Багатовимірна статистика

Зміст

Вступ

Наш курс складається з декількох частин. В першій частині ми будемо вивчати багатовимірні моделі лінійної регресії та пов'язані з ними задачі оцінювання параметрів та перевірки гіпотез про ці параметри. На відміну від одновимірного випадку, параметри в багатовимірній лінійній регресії – це вектори або матриці.

Другу частина курсу присвячено задачам класифікації та проблемі вибору оптимальної кількості важливих факторів, які із достатньою точністю описують нашу модель. А саме: ми розберемо, що таке дискримінантний аналіз, метод головних компонент, кластерний та факторний аналіз.

В нашому курсі ми будемо використовувати пакет R з тих міркувань, що мова R є простою для розуміння і для ілюстрації математичних моделей.

Курс лекцій базується на підручниках Алвіна Ренчера [R02] та Алвіна Ренчера в співавторстві з Брюсом Шаальє [RS08].

1 Лінійна алгебра. Нормальний розподіл. Відстань Махаланобіса

1.1 Відомості з лінійної алгебри

У цьому розділі ми пригадаємо деякі відомості з лінійної алгебри.

Нехай A - квадратна матриця розмірності $n \times n$. Існує обернена матриця A^{-1}, якщо $\operatorname{rank} A = n$; це означає, що A не сингулярна. Розглянемо деякі властивості.

1. Якщо матриці A та B не сингулярні та однієї розмірності, то
$$(AB)^{-1} = A^{-1}B^{-1}.$$

2. Якщо матриця B не сингулярна, то

$$AB = CB \quad \Longrightarrow \quad A = C.$$

3. Нехай A' - матриця транспоновона до A, тоді

$$(A')^{-1} = (A^{-1})'.$$

4. Матриця A є *додатно визначеною (відповідно, напіввизначеною)*, якщо

$$\forall x \in \mathbb{R}^n, \quad x \neq 0: \quad x'Ax > 0 \qquad (x'Ax \geq 0).$$

Діагональні елементи додатно визначеної (напіввизначеної) матриці A є додатніми (невід'ємними).

23p154 **Твердження 1.** *Якщо B є матрицею розмірності $n \times p$, $p \leq n$, то матриця $A = B'B$ є додатно визначеною.*

23p154 **Твердження 2.** *Якщо матриця A є додатно визначеною, то існує не сингулярна нижня трикутна матриця B така, що $A = B'B$.*

Останнє представлення має назву *розклад Холецького*, причому матриця B будується явно.

5. Позначимо через $|A| = \det A$ визначник матриці A. Тоді:

   - визначник добутку двох матриці дорівнює добутку визначників відповідних матриць:
   $$|AB| = |A||B|;$$
   - визначники матриці та відповідної їй транспонованої матриці рівні між собою:
   $$|A| = |A'|;$$
   - якщо матриця A така, що до неї існує обернена A^{-1}, то
   $$|A^{-1}| = \frac{1}{|A|}.$$

6. Матриця $A^{\frac{1}{2}}$ визначається як $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$. Тоді її визначник дорівнює

$$|A^{\frac{1}{2}}| = |A|^{\frac{1}{2}}.$$

   В більш загальному випадку, для матриці A^k, $k \geq 1$, маємо

$$|A^k| = |A|^k$$

7. *Власними числами* матриці A називаються корені характеристичного рівняння

$$|A - \lambda I| = 0.$$

   Власними векторами матриці A називаються такі вектори $v \neq 0$, що

$$Av = \lambda v.$$

23p154 **Твердження 3.** *Якщо матриця A розмірності $n \times n$ є додатно визначеною, то $\lambda_i > 0$, $i = 1, \ldots, n$. Якщо матриця A є додатно напіввизначеною, то $\lambda_i \geq 0$, та кількість додатніх λ_i дорівнює рангу A, тобто $\#\{\lambda_i \colon \lambda_i > 0\} = \operatorname{rank} A$.*

23p154 **Твердження 4.** *Власні вектори λ_i додатно визначеної симетричної матриці A є ортогональними.*

Нехай C - матриця, складена з власних вектрів v_i, $i = 1, \ldots, n$ матриці A:

$$C = (v_1, \ldots, v_n) = \begin{pmatrix} v_{11} & v_{12} & \ldots & v_{1n} \\ v_{21} & v_{22} & \ldots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \ldots & v_{nn} \end{pmatrix}.$$

Нехай матриця C є нормованою, тобто $CC' = I$. Тоді

$$A = ACC' = (Av_1, \ldots, Av_n)\, C' = (\lambda v_1, \ldots, \lambda v_n)\, C' = CDC',$$

де D - діагональна матриця вигляду

$$\begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix}.$$

Отже, отримали, що $A = CDC'$. Останнє представлення має місце для додатно визначеної симетричної матриці та називається спектральним розкладом матриці A.

9. Спектральний розклад дозволяє легко рахувати степені матриці. Наприклад, для матриці $A^{\frac{1}{2}}$ маємо

$$A^{\frac{1}{2}} = CD^{\frac{1}{2}}C',$$

де

$$D^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{\lambda_n} \end{pmatrix}.$$

1.2 Багатовимірний нормальний розподіл

Нехай випадковий вектор X має багатовимірний нормальний розподіл в $\mathbb{R}^p$ з вектором середніх $\boldsymbol{\mu}$ та матрицею коваріацій $\boldsymbol{\Sigma}$ (позначення: $X \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). Тоді його щільність розподілу має вигляд

$$g(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2}$$

23p154 **Означення 1.** Величина $(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) =: \Delta^2$ називається *відстаню Махаланобіса* між векторами $\boldsymbol{x}$ та $\boldsymbol{\mu}$ в $\mathbb{R}^p$.

В одновимірному випадку, тобто при $p = 1$, щільність розподілу має вигляд

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|x-\mu|^2}{\sqrt{2\pi\sigma^2}}},$$

а відстаність Махалонобіса $\Delta^2 = \frac{|x-\mu|^2}{\sqrt{2\pi\sigma^2}}$. Зафіксуємо її. Нехай $\Delta^2 = R^2$, тоді

$$|x-\mu|^2 = \sigma^2 R^2,$$

що геометрично означає коло радіуса $\sigma^2 R^2$.

Таким чином, якщо зафіксувати Δ^2, то можна оцінити ймовірність потрапляння елементів вибірки $(\xi_1, \ldots, \xi_n)$ в коло радіуса $\sigma^2 R^2$, тобто

$$\mathbb{P}\left(|\hat{\mu}_n - \mu|^2 \leq \sigma^2 R^2\right) = 1 - \alpha.$$

Аналогічно до того як в одновимірному випадку дисперсія σ^2 вказує на розкид значень навколо середнього μ, визначник коваріаційної матриці $|\mathbf{\Sigma}|$ вказує на розкид векторів $\boldsymbol{x}$ навколо $\boldsymbol{\mu}$ в p-вимірному просторі. Так, при малих значеннях $|\mathbf{\Sigma}|$ вектори $\boldsymbol{x}$ сконцентровані ближче до $\boldsymbol{\mu}$, і навпаки.

Зазначимо, що матриця $\mathbf{\Sigma}$ є симетричною (оскільки вона є матрицею коваріацій) та додатно визначеною, тому можемо записати її спектральний розклад

$$\mathbf{\Sigma} = CDC',$$

де C - матриця з власних векторів $\mathbf{\Sigma}$, а D - діагональна матриця з власних чисел $\mathbf{\Sigma}$. Обчислимо визнаник матриці $\mathbf{\Sigma}$.

$$|\mathbf{\Sigma}| = |CDC'| = |C||D||C'| = |CC'||D| = |D| = \prod_{k=1}^{p} \lambda_k.$$

Оскільки $\mathbf{\Sigma}$ додатно визначена, то $\lambda_k > 0, \quad k = \overline{1,p}$. Помітимо, що в такому разі малим значенням $|\mathbf{\Sigma}|$ відповідають малі значення власних чисел матриці $\mathbf{\Sigma}$.

Розглянемо властивості багатовимірного нормального розподілу.

lem12 **Лема 1.** *1. Нехай $a = (a_1, \ldots, a_p)$, тоді $a'X \sim N_p\left(a'\boldsymbol{\mu}, a'\mathbf{\Sigma}a\right)$.*

2. Якщо A - матриця розмірності $q \times p, q \leq p$, то $AX \sim N_q\left(A\boldsymbol{\mu}, A\mathbf{\Sigma}A'\right)$.

23p154 **Наслідок 1.** *Якщо $Z \sim N_p\left(0, I_p\right)$, тоді $X = \boldsymbol{\mu} + BZ \sim N_p\left(\boldsymbol{\mu}, \mathbf{\Sigma}\right)$, де $\mathbf{\Sigma} = BB'$ - розклад Холецького.*

Навпаки, маючи зображення $X = \boldsymbol{\mu} + BZ$, можна знайти Z:

$$X = \boldsymbol{\mu} + BZ \Rightarrow X - \boldsymbol{\mu} = BZ \Rightarrow Z = B^{-1}\left(X - \boldsymbol{\mu}\right).$$

Формально $B = \mathbf{\Sigma}^{\frac{1}{2}}$.

Сформулюємо інші властивості.

3. Нехай $Z \sim N_p\left(0, I_p\right)$, тоді

$$Z'Z = (z_1, \ldots, z_n) \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = \sum_{i=1}^{p} z_i^2 \sim \chi_p^2.$$

Вище було отримано, що $Z = B^{-1}\left(X - \boldsymbol{\mu}\right) = \mathbf{\Sigma}^{-\frac{1}{2}}\left(X - \boldsymbol{\mu}\right)$, тому

$$\begin{aligned} Z^T Z &= \left(\mathbf{\Sigma}^{-\frac{1}{2}}\left(X - \boldsymbol{\mu}\right)\right)^T \mathbf{\Sigma}^{-\frac{1}{2}}\left(X - \boldsymbol{\mu}\right) = \left(X - \boldsymbol{\mu}\right)^T \left(\mathbf{\Sigma}^{-\frac{1}{2}}\right)^T \mathbf{\Sigma}^{-\frac{1}{2}}\left(X - \boldsymbol{\mu}\right) \\ &= \left(X - \boldsymbol{\mu}\right)^T \mathbf{\Sigma}^{-1}\left(X - \boldsymbol{\mu}\right) = \Delta^2 \sim \chi_p^2. \end{aligned}$$

Якщо X - випадковий вектор розмірності p, то відстань Махаланобіса від X до $\boldsymbol{\mu}$ не перевищує $X_{p,\alpha}$ з ймовірністю $1-\alpha$:

$$\mathbb{P}\left(\left(X-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(X-\boldsymbol{\mu}\right)\leq X_{p,\alpha}\right)=1-\alpha,$$

де $X_{p,\alpha}$ квантиль рівня $1-\alpha$ розподілу χ^2_p.

4. Розглянемо вектори X в $\mathbb{R}^p$ та Y в $\mathbb{R}^q$ такі, що

$$\begin{pmatrix}X\\Y\end{pmatrix}\sim \mathrm{N}_{p+q}\left(\begin{pmatrix}\boldsymbol{\mu}_x\\\boldsymbol{\mu}_y\end{pmatrix},\begin{pmatrix}\boldsymbol{\Sigma}_{xx}&\boldsymbol{\Sigma}_{xy}\\\boldsymbol{\Sigma}_{yx}&\boldsymbol{\Sigma}_{yy}\end{pmatrix}\right).$$

У випадку, коли $p=q=1$, та $X\sim N\left(\mu_x,\sigma_x^2\right)$, $Y\sim N\left(\mu_y,\sigma_y^2\right)$, маємо

$$\begin{pmatrix}X\\Y\end{pmatrix}\sim \mathrm{N}_2\left(\begin{pmatrix}\mu_x\\\mu_y\end{pmatrix},\begin{pmatrix}\sigma_x^2&\rho\sigma_x\sigma_y\\\rho\sigma_x\sigma_y&\sigma_y^2\end{pmatrix}\right),$$

де $\rho=\frac{\mathrm{cov}(X,Y)}{\sqrt{\sigma_x^2\sigma_y^2}}$.

5. В одновимірному випадку, при $p=q=1$, регресією Y на X є умовне сподівання Y, якщо відоме X:

$$\mathbb{E}\left(Y|X\right)=\min_g\mathbb{E}\left(Y-g(X)\right)^2,$$

де функція $g=g(X)$ - борелева.

Функція регресії має вигляд

$$Y=\mu_y+\rho\frac{\sigma_y}{\sigma_x}\left(X-\mu_x\right).$$

Аналогічно в $\mathbb{R}^{p+q}$ регресією Y на X є

$$\mathbb{E}\left(Y|X\right)=\boldsymbol{\mu}_y+\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{xx}^{-1}\left(X-\boldsymbol{\mu}_x\right),$$

та регресією X на Y є

$$\mathbb{E}\left(X|Y\right)=\boldsymbol{\mu}_x+\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\left(Y-\boldsymbol{\mu}_y\right).$$

1.3 Приклади

Пригадаємо спочатку деякі базові операціх в R.

Задомо вектор x:

```
x<-c(1,2,3,4,5,6,7,8,9)   # вектор x
matrix(x,nrow=3)          # 3 рядки, заповеннння по стовбчикам
```

```
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

Якщо ви хочимо заповнювати матрицю по рядкам, то це можна зробити наступним чином:

```
matrix(x,nrow=3,byrow= TRUE)
```

Позначимо тепер цю матрицю через A, та порахуємо її детермінант:

```
A<-matrix(x,nrow=3)   # визначаємо матрицю A
det(A)                # порахуємо детермінант
```

На жаль, детермінант цієї матриці дорівнює 0.

Пригадаємо деякі операції з матрицями. Задамо також матрицю, детермінант якої не дорівнює 0.

