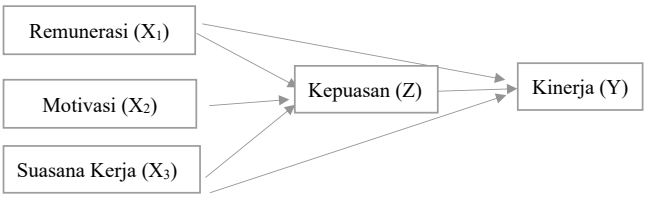
$$Z = 0{,}335X_1 + 0{,}129X_2 + 0{,}287X_3 + e$$
$$Y = 0{,}474Z + e$$

3. Koefisien Determinasi

Nilai *an adjusted* pada persamaan di atas (kepuasan) sebesar 0,351, yang menunjukkan bahwa, secara kolektif, Remunerasi, Motivasi dan Suasana Kerja menjelaskan 36,9% varians dalam variabel Kepuasaan. Adapun nilai *an adjusted* pada persamaan 2 (Kinerja) diperoleh sebesar 0,198, yang menunjukkan bahwa variabel Kepuasaan menjelaskan 18,6% varians dalam variabel Kinerja.

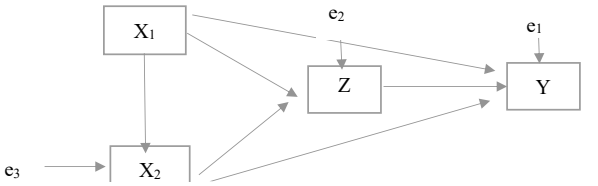
13.5 Latihan

1. Dengan menggunakan data yang sama **"Kinerja.csv"**. Buatlah analisis jalur dengan bentuk diagram path seperti di bawah ini.



   Analisis hasilnya!

2. Seorang peneliti pendidikan ingin menganalisis faktor-faktor yang mempengaruhi prestasi belajar siswa (Y). Dalam penelitian ini digunakan dua variabel eksogen, yaitu motivasi belajar (X_1) dan lingkungan belajar (X_2). Peneliti menduga bahwa kedisiplinan belajar (Z) berperan sebagai variabel intervening yang menengahi pengaruh motivasi belajar dan lingkungan belajar terhadap prestasi belajar siswa. Pada kasus ini, diagram jalur dijelaskan di bawah ini:



   Hasil survei terhadap 20 orang siswa diperoleh hasil seperti di bawah ini.

Responden	Y	X_1	X_2	Z
1	34	35	35	35
2	24	25	24	25
3	17	17	15	15
4	25	26	24	20
5	28	34	30	25
6	18	21	17	16
7	36	36	35	35
8	30	30	27	28

Responden	Y	X_1	X_2	Z
9	31	30	31	29
10	15	16	17	13
11	39	38	38	39
12	23	22	22	23
13	25	23	28	24
14	20	22	21	17
15	17	17	15	21
16	25	26	24	21
17	30	30	27	28
18	31	30	31	29
19	15	16	17	13
20	28	28	29	28

Lakukan analisis jalur dengan data di atas!

BAB

XIV

MODEL PERSAMAAN STRUKTURAL (STRUCTURAL EQUATION MODELLING/SEM)

14.1 Pendahuluan

Structural Equation Modeling (SEM) merupakan metode analisis statistik multivariat yang berkembang sebagai respons terhadap kebutuhan untuk menganalisis hubungan antarvariabel yang bersifat kompleks dan saling terkait. Secara historis, SEM berakar dari dua tradisi keilmuan utama, yaitu analisis jalur (*path analysis*) yang diperkenalkan oleh Sewall Wright pada awal abad ke-20, serta analisis faktor yang berkembang dalam bidang psikometri untuk mengukur konstruk laten. Analisis jalur berfokus pada pengujian hubungan kausal antarvariabel teramati, sedangkan analisis faktor digunakan untuk mengidentifikasi struktur laten di balik sejumlah indikator teramati.

Tabel 27 Perbedaan Analisis Jalur dengan SEM

Aspek	Analisis Jalur (*Path Analysis*)	*Structural Equation Modeling* (SEM)
Jenis variabel	Hanya menggunakan variabel teramati (observed variables)	Dapat menggunakan variabel laten dan variabel teramati
Model pengukuran	Tidak ada	Ada (menghubungkan indikator dengan variabel laten)
Kesalahan pengukuran	Tidak dimodelkan secara eksplisit	Dimodelkan secara eksplisit
Kompleksitas model	Relatif sederhana	Lebih kompleks dan fleksibel
Uji kesesuaian model	Terbatas	Lengkap (CFI, TLI, RMSEA, SRMR, dll.)
Pendekatan analisis	Pengembangan dari regresi linier berganda	Pengembangan dari analisis jalur dan analisis faktor
Ukuran sampel	Relatif kecil	Umumnya memerlukan ukuran sampel lebih besar

Perkembangan SEM semakin pesat pada dekade 1960-an dan 1970-an, seiring dengan kemajuan teori matriks dan peningkatan kapasitas komputasi. Pada periode ini, para peneliti mulai mengintegrasikan analisis jalur dan analisis faktor ke dalam satu kerangka pemodelan yang utuh, sehingga

memungkinkan pengujian model pengukuran dan model struktural secara simultan. Kontribusi penting dalam pengembangan SEM diberikan oleh tokoh-tokoh seperti Jöreskog, yang memperkenalkan pendekatan *covariance structure analysis* dan mengembangkan perangkat lunak LISREL, serta Bentler, yang berperan dalam pengembangan metode estimasi dan indeks kesesuaian model.

Saat ini, SEM telah menjadi salah satu metode analisis yang sangat populer dan banyak digunakan dalam berbagai bidang ilmu, termasuk pendidikan, psikologi, sosiologi, ekonomi, dan manajemen. Didukung oleh perangkat lunak modern seperti LISREL, AMOS, Mplus, dan Lavaan pada R. SEM terus berkembang sebagai alat analisis yang fleksibel dan kuat untuk menguji serta mengembangkan teori berbasis data empiris.

14.2 Variabel dalam *Structural Equation Modelling*

Variabel-variabel pada SEM masing-masing saling mempengaruhi. Variabel-variabel yang terdapat dalam SEM meliputi:

1. Variabel laten (*Latent Variable*)

Variabel laten atau konstruk laten adalah variabel yang tidak dapat diukur secara langsung, tetapi diukur melalui indikator-indikator teramati, sebagai contoh perilaku, sikap, perasaan, dan motivasi.

Variabel laten terdapat dua jenis, yaitu:

a. Variabel Eksogen: Variabel laten eksogen dinotasikan dengan huruf Yunani yaitu ξ ("ksi"). Disebut juga variabel bebas (*independent latent variable*) pada semua persamaan yang ada pada SEM, dengan simbol lingkaran dengan anak panah menuju keluar.
b. Variabel Endogen dinotasikan dengan huruf Yunani η ("eta"). Disebut juga variabel terikat (*dependent latent variable*) pada paling sedikit satu persamaam dalam model, dengan simbol lingkaran dengan anak panah menuju keluar dan satu panah ke dalam. Simbol anak panah untuk menunjukkan adanya hubungan kausal (ekor anak panah untuk hubungan penyebab dan kepala anak panah untuk variabel akibat).

2. Variabel teramati (*Observed atau Measured atau Manifest Variable*)

Variabel teramati adalah variabel yang dapat diamati atau dapat diukur secara empiris dan disebut sebagai indikator. Variabel teramati merupakan efek atau ukuran dari variabel laten. Variabel teramati yang berkaitan dengan variabel eksogen diberi notasi matematik dengan label *X*, sedangkan yang berkaitan dengan dengan variabel laten endogen diberi label *Y*. Disimbolkan dengan bujur sangkar atau kotak, variabel ini merupakan indikator. Pemberian nama variabel teramati pada diagram lintasan bisa mengikuti notasi matematiknya, atau nama/kode dari pertanyaan-pertanyaan pada kuisioner.

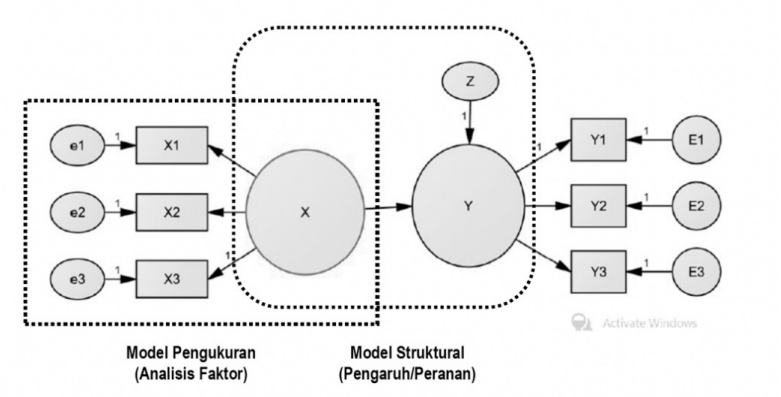
14.3 Model Persamaan Struktural

Structural Equation Modeling (SEM) adalah metode analisis statistik multivariat yang digunakan untuk menguji dan mengembangkan model konseptual yang kompleks, seringkali dalam bentuk model sebab akibat.

Structural Equation Modelling (SEM) memiliki dua jenis model yaitu model struktural dan model pengukuran.

1. **Model Pengukuran (*Measurement Model*)**
   Model ini menjelaskan bagaimana variabel laten diukur oleh indikator-indikator teramati. Biasanya, analisis faktor konfirmatori (CFA) digunakan untuk menguji model pengukuran.
2. **Model Struktural (*Structural Model*)**
   Model ini menjelaskan hubungan sebab akibat antara variabel laten dan/atau variabel teramati. Model ini dapat berupa model jalur (*path analysis*) atau model yang lebih kompleks.

Apabila digambarkan bentuk kedua model tersebut, seperti contoh gambar di bawah ini :



Gambar 18. Contoh Model SEM

Berikut penjelasan untuk masing-masing model dalam SEM

a. Persamaan struktural (*structural equation*) atau disebutkan *outer* model dirumuskan seperti di bawah ini.

$$\boldsymbol{\eta} = \boldsymbol{\beta\eta} + \boldsymbol{\Gamma\xi} + \boldsymbol{\zeta} \qquad (14.1)$$

dengan :

$\boldsymbol{\eta}$ = vektor variabel laten endogen
$\boldsymbol{\xi}$ = vektor variabel laten eksogen
$\boldsymbol{\beta}$ = matriks koefisien
$\boldsymbol{\Gamma}$ = matriks koefisien
$\boldsymbol{\zeta}$ = vektor galat untuk persamaan struktural

b. Persamaan pengukuran (*measurement equation*) atau disebut *inner model* dirumuskan seperti di bawah ini.

$$\boldsymbol{Y} = \boldsymbol{\lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \qquad (14.2)$$

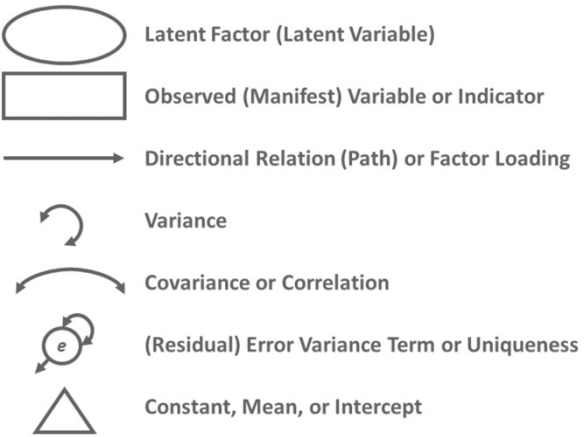
$$\boldsymbol{X} = \boldsymbol{\lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \qquad (14.3)$$

dengan :

$\boldsymbol{\eta}$ = vektor variabel laten endogen
$\boldsymbol{\xi}$ = vektor variabel laten eksogen
$\boldsymbol{\lambda}_y$ = matriks koefisien terhadap

$\boldsymbol{\lambda}_x$ = matriks koefisien terhadap
$\boldsymbol{\varepsilon}$ = vektor galat untuk model pengukuran
$\boldsymbol{\delta}$ = vektor galat untuk model pengukuran

Model SEM sering divisualisasikan menggunakan diagram jalur. Diagram jalur menampilkan spesifikasi parameter model dan juga dapat mencakup estimasi parameter. Simbol diagram jalur ditunjukkan pada gambar berikut ini.



Gambar 19. Simbol Diagram Jalur dan Maknanya

Secara umum prosedur SEM mengandung tahap-tahap seperti di bawah ini.

1. Spesifikasi model (*model specification*). Tahap ini berkaitan dengan pembentukan model awal persamaan struktural. Model awal ini diformulasikan suatu teori atau penelitian sebelumnya.
2. Identifikasi (*identification*). Tahap ini berkaitan dengan pengkajian tentang kemungkinan diperolehnya nilai yang unik untuk setiap parameter yang ada di dalam model dan kemungkinan persamaan simultan tidak ada solusinya.
3. Estimasi (*estimation*). Tahap ini berkaitan dengan estimasi terhadap model untuk menghasilkan nilai-nilai parameter menggunakan salah satu metode estimasi yang tersedia. Pemilihan metode estimasi

yang digunakan seringkali ditentukan berdasarkan karakteristik dari variabel-variabel yang dianalisis.

4. Uji kecocokan (*testing fit*). Tahap ini berkaitan dengan pengujian kecocokan antara model dengan data. Beberapa kriteria ukuran kecocokan atau *Goodness of Fit* (GOF) dapat digunakan untuk melaksanakan langkah ini.
5. Respesifikasi (*Respecification*). Tahap ini dapat juga disebut modifikasi yang berkaitan dengan respesifikasi model berdasarkan hasil uji kecocokan pada tahap

14.4 Spesifikasi Model

Tahap spesifikasi model merupakan pembentukan hubungan antara variabel laten yang satu dengan variabel laten lainnya dan juga hubungan antara variabel laten dengan variabel manifest didasarkan pada teori yang berlaku.