```
x<-c(1,2,3,4,5,6,7,8,8)  # змінимо матрицю
B<-matrix(x,nrow=3)
det(B)                   # тепер вже не 0
t(B)                     # транспонована матриця
TrB<-sum(diag(B))        # слід матриці
A\%*\%B                  # добуток матриць
solve(B)                 # обернена до B
```

Також, пригадаємо, як в R можна згенерувати вибірку з первним (заданим) розподілом, знайти квантілі

Якщо ви щось забули, можна викликати справку за допомогою знаку питання ?, тобто наступний виклик дає посилання на відповідну сторінку документації:

```
?dnorm      #задати запитання, що таке dnorm
```

```
dnorm(x, mean = 0, sd = 1, log = FALSE)  # щільність в точці x нормального
розподілу з середнім 0 і стандартним відхиленням 1;

pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE) # значення
в точці x функції розподілу нормального розподілу з середнім 0 і стандартним
відхиленням 1;

qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)  # квантіль
рівня p нормального розподілу з середнім 0 і стандартним відхиленням 1.
```

Якщо обрати lower.tail = FALSE, то ми отримаємо квантіль рівня $1 - p$.

Для того, щоб вивести на екран перші 6 значень згенерованої вибірки, використаємо функцію head:

```
N<-rnorm(1000, 0, 1)
head(N)
```

```
-0.1535418  0.9394166  1.7203574 -1.6493087  0.3559709  0.1189281
```

Тобто, ми згенерували 1000 стандартних нормально розподілених випадкових величин, але вивели на екран 6.

Ще одна базова функція, яку ми будемо части використовувати, це hist, яка дозволяє побудувати гістограму розподілу. В залежності від потреби, можна побудувати гістограму частот, або гістограму гістограму щільності. Відповідно, у першому випадку по осі будуть відкладені абсолютні частоти, а у другому– відносні.

```
hist(x,breaks=150,xlim=c(0,20),freq=FALSE)   # графік щільності
hist(x,breaks=150,xlim=c(0,20),freq=TRUE)    # графік частот
```

Подивимось тепер, що відбувається в багатовимірному випадку, тобто коли у нас є нормальний випадковий вектор в $\mathbb{R}^p$. Для цього нас знадобляться наступні пакети і бібліотеки:

```
install.packages("mvtnorm")  # Завантажуємо пакети
install.packages("MASS")
library(mvtnorm)             # відкриваєте бібліотеки
library(MASS)
```

Задамо тепер вектор середніх значень, коваріаційну матрицю, і за допомогою побудуємо проекцію двовимірного розподілу на площину XOY.

```
mean<-c(0,0)                                  # вектор середніх
sig<-matrix(c(1, .5, .5, 1), nrow =2, byrow= FALSE)  # коваріація
mv<-rmvnorm(1000, mean, sig)                  # генеруємо нормальний розподіл з
цими середнім і коваріацією
mv.kde <- kde2d(mv[,1], mv[,2], n = 50)  # генеруємо щільність. kde = kernel
density estimation
image(mv.kde)                                 # малюємо картинку в 2d
contour(mv.kde, add = TRUE)
box()                                         # малює "бокс" навколо малюнку
title(main = "Проекція на XY", font.main = 4)
```

На Рисунку 1 наведені контури, які відповідають значенням щільності нормального розподілу, які спроектовані на $X0Y$.

Розглянемо тепер двовимірні довірчі інтервали. На площині двовимірним довірчим інтервалом буде еліпс. Для графічного зображення нам необхідно завантажити пакет *ellipse*.

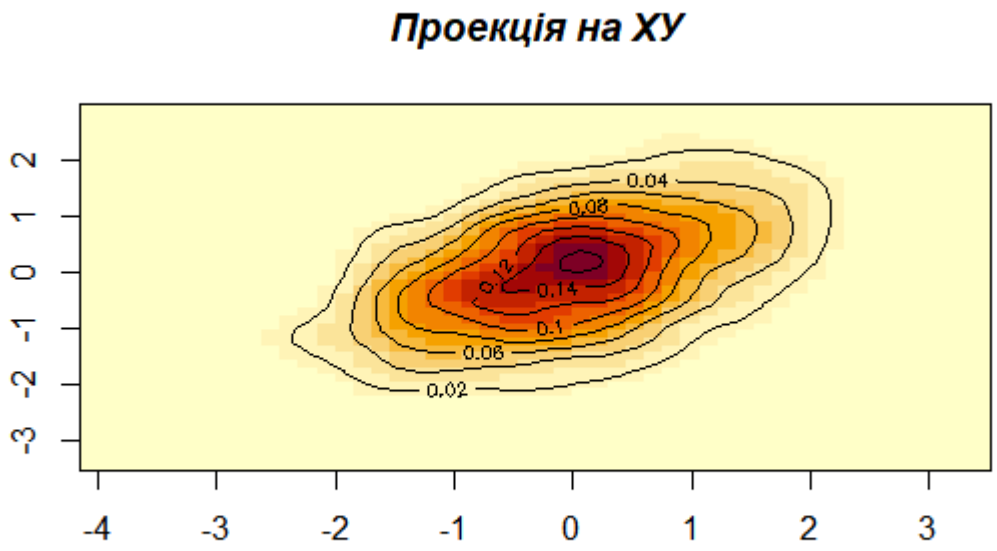


fig1-1

Рис. 1: Проекція двовимірної щільності розподілу на площину XOY

```
install.packages('ellipse')
library(ellipse)
rho <- cor(mv)
y_on_x <- lm(mv[,2] ~ mv[,1])  # регресія Y на X
x_on_y <- lm(mv[,1] ~ mv[,2])  # регресія X на Y
lines(ellipse(rho),  level = .95), col="red")
lines(ellipse(rho, level = .99), col="green")
lines(ellipse(rho, level = .90), col="blue")
abline(y_on_x, col="magenta") # малюємо лінію регресії y на x
abline(x_on_y, col="blue")    # малюємо лінію регресії x на y
```

На Рисунку 2 зеленим кольором позначено еліпс, який відповідає 99%-му двовимірному довірчому інтервалу (тобто точка має попасти в в середину еліпсу з ймовірністю 0.99). Відповідно, червоним та синім кольорами позначено 95%й та 90%й довірчі еліпси. Рожевим кольором зображена лінія регресії Y на X, та синім- X на Y. Зауважимо, що лінії регресії не співпадають з головними осями еліпсу!

Розглянемо більш делатьно вираз $(x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})$.

- Щільність розподілу в тих точках x, для яких має місце рівність $(x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) = c$ для деякої сталої c, приймає однакові значення;
- Значення $(x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})$ збільшується, якщо збільшується відстань між x та $\boldsymbol{\mu}$;
- Величина $d^2 := (x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})$ має χ^2 розподіл з p ступенями свободи.
- При фіксованому x величина d^2 називається відстанню Махаланобіса вектору x від вектору $\boldsymbol{\mu}$.

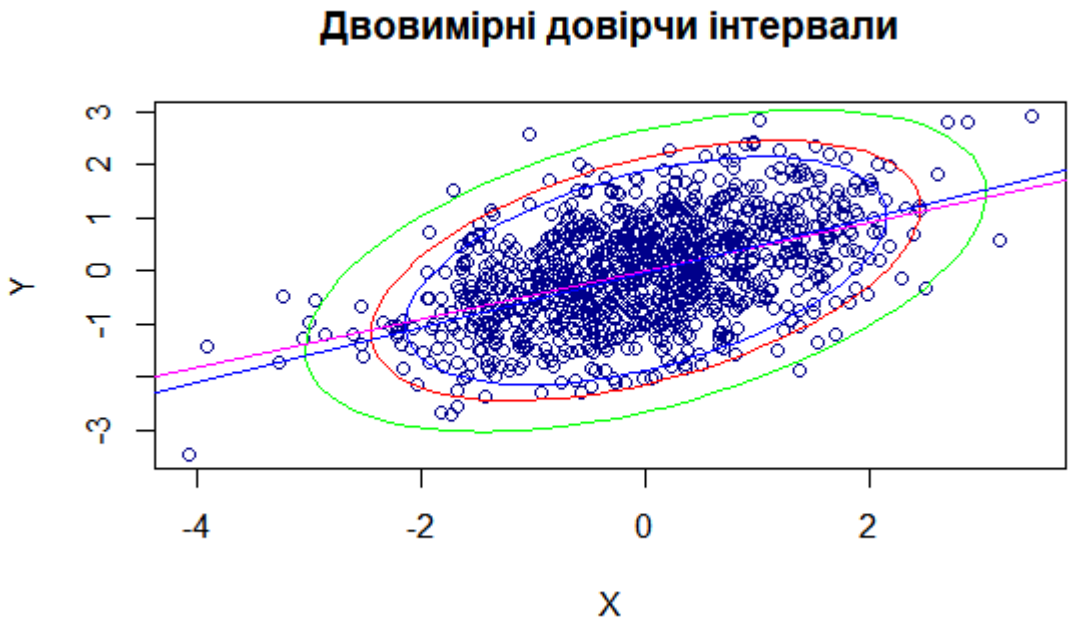


fig1-2

Рис. 2: Еліпсоїди розсіювання та лінії регресії

За допомогою відстанні Махаланобіса можна знайти ймовірність того, що випадкова величина X попаде в еліпсоїд $\{x : (x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) \le c\}$ з ймовірністю $1-\alpha$. Дійсно, нехай $\chi^2_{p,\alpha}$ – квантіль рівня $1-\alpha$. Тоді

$$\mathbb{P}((X-\boldsymbol{\mu})'\Sigma^{-1}(X-\boldsymbol{\mu}) \le \chi^2_{p,\alpha}) = 1-\alpha. \tag{1}$$

1-prb

2 Тести для перевірки нормальності розподілу

В наступних 2-х розділах ми будемо перевіряти гіпотези про вектори середніх значень та коваріаційні матриці. Але перш за все треба переконатися, що наша вибірка дійсно нормальна.

Розглянемо деякі тести на розподіл.

2.1 Тест Колмогорова-Смірнова (Kolmogorov-Smirnoff test)

Література: [K07].

Тест Колмоговора-Смірнова перевіряє, чи збігається функція розподілу окремих спостережень вибірки X з заданою (неперервною) функцією розподілу $F(x)$ на $\mathbb{R}$.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\xi_k<x}.$$

Тестова статистика має вигляд

$$\hat{\kappa}_n(X) = \sqrt{n}\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|.$$

Перевіряємо нульову гіпотезу

$$H_0: \quad \mathbb{P}(\xi_1 < x) = F(x) \quad \forall x \in \mathbb{R},$$

проти альтернативної

$$H_1: \quad \mathbb{P}(\xi_1 < x) = G(x) \quad \forall x \in \mathbb{R}, \quad G(x) \neq F(x), \quad G \in C(\mathbb{R}).$$

За умови гіпотези статистика Колмогорова $\hat{\kappa}_n$ слабко збігається до випадкової величини κ, яка має розподіл Колмогорова, тобто

$$\mathbb{P}(\kappa < x) = K(x) = \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2x^2).$$

Зауважимо, що це одновимірний тест, хоча існують узагальнення на багатовимірний випадок. Причина полягає у складності визначення у багатовимірному випадку емпіричної функції розподілу $\hat{F}_n$.

В R є вбудована функція, яка перевіряє гіпотезу H_0 за допомоги критерію Колмогорова-Смірнова:

```
ks.test(x, y, alternative = c("two.sided", "less", "greater"))
```

Тут x, y- вектори даних, причому y генерується за допомогою функції розподілу F:

```
y<-rnorm(50, -1, 1)
```

а можна задати розподіл, наприклад, $\Gamma(3, 2)$:

```
x<-rnorm(50)
ks.test(x, "pgamma", 3, 2)
```

2.2 Тест Шапіро-Уілка (Shapiro-Wilk test)

Література: [SW65] ($p = 1$), [VAG09] ($p \geq 1$), [AD54].

Тест Шапіро-Вілка є тестом одновимірної вибірки $p = 1$. Розглянемо впорядковану виборку $X = (x_1, \ldots, x_n)$. За умови нормальності маємо $X \sim \mathrm{N}_n(0, \mathbb{1}_n)$.

Розглянемо вектор середніх та коваріаційну матрицю:

$$m = \mathbb{E}X = (m_1, \ldots, m_n), \quad V = (\mathrm{cov}(x_i, x_j)_{i,j=1}^n.$$

Нехай тепер $y = (y_1, \ldots, y_n)$ - інша впорядкована виборка. За умови $y_i \sim \mathrm{N}(\mu, \sigma^2)$, координати можна зобразити наступним чином:

$$y_i = \mu + \sigma x_{(i)}, \quad i = 1, \ldots, n,$$

для деяких μ та σ^2. Оцінимо μ та σ^2 методом наймених квадратів, тобто мінімізуємо

$$(y - \mu_1 - \sigma m)' V (y - \mu_1 - \sigma m) \mapsto \min,$$

де $1 = (1, \ldots, 1)$. Отримаємо наступні оцінки:

$$\hat{\mu} = \frac{m'V^{-1}(m1' - 1'm)V^{-1}}{1'V^{-1}1m'V^{-1}m - (1'V^{-1}m)^2}$$

За умови симетричности, тобто $1'V^{-1}m = 0$, ці оцінки можна спростити та отримати

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}^2 = \frac{m'V^{-1}y}{m'V^{-1}m}.$$

Розглянемо W-статистику Шапіро-Уілка:

$$W = \frac{(\sum_{i=1}^{n} y_i a_i)^2}{\sum_{i=1}^{n}(y_i - y)^2},$$

де $a = (a_1, \ldots, a_n) = \frac{m'V^{-1}}{(m'V^{-1}V^{-1}m)^{1/2}}$. За умови нормальності вибірки X, розподіл цієї статистики залежить лише від n і є табульованим (хоча точне аналітичне зображення невідоме). Тому цю статистику можна використовувати для перевірки гіпотези про нормальність. В залежності від того, скільки точок попадає в "допустимий інтервал", таке буде значення статистики W. Критичними значеннями є такі значення емпіричної W, яка менша за табличне значення. У цьому випадку нульова гіпотеза відхиляється. Іншими словами, H_0 відхиляємо, якщо $W < c_{\alpha,n}$, де

$$\alpha = \mathbb{P}(W < c_{\alpha,n}|H_0).$$

В R тест Шапіро-Вілка реалізується, наприклад, так (для наглядності протестуємо на нормальній вибірці):

```
shapiro.test(x)
```

Тут x є вектором, наприклад, $x = rnorm(100, 0, 1)$.

Існує багатовимірний аналог тесту Шапіро-Уілка. Нехай $X = (X_1, \ldots, X_n)$ вибірка випадкових векторів з $\mathbb{R}^p$.

Розглядається нульова гіпотеза H_0: $X_i \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$, $i = 1, \ldots, n$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ - невідомі.

Нехай $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$. Розглянемо нормовані вектори

$$Z_j^* = S^{-1/2}(X_j - \bar{X}), \quad Z_j^* = (Z_{1j}, \ldots, Z_{pj}).$$

В якості узагальненої статистика Шапіро Уілкса можна взяти

$$W^* = \frac{1}{p}\sum_{i=1}^{p} W_{Z_i},$$

де W_{Z_i}- статистика Шапіро-Уілкса для (одновимірних!) спостепежень $(Z_{i1}, \ldots, Z_{in})$, $i = 1, \ldots, p$. Як і в одновимірному випадку, H_0 відхиляємо, якщо $W^* < c_{\alpha,p,n}$, де

$$\alpha = \mathbb{P}(W^* < c_{\alpha,p,n}|H_0).$$

Для перевірки нормальності випадкових векторів застосовується, наприклад, пакет *mvnormtest*, який потрібно встановити за завантажити відповідну бібліотеку. Для перевірки нормальності застосовується функція (тут $DataX$ у форматі матриці.)