Langkah-langkah untuk membuat model yang diinginkan, yaitu :

1. Spesifikasi model pengukuran
   a. Mendefinisikan variabel-variabel laten yang ada di dalam penelitian
   b. Mendefinisikan variabel-variabel yang teramati
   c. Mendefinisikan hubungan di antara variabel laten dengan variabel yang teramati.
2. Spesifikasi model struktural
   Mendefinisikan hubungan kausal di antara variabel-variabel laten tersebut.
3. Menggambarkan diagram jalur dengan *hybrid model* yang merupakan kombinasi dari model pengukuran dan model struktural, jika diperlukan (bersifat opsional).

Identifikasi Model

Secara garis besar ada 3 kategori dalam persamaan secara simultan, yaitu:

1. *Under-identified* model: model dengan jumlah parameter yang diestimasi lebih besar dari jumlah data yang diketahui (data

tersebut merupakan varians dan kovarians dari variabel-variabel yang teramati).

2. *Just-identified* model: model dengan jumlah parameter yang diestimasi sama dengan jumlah data yang diketahui.
3. *Over-identified* model: model dengan jumlah parameter yang diestimasi lebih kecil dari jumlah data yang diketahui.

Cara melihat ada tidaknya problem identifikasi adalah dengan melihat hasil estimasi yang meliputi :

1. Adanya nilai standart *error* yang besar untuk 1 atau lebih koefisien;
2. Ketidakmampuan program untuk invert information matrix;
3. Nilai estimasi yang tidak mungkin *error* variance yang negatif; dan
4. Adanya nilai korelasi yang tinggi (> 0,90) antar koefisien estimasi.

Jika diketahui ada problem identifikasi maka ada tiga hal yang harus dilihat:

1. Besarnya jumlah koefisien yang diestimasi relatif terhadap jumlah kovarian atau korelasi, yang diindikasikan dengan nilai *degree of freedom* yang kecil;
   (df= jumlah data yang diketahui-jumlah parameter yang diestimasi) < 0)
2. Digunakannya pengaruh timbal balik atau respirokal antarkonstruk (model non *recursive*); atau
3. Kegagalan dalam menetapkan nilai tetap (fix) pada skala konstruk.

14.5 Uji Kecocokan Model

Setelah melakukan metode estimasi terhadap model, langkah selanjutnya adalah melakukan uji kecocokan model. Uji kecocokan model dilakukan untuk menguji apakah model yang dihipotesiskan merupakan model yang baik untuk merepresentasikan hasil penelitian. Menurut Hair dkk sebagaimana dikutip oleh Wijanto (2008), evaluasi terhadap tingkat kecocokan data dengan model dilakukan melalui beberapa tahapan, yaitu:

1. Kecocokan Keseluruhan Model (*Overall Model Fit*);
2. Kecocokan Model Pengukuran (*Measurement Model Fit*); dan
3. Kecocokan Model Struktural (*Structural Model Fit*).

14.6 Prosedur Analisis SEM dengan R

Structural Equation Modeling (SEM) adalah metode analisis statistik multivariat yang digunakan untuk menguji dan mengembangkan model konseptual yang kompleks, seringkali dalam bentuk model sebab akibat. Dalam R, SEM dapat dilakukan dengan menggunakan berbagai paket, seperti `lavaan` dan `sem`.

Pada modul praktikum ini, analisis SEM menggunakan perintah `sem()` yang tersedia dalam *package* `lavaan`. Berikut sintak dari fungsi `lavaan()`

```
sem(model = NULL, data = NULL, ordered = NULL, sampling.weights
= NULL,
 sample.cov = NULL, sample.mean = NULL, sample.th = NULL,
 sample.nobs = NULL,r group = NULL, cluster = NULL,
 constraints = "", WLS.V = NULL, NACOV = NULL, ov.order =
"model",

Keterangan
- model = A description of the user-specified model
- data = An optional data frame containing the observed
  variables used in the model
- ordered = Character vektor. Only used if the data is in a
  data.frame
- sampling.weights = A variable name in the data frame
  containing sampling weight information.
- sample.cov = A sample variance-covariance matrix
- sample.mean sample.mean = A sample mean vektor.
- sample.th vektor of sample-based thresholds.
- sample.nobs = Number of observations if the full data frame
  is missing and only sample moments are given.
- Group = A variable name in the data frame defining the groups
  in a multiple group analysis.
- Cluster = A (single) variable name in the data frame defining
  the clusters in a two-level dataset
- Constraints = Additional (in)equality constraints not yet
  included in the model syntax
- WLS.V = A user provided weight matrix to be used by estimator
  "WLS"; if the estimator is "DWLS", only the diagonal of this
  matrix will be used.
- NACOV = A user provided matrix containing the elements of
  (N times) the asymptotic variance-covariance matrix of the
  sample statistics
```

```
- ov.order = If "model" (the default), the order of the
  observed variable names (as reflected for example in the
  output of lavNames()) is determined by the model syntax.
```

Sebelum menjalankan SEM, berikut beberapa symbol/sintaksis yang

paling sering digunakan dalam `lavaan`.

1. ~ predict, digunakan untuk meregresikan variabel endogenous dengan hasil teramati variabel exogenous (misalnya, y ~ x)
2. =~ indicator, digunakan untuk menghubungkan variabel laten terhadap indikator teramati dalam model pengukuran analisis faktor (misalnya, f =~ q + r + s)
3. ~~ kovarians, digunakan untuk menunjukkan hubungan covariansi atau korelasi (misalnya, x ~~ x)
4. ~1 intersep atau rata-rata (misalnya, x ~ 1 mengestimasi rata-rata variabel x)
5. 1* menetapkan parameter atau pemuatan menjadi satu (misalnya, f =~ 1*q)
6. NA* membebaskan parameter atau pemuatan (berguna untuk mengganti metode penanda default, (misalnya, f =~ NA*q)
7. a* memberi label parameter 'a', digunakan untuk batasan model (misalnya, f =~ a*q)

Contoh

Pihak sekolah menengah berupaya meningkatkan prestasi belajar siswa sebagai salah satu indikator keberhasilan proses pembelajaran. Prestasi belajar siswa tidak hanya ditentukan oleh kemampuan kognitif semata, tetapi juga dipengaruhi oleh faktor non-kognitif, seperti motivasi belajar, serta faktor eksternal berupa lingkungan belajar. Motivasi belajar berperan dalam mendorong siswa untuk terlibat aktif dalam proses pembelajaran, sedangkan lingkungan belajar yang kondusif dapat mendukung terciptanya suasana belajar yang efektif.

Berdasarkan pertimbangan tersebut, dilakukan penelitian untuk menganalisis pengaruh motivasi belajar dan lingkungan belajar terhadap prestasi belajar siswa menggunakan pendekatan *Structural Equation Modeling* (SEM). Metode SEM dipilih karena variabel-variabel yang diteliti bersifat laten dan tidak dapat diukur secara langsung, melainkan direpresentasikan oleh beberapa indikator teramati.

Dalam penelitian ini, prestasi belajar (η□) diperlakukan sebagai variabel

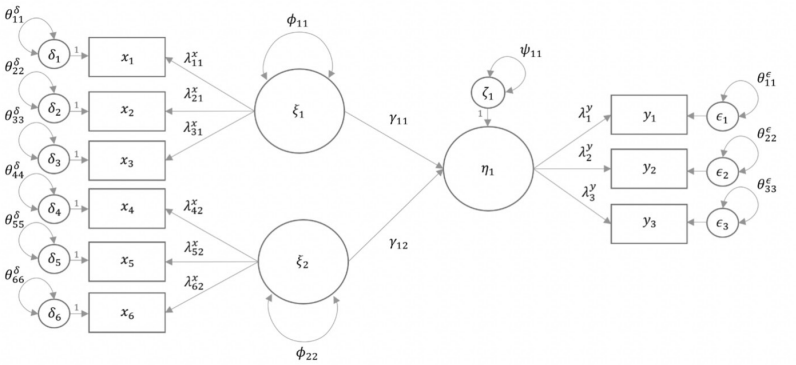
laten endogen yang diukur melalui tiga indikator, yaitu kemampuan membaca (baca/y□), kemampuan aritmetika (arith/y□), dan kemampuan ejaan (eja/y□). Sementara itu, motivasi belajar (ξ□) sebagai variabel laten eksogen diukur melalui indikator minat belajar (minat/x□), ketekunan (tekun/x□), dan orientasi tujuan (tujuan/x□). Selanjutnya, lingkungan belajar (ξ□) sebagai variabel laten eksogen lainnya diukur melalui dukungan guru (guru/x□), fasilitas belajar (fas/x□), dan suasana kelas (kelas/x□).

Variabel penelitian beserta indikatornya dijelaskan pada tabel berikut:

Tabel 28. Data Siswa beserta Variabel Konstruk

No	Variabel Laten	Indikator	Symbol
1	Prestasi/ Achievment	Kemampuan Membaca	baca (y_1)
		Kemampuan Aritmetika	arith (y_2)
		Kemampuan Ejaan	eja (y_3)
2	Motivasi/ Motivation	Minat Belajar	minat (x_1)
		Ketekunan	tekun (x_2)
		Orientasi Tujuan	tujuan (x_3)
3	Lingkungan Belajar	Dukungan Guru	guru (x_4)
		Fasilitas Belajar	fas (x_5)
		Suasana Kelas	kelas (x_6)

Model struktural yang dibangun dalam penelitian ini bertujuan untuk menguji sejauh mana motivasi belajar dan lingkungan belajar secara simultan mempengaruhi prestasi belajar siswa. Dengan menggunakan SEM, peneliti dapat mengevaluasi kesesuaian model secara keseluruhan serta menguji kekuatan hubungan antarvariabel laten berdasarkan data empiris. Model SEM pada kasus ini digambarkan sbb :



Gambar 20. Diagram Jalur SEM

1. Persamaan Model Struktural (*Structural Model*)

Model struktural menggambarkan hubungan kausal antarvariabel laten :

$$\eta\square = \gamma\square\square\xi\square + \gamma\square\square\xi\square + \zeta\square$$

dengan :

$\eta\square$: Prestasi Belajar
$\xi\square$: Motivasi Belajar
$\xi\square$: Lingkungan Belajar
$\gamma\square\square$: Koefisien pengaruh langsung
$\zeta\square$: Error struktural

2. Persamaan Model Pengukuran Variabel Endogen (Measurement Model-Y)

Prestasi Belajar ($\eta\square$) diukur oleh tiga indikator :

$$Y_1 = \lambda_{y1}\eta_1 + \varepsilon_1 \quad \text{(Kemampuan Membaca)}$$
$$Y_2 = \lambda_{y2}\eta_2 + \varepsilon_2 \quad \text{(Kemampuan Aritmetika)}$$
$$Y_3 = \lambda_{y3}\eta_3 + \varepsilon_3 \quad \text{(Kemampuan Ejaan)}$$

dengan :

λ_{y1} : factor loading indikator ke-i
ε_1 : error pengukuran indikator Y

3. Persamaan Model Pengukuran Variabel Eksogen (Measurement Model – X)

a. Motivasi Belajar ()

$X_1 = \lambda_{x1}\xi_1 + \delta_1$ (Minat Belajar)

$X_2 = \lambda_{x2}\xi_1 + \delta_2$ (Ketekunan)

$X_3 = \lambda_{x3}\xi_1 + \delta_3$ (Orientasi Tujuan)

b. Lingkungan Belajar ()

$X_1 = \lambda_{x1}\xi_1 + \delta_1$ (Dukungan Guru)

$X_2 = \lambda_{x2}\xi_1 + \delta_2$ (Fasilitas Belajar)

$X_3 = \lambda_{x3}\xi_1 + \delta_3$ (Suasana Kelas)

dengan :

λ_{x1} : factor loading indikator

δ_1 : error pengukuran indikator

Berdasarkan pengumpulan data, peneliti memperoleh 500 responden. Data disimpan dengan nama "Siswa.csv". Data dapat diakses pada tautan ini: https ://drive.google.com/file/d/1HGxIfZWo07us1ZI0M7CBeh5-Lk6P_hDG/view?usp=sharing

Untuk melihat data pada kasus ini, dapat dilakukan melalui perintah berikut ini :

```
#menginstal package readr apabila belum tersedia,
>install.packages("readr")
#memanggil package readr
>library(readr)
#memasukkan data ke dalam R dan menyimpannya dalam
sebuah data frame bernama data_siswa
> data_siswa <- read.csv("Siswa.csv")
#menampilkan data dalam bentuk tabel.
> View (data_siswa)
```

baca	arith	ejaan	minat	tekun	tujuan	guru	fas	kelas
-3.488981	-9.989121	-6.567873	-7.907122	-5.075312	-3.138836	-17.800210	4.766450	-3.633360
-4.520552	8.196238	8.778973	1.751478	-4.155847	3.520752	7.009367	-6.048681	-7.693461
-4.818170	7.529984	-5.688716	14.472570	-4.540677	4.070600	23.734260	-16.970670	-3.909941
-3.138441	5.730547	-2.915676	-1.165421	-5.668406	2.600437	1.493158	1.396363	21.409450
-2.009742	-0.623953	-1.024624	-4.222899	-10.072150	-6.030737	-5.985864	-18.376400	-1.438816
0.822650	5.045174	0.904154	4.868769	3.029841	-7.648277	14.668790	-2.235039	-6.826892
-5.672760	8.638372	-1.525859	10.367370	5.039368	6.031902	-0.952449	-9.258073	8.485556
-11.338880	3.753714	-7.449076	-1.861007	0.398970	-1.041958	-14.569340	-15.998880	-0.763939
2.397864	-8.260820	2.343741	-13.452220	-15.333230	-10.938220	-5.084801	-3.269256	-1.157033
-5.819805	-1.407573	-2.550445	2.852636	2.202829	-3.961495	11.203110	0.255099	-11.238150
-9.586946	-5.120381	-10.116000	-2.135480	-4.988594	-5.197804	18.165050	-8.722278	-22.846290
-3.136714	-9.810998	-5.178977	-13.271710	-14.716500	-22.237840	-13.617300	-2.468196	3.227680
-11.526480	-5.075179	1.392740	0.074682	-3.900125	-0.681536	13.456060	-1.117470	-3.047351
-0.894078	3.111883	4.711888	-1.234672	-7.602469	-11.120770	9.088286	-8.614449	-2.908931

Untuk mengestimasi model *Structural Equation Modeling* (SEM), digunakan fungsi `sem()` yang tersedia dalam *package* `lavaan` pada perangkat lunak R. Prosedur analisis SEM dilakukan melalui beberapa tahapan yang terstruktur.