```
mshapiro.test(t(DataX))
```

Існує багато різних тестів розподілу вибірки, наприклад тест Андерсона-Дарлінга (Anderson, Darling (1954), тест можна застосовувати також для інших розподілів), тест Маріди (Mardia (1970)), Генце і Циклера (Henze and Zirkler (1990)). Більш детально про ці тести можна прочитати в літературі, що наведена напочатку підрозділу.

3 Оцінки максимальної вирогідності (OMB) Тестування середнього

[PP12, (108)]

Нехай $X = (\xi_1, \ldots, \xi_n)$ – нормально розподілена вибірка розміру n з середнім $\boldsymbol{\mu}$ та дисперсією σ^2. Оцінки для параметрів $\boldsymbol{\mu}$ та σ можна отримати за допомогою методу максимальної вірогідності. Так, OMB для $\boldsymbol{\mu}$ є вибіркове середнє

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{i=1}^{n}\xi_i,$$

для σ^2 зміщеною оцінкою є

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \hat{\boldsymbol{\mu}}_n)^2$$

та незміщеною

$$\hat{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \hat{\boldsymbol{\mu}}_n)^2.$$

В багатовимірному випадку, тобто, коли $X \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, також можна знайти OMB для невідомого параметру $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Запишемо функцію вірогідності:

$$\begin{aligned}
L(X,\theta) &= \prod_{i=1}^{n} f(\xi_i,\theta) \\
&= \prod_{i=1}^{n} \frac{1}{(2\pi)^p|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\xi_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\xi_i-\boldsymbol{\mu})} \\
&= \frac{1}{(2\pi)^{np}|\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\xi_i-\boldsymbol{\mu}\pm\hat{\mu}_n)'\boldsymbol{\Sigma}^{-1}(\xi_i-\boldsymbol{\mu}\pm\hat{\mu}_n)} \\
&= \frac{1}{(2\pi)^{np}|\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\xi_i-\boldsymbol{\mu}\pm\hat{\mu}_n)'\boldsymbol{\Sigma}^{-1}(\xi_i-\boldsymbol{\mu}\pm\hat{\mu}_n)} \\
&= \frac{1}{(2\pi)^{np}|\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{n}(\xi_i-\hat{\mu}_n)'\boldsymbol{\Sigma}^{-1}(\xi_i-\hat{\mu}_n)+\sum_{i=1}^{n}(\xi_i-\hat{\mu}_n)'\boldsymbol{\Sigma}^{-1}(\hat{\mu}_n-\boldsymbol{\mu})\right]} \times \\
&\quad \times e^{-\frac{1}{2}\left[\sum_{i=1}^{n}(\hat{\mu}_n-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\xi_i-\hat{\mu}_n)+\sum_{i=1}^{n}(\hat{\mu}_n-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\hat{\mu}_n-\boldsymbol{\mu})\right]} \\
&= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}^{n/2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{n}(\xi_i-\hat{\mu}_n)'\boldsymbol{\Sigma}^{-1}(\xi_i-\hat{\mu}_n)+n(\hat{\mu}_n-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\hat{\mu}_n-\boldsymbol{\mu})\right]}.
\end{aligned}$$

Оскільки матриця $\boldsymbol{\Sigma}$ є додатно визначеною, то $\boldsymbol{\Sigma}^{-1}$ також додатно визначена, тому

$$(\hat{\mu}_n - \boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\hat{\mu}_n - \boldsymbol{\mu}) > 0.$$

Тоді

$$\begin{aligned} L(X,\theta) =& \frac{1}{(2\pi)^{np/2}|\mathbf{\Sigma}|^{n/2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{n}(\xi_i-\hat{\mu}_n)'\mathbf{\Sigma}^{-1}(\xi_i-\hat{\mu}_n)+n(\hat{\mu}_n-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\hat{\mu}_n-\boldsymbol{\mu})\right]} \\ \leq& \frac{1}{(2\pi)^{np/2}|\mathbf{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\xi_i-\hat{\mu}_n)'\mathbf{\Sigma}^{-1}(\xi_i-\hat{\mu}_n)}. \end{aligned}$$

Отже, максимум функції $L(X,\theta)$ досягається при $\hat{\mu}_n = \boldsymbol{\mu}$.

Для того, щоб знайти ОМВ для матриці коваріацій, вткористаємо наступні результати з лінійної алгебри.

t1 **Теорема 1.** *Нехай $X \in \mathbb{R}^p$, A - матриця розмірності $p \times p$, тоді*

$$X'AX = \operatorname{trace}[X'AX] = \operatorname{trace}[AXX'].$$

t2 **Теорема 2.** *Похідна сліду добутку двох матриць дорівнює*

$$\frac{\partial}{\partial A}\operatorname{trace}[AB] = B'.$$

Дійсно,

$$\frac{\partial}{\partial a_{ij}}\operatorname{trace}[AB] = \frac{\partial}{\partial a_{ij}}\sum_k\sum_l a_{kl}b_{lk} = b_{ji}.$$

t3 **Теорема 3.** *Похідна визначника матриці по цій матриці дорівнює*

$$\frac{\partial}{\partial A}|A| = |A|\left(A^{-1}\right)'.$$

Застосовуючи останній результат, можемо обчислити похідну від логарифму:

$$\frac{\partial}{\partial A}\ln|A| = \frac{1}{|A|}\frac{\partial}{\partial A}|A| = \frac{1}{|A|}|A|\left(A^{-1}\right)' = \left(A^{-1}\right)' = \left(A'\right)^{-1}.$$

Тепер ми можемо отримати ОМВ для коваріаційної матриці.

Використовуючи твердження Теореми 1 та підставляючи ОМВ для середнього $\boldsymbol{\mu} = \hat{\mu}_n$, маємо

$$\begin{aligned} L(X,\theta) =& \frac{1}{(2\pi)^{np/2}|\mathbf{\Sigma}|^{n/2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{n}(\xi_i-\hat{\mu}_n)'\mathbf{\Sigma}^{-1}(\xi_i-\hat{\mu}_n)+n(\hat{\mu}_n-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\hat{\mu}_n-\boldsymbol{\mu})\right]} \\ =& \frac{1}{(2\pi)^{np/2}|\mathbf{\Sigma}|^{n/2}} e^{-\frac{n}{2}\sum_{i=1}^{n} trace\left[\mathbf{\Sigma}^{-1}\frac{1}{n}(\xi_i-\hat{\mu}_n)(\xi_i-\hat{\mu}_n)'\right]} \\ =& \frac{1}{(2\pi)^{np/2}|\mathbf{\Sigma}|^{n/2}} e^{-\frac{n}{2}\operatorname{trace}\left[\mathbf{\Sigma}^{-1}S_n\right]}, \end{aligned}$$

де

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(\xi_i-\boldsymbol{\mu})(\xi_i-\boldsymbol{\mu})'.$$

Знайдемо максимум функції вірогідності $L(X,\theta)$, продиференціювавши її логарифм по $\mathbf{\Sigma}^{-1}$. За теоремами 2 та 3 отримаємо наступне:

$$\begin{aligned} \frac{\partial}{\partial\mathbf{\Sigma}^{-1}}\ln L(X,\theta) =& \frac{\partial}{\partial\mathbf{\Sigma}^{-1}}\ln\left(\frac{1}{(2\pi)^{np/2}|\mathbf{\Sigma}|^{n/2}} e^{-\frac{n}{2}\operatorname{trace}\left[\mathbf{\Sigma}^{-1}S_n\right]}\right) \\ =& \frac{n}{2}\frac{\partial}{\partial\mathbf{\Sigma}^{-1}}\ln\left|\mathbf{\Sigma}^{-1}\right| - \frac{n}{2}\frac{\partial}{\partial\mathbf{\Sigma}^{-1}}\operatorname{trace}\left[\mathbf{\Sigma}^{-1}S_n\right] \\ =& \frac{n}{2}\left[\left(\left(\mathbf{\Sigma}^{-1}\right)^{-1}\right)' - S_n'\right] = \frac{n}{2}\left[\mathbf{\Sigma} - S_n'\right] = 0. \end{aligned}$$

Оскільки матриця S_n є симетричною, то $S_n = S_n'$, тоді з останнього співвідношення $\mathbf{\Sigma} = S_n$. Таким чином, отримали наступні ОМВ для $\theta = (\boldsymbol{\mu}, \mathbf{\Sigma})$:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i, \quad \hat{\mathbf{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^{n} (\xi_i - \hat{\mu}_n)(\xi_i - \hat{\mu}_n)'.$$

3.1 Розподіли $\hat{\mu}_n$ та S_n

В одновимірному випадку $\hat{\mu}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

$$\mathbb{E}\hat{\mu}_n = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = \mu, \quad \mathbb{D}\hat{\mu}_n = \mathbb{D}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}\xi_i = \frac{\sigma^2}{n}.$$

Аналогічно, для довільного $p > 1$

$$\hat{\mu}_n \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\mathbf{\Sigma}\right).$$

Для дисперсії σ^2 в одновимірному випадку маємо:

$$\sum_{i=1}^{n}\left(\frac{\xi_i - \mu}{\sigma}\right)^2 \sim \chi^2_n.$$

Аналогічно, в багатовимірному випадку

$$\sum_{i=1}^{n}(\xi_i - \boldsymbol{\mu})(\xi_i - \boldsymbol{\mu})' \sim W_p(n, \mathbf{\Sigma}),$$

де $W_p(n, \mathbf{\Sigma})$ – розподіл Уішарта з n ступенями свободи та відомою матрицею коваріацій.

Якщо μ – невідоме, то при $p = 1$

$$\sum_{i=1}^{n}\left(\frac{\xi_i - \hat{\mu}_n}{\sigma}\right)^2 \sim \chi^2_{n-1}$$

та при $p > 1$

$$\sum_{i=1}^{n}(\xi_i - \hat{\mu}_n)(\xi_i - \hat{\mu}_n)' \sim W_p(n-1, \mathbf{\Sigma}).$$

Отже, $nS_n = \sum_{i=1}^{n}(\xi_i - \hat{\mu}_n)(\xi_i - \hat{\mu}_n)' \sim W_p(n-1, \mathbf{\Sigma})$.

23p154 **Лема 2.** (Фішера) *Нехай $X \sim N(\mu, \sigma^2)$, тоді*
1) вибіркове середнє $\hat{\mu}_n$ та вибіркова дисперсія $\hat{s}^2_n$ – незалежні;
2) нормоване відношення вибіркового середнього $\hat{\mu}_n$ та вибіркової дисперсії $\hat{s}^2_n$ має розподіл Стьюдента з $n-1$ ступенями свободи:

$$\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sqrt{\hat{s}^2_n}} \sim t_{n-1}.$$

23p154 **Лема 3.** (Фішера) *Нехай $X \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, тоді*
1) вибіркове середнє $\hat{\mu}_n$ та вибіркова матриця коваріацій S_n – незалежні,

$$nS_n \sim W_p(n-1, \mathbf{\Sigma}), \quad \hat{\mu}_n \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\mathbf{\Sigma}\right);$$

2) нормоване відношення вибіркового середнього $\hat{\mu}_n$ та вибіркової матриці коваріацій S_n має розподіл Хоттелінга з $n-1$ ступенями свободи:

$$\frac{\sqrt{n}(\hat{\mu}_n - \boldsymbol{\mu})}{\sqrt{S_n}} \sim T^2.$$

$$\frac{\partial}{\partial X} X'AX = \frac{\partial}{\partial X} \operatorname{trace}(X'AX) = BX + B'X. \tag{2}$$ der2

$$L(X, \hat{\mu}, \Sigma) = (2\pi)^{-nd/2} |S|^{-n/2} \exp\big(-\frac{n}{2} \operatorname{trace}(S\Sigma^{-1}))\big). \tag{3}$$ OMV-1

4 Перевірка гіпотези про середнє

У цьому розділі ми будемо перевіряти гіпотезу про середнє значення вектору нормального розподілу у випадку, коли коваріаційна матриця а) відома б) невідома. Також, ми перевіримо гіпотезу про те, що середні двох виборок співпадають.

4.1 Гіпотеза про середнє. Відома коваріаційна матриця

Нехай є нормальна вибірка $X \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, де відома коваріаційна матриця. Перевіримо гіпотезу H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ проти альтернативи H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Зауважимо, що тут $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}$ є векторами, тобто H_0 виконано, коли всі координати вектора $\boldsymbol{\mu}$ співпадають з координатами вектора $\boldsymbol{\mu}_0$, а H_1- принаймні 2 координати відрізняються.

Використаємо наступну статистику:

$$Z^2 = n(\bar{X} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{X} - \boldsymbol{\mu}_0) = n\Delta^2 \tag{4}$$ Z

(нагадаємо, що є квадратом відстані Махаланобіса, а).

За умови виконанння H_0,

$$Z^2 \sim \chi^2_p. \tag{5}$$ Z2

Якщо $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, то значення Δ^2 великі, оскільки середнє $\bar{X}$ буде відрізнятися від $\boldsymbol{\mu}_0$, а отже, Z^2 велике, іншими словами, критични ми значеннями є великі значення Z^2. Отже, алгоритм перевірки гіпотези H_0 є наступним.