1. Pertama, *package* Lavaan diinstal dan dipanggil ke dalam lingkungan kerja R.
2. Kedua, data penelitian dibaca dan disiapkan dalam bentuk *data frame*.
3. Ketiga, model SEM didefinisikan secara eksplisit dalam bentuk sintaks yang mencakup model pengukuran (hubungan antaravariabel laten dan indikatornya) serta model struktural (hubungan kausal antarvariabel laten).
4. Selanjutnya, model yang telah dispesifikasikan, diestimasi menggunakan fungsi `sem()` dengan metode estimasi *Maximum*

Likelihood.

5. Hasil estimasi kemudian dievaluasi melalui ringkasan *output* yang meliputi nilai koefisien parameter, uji signifikansi, koefisien determinasi (*R-square*), serta berbagai indeks kelayakan model (*goodness-of-fit indices*) untuk menilai kesesuaian model dengan data empiris.

Syntax SEM

```
>install.packages("lavaan")
>library(lavaan)

model_sem <- '
 # Model pengukuran (measurement model)
 Prestasi =~ baca + arith + ejaan
 Motivasi =~ minat + tekun + tujuan
 Lingkungan =~ guru + fas + kelas

 # Model struktural (structural model)
 Prestasi ~ Motivasi + Lingkungan
'

fit_sem <- sem(model_sem, data = data_siswa)

# Menampilkan ringkasan hasil SEM
summary(fit_sem,
    fit.measures = TRUE,     # indeks kelayakan model
    standardized = TRUE,     # koefisien terstandarisasi
    rsquare = TRUE)          # nilai R-square
```

Hasilnya

```
lavaan 0.6-20 ended normally after 148 iterations

 Estimator                                         ML
 Optimization                           method NLMINB
 Number of model parameters                        21

 Number of observations                           500

Model Test User Model :

 Test statistic                                78.654
 Degrees of freedom                            24
 P-value (Chi-square)                          0.000

Model Test Baseline Model :

 Test statistic                              1959.646
 Degrees of freedom                            36
 P-value                                       0.000

User Model versus Baseline Model :

 Comparative Fit Index (CFI)                   0.972
 Tucker-Lewis Index (TLI)                      0.957

Loglikelihood and Information Criteria :

 Loglikelihood user model (H0)          -16714.873
 Loglikelihood unrestricted model (H1)  -16675.547

 Akaike (AIC)                            33471.747
 Bayesian (BIC)                          33560.254
 Sample-size adjusted Bayesian (SABIC)   33493.598

Root Mean Square Error of Approximation :

 RMSEA 0.067
 90 Percent confidence interval - lower        0.051
 90 Percent confidence interval - upper        0.084
 P-value H_0: RMSEA <= 0.050                   0.039
 P-value H_0: RMSEA >= 0.080                   0.115

Standardized Root Mean Square Residual :

 SRMR                                          0.037
```

```
Parameter Estimates :

 Standard errors                           Standard
 Information                               Expected
 Information saturated (h1) model       Structured

Latent Variables :
              Estimate  Std.Err  z-value  P(>|z|)  Std.lv  Std.all
 Prestasi =~
   baca          1.000    9.463   0.947
   arith         0.828    0.034   24.286    0.000   7.833    0.784
   ejaan         0.965    0.028   34.057    0.000   9.134    0.914
 Motivasi =~
   minat         1.000    9.024    0.504
   tekun         0.911    0.094    9.662    0.000   8.220    0.823
   tujuan        0.762    0.080    9.501    0.000   6.875    0.688
 Lingkungan =~
   guru          1.000    5.831    0.169
   fas          -0.943    0.296   -3.182    0.001  -5.497   -0.550
   kelas        -1.252    0.388   -3.228    0.001  -7.301   -0.731

Regressions :
              Estimate  Std.Err  z-value  P(>|z|)  Std.lv  Std.all
 Prestasi ~
   Motivasi      0.301    0.065    4.608    0.000   0.287    0.287
   Lingkungan   -1.042    0.332   -3.142    0.002  -0.642   -0.642

Covariances :
              Estimate  Std.Err  z-value  P(>|z|)  Std.lv  Std.all
 Motivasi ~~
   Lingkungan  -25.266    8.717   -2.899    0.004  -0.480   -0.480

Variances :
              Estimate  Std.Err  z-value  P(>|z|)  Std.lv  Std.all
  .baca          10.254    1.634    6.274    0.000  10.254    0.103
  .arith         38.445    2.715   14.160    0.000  38.445    0.385
  .ejaan         16.375    1.754    9.336    0.000  16.375    0.164
  .minat        239.707   16.749   14.312    0.000 239.707    0.746
  .tekun         32.227    4.939    6.524    0.000  32.227    0.323
  .tujuan        52.532    4.669   11.251    0.000  52.532    0.526
  .guru        1161.770   74.215   15.654    0.000 1161.770   0.972
  .fas           69.585    5.218   13.336    0.000  69.585    0.697
  .kelas         46.497    5.670    8.201    0.000  46.497    0.466
  .Prestasi      29.391    4.429    6.636    0.000   0.328    0.328
   Motivasi      81.428   15.276    5.330    0.000   1.000    1.000
   Lingkungan    33.999   20.727    1.640    0.101   1.000    1.000
```

```
R-Square :
             Estimate
 baca           0.897
 arith          0.615
 ejaan          0.836
 minat          0.254
 tekun          0.677
 tujuan         0.474
 guru           0.028
 fas            0.303
 kelas          0.534
 Prestasi       0.672
```

Evaluasi Kecocokan Model (Goodness of Fit)

Hasil estimasi menunjukkan bahwa model SEM berhasil dikonvergensi secara normal setelah 148 iterasi menggunakan metode estimasi *Maximum Likelihood* (ML) dengan jumlah sampel sebanyak 500 responden. Uji Chi-square menghasilkan nilai $X^2 = 78{,}654$ dengan derajat bebas 24 dan *p-value* = 0,000. Meskipun uji Chi-square signifikan (sensitif terhadap ukuran sampel besar), evaluasi kelayakan model lebih lanjut dilakukan menggunakan indeks fit lainnya.