- Обчислюємо Z^2;
- Обчислюємо квантіль χ^2_p розподілу рівня α, тобто $\chi^2_{\alpha,p}$, де α задане, наприклад, $\alpha = 0.05$.
- Якшо $Z^2 > \chi^2_{\alpha,p}$, то відхиляємо H_0.

4.2 Гіпотеза про середнє. Невідома коваріаційна матриця

Нехай тепер матриця $\boldsymbol{\Sigma}$ невідома. У цьому випадку використаємо для перевірки гіпотези H_0 статистику Хотеллінга:

$$T^2 := n(\bar{X} - \boldsymbol{\mu}_0)'\tilde{S}^{-1}(\bar{X} - \boldsymbol{\mu}_0), \tag{6}$$ T2

$$\tilde{S} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})'(X_i - \bar{X}).$$

За цмови виконання H_0,

$$T^2 \sim T_{p,n-1} \tag{7}$$ T21

де $T_{p,n-1}$ є розподілом Хотеллінга з p та $n-1$ ступенями свободи. Так само, критичними значеннями статистики є велиці значення.

Алгоритм перевірки H_0 є наступним.

- Обчислюємо T^2.

- ОБчислюємо квантіль $T^2_{\alpha,p,n-1}$ рівня α розподілу $T^2_{p,n-1}$.
- Якщо $T^2 > T^2_{\alpha,p,n-1}$, то відхиляємо H_0.

Розглянемо ще один варіант гіпотези H_0, а саме,

H_0: $\mu_1 = \mu_2 = \cdots = \mu_p$, проти альтернативи H_1 $\exists i, j$: $\mu_i \neq \mu_j$. У векторному випадку, гіпотезу H_0 можна сформулювати наступним чином:

$$H_0: \qquad \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \\ \cdots \\ \mu_{p-1} - \mu_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}.$$

Альтернативно, гіпотезу H_0 можна задати наступним чином:

$$H_0: \qquad \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \cdots \\ \mu_1 - \mu_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}$$

Позначимо $J = \underbrace{(1, \ldots, 1)}_{p}$. Тоді гіпотезу H_0 можна сформулювати наступним чином: $CJ = 0$, де C така матриця, сума в кожному рядку якої дорівнює 0 та $\operatorname{rank} C = p - 1$. Така матриця називається матрицею контрастів. Таке перетворення можна також використати при перевірці гіпотези про коваріаційну матрицю, див.

4.3 Гіпотеза про рівність середніх двох виборок

Перевіримо гіпотезу про рівність середніх 2х виборок. Для простоти розглянемо спочатку одновимірний випадок.

Нехай є 2 виборки $Y_1 = (y_{11}, y_{12}, \ldots, y_{1n_1})$ та $Y_2 = (y_{21}, y_{22}, \ldots, y_{2n_2})$, $y_{1i} \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $1 \leq i \leq n_1$, $y_{1i} \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $1 \leq j \leq n_2$. Припустимо, що виборки незалежні та $\sigma_1^2 = \sigma_2^2 = \sigma^2$ невідоме.

Оскільки виборки можуть бути різного розміру, розглянемо зважену коваріаційну матрицю (Eng.: pooled covariance):

$$s^2_{pl} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_i^2 := \sum_{k=1}^{n_i} \frac{(y_{ik} - \bar{y}_i)^2}{n_i - 1}, \quad \bar{y}_i := \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}, \quad i = 1, 2.$$

Тоді s^2_{pl} –незміщена оцінка дисперсії σ^2: $\mathbb{E}s^2_{pl} = \sigma^2$.

Перевіримо основну гіпотезу H_0: $\mu_1 = \mu_2$ проти альтернативи H_1: $\mu_1 \neq \mu_2$.

$$t = \frac{\bar{y}_1 - \bar{y}_2}{S_{pl}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

За умови H_0, ця статистика має розподіл Стьютедта $t_{n_1+n_2-2}$ з $n_1 + n_2 - 2$ ступенями свободи. Тому для перевірки гіпотези можна використати тестову статистику $|t|$. Якщо $|t| > t_{\alpha/2,n-1}$, де $t_{\alpha/2,n-1}$ квантіль рівня $1 - \alpha/2$ розподілу Стьюдента з $n - 1$ ступенем свободи, то H_0 відхиляємо.

Альтернативно, використовуючи наступне співвідношення між розподілом Стьюдента та розподілом Фішера $t^2_{n-1} = F_{1,n-1}$, можна перевіряти гіпотезу H_0, обчисливши відповідну статистику та знайшовши квантіль рівня розподілу Фішера

Розглянемо багатовимірний випадок. Тепер елементи y_{ik} виборок Y_1, Y_2 – це вектори, які мають нормальний розподіл $\mathrm{N}_p(\mu_i, \Sigma_i)$, $i = 1, 2$, відповідно. Припустимо, що коваріційні матриці співпадають: $\Sigma_1 = \Sigma_2$ (трохи згодом ми розглянемо, як перевіряти гіпотезу про рівність коваріаційних матриць).

Ми перевіримо основну гіпотезу H_0: $\mu_1 = \mu_2$ проти альтернативи H_1: $\mu_1 \neq \mu_2$. Позначимо через

$$S_i := \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_{ik})(Y_{ik} - \bar{Y}_{ik})', \quad i = 1, 2,$$

вибіркові коваріційні матриці у першій і другів вибірці, відповідно.

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

Тестовою статистикоб є статистика Хотеллінга

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{Y}_1 - \bar{Y}_2)' S_{pl}^{-1} (\bar{Y}_1 - \bar{Y}_2).$$

Зауважимо, що вибіркова коваріційна матриця завжди додатньо напіввизначена, а отже, статистика T^2 додатня. Якщо справедлива гіпотеза H_0, то статистика T^2 має розподіл Хотеллінга T^2_{p,n_1+n_2-2} з n_1 та n_2 ступенями свободи. Ми відхиляємо гіпотезу H_0, якщо $T^2 < T^2_{\alpha,p,n_1+n_2-2}$, де T^2_{α,p,n_1+n_2-2} квантіль розподілу T^2_{p,n_1+n_2-2} рівня $1 - \alpha$.

Як і в одновимірному випадку, можна перевіряти гіпотезу H_0, використовуючи наступне співвідношення між статистикою Хотеллінга та розподілом Фішера (див. [R02, (5.7)]):

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2_{p,n_1+n_2-2} = F_{p,n_1+n_2-p-1}.$$

Якщо гіпотеза H_0 не має місце для векторного випадку, тоді має сенс перевірити одновимірний варіант цієї гіпотези покоординатно.

5 Перевірка гіпотези про коваріаційну матрицю

covM

5.1 Статистика відношння вирогідностей. Тест про сферичність

Ми скористаємося відношенням вирогідностей (3).

Підставивиши в (3) $\hat{\mathbf{\Sigma}} = S$, отримаємо

$$L(X, \hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}) = (2\pi)^{-np/2} |S|^{-n/2} \exp\left(-\frac{np}{2})\right).$$

З іншого боку, підставимо в (3) $\mathbf{\Sigma} = \mathbf{\Sigma}_0$ і розглянемо відношення правдоподібності (likelihood relation)

$$LR := \frac{L(X, \hat{\boldsymbol{\mu}}, S)}{L(X, \hat{\boldsymbol{\mu}}, \mathbf{\Sigma}_0)} = \left(\frac{|\mathbf{\Sigma}_0|}{|S|} \exp\left(\mathrm{trace}(S\mathbf{\Sigma}_0^{-1}) - p\right) \right)^{\frac{n}{2}}. \tag{8}$$

LR2

Розглянемо модифіковані статистики відношення вирогідностей:

$$U := (n - 1) \left[\ln |\mathbf{\Sigma}_0| - \ln |S| + \mathrm{trace}(S\mathbf{\Sigma}_0^{-1}) - p \right],$$

$$\tilde{U} := (n - 1) \left[\ln |\mathbf{\Sigma}_0| - \ln |S| + \mathrm{trace}(S\mathbf{\Sigma}_0^{-1}) - p \right],$$

Сформулюємо без доведення наступні твердження (див. доведення в ????)

sigma1 **Твердження 5.** *Припустимо, що гіпотеза* $H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_0$ *має місце.*

- *Для великих* n *статистика* U *має наближено* $\chi^2_{\frac{p(p+1)}{2}}$ *розподіл.*
- *Для вибірку середнього розміру* n *статистика* $\tilde{U}$ *має наближено* $\chi^2_{\frac{p(p+1)}{2}}$ *розподіл.*

Розглянемо частковий випадок, а саме, гіпотезу про сферичність:

H_0: $\Sigma = \sigma^2 I_p$ проти альтернативи H_1: $\Sigma \neq \sigma^2 I_p$. Тут σ^2 є невідомою, тобто H_0- це гіпотеза про форму матриці Σ. При цьому, якщо $(X-\mu)'\Sigma^{-1}(X-\mu) = c^2$, де $X = (\xi_1, \ldots, \xi_p)$, є рівнянням елліпсу, то за гіпотези сферичності маємо $(X-\mu)'(X-\mu) = \sum_{i=1}^d (\xi_i - \mu_i)^2 = c^2\sigma^2$, що є рівнянням еліпсу.

Знайдемо за умови H_0 відношення вирогідностей. Якщо $\Sigma = \sigma^2 I_p$, то $\operatorname{trace}(S\Sigma^{-1}) = \sigma^{-2}\operatorname{trace} S$, а $|\Sigma| = (\sigma^2)^p = (\operatorname{trace}\Sigma/p)^p$. За гіпотези H_0, $\operatorname{trace}(S) \approx p\sigma^2$, тому

$$\exp\left(\operatorname{trace}(S\Sigma_0^{-1}) - p\right) = \exp\left(\sigma^{-2}\operatorname{trace}(S) - p\right) \approx 1.$$

Тому розглядаємо відношення вирогідностей у вигляді

$$LR = \left(\frac{(\operatorname{trace}(S)/p)^p}{|S|}\right)^{\frac{n}{2}}.$$

Визначимо тестову статистику u наступним чином:

$$2\ln LR = n\ln\left(\frac{(\operatorname{trace}(S)/p)^p}{|S|}\right) = -\ln u,$$

або

$$u = \frac{p^p|S|}{(\operatorname{trace}(S))^p}$$

Також розглянемо модифікацію цієї статистики:

$$u' = -\left(n - 1 - \frac{2p^2+p+2}{6p}\right)\ln u$$

sigma2 **Твердження 6.** *Припустимо, що гіпотеза* H_0 *має місце.*

- *Для великих* n *статистика* u *має наближено* $\chi^2_{\frac{p(p-1)}{2}}$ *розподіл.*
- *Для вибірку середнього розміру* n *статистика* u' *має наближено* $\chi^2_{\frac{p(p+1)}{2}-1}$ *розподіл.*

Статистику було вперше отримано в робоі [Ma40], тому відповідний статистичний тест носить назву тест Маулчи (Maulchy test).

Як і в розділі „, можна перевіряти гіпотезу про коваріаційну матрицю не для самої вибірки X, а для її перетворення. Розглянемо наступне перетворення X: $Z = CX$. При цьому Z має розмірність $p-1$, $Z = (z_1, \ldots z_p) \sim \mathrm{N}_{p-1}(0, C\mathbf{\Sigma}C')$, $\bar{Z} = C\bar{X}$, а матриця $S_Z = CSC'$ має розмірність $(p-1)\times(p-1)$. Маємо:

$$\bar{Z} \sim \mathrm{N}_{p-1}\left(0, \frac{C\mathbf{\Sigma}C'}{n}\right), \quad T^2 = (C\bar{X})'\left(\frac{CSC'}{n}\right)(CX).$$

Тоді можна перевіряти гіпотезу H_0: $C\Sigma C' = \sigma^2 I_{p-1}$, проти альтернативи H_1: $C\Sigma C' \neq \sigma^2 I_{p-1}$. Тестовою статистикою для цього є вищенаведена статистика T^2.

5.2 Тест на рівність коваріаційних матриць.

Розглянемо спочатку одновимірний випадок. В одновимірному випадку, коли є 2 вибірки ($k = 2$) розміру n_1 та n_2 відповідно, розглянемо основну гіпотезу

H_0: $\sigma_1 = \sigma_2$ проти альтернативи H_1: $\sigma_1 \neq \sigma_2$.

Тестовою статистикою у цьому випадку є [K07]

$$F = \frac{s_1^2}{s_2^2},$$

де s_1^2 та s_2^2 є незміщеними оцінкам дисперсій. Нагадаємо, що s_1 та s_2 є незалежними випадковими величинами (оскільки серії спостережень є незалежними). За умови H_0 статистика F має розподіл Фішера F_{n_1-1,n_2-1} з $n_1 - 1$ та $n_2 - 1$ ступенями свободи.

У випадку, коли є k серій незалежних спостережень, для перевірки гіпотезі

$$H_0 = \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2,$$

використовується тест Барлетта однорідності.

Для перевірки гіпотози H_0 використовується статистика

$$\frac{m}{c} \approx \chi^2_{k-1},$$

де

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{\nu_i} - \frac{1}{\sum_{i=1}^{k}\nu_i}\right], \quad m = \ln\left[\frac{(s^2)^{\sum_{i=1}^{k}\nu_i}}{(s_1^2)^{\nu_1}(s_2^2)^{\nu_2}\dots(s_k^2)^{\nu_k}}\right], \tag{9}$$