Nilai CFI = 0,972 dan TLI = 0,957 menunjukkan bahwa model memiliki kecocokan yang baik, karena keduanya melebihi batas rekomendasi 0,90. Nilai RMSEA = 0,067 dengan interval kepercayaan 90% (0,051–0,084) menunjukkan tingkat kesalahan aproksimasi yang masih dapat diterima (*acceptable fit*), sedangkan nilai SRMR = 0,037 (< 0,08) mengindikasikan kesesuaian model yang sangat baik. Secara keseluruhan, model SEM dinyatakan layak untuk dianalisis lebih lanjut.

Model Pengukuran (Measurement Model)

Variabel laten Prestasi diukur oleh tiga indikator, yaitu kemampuan membaca (*baca*), aritmetika (*arith*), dan ejaan (*ejaan*). Seluruh indikator memiliki *loading factor* terstandarisasi yang tinggi (0,784–0,947) dan signifikan ($p < 0{,}001$), sehingga dapat disimpulkan bahwa indikator-indikator tersebut merupakan pengukur yang valid bagi konstruk Prestasi.

Variabel laten Motivasi diukur oleh indikator minat belajar, ketekunan,

dan orientasi tujuan. Nilai *loading factor* terstandarisasi berkisar antara 0,504 hingga 0,823 dan seluruhnya signifikan ($p < 0,001$), menunjukkan bahwa indikator-indikator tersebut secara memadai merefleksikan konstruk Motivasi.

Variabel laten Lingkungan Belajar diukur oleh dukungan guru, fasilitas belajar, dan suasana kelas. Meskipun seluruh indikator signifikan secara statistik ($p < 0,01$), dua indikator (fasilitas dan suasana kelas) memiliki arah hubungan negatif, yang mengindikasikan kemungkinan perbedaan skala pengukuran atau perlunya peninjauan ulang pada pengkodean indikator.

Model Struktural (Structural Model)

Hasil analisis struktural menunjukkan bahwa Motivasi belajar berpengaruh positif dan signifikan terhadap Prestasi belajar dengan koefisien terstandarisasi sebesar 0,287 ($p < 0,001$). Hal ini menunjukkan bahwa peningkatan motivasi siswa akan diikuti oleh peningkatan prestasi belajar.

Sebaliknya, Lingkungan belajar berpengaruh negatif dan signifikan terhadap Prestasi belajar dengan koefisien terstandarisasi sebesar –0,642 ($p = 0,002$). Temuan ini mengindikasikan adanya hubungan yang berlawanan arah, yang dapat disebabkan oleh karakteristik data, persepsi siswa terhadap lingkungan belajar, atau masalah pengukuran indikator.

Selain itu, terdapat kovarians negatif yang signifikan antara Motivasi dan Lingkungan belajar ($r = –0,480$; $p = 0,004$), yang menunjukkan bahwa kedua konstruk tersebut saling berkaitan namun bergerak dalam arah yang berlawanan.

Koefisien Determinasi (R-Square)

Nilai R^2 untuk variabel Prestasi sebesar 0,672, yang berarti bahwa 67,2% variasi Prestasi belajar siswa dapat dijelaskan oleh Motivasi dan Lingkungan belajar. Sisanya sebesar 32,8% dijelaskan oleh faktor lain di luar model. Nilai R^2 yang relatif tinggi ini menunjukkan bahwa model memiliki daya jelas (*explanatory power*) yang kuat.

Kesimpulan

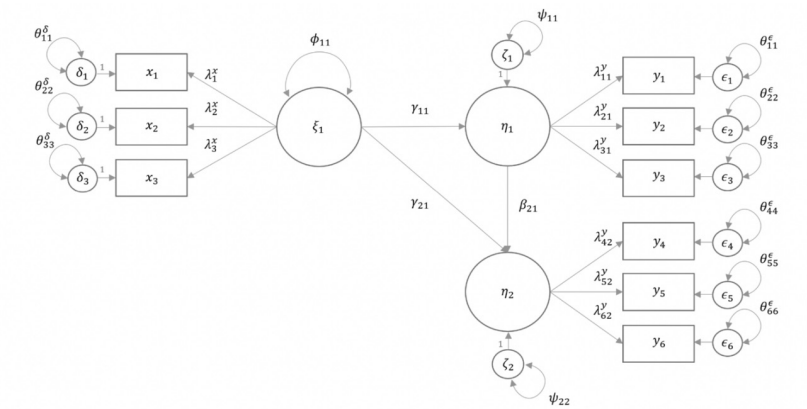
Berdasarkan hasil analisis Structural Equation Modeling (SEM), dapat

disimpulkan bahwa model yang dibangun memiliki kecocokan yang baik dengan data empiris. Hal ini ditunjukkan oleh nilai indeks kelayakan model seperti CFI = 0,972, TLI = 0,957, RMSEA = 0,067, dan SRMR = 0,037, yang secara umum memenuhi kriteria *good fit*. Dengan demikian, model layak digunakan untuk menjelaskan hubungan antarvariabel yang diteliti.

Secara struktural, hasil penelitian menunjukkan bahwa motivasi belajar berpengaruh positif dan signifikan terhadap prestasi belajar siswa ($p < 0{,}001$), sedangkan lingkungan belajar berpengaruh signifikan terhadap prestasi belajar ($p = 0{,}002$). Kedua variabel tersebut secara bersama-sama mampu menjelaskan 67,2% variasi prestasi belajar siswa ($R^2 = 0{,}672$). Temuan ini menegaskan bahwa selain faktor kognitif, faktor non-kognitif dan konteks belajar memiliki peran penting dalam meningkatkan prestasi belajar siswa.

14.7 Latihan

Dengan menggunakan data yang sama *"Siswa.csv"*. Buatlah model SEM dengan konsep model seperti gambar di bawah ini



Dalam penelitian ini motivasi belajar (η_1) dan prestasi belajar (η_2) merupakan variabel endogen sedangkan lingkungan belajar (ξ_2) sebagai variabel laten eksogen yang diduga berpengaruh langsung maupun tidak langsung terhadap prestasi belajar melalui motivasi belajar.

1. Berdasarkan gambar di atas, tuliskan :
    a. Persamaan model pengukuran (*measurement model*) untuk

variable laten eksogen dan endogen.

b. Persamaan model struktural (*structural model*) yang menunjukkan hubungan kausal antarvariabel laten.
(Gunakan notasi SEM yang baku (λ, γ, β, δ, ε, ζ).

2. Estimasi model SEM menggunakan R serta analisis hasilnya!.

Daftar Pustaka

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). Wiley.

Aggara, Y dan Supandi, E.D. (2021). Perbandingan Analisis Faktor Klasik dan Robust secara Empiris dan Simulasi (Studi kasus: Mobilitas Aktivitas Masyarakat Setiap Provinsi Di Masa Covid-19). *Jurnal Fourier*. Vol 10 No 1, 89 – 98.

Beaujean, A. A. (2014). *Latent variable modeling using R*. New York: Routledge

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

Borg, I., and Groenen, P. J. F. (2005), *Modern multidimensional scaling: Theory and applications (2nd ed.)*. Springer.