$$s^2 = \frac{\sum_{i=1}^{k}\nu_i s_i^2}{\sum_{i=1}^{k}\nu_i}, \quad \nu_i = n_i - 1, \quad i = 1, \dots, k. \tag{10}$$

Якщо H_0 справедлива, то дріб під знаком логарифму у зображенні m близький до 1, тому критичними значеннями статистики $\frac{m}{c}$ є великі значення. Ми відхиляємо гіпотезу H_0, якщо $\frac{m}{c} > \chi^2_{\alpha,k-1}$, де $\chi^2_{\alpha,k-1}$ квантіль рівня $1-\alpha$ розподілу χ^2_{k-1}.

Справедливість гіпотези H_0 можна також перевіряти за допомогою розподілу Фішера. Для цього статистику m треба перетворити.

Нехай

$$a_1 = k - 1, \quad a_2 = \frac{k+1}{(c-1)^2}, \quad b = \frac{a_2}{2 - c + 2/a_2}.$$

Тоді статистика має наближено розподіл Фішера F_{a_1,a_2} з a_1 та a_2 ступенями свободи:

$$F := \frac{a_2 m}{a_1(b-m)} \approx F_{a_1,a_2}.$$

Критичними значеннями статистики F є також великі значення, тому ми відхиляємо H_0, якщо $F > F_{\alpha,a_1,a_2}$, де F_{α,a_1,a_2} є кавнтілем рівня $1-\alpha$ розподілу F_{a_1,a_2}.

Існують і інші тести на рівність дисперсій. Тест Левена (Levene test [?], див. також с.130) є більш стійким до відхилення від нормального розподілу.

В багатовимірному випадку ситуація схожа, але для перевірки основної гіпотези H_0 ми порівнюємо детермінанти матриць вибіркових коваріацій S_i, $i = 1, \dots, k$.

Нехай є k виборок нормального розподілу, причому i-та вибірка $\mathrm{N}_p(\mu_i, \Sigma_i)$ має розмір n_i. Розглянемо основну гіпотезу

H_0: $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ H_1: $\exists i, j$: $\Sigma_i = \Sigma_j$.

Розглянемо наступну статистику:

$$M := \frac{|S_1|^{\nu_1/2}|S_2|^{\nu_2/2}\dots|S_k|^{\nu_k/2}}{|S_{pl}|^{\sum_{i=1}^{k}\nu_i/2}},$$

де $\nu_i = n_i - 1$ та $\nu_E = \sum_{i=1}^{k}\nu_i = \sum_{i=1}^{k} n_i - k$.

Теорема 4. *Тест Бокса (Box M-method test) Нехай*

$$c_1 := \left[\sum_{i=1}^{k} \frac{1}{\nu_i} - \frac{1}{\sum_{i=1}^{k} \nu_i}\right] \left(\frac{2p^2+3p-1}{6(p+1)(k-1)}\right).$$

Тоді наближено

$$-2(1-c_1)\ln M \sim \chi^2_{\frac{1}{2}(k-1)p(p+1)}. \tag{11}$$

У випадку, коли $\nu_1 = \nu_2 = \ldots \nu_k = \nu$, константа виглядає наступним чином:

$$c_1 = \frac{(k+1)(2p^2+3p-1)}{6k\nu(p+1)}.$$

5.3 Тест про незалежність двох субвекторів

cov3

Нехай вектор спостережень рокпадається на 2 групи y та x розмірністю $p \times 1$ та $q \times 1$, відповідно. При цьому коваріаційну та вибіркову коваріаційну та кореляційну матриці Σ, S та R можна розкласти наступним чином:

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{yy} & \Sigma_{yx} \\ \hline \Sigma_{xy} & \Sigma_{xx} \end{array}\right), \qquad S = \left(\begin{array}{c|c} S_{yy} & S_{yx} \\ \hline S_{xy} & S_{xx} \end{array}\right), \qquad R = \left(\begin{array}{c|c} R_{yy} & R_{yx} \\ \hline R_{xy} & R_{xx} \end{array}\right) \tag{12}$$

RS

Перевіримо нульову гіпотезу H_0:

$$\Sigma = \Sigma_0 = \left(\begin{array}{c|c} \Sigma_{yy} & 0 \\ \hline 0 & \Sigma_{xx} \end{array}\right) \tag{13}$$

block

проти альтернативи, що коваріаційна матриця не має блокової структури (14). Основна гіпотеза означає, що вектори X і Y є некорельованими, а отже, за припущення нормальності, незалежними.

Розглянемо статистику (8) з Σ_0,

$$S_0 = \left(\begin{array}{c|c} S_{yy} & 0 \\ \hline 0 & S_{xx} \end{array}\right) \tag{14}$$

block

За нульової гіпотези експонента в статистиці (8) має порядок $1 + o(1)$, тому

$$LR \approx \left(\frac{|S_{xx}||S_{yy}|}{|S|}(1+o(1))\right)^{\frac{n}{2}}.$$

Виходячи з цього, розглянемо статистику Уілкса

$$\Lambda := \frac{|S|}{|S_{xx}||S_{yy}|} = \frac{|R|}{|R_{xx}||R_{yy}|}.$$

Детермінант матриці можна перетворити наступним чином:

$$|S| = |S_{xx}(S_{yy} - S_{yx}S_{xx}^{-1}S_{xy})| = |S_{xx}||S_{yy} - S_{yx}S_{xx}^{-1}S_{xy}|,$$

а отже, статистику можна перетворити наступним чином:

$$\Lambda = \frac{|E|}{|E+H|}, \tag{15}$$

Lam

де

$$E := S_{yy} - S_{yx}S_{xx}^{-1}S_{xy}, \quad H := S_{yx}S_{xx}^{-1}S_{xy},$$

є матрицями похибок (Error matrix) та гіпотези (Hypothesis matrix). Якщо матриця H є близькою до нульової матриці, то $\Lambda \asymp 1$. Отже, "хорошими значеннями" Λ є значення, близькі до 1.

За умови H_0 статистика має розподіл Уілкса, який визначається наступним чином. Нехай є два незалежних розподіли Уішарта $A \sim W_p(\Sigma, m)$ та $B \sim W_p(\Sigma, n)$, та $m \geq p$. Тоді розподіл випадкової величини

$$\lambda = \frac{|A|}{|A+B|} = \frac{1}{|I_p + A^{-1}B|} \sim \Lambda(p, m, n).$$

називається розподілом Уілкса p, m та n. Можна порівнювати значення статистики Уілкса зі значеннями квантілю рівня розподілу Уілклса, але на практиці це не так просто зробити, якщо великі. За умови H_0 статистику Λ можна апроксимувати за допомоги розподілу Фішера $F_{a,b}$ з певними параметрами a і b, а саме [R02, (6.15)]:

$$F = \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{df_2}{df_1} \sim F_{df_1, df_2}, \tag{16}$$ F

де p - кількість змінних, ν_H -кількість ступенів свободи матриці H, ν_E кількість ступенів свободи матриці E,

$$df_1 = p\nu_H, \quad df_2 = wt - \frac{1}{2}(p\nu_H - 2), \tag{17}$$ df

$$w = \nu_E + \nu_H - \frac{1}{2}(p + \nu_H + 1), \quad t = \sqrt{\frac{p^2\nu_H^2 - 4}{p^2 + \nu_H^2 - 5}}.$$

5.4 Приклади

6 ANOVA: одновимірний дисперсійний аналіз (однофакторна модель)

ANOVA

(Univariate One-Way Analysis of Variance)

Література: [Ma07, §2.5], [R02, Sec.6]

6.1 ANOV1 ANOVA модель

Нехай є k груп незалежних спостережень нормально розподілених випадкових величин. Дисперсії в кожній групі вважаємо однаковими і рівними σ^2, середені значення є невідомими. Задамо ці спостерження в Таблиці 1. Тут $\bar{y}_{i\bullet} = \sum_{j=1}^{n} y_{ij}/n$. Ми застосуємо одновимірний дисперсійний аналіз (ANOVA) для того, щоб перевірити гіпотезу H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ проти альтернативи H_1: $\mu_i \neq \mu_j$ для деяких i, j. Якщо має місце H_0, то всі дані y_{ij} належать одній популяції. Для оцінювання σ^2 можна застосувати 2 типи оцінок: перший базується на застосуванні вибіркових дисперсій s_1^2, s_2^2, ..., s_k^2 та є їх середнім

$$s_e^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} \frac{(y_{ij} - \bar{y}_{i\bullet})^2}{k(n-1)} =: \frac{SSE}{k(n-1)}, \tag{18}$$ SSE

Табл. 1: ANOVA

	Вибірки			
	$N(\mu_1,\sigma^2)$	$N(\mu_2,\sigma^2)$	$\dots$	$N(\mu_k,\sigma^2)$
	y_{11}	y_{12}	$\dots$	y_{1k}
	y_{21}	y_{22}	$\dots$	y_{2k}
	$\dots$	$\dots$	$\dots$	$\dots$
	y_{n1}	y_{n2}	$\dots$	y_{nk}
Сумма	$y_{1\bullet}$	$y_{2\bullet}$	$\dots$	$y_{k\bullet}$
Середнє	$\bar y_{1\bullet}$	$\bar y_{2\bullet}$	$\dots$	$\bar y_{k\bullet}$
Дисперсія	s_1^2	s_1^2	$\dots$	s_k^2

ANOVA

а інший– на застосуванні вибіркової дисперсії

$$s_{\bar y}^2 := \sum_{i=1}^k \frac{(\bar y_{i\bullet} - \bar y_{\bullet\bullet})^2}{k-1} =: \frac{SSH}{k-1}, \quad \bar y_{\bullet\bullet} = \sum_{i=1}^k \frac{\bar y_{i\bullet}}{k}. \tag{19}$$

SSH1

Припустимо, що наші дані можна зобразити у вигляді

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1,\dots,k, \quad j = 1,\dots,n, \tag{20}$$

reg1

де $\epsilon_{ij} \sim \mathrm{N}(0,\sigma^2)$ є незалежними, а α_i, $1 \le i \le k$, є константами. Тобто, на y_{ij} впливає один фактор, який відображається у значенні α_i, Тому така модель називається **однофакторною**. Модель (20) можна переписати у наступному вигляді:

$$Y = X\beta + \mathcal{E}, \tag{21}$$

reg2

або

$$\begin{pmatrix} y_{11}\\ y_{12}\\ \dots\\ y_{21}\\ y_{22}\\ \dots\\ y_{31}\\ y_{3n}\end{pmatrix} = \begin{pmatrix} 1 & 0 & 0\\ 1 & 0 & 0\\ & \dots & \\ 0 & 1 & 0\\ 0 & 1 & 0\\ & \dots & \\ 0 & 0 & 1\\ 0 & 0 & 1\end{pmatrix} \begin{pmatrix}\beta_0\\ \beta_1\\ \dots\\ \beta_{k-1}\end{pmatrix} + \begin{pmatrix}\epsilon_1\\ \epsilon_2\\ \dots\\ \dots\\ \dots\\ \dots\\ \dots\\ \epsilon_{3n}\end{pmatrix}, \tag{22}$$

reg3

де $\beta_i = \mu + \alpha_{i+1}$, $i = 0,\dots,k-1$. В матриці X кожна 1 повторюється n разів. Ця модель є частковим випадком моделі лінійної регресії, яку ми розглянемо в Розділі 8

Оцінку $s_{\bar y}^2$ можна переписати у вигляді

$$s_{\bar y}^2 = \frac{1}{n(k-1)}\left(\sum_{i=1}^k \bar y_{i\bullet}^2 - \frac{y_{\bullet\bullet}^2}{kn}\right) =: \frac{SSH}{n(k-1)}, \tag{23}$$

SSH2

де $y_{\bullet\bullet}^2 = \sum_{i=1}^k \sum_{j=1}^n y_{ij}$. За умови виконання H_0, $ns_{\bar y}^2$ є оцінкою σ^2: оскільки $\bar y_{\bullet\bullet}$ є оцінкою $\bar y_{i\bullet} =: \bar y$ (тобто $\bar y_{i\bullet}$ однакові за умови H_0), то

$$\mathbb{E}s_{\bar y}^2 = \mathbb{D}\bar y = \frac{\sigma^2}{n}.$$

Зауважимо, що у будь-якому випадку

$$\mathbb{E}s_e^2 = \frac{1}{k}\sum_{i=1}^k \mathbb{E}s_i^2 = \sigma^2.$$

З іншого боку, s_e^2 можна переписати наступним чином:

$$s_e^2 = \frac{\sum_{i=1}^k \sum_{j=1}^n y_{ij}^2 - \sum_{i=1}^k y_{i\bullet}^2/n}{k(n-1)} =: \frac{SSE}{k(n-1)}. \tag{24}$$ SSE2

Значення SSH (англ.: Sum of Squares Hypothesis, "between", тому що просумували по всім значенням) є сумою квадратів між вибірками, в той час як SSE є сумою квадратів у вибірці (англ.: Sum of Squares Error, "within", тому що просумували всередині кожної вибірки). [1]

Якщо гіпотеза H_0 не виконується, то (**довести!**)

$$\mathbb{E}[ns_{\bar{y}}^2] = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^k \alpha_i^2, \tag{25}$$ sn1

де $\alpha_i \neq 0$ для деяких i (інакше має місце H_0). Отже, дисперсія в цьому випадку більша за σ^2. За умови H_0, статистики SSE та SSH є незалежними [2], та їх відношення має розподіл Фішера з $k-1$ та $k(n-1)$ ступенями свободи.

Запишемо тестову статистику:

$$F = \frac{ns_{\bar{y}}^2}{s_\ell^2} = \frac{SSH/(k-1)}{SSE/(k(n-1))}. \tag{26}$$ F20

За умови H_0 маємо $F \sim F_{k-1,k(n-1)}$. Якщо $F > F_{\alpha,k-1,k(n-1)}$, де $F_{\alpha,k-1,k(n-1)}$- квантіль рівня $1-\alpha$ розподілу Фішера $F_{k-1,k(n-1)}$, відхиляємо H_0.

Хорошою характеристикою адекватності моделі є коефіцієнт детермінації

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSH}{SST}, \tag{27}$$ R

де

$$SST = SSH + SSE = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{\bullet\bullet})^2. \tag{28}$$ SST

де SST ("total") є повною сумою квадратів. Якщо нульова гіпотеза адекватно описує модель, то $R^2 \asymp 1$, тобто $SSH \asymp SST$.

6.2 Приклади

NOVexa

У наведених нижче прикладах нам знадобляться наступні бібліотеки:

```
library(ggplot2)
library(car)
```

Розглянемо такий приклад[3] Порівнюють вагу рослини в залежності від групи, до якої ця рослина належить.

[1] В літературі значення ще позначають як SSB ("between") або SSR при $k = 1$ ("residual", у моделі лінійної регресії, див. (8)).

[2] Це аналог леми Фішера. Ми доведемо це твердження в Розділі 8, де будемо розглядати моделі лінійної регресії.

[3] https://rpubs.com/aaronsc32/anova-compare-more-than-two-groups https://www.scribbr.com/statistics/anova-in-r/

```r
data("PlantGrowth")      # Дані з пакету R
PlantGrowth
```

```
   weight group
1    4.17  ctrl
2    5.58  ctrl
3    5.18  ctrl
4    6.11  ctrl
5    4.50  ctrl
6    4.61  ctrl
7    5.17  ctrl
8    4.53  ctrl
9    5.33  ctrl
10   5.14  ctrl
11   4.81  trt1
12   4.17  trt1
13   4.41  trt1
14   3.59  trt1
15   5.87  trt1
16   3.83  trt1
17   6.03  trt1
18   4.89  trt1
19   4.32  trt1
20   4.69  trt1
21   6.31  trt2
22   5.12  trt2
23   5.54  trt2
24   5.50  trt2
25   5.37  trt2
26   5.29  trt2
27   4.92  trt2
28   6.15  trt2
29   5.80  trt2
30   5.26  trt2
```

Як ми бачимо, є три групи рослин та 30 спостережень (по 10 в кожній групі), тобто в описаній вище моделі $k = 3$ та $n = 10$.

Ми можемо застосувати модель для того, щоб перевірити гіпотезу H_0: $\beta_0 = \beta_1 = \beta_2$ (див. (22)). Для цього ми спочатку будуємо лінійну модель за допомоги функції lm.

```r
lm(weight ~ group, data = PlantGrowth)
```

де group -це незалежна змінна, а weight- залежна. В результаті отримаємо оцінки на коефіцієнти регресії $(\beta_0, \beta_1, \beta_2)$ за формулою (47), див. Розділ (8.1).

```
Call:
lm(formula = weight ~ group, data = PlantGrowth)

Coefficients:
(Intercept)    grouptrt1     grouptrt2
5.032          -0.371        0.494
```

В принципі, ми задаємо тим самим модель лінійної регресії, але зараз нас цікавить не оцінка коефіцієнтів регресії (що буде зроблено пізніше в Розділі 8), а аналіз сум квадратів та перевірка гіпотези про модель.

```
plant.aov <- anova(lm(weight ~ group, data = PlantGrowth))
plant.aov
```

```
Analysis of Variance Table

Response:  weight
           Df  Sum Sq   Mean Sq  F value  Pr(>F)
group      2   3.7663   1.8832   4.8461   0.01591 *
Residuals  27  10.4921  0.3886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Для зручності, пояснемо дані за допомогою наступної таблиці.

Табл. 2: Пояснення до ANOVA таблиці

	Df	Sum Sq	Mean Sq	F value	Pr($F >$)
Group	$k-1$	SSH	$SSH/(k-1)$	$\frac{SSH/(k-1)}{SSE/(k(n-1))}$	p-value
Residuals	$k(n-1)$	SSE	$SSE/(k(n-1))$		

A-Tab2

У нас $k-1=3-1=2$, $k(n-1)=3*9=27$, а суми SSH та SSE обчислюються, відповідно, за (23) та (50). Оскільки p-value<0.05, то відхиляємо гіпотезу про рівність середніх.

Альтернативно, можна зробити аналіз дисперсій за допомогою функції aov.

```
aov(lm(weight ~ group, data = PlantGrowth))
```

```
Call:
aov(formula = lm(weight ~ group, data = PlantGrowth))

Terms:
```

```
group Residuals
Sum of Squares   3.76634  10.49209
Deg. of Freedom        2        27

Residual standard error: 0.6233746
Estimated effects may be unbalanced
```

Попередження "Estimated effects may be unbalanced" означає, що дані можуть містити різну кількість спостепежень (модель використовується, якщо кількість спостережень однакова). У нашому випадку в кожній групі по 10 спостережень, отже, модель є сбалансованою.

7 MANOVA: багатовимірний дисперсійний аналіз (однофакторна модель)

Література: [Ma07, §2.6], [R02, Sec.6]

7.1 MANOVA модель

Розглянемо однофакторну модель багатовимірного дисперсійного аналізу (Multivariate One-Way Analysis of Variance, або MANOVA). Розглянемо спочатку випадок, коли всі вибірки одного розміру. За винятком рядочку "дисперсія", спостереження можна записати у вигляді Таблиці 1, тільки тепер всі значення цієї таблиці– це вектори розмірності $p \times 1$, які мають багатовимірний нормальний розподіл $\mathrm{N}_p(\mu_i, \Sigma)$, $1 \le i \le k$ (відповідно до номеру групи).

Аналогами значень тепер є матриці

$$H := n \sum_{i=1}^{k} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})' = \sum_{i=1}^{k} \frac{1}{n} \bar{y}_{i\bullet} \bar{y}'_{i\bullet} - \frac{1}{kn} y_{\bullet\bullet} y'_{\bullet\bullet} \tag{29}$$ H

та

$$E := \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet})' = \sum_{ij} y_{ij} y'_{ij} - \sum_{i=1}^{k} \frac{1}{n} \bar{y}_{i\bullet} \bar{y}'_{i\bullet}. \tag{30}$$ E

Замість статистики F (див. (26)) будемо використовувати статистику Уілкса Λ, яка має вигляд (15), а "матриця гіпотези" і "матриця похибок" мають ступені свободи, відповідно,

$$\nu_H = k - 1, \quad \nu_E = k(n-1). \tag{31}$$ nu1

(порівняйте з однофакторною моделлю у попередній главі!). Зважена коваріаційна матриця у цьому випадку має вигляд

$$S_{pl} = \frac{E}{k(n-1)} = \frac{E}{\nu_E}. \tag{32}$$ Spl

Перевіримо нульову гіпотезу H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ проти альтернативи $\mu_i \ne \mu_j$ для деяких i, j. Оскільки ми розглядаємо векторний випадок, нульова гіпотеза означає, що

$$\begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \cdots \\ \mu_{1p} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \cdots \\ \mu_{2p} \end{pmatrix} = \cdots = \begin{pmatrix} \mu_{k1} \\ \mu_{k2} \\ \cdots \\ \mu_{kp} \end{pmatrix}.$$

У випадку, коли вибірки різного розміру (тобто модель не є сбалансованою), тобто вибірка i має розмір n_i,

$$y_i = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i.$$

Відповідно,

$$\bar{y}_{i\bullet} = \sum_{j=1}^{n_i} y_{ij}/n_i, \quad y_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}, \quad \bar{y}_{\bullet\bullet} = y_{\bullet\bullet}/N, \quad \text{де} \quad N = \sum_{i=1}^{k} n_i.$$

Тоді

$$H := \sum_{i=1}^{k} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})' \tag{33}$$ H2

та

$$E := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet})'. \tag{34}$$ E2

Цю (незбалансовану) модель природним чином можна спроектувати на одновимірний випадок (ANOVA). Відповідно, замість скалярних добутків ми будемо мати квадрати $(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$ та $(y_{ij} - \bar{y}_{i\bullet})$, відповідно.

Статистика Λ така сама, як і в одновимірному випадку, а ступені свободи матриць H і E наступні:

$$\nu_H = k - 1, \quad \nu_E = N - k = \sum_{i=1}^{k} n_i - k. \tag{35}$$ nu2

При цьому,

$$\operatorname{rank} H = \min(p, k-1) =: s, \quad \operatorname{rank} E = \min(p, (k-1)n). \tag{36}$$ rank

Для перевірки гіпотези розглянемо теж статистику $\Lambda = \frac{|E|}{|E+H|}$, див. (15). Як і разіше (див. розділ 5.3), ми відхиляємо нульову гіпотезу, якщо відповідна статисика F (cf. (16)) перевищує $F_{\alpha,df1,df2}$, див. визначення $df1$ та $df2$ в (17).

Статистика Роя (Roy's statistics)

Розглянемо наступне перетворення вектору y_{ij}: $z_{ij} = a'y_{ij}$, де $a \in \mathbb{R}^p$; тобто, ми проектуємо y_{ij} на пряму. Застосуємо тепер до z_{ij} статистику F з (26):

$$\begin{aligned} F &= \frac{n\sum_{i=1}^{k}(\bar{z}_{i\bullet} - \bar{z}_{\bullet\bullet})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n}(z_{ij} - \bar{z}_{i\bullet})^2/(k(n-1))} = \frac{SSH_z/(k-1)}{SSE_z/(k(n-1))} \\ &= \frac{a'Ha/(k-1)}{a'Ea/(k(n-1))}. \end{aligned} \tag{37}$$ F30

(тут індекс z означає, що величини SSE та SSH записані для значень z, а не для значень y, для яких вони не мають сенсу). Оберемо тепер в якості a власний вектор матриці $E^{-1}H$, який відповідає найбільшому власному числу λ_1. Тоді на цьому значенні a_1 досягається $\max F$, оскільки λ_1- це максимальний власний вектор:

$$\max_{a\in\mathbb{R}^p} F = \frac{a_1'EE^{-1}Ha_1/(k-1)}{a_1'Ea_1/(k(n-1))} = \frac{a'E\lambda_1 a_1/(k-1)}{a_1'Ea_1/(k(n-1))} = \frac{\lambda_1 k(n-1)}{k-1}.$$

Зауважимо, що на відміну від F величина $\max_{a\in\mathbb{R}^p} F$ не є розподіленою за Фішером. Для перевірки гіпотези H_0 застсуємо тест Роя (англ.: Roy's intersection-union test, або Roy's largest number test):

$$\theta = \frac{\lambda_1}{1+\lambda_1}. \tag{38}$$ Pilai

В загальному випадку точний розподіл θ невідомий, але розглядають верхню межу

$$F := \frac{(\nu_E - d - 1)\lambda_1}{d} \sim F_{d,\nu_E-d-1}, \quad d = \max(p, \nu_H).$$

Термін "верхня межа" означає, що реальне значення F більше за F_{d,ν_E-d-1}. Тому ми відхиляємо H_0 якщо $F < F_{d,\nu_E-d-1}$.

Розглянемо ще дві статистики.

Статистика Пілая (Pillai statistics):

$$V^{(s)} = \text{trace}[(E+H)^{-1}H] = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}. \tag{39}$$ Vs

За умови H_0,

$$F := \frac{(2n+s+1)V^{(s)}}{(2m+s+1)(s-V^{(s)})} \sim F_{s(2m+s+1),s(2n+s+1)}, \tag{40}$$ VsF

де $m := (|p-\nu_H|-1)/2$, $n := (\nu_E - p - 1)/2$.

Статистика Лаули-Хотеллінга (Lawley-Hotelling statistics):

$$U^{(s)} = \text{trace}[E^{-1}H] = \sum_{i=1}^{s} \lambda_i. \tag{41}$$ Vs

Цю статистику також можна трансформувати в величину, яка за виконання нульової гіпотези "наближено" має розподіл Фішера з певними ступенями свободи, див. [R02, (6.29)–(6.31)].

Всі ці тести є еквівалентними при $\nu_H = 1$. В іншому випадку результати тестів можуть розрізнятися завдяки різному впливу факторів на статистики. Тому для більш повного аналізу краще перевіряти гіпотезу декількома тестами.