Cindi, C. Setyawati, I dan Supandi. E.D. (2024). Analisis Structural Equation Modeling Partial Least Square Terhadap Faktor Yang Mempengaruhi Keputusan Pembelian Produk AMDK Pada Mahasiswa PTS X di Sleman. *Performa: Media Ilmiah Teknik Industri*. Vol. 23 (1). pp: 55-61

Collier, J. E. (2020). *Applied Structural Equation Modeling Using AMOS*. Routledg

Cox, T. and Cox M. (2001). *Multidimensional Scaling* (Second Edition). Chapman & Hall/CRC. United States of America.

Cox, M.A., dan Cox, T.F. (2008). Multidimensional Scaling. In *Handbook of data visualization*, 315–347. Springer, Berlin, Heidelberg.

Cui, J. (2024). Variable Selection in Multivariate Regression Models with Covariates Subject to Measurement Error. *Journal of Multivariate Analysis, 205,* 105299. https ://doi.org/10.1016/j.jmva.2024.105299

Dalmijer, E.D., Nord, C.L. and Astle D.E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics* . Vol (23) :205.

Dentrinidad, E., and Lopez-Ruiz, V. (2024), *The interplay of happiness and sustainability: A multidimensional scaling and K-Means cluster approach, Sustainability,* 16, 10068.

Dong, X. (2024). Adaptive Analysis of the Heteroscedastic Multivariate Regression. *Journal of Statistical Computation and Simulation, 94*(16), 3379–3399. https ://doi.org/10.1080/00949655.2024.2375396

Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70, 927–939.

Everitt, B. and, Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer. London.

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2011), *Cluster analysis (5th ed.).* Wiley.

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some Implementations of the Boxplot. *The American Statistician*, *43*(1), 50–54

Ghozali, I. (2021). *Analisis Multivariat dengan Program IBM SPSS 26*. Badan Penerbit Universitas Diponegoro.

Grimm & Yarnold, PR. (1995). Reading and understanding multivariate statistics. *American Psychological Association*.

Groenen, P. and Ingwer B. (2005). *Modern Multidimensional Scaling Theory and Application* (Second Edition). Springer. United States of America

Handaningrum, E. Y., Safitri, D., & Ispriyanti, D. (2014). Analisis jalur untuk mengetahui hubungan antara usia ibu, kadar hemoglobin, dan masa gestasi terhadap berat bayi lahir. *Jurnal Gaussian, Vol 3(1), 71-80.*

Hair, J.F. (1998). *Multivariate Analysis.* Fifth Edition, New Jersey: Prentice Hall International

Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics—Theory and*

Methods, 19(10), 3595–3617.

Hoyle, R. H. (Ed.). (2023). *Handbook of Structural Equation Modeling* (2nd ed.). Guilford Press

Huberty, C.J., dan Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. John Wiley & Sons.

Inamete, E. N., Happines, & Emmanuel, B. O. (2020). Two-way Multivariate Analysis of Variance (MANOVA) of the Effect of Parity on Body Mass Index and Body Surface Area of Pregnant Women. *Asian Journal of Physical and Chemical Sciences*, *8*(3), 38–48.

Iqbal, M., Salsabila, I.,Syahbani, DA., Douw, J., Marzuki, Rusyana, A. (2020). Analisis MANOVA Satu Arah untuk Melihat Perbedaan Status Gizi Balita Berdasarkan Wilayah Pembangunan Utama di Indonesia Tahun 2017. *Journal of Data Analysis*, Vol 3(1), 50-61.

Jain AK, Murty MN, Flynn PJ. (1999). *Data clustering: a review*. ACM Comput Surv CSUR. 1999;31(3) :264–323.

Jain, A. K. (2010), *Data clustering: 50 years beyond K-Means. Pattern Recognition Letters,* 31(8), 651–666.

Johnson R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*. Sixth Edition. Pearson Prentice Hall. New Jersey.

Jolliffe, I.T. and Cadima J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of The Royal Society A. Mathematical, Fhysical and Engineering Sciences*.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.

Kruskal, J. B., and Wish, M. (1978), *Multidimensional scaling*. Sage Publications.

Krzanowski, W. (2000). *Principles of multivariate analysis*. Oxford: OUP Oxford.

Kodinarya, T. M., and Makwana, P. R. (2013), *Review on determining the number of clusters in K-Means clustering, International Journal*, 1(6),

90–95

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill.

Leandre R.F. dan Duane T.W. (2012) *Exploratory Factor Analysis*. Oxford University Press, Inc.

Lestari, S.P, Supandi, E.D. dan Pratiwi, P. (2018). Pengklasteran Kabupaten/ Kota di Jawa Tengah berdasrkan Tenaga Kesehatan dengan Menggunakan Metode Ward dan K-Means. *Journal Fourier*. Jurnal Matematika dan Pembelajaran. Vol. 7(2), 103 – 109

Maindonald, JH. (2008). *Using R for Data Analysis and Graphics. Introduction, Code and Commentary*. Centre for Mathematics and Its Applications. Australian National University

Morrison, D. F. (2004). *Multivariate Statistical Methods* (4th ed.). Duxbury Press.

Okeke, E. N., Okeke, J. U., & Adashu, D. (2018). Multivariate Analysis of Variance of University Students' Academic Performance. *International Journal of Applied Mathematics and Statistical*, *7*(1), 13–20.

Raltan, P. dan Renhard M. (2014). Analisis Jalur (Path Analysis): Teori dan Aplikasi dalam Riset Bisnis. Jakarta: Rineka Cipta.

Riduwan dan Kuncoro, E.A. (2014). Cara Mudah Menggunakan dan Memaknai Path Analysis (Analisis Jalur). Bandung: Alfabeta.

R Core Team. (2020). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing

Rencher, A. C., & Schaalje, G. B. (2008). *Linear Models in Statistics* (2nd ed.). Wiley.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1–36.

Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro–Wilk W. *Applied Statistics*, 32(2), 121–133.

Safitri D. dan Paramitra I. (2009). Analisis Korelasi Kanonik pada Perilaku

Kesehatan dan Karakteristik Sosial Ekonomi di Kota Pati Jawa Tengah. *Media Statistika,* 2(1), 39–48.

Santoso, S. (2018). *Statistik Multivariat dengan SPSS*. PT Elex Media Komputindo.

Simamora, B. (2005). *Analisis Multivariat Pemasaran*. Jakarta: PT. Gramedia Pustaka Utama

Simar, L. and Wolfgang H. (2007). *Applied Multivariate Statistical Analysis* (Second Edition). Springer. United states of America.

Suhartono. (2008). *Analisis Data dengan R*. Jurusan Statistika. Institut Teknologi Sepuluh Nopember. Surabaya

Supandi, E.D (2020). Structural Equation Modeling with Generalized Structured Component Analysis on the Relationship between Renumeration and Motivation on Employee Performamce at UIN Sunan Kalijaga Yogyakarta. *Media Statistika*, Vol 13, 116 – 224.

Supandi, E.D. dan Anggara Y. (2023). Analisis Klaster dalam Pembentukan Portofolio Robust Mean-Variance. *Jurnal Sains Matematika dan Statistika*. Vol 9 No 1, 37- 47

Supranto, J. (2004). *Analisis Multivariat: Arti dan Interprestasi*. Jakarta: PT. Rineka Cipta

Venables, W.N. and Smith, D.M. (2007) *An Introduction to R*. The R Development Core Team.