7.2 Приклади

Розглянемо Приклад 6.1.7 з [R02]. Наведений нижче код запропоновано на сторінці [Sh], in particular, `https://rpubs.com/aaronsc32/manova`. Зчитаємо Таблицю 6.2 з [R02]:

```
root <- read.table('c:/your_root/T6_2_ROOT.DAT',
                   col.names = c('V1', 'V2', 'V3', 'V4', 'V5'))
```

Ми позначили через колонки наступні величини (для зручності, наведемо оригінальні назви, використані в [R02]):

$V1$: Номер дерева ('Tree.Number');

$V2$: Обхват дерева через 4 роки ('Trunk.Girth.4.Years')

$V3$: Висота через 4 роки ('Ext.Growth.4.Years')

$V4$: Обхват дерева через 15 років ('Trunk.Girth.15.Years')

$V5$: Вага над грунтом через 15 років. ('Weight.Above.Ground.15.Years')

Наведемо перші 6 значень цієї таблиці (в першій колонці у нас поки лише результати для дерев з першої групи, усього в таблиці 6 груп дерев, по 8 у кожній групі, тобто $k = 6$, $n = 8$; також, у нас 4 змінні, а отже, $p = 4$):

```
   V1   V2     V3      V4     V5
1  1    1.11   2.569   3.58   0.760
2  1    1.19   2.928   3.75   0.821
3  1    1.09   2.865   3.93   0.928
4  1    1.25   3.844   3.94   1.009
5  1    1.11   3.027   3.60   0.766
6  1    1.08   2.336   3.51   0.726
```

Далі ми подилимо дані на групи відповідно до номеру дерева та позничимо залежні змінні. Після цього ми можемо застосувати функцію aov.

```
root$V1 <- as.factor(root$V1)
dependent.vars2 <- cbind(root$V2, root$V3, root$V4, root$V5)
aov(dependent.vars2 ~ root$V1)
```

```
Call:
aov(formula = dependent.vars2 ~ root$V1)

Terms:
                 root$V1    Residuals
resp 1          0.073560    0.319987
resp 2          4.199662   12.142790
resp 3          6.113935    4.290813
resp 4          2.493091    1.722525
Deg. of Freedom         5          42

Residual standard errors: 0.08728545 0.5376933 0.3196282 0.2025154
Estimated effects may be unbalanced
```

Такий самий результат дає застосування функції manova:

```
manova(dependent.vars2 ~ root$V1)
```

Або, у розгорнотому вигляді,

```
summary(aov(dependent.vars ~ root$V1))
```

```
Response 1 :
             Df    Sum Sq    Mean Sq    F value    Pr(>F)
root$V1      5     0.07356   0.0147121  1.931      0.1094
Residuals    42    0.31999   0.0076187

Response 2 :
             Df    Sum Sq    Mean Sq    F value    Pr(>F)
```

```
root$V1       5     4.1997    0.83993     2.9052     0.0243 *
Residuals    42     12.1428   0.28911
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Response 3 :
             Df     Sum Sq    Mean Sq     F value    Pr(>F)
root$V1       5     6.1139    1.22279     11.969     3.112e-07 ***
Residuals    42     4.2908    0.10216
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Response 4 :
             Df     Sum Sq    Mean Sq     F value    Pr(>F)
root$V1       5     2.4931    0.49862     12.158     2.587e-07 ***
Residuals    42     1.7225    0.04101
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Аналіз проводиться для кожної змінної окремо (Response 1–Response 4).

Розберемо, що знаходиться в цій таблиці. Перша колонка- це кількість ступенів свободи: $\nu_H = k - 1 = 5$ та $\nu_E = k(n-1) = 6 * 7 = 42$ (див. (31)). Друга колонка- це діагональні елементи матриць H та E, відповідно. Ці матриці можна отримани з MANOVA наступним чином:

```
root.manova <- summary(manova(dependent.vars2 ~ root$V1))
H<-root.manova$SS[1]
E<-root.manova$SS[2]
```

Як результат, отримаємо матриці H

```
$‘root$V1‘
[,1]      [,2]      [,3]     [,4]
[1,] 0.07356042 0.5373852 0.3322646 0.208470
[2,] 0.53738521 4.1996619 2.3553885 1.637108
[3,] 0.33226458 2.3553885 6.1139354 3.781044
[4,] 0.20847000 1.6371084 3.7810437 2.493091
```

та E:

```
$Residuals
[,1]      [,2]      [,3]     [,4]
[1,] 0.3199875  1.696564 0.5540875 0.217140
[2,] 1.6965637 12.142790 4.3636125 2.110214
[3,] 0.5540875  4.363612 4.2908125 2.481656
[4,] 0.2171400  2.110214 2.4816562 1.722525
```

Ми наведемо також код для обчислення цих матриць, див. [Sh].

```
root.group <- split(root[,2:5], root$V1)

root.means <- sapply(root.group, function(x) {apply(x, 2, mean)},
        simplify = 'data.frame')

root.means
n <- dim(root)[1] / length(unique(root$V1))

total.means <- colMeans(root[,2:5])
total.means


H = matrix(data = 0, nrow = 4, ncol = 4)  # Обчислення H
for (i in 1:dim(H)[1]) {
        for (j in 1:i) {
                H[i,j] <- n * sum((root.means[i,] - total.means[i])
                                  * (root.means[j,]-total.means[j]))
                H[j,i] <- n * sum((root.means[j,] - total.means[j])
                                  * (root.means[i,]-total.means[i]))
        }
}

E = matrix(data = 0, nrow = 4, ncol = 4)    # Обчислення E
for (i in 1:dim(E)[1]) {
        for (j in 1:i) {
                b <- c()
                for (k in root.group) {
                  a <- sum((k[,i] - mean(k[,i])) * (k[,j] - mean(k[,j])))
                  b <- append(b, a)
                }
                E[i,j] <- sum(b)
                E[j,i] <- sum(b)
        }
}
```

Числа в третій колонці- це числа у другій колонці, поділені на кількість ступенів свободи, тобто на відповідні числа у першій колонці. Числа у четвертій колонці- це відношення зваженої суми квадратів, що відповідає змінній, і суми квадратів, що відповідає залишкам, тобто, наприклад, для першої змінної $0.0147121/0.0076187 = 1.931051$. Як і в попередньому розділі, за умови гіпотези H_0 – це відношення має розподіл Стьюдента. Остання колонка– це p-value.

З іншого боку, можна застосувати summary(manova) для того, щоб перевірити гіпотезу H_0 про те, всі групи мають однакові середні. Для цього треба подивитися результат root.manova:

```
           Df   Pillai   approx F  num Df   den Df    Pr(>F)
root$V1    5    1.3055   4.0697     20      168       1.983e-07 ***
Residuals 42
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Тобто, використання root.manova дає нам значення статистики Пілая (41). Маємо: $\nu_H = k - 1 = 5$, $\nu_E = k(n-1) = 42$, $m = 0$, $2n = 37$. Випадкова величина F з (41) має за виконання гіпотези розподіл Фішера $F_{20,168}$.

Для того, щоб обчислити це значення вручну, застосуємо матриці E та H. Вище ми зчитали матриці E та H як таблиці. Перетворимо їх у матриці та обчислимо стетистику Пілая:

```
E1 <- as.matrix(E$Residuals)
H1 <- as.matrix(H$`root$V1`)
Vs <- sum(diag(solve(E1 + H1) %*% H1))
```

Отримаємо $V^{(s)} = 1.305472$. Пряме обчислення (див. (41)) приводить до значення $F = 4.069846$, що лежить в критичній області (ми бачимо, що p-value дуже мале).

Можна також використати інші тести, які вбудовані у функцію manova:

```
summary(manova(dependent.vars2 ~ root$Tree.Number), test="Wilks")
summary(manova(dependent.vars2 ~ root$Tree.Number), test="Roy")
summary(manova(dependent.vars2~ root$Tree.Number),
                                        test="Hotelling-Lawley")
```

Так само, треба обчислити відповідну статистику, і використати те, що після певного перетворення статистики θ, $U^{(s)}$ $V^{(s)}$ мають наближено розподіл Фішера.

8 Багатовимірна лінійна регресія

LR

Розглянемо наступні моделі лінійної регресії.

1. Проста ланайна регресія: маємо одну змінну регресії x і один відгук y. Наприклад, наша задача спрогнозувати серендній бал студента вузу, маючи середній бал цього студента під час навчання в школі.

2. Множинна лінійна регресія: маєму один відгук y і декілька змінних x. Така задача виникає, наприклад, коли треба спрогнозувати середній бал у вузі, виходячи з оцінок з певних предметів у школі.

3. Багатовимірна ланайна регресія: маємо декількі відгуків y та декілька змінніх x регресії. Продовжуючи попередні приклади, така задача виникає, коли ми пригнозуємо бали по певним предметам у вузі, маючи бали по певним предметам у школі.

Незалежні змінні можуть бути фіксованими або випадковими. У попередніх прикладах всі були випадковими, оскільки ми випадковим чином обираємо студента. Якщо, наприклад, ми дослуждуємо вплив певних ліків на рівень холестерину в крові, можна зафіксувати кількість медикаментів та спостерігати зміни в рівні холестирину.

Ми будемо розглядати модель множинної , в яких змінні регресії x є фіксованими.

LR1

8.1 Проста лінійна регресія

Розглянемо наступну (одновимірну) модель:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_q x_{1q} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_q x_{2q} + \epsilon_2 \\ &\cdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_q x_{nq} + \epsilon_n. \end{aligned} \tag{42}$$

mulr

В цій моделі у нас x є вектором розмірності q (тобто є q змінних). Величини β_i, $i = 0, \ldots, q$, називаються коефіцієнтами регресії. Похибки ϵ_i є незалежними нормальними однаково розподіленими випадковими величинами. Припустимо, що виконані наступні умови:

1. $\mathbb{E}\epsilon_i = 0$ $i = 1, \ldots, n$ (тобто похибки центровані);
2. $\mathbb{E}\epsilon_i^2 = \sigma^2$, $i = 1, \ldots, n$ (похибки мають однакову дисперсію);
3. $\operatorname{cov}(\epsilon_i, \epsilon_j) = 0$, $i \neq j$ (похибки некорельовані).

Якщо x_i є фіксованими, то

$$\mathbb{E}y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}.$$

При цьому, з 2) випливає, що $\mathbb{D}y_i = \sigma^2$, та з 3) отримаємо $\operatorname{cov}(y_i, y_j) = 0$.

У матричному вигляді ця модель має наступний вигляд:

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1q} \\ 1 & x_{21} & x_{22} & \ldots & x_{2q} \\ & \cdots & & \cdots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{pmatrix}, \tag{43}$$

або

$$Y = X\beta + \mathcal{E}, \tag{44}$$

lin1

де $Y = (y_1, y_2, \ldots, y_n)$, $X = (1, x_1, \ldots, x_q)$ є матрицею $n \times (1 + q)$, $\mathbb{E}\mathcal{E} = 0$, $\mathbb{E}Y = X\beta$, $\operatorname{cov}\mathcal{E} = \sigma^2 I$, $\operatorname{cov} Y = \sigma^2 I$. Надалі ми будемо припускати, що $n > q + 1$, тобто спостережень досить багато; тоді матриця $X'X$ не є сингулярною.

Наша задача- побудувати оцінки коефіцієнтів регресії β_i, $i = 0, \ldots, q$.

$$\hat{y}_i = \mathbb{E}y_i = \hat{\beta}_0 + \hat{\beta}_i x_{i1} + \cdots + \hat{\beta}_q x_{iq}, \tag{45}$$

haty

Зауважимо, що $\hat{y}_i$ є оцінкою $\mathbb{E}y_i$, а не y_i. Побудуємо оцінку методу найменших квадратів (least squares estimate, LSE) векторного пареметру β:

$$\hat{\beta} = \arg\min_\beta |Y - \hat{Y}|^2 = \arg\min_\beta \sum_{i=1}^n |y_i - \hat{y}_i|^2 = \arg\min_\beta SSE_\beta,$$

де SSE є сумою квадратів залишків регресії SSE (error sum of squares) (див. також Розділ 6.1; тут у нас лише обна група випробувань, тобто $k = 1$). Іншими словами,

$$\hat{\beta} = \arg\min_\beta \sum_{i=1}^n \hat{\epsilon}_i^2. \tag{46}$$

LSE **Теорема 5.** *Припустимо, що $n > q+1$ та $(x_j)_{j=1}^q$ є лінійно незалежними. Тоді оцінкою β методу найменших квадратів є*

$$\hat\beta = (X'X)^{-1}X'Y. \tag{47}$$ LSE-eq

Доведення. Розглянемо $SSE_b = (Y - Xb)'(Y - Xb)$ Добавимо та віднімимо $X\hat\beta$, де $\hat\beta$ визначено в (47). Маємо:

$$\begin{aligned} SSE_b &= (Y - X\hat\beta + X(\hat\beta - b))'(Y - X\hat\beta + X(\hat\beta - b)) \\ &= (Y - X\hat\beta)'(Y - X\hat\beta) + (X(\hat\beta - b))'(X(\hat\beta - b)) \\ &\quad + (Y - X\hat\beta)'(X(\hat\beta - b)) + (X(\hat\beta - b))'(Y - X\hat\beta) \\ &= I_1 + I_2 + I_3 + I_4. \end{aligned}$$

Підставивши $\hat\beta = (X'X)^{-1}X'Y$, отримаємо $I_3 = I_4 = 0$. Наприклад,

$$I_3 = (Y - X\hat\beta)'(X(\hat\beta - b)) = (Y'X - Y'X(X'X)^{-1}(X'X))(\hat\beta - b) = 0.$$

Оскільки $I_2 \geq 0$, SSE_b набуває мінімуму при $b = \hat\beta$, тобто коли $I_2 = 0$ □

Надалі ми будемо позначати $SSE \equiv SSE_{\hat\beta}$. Зауважимо, що (див. Лема 1)

$$\mathbb{E}\hat\beta = (X'X)^{-1}X'\mathbb{E}Y = (X'X)^{-1}X'X\beta = \beta, \tag{48}$$ Eb

$$\operatorname{Var}(\hat\beta) = (X'X)^{-1}X' \operatorname{Var} Y X(X'X)^{-1} = \sigma^2(X'X)^{-1}, \tag{49}$$ varb

де ми використали $\operatorname{Var} Y = \sigma^2 I_n$. З теореми 5 випливає, що

$$\begin{aligned} SSE &= Y'Y - Y'X(X'X)^{-1}X'Y - Y'X(X'X)^{-1}X'Y + Y'X(X'X)^{-1}(X'X)(X'X)^{-1}X'Y \\ &= Y'Y - Y'X(X'X)^{-1}X'Y \\ &= Y'Y - \hat\beta'X'Y. \end{aligned}$$

(50) SSE2

Обчислимо математичне сподівання SSE. Оскільки $Y'Y = \operatorname{trace}(Y'Y) = \sum_{i=1}^n y_i^2$ та $\mathbb{E}y_i^2 = \sigma^2$, отримаємо $\mathbb{E}Y'Y = n\sigma^2$. Далі,

$$\begin{aligned} \hat\beta'X'Y &= Y'X(X'X)^{-1}X'Y = \operatorname{trace}(Y'X(X'X)^{-1}X'Y) \\ &= \operatorname{trace}(Y'XAX'Y) = \operatorname{trace}(AX'YY'X), \end{aligned}$$

де $A = (X'X)^{-1}$. Матриця A має розмірність $(q+1)\times(q+1)$, а $Y'X$ є вектором розмірності $1 \times (q+1)$. Зауважимо також, що за означенням, якщо всі y_i однаково розподілені (а це можливо, якщо коефіцієнти при x_{ij} нульлві). Використовуючи лінійність, обчислимо математичне сподівання:

$$\mathbb{E}\operatorname{trace}(AX'YY'X) = \operatorname{trace}(AX'\mathbb{E}(YY')X) = \sigma^2 \operatorname{trace}(AX'X) = \sigma^2(1+q).$$

Отже,

$$\mathbb{E}SSE = \sigma^2(n-1-q),$$

та $SSE/(n-1-q)$ є незміщеною оцінкою дисперсії σ^2.