Vera, J. F., and Macías, R. (2021), *On the behaviour of K-Means clustering of a dissimilarity matrix by means of full multidimensional scaling, Psychometrika*, 86, 489–513.

Wahyuni, C., Bagus Sumargo, & Qorry Meidianingsih. (2023). Penerapan Analisis Jalur (Path Analysis) dalam Menentukan Faktor-faktor yang Mempengaruhi Angka Harapan Hidup di Wilayah Indonesia Bagian Tengah. *Jurnal Statistika Dan Aplikasinya*, *7*(1), 74–83. https ://doi. org/10.21009/JSA.07107

Thomson B. (2004). *Exploratory and Confirmatory Factor Analysis Understanding Concepts and Applications*. American Psychological Association.

Williams B., Onsman A., Brown T. (2010). Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Health Care (JEPHC)*. 8(3), 990399.

Yong A.G. dan Pearce S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*. 9(2), 79–94.

Zuva , A.Z, Mussadi, N.S.M dan Supandi. E.D. (2025). Cluster Analysis of K-Means and Ward Method in Forming a Robust Portfolio: An Empirical Study of Jakarta Islamic Index. *Barekeng: Journal of Mathematics and Its Application*. Vol 19 (1). 537-546

Indeks

Biodata Penulis

Epha Diana Supandi lahir di Tasikmalaya, 12 September 1975. Pendidikan SD, SMP dan SMA di kabupaten Tasikmalaya Jawa Barat. Pada tahun 1994 melanjutkan pendidikan di Institut Pertanian Bogor di Departemen Statistika. Setelah lulus S1, penulis sempat bekerja sebagai peneliti di Lembaga Penelitian Ekonomi dan Masyarakat (LPEM), Fakultas Ekonomi Universitas Indonesia.

Pada tahun 2003, penulis mendapat kesempatan melanjutkan program Magister di Institut Penyelidikan Matemateika (INSPEM), Universiti Putra Malaysia dengan bidang Statistika. Sedangkan, program Doktor ditempuh pada tahun 2017 di program studi Matematika, FMIPA Universitas Gadjah Mada, Yogyakarta dengan minat ilmu Statistika.

Sejak tahun 2008, penulis menjadi dosen tetap di program Studi Matematika, Fakultas Sains dan Teknologi, UIN Sunan Kalijaga Yogyakarta. Mata kuliah yang pernah diampu antara lain Metode Statistika, Analisis Data, Statistika Industri, Statistika Biologi, Statistika Psikologi, Statistika dan Probabilitas, Analisis Multivariat dan lain-lain.

Selain aktif mengajar dan meneliti, penulis juga berkecimpung dalam penjaminan mutu dan menjadi narasumber workshop kurikulum, sistem penjaminan mutu internal, penyusunan standar mutu dan lain-lain. Pengalaman dalam penjaminan mutu diperoleh dari tahun 2008-2012 saat menjadi Koordinator Data dan Statistika Unit Penjaminan Mutu dan tahun 2020-2024 sebagai Kepala Pusat Pengembangan Standar Mutu Akademik, Lembaga Penjaminan Mutu UIN Sunan Kalijaga Yogyakarta. Penulis memiliki pengalaman sebagai Asesor BAN-PT, Asesor LAMDIK dan Asesor

LAMSAMA. Penulis dapat dihubungin melalui email epha.supandi@uin-suka.ac.id

Akhmad Fauzy lahir di Tegal pada 8 Juli 1970. Pendidikan dasar hingga menengah diselesaikannya di Kabupaten Tegal, Jawa Tengah. Pendidikan tinggi ditempuh pada Program Studi Statistika, FMIPA Universitas Gadjah Mada, tempat ia meraih gelar Sarjana (S1) pada tahun 1994. Selanjutnya, gelar Magister (S2) diperoleh dari Program Studi Statistika, FMIPA Institut Pertanian Bogor pada tahun 1998. Pendidikan doktoral (S3) diselesaikannya di Jabatan Matematika, Fakultas Sains, Universiti Putra Malaysia pada tahun 2005.

Sejak tahun 1995, ia menjadi dosen tetap pada Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia (UII) Yogyakarta. Mata kuliah yang pernah diampu antara lain Statistika Nonparametrik, Metode Sampling, Statistika Industri, serta Metodologi Penelitian.

Dalam perjalanan karier akademiknya, Akhmad Fauzy aktif menulis berbagai buku teks dan buku referensi, khususnya di bidang statistika, metode penelitian, dan statistika industri. Karya-karyanya banyak digunakan sebagai rujukan oleh mahasiswa maupun praktisi. Beberapa buku yang telah diterbitkan antara lain *Statistik Industri*, *Metode Sampling*, dan *Pengantar Statistika Nonparametrik*.

Selain mengajar dan meneliti, ia juga aktif sebagai reviewer penelitian nasional, reviewer Pengabdian kepada Masyarakat (PkM), serta narasumber dalam berbagai kegiatan akademik dan profesional. Pengalaman struktural yang pernah diemban antara lain sebagai Dekan FMIPA UII, Direktur DPPM UII, serta anggota Dewan Eksekutif BAN-PT. Penulis dapat dihubungin melalui surel akhmad.fauzy@uii.ac.id.

ANALISIS MULTIVARIAT DENGAN R

Konsep dan Implementasi

Analisis multivariat merupakan salah satu teknik statistik yang digunakan untuk memahami struktur data dalam dimensi tinggi. Analisis multivariat memiliki kemampuan untuk memahami hubungan kompleks antar banyak variabel secara bersamaan. Kemampuan ini sangat relevan dalam dunia nyata karena masalah yang dihadapi sering kali melibatkan lebih dari dua variabel dan saling berhubungan. Oleh karena itu, mempelajari analisis multivariat adalah kunci untuk mengelola dan menafsirkan data yang kompleks dan membuat keputusan yang cerdas, menemukan peluang, serta mengatasi masalah di berbagai sektor.

Dalam analisis multivariat diperlukan suatu pengolah data yang *powerfull* karena sering melibatkan banyak variabel dan hubungan yang kompleks. Salah satu *software* statistik yang banyak digunakan saat ini adalah *software* R. Buku ini memberikan pengetahuan berbagai metode dalam analisis multivariat baik secara konsep dan implementasinya pada berbagai kasus dengan menggunakan bantuan *software* R. Oleh karena itu, untuk memahami buku ini, pengetahuan dasar *software* R sangat diperlukan.

SUKA PRESS
Sunan Kalijaga State Islamic University
Gedung KH. Abdul Wahab Hasbullah Lt. 3 UIN Sunan Kalijaga Yogyakarta
Jl. Laksda Adisucipto, Sleman, Daerah Istimewa Yogyakarta
suka.press.uin-suka.ac.id suka.press@uin-suka.ac.id Suka Press sukapress

ISBN 978-634-7179-10-4
9 786347 179104