Розглянемо загальну суму квадратів (total sum of squares):

$$SST := \sum_{i=1}^n (y_i - \bar y)^2 = Y'Y - n\bar Y^2.$$

Зауважимо, що

$$\mathbb{E}SST = \sigma^2(n-1),$$

та оскільки всі y_i однаково розподілені та нормальні,

$$\frac{SST}{\sigma^2} \sim \chi^2_{n-1},$$

а отже, цю суму можна розкласти в суму квадратів залишків регресії SSE та залишкової суми квадратів SSR (regression sum of squares), що пояснюється регресією:

$$SST = (Y'Y - \hat{\beta}'X'Y) + (\hat{\beta}'X'Y - n\bar{Y}^2) = SSE + SSR.$$

Покажемо, що за умови $H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$ статистики SSE/σ^2 та SSR/σ^2 мають хі-квадрат розподіл з $n-q-1$ та q ступенями свободи, відповідно.

Запишемо SSE та SSR у вигляді

$$SSE = Y'(I - X(X'X)^{-1}X')Y = Y'Y - Y'HY = Y'Y - Y'H'HY = \|Y(I-H)\|^2,$$

$$SSR = Y'HY - (B_nY)'(B_nY) = \|(H-B_n)Y\|^2.$$

де B_n - це матриця, всі елементами якої є $1/n$ (зауважимо, що $B_n'B_n = B_n = B_n'$ та $H'H = H = H'$).

Зауважимо, що матриці $I-H$ і $H-B_n$ ортогональні:

$$(I-H)(H-B_n) = 0.$$

Оскільки $Y \sim \mathrm{N}_n(0, \sigma^2 I_n)$ та $\mathrm{cov}((I-H)Y, (H-B_n)Y) = 0$, це означає, що $(I-H)Y$ та $(H-B_n)Y$ некорельовані (перевірити!), а отже, незалежні. Більш того, як лінійні перетворення нормальних випадкових величин, $(I-H)Y$ та $(H-B_n)Y$ також нормально розподілені. А отже, після нормування, SSE та SSR мають хі-квадрат розподіл як квадрат норми нормально розподілених випадкових величин:

$$SSE/\sigma^2 \sim \chi^2_{n-1-q}, \quad SSR/\sigma^2 \sim \chi^2_q. \tag{51}$$ SS

Перевіримо тепер гіпотезу H_0: $\beta_1 = \beta_2 = \cdots = \beta_q = 0$ (зауважимо, що ми не робимо припущень відносно β_0!). З (51) випливає, що

$$F := \frac{SSR/q}{SSE/(n-q-1)} \sim F_{q,n-q-1}. \tag{52}$$ Fr

Подивимось, які значення є критичними для статистики F. За умови виконання гіпотези H_0,

$$\begin{aligned}
\hat{Y}'Y = \hat{\beta}'X'Y &= \begin{pmatrix} \hat{\beta}_0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{1n} \\ & & \dots & \\ x_{1q} & x_{2q} & \dots & x_{nq} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \\
&= \begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_0 & \dots & \hat{\beta}_0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \\
&= \hat{\beta}_0 \sum_{i=1}^n y_i = \hat{\beta}_0 n\bar{Y}.
\end{aligned} \tag{53}$$

З іншого боку, за умови H_0 маємо $\hat{\beta}_0 = \bar{Y}$. Отже, якщо виконується H_0, то $SSR = 0$, а отже, $SST = SSE = \sum_{i=1}^n y_i^2 - n(\bar{Y})^2$. Це означає, що критичними значеннями для F є великі значення. Тобто, фіксучи рівень надійності α, ми відхиляємо H_0, якщо $F > F_{\alpha,q,n-q-1}$.

По аналогії з (27), відношення SSR до SST є частиною дисперсії, яка пояснюється регресією, називається коефіцієнтом детермінації і позначається R^2:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \in (0,1). \tag{54}$$ R2

Якщо справедлива H_0, то коефіцієнт детермінації має були близьким до 0, тобто y не залежить від x. Іншими словами, лінійна регресія погано пояснює значення y.

Статистику F також можна зобразити за допомогою R^2:

$$F = \frac{n-q-1}{q} \frac{R^2}{1-R^2}.$$

8.2 Багатовимірна регресія

У багатовимірному випадку маємо спостереження

$$Y = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1p} \\ y_{21} & y_{22} & \dots & y_{2p} \\ & & \dots & \\ y_{n1} & y_{n2} & \dots & y_{np} \end{pmatrix} = \begin{pmatrix} y_1' \\ y_2' \\ \dots \\ y_n' \end{pmatrix}.$$

Позначимо

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1q} \\ 1 & x_{21} & x_{22} & \dots & x_{2q} \\ & \dots & & \dots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nq} \end{pmatrix}, \quad B = \begin{pmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1p} \\ & \dots & \dots & \\ \beta_{q1} & \beta_{q2} & \dots & \beta_{qp} \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \dots & \epsilon_{1p} \\ \epsilon_{21} & \epsilon_{22} & \dots & \epsilon_{2p} \\ & \dots & \dots & \\ \epsilon_{n1} & \epsilon_{n2} & \dots & \epsilon_{np} \end{pmatrix}.$$

Тоді модель багатовимірної лінійної регресії можна записати у наступному вигляді:

$$Y = XB + \mathcal{E}. \tag{55}$$ multreg

Аналогічно одновимірному випадку, $\mathcal{E} \sim \mathrm{N}_p(0, \Sigma)$,

1. $\mathbb{E}Y = XB$,
2. $\mathrm{cov}(Y_i) = \Sigma$, $i = 1, \dots, n$.
3. $\mathrm{cov}(Y_i, Y_j) = 0$ при $i \neq j$.

Наступна теорема є узагальненням Теореми 5.

Теорема 6. *Оцінкою методу найменших квадратів матриці B є*

$$\hat{B} = (X'X)^{-1}X'Y, \tag{56}$$ B

тобто

$$\hat{B} = \arg\min E = \arg\min \mathcal{E}'\mathcal{E} = (Y - X\hat{B})'(Y - X\hat{B}) =: E.$$

Ця оцінка має наступні властивості:

1. $\mathbb{E}B = B$ (оцінка є незміщеною);
2. дисперсія $\mathbb{D}\hat{\beta}_{ij}$ є мінімальною у класі всіх незміщених оцінок (теорема Гауса-Маркова);
3. величини $\hat{\beta}_{ij}$ є корельованими.

Оскільки стовбчики в B є корельваними, ми не можемо використовувати F-тести як в простій регресії, тому потрібен інший механізм тестування.

Аналогом матриці SSE є

$$E = (Y - X\hat{B})'(Y - X\hat{B}). \tag{57}$$ E1

Позначимо

$$S_e := \frac{E}{n - q - 1}. \tag{58}$$

Ця оцінка є незміщеною, $\mathbb{E}S_e = \Sigma$. Зобразимо матрицю B у вигляді

$$B = \begin{pmatrix} \beta_0' \\ B_1 \end{pmatrix} = \begin{pmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1p} \\ & \dots & \dots & \\ \beta_{q1} & \beta_{q2} & \dots & \beta_{qp} \end{pmatrix},$$

та перевіримо гіпотезу H_0: $B_1 = 0$, проти альтернативи H_1: $B_1 \neq 0$. Розкладемо сумарну похибку у суму матриць E та H:

$$Y'Y - n\bar{Y}\bar{Y}' = (Y'Y - \hat{B}'X'Y) + (\hat{B}'X'Y - n\bar{Y}\bar{Y}') = E + H.$$

Використовуючи матриці E та H, обчислимо статистику Уілкса $\Lambda = \frac{|E|}{|E+H|}$. За умови виконання гіпотези H_0, матриця має бути близькою до нульової, а отже, критичні значення Λ - це малі значення. Як і в Розділі 5, можна використати статистику (16) для перевірки гіпотези H_0.

Приклад 1. Розглянемо приклад з $p = 2$, $q = 3$. Якщо $B_1 = 0$, то $b_{ij} = 0$, $i \neq 0$, а тоді

$$\hat{B} = \begin{pmatrix} \hat{\beta}_{01} & \hat{\beta}_{02} \\ 0 & 0 \\ \dots & \dots \\ 0 & 0 \end{pmatrix}.$$

Далі,

$$(\hat{X}\hat{B})'Y = \begin{pmatrix} \hat{\beta}_{01} & \hat{\beta}_{01} & \dots \hat{\beta}_{01} \\ \hat{\beta}_{02} & \hat{\beta}_{02} & \dots \hat{\beta}_{02} \end{pmatrix} Y = \begin{pmatrix} \hat{\beta}_{01} \sum_{i=1}^n y_{i1} & \hat{\beta}_{01} \sum_{i=1}^n y_{i2} \\ \hat{\beta}_{02} \sum_{i=1}^n y_{i1} & \hat{\beta}_{02} \sum_{i=1}^n y_{i2} \end{pmatrix}.$$

З іншого боку,

$$\begin{aligned} \bar{Y}\bar{Y}' &= \begin{pmatrix} \frac{1}{n}\sum_{k=1}^n y_{k1} \\ \frac{1}{n}\sum_{k=1}^n y_{k2} \end{pmatrix} \begin{pmatrix} \frac{1}{n}\sum_{k=1}^n y_{k1} & \frac{1}{n}\sum_{k=1}^n y_{k2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{n^2}\left(\sum_{k=1}^n y_{k1}\right)^2 & \frac{1}{n^2}\left(\sum_{k=1}^n y_{k1}\right)\left(\sum_{k=1}^n y_{k2}\right) \\ \frac{1}{n^2}\left(\sum_{k=1}^n y_{k1}\right)\left(\sum_{k=1}^n y_{k2}\right) & \frac{1}{n^2}\left(\sum_{k=1}^n y_{k2}\right)^2 \end{pmatrix} \end{aligned}$$

За умови виконання гіпотези H_0,

$$\begin{pmatrix} y_{1i} \\ y_{2i} \\ \dots \\ \dots \\ y_{ni} \end{pmatrix} \approx \begin{pmatrix} \beta_{0i} \\ \beta_{0i} \\ \dots \\ \beta_{0i} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \dots \\ \dots \\ \epsilon_{ni} \end{pmatrix}, \quad i = 1, 2,$$

де вектор-стовбчик $(\beta_{0i}, \dots, \beta_{0i})'$ має розмірність $(q+1)\times 1$. Отже, $\hat{\beta}_{0i} = \frac{1}{n}\sum_{k=1}^{n} y_{ki}$, $i = 1, 2$. Тоді $n\bar{Y}\bar{Y}' \asymp (\hat{X}\hat{B})'Y$, звідки $H \asymp 0$. Отже, ми приймаємо H_0, якщо $\Lambda \approx 1$.

8.3 Приклади

Розглянемо наступний приклад, в якому дані взято з так званого "квартету Анскомбе"[4](Anscombe's Quartet). Ми розглянемо квартет Анскомбе у Додатку **??**. Квартет Анскомбе складається з чотирьох послідовностей, які задовольняють моделі лінійної регресії, але їхні графіки істотно відрізняються. **Зауважимо, що наведені нижче обчислення залежать від згенерованих нормально розподілених випадкових величин, тому мають суто ілюстративну мету.**

```
y=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68)
x1=c(10,8,13,9,11,14,6,4,12,7,5)

x2=sqrt(y)+rnorm(length(y))

model=lm(y~x1+x2)     # задаємо модель лінійної регресії
model                 # отримаємо коротку інформацію про модель
```

```
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)          x1          x2
0.9563           0.4800      0.8345
```

Тобто, ми отримали оцінки на вільний доданок (Intercept), та на коефіцієнти при $x1$ та $x2$ (нижній рядок звіту). Переконаємося, що це дійсно оцінки, які задаються формулою (47):

```
X<-cbind(1,x1,x2)
est_b<-solve(t(X)%*% X) %*% (t(X)%*%y)
est_b
```

[4] http://www.learnbymarketing.com/tutorials/explaining-the-lm-summary-in-r/

```
  [,1]
0.9563469
x1 0.4800062
x2 0.8344520
```

Викликати інформацію про кожну колонку можна за допомогою операцій, відповідно

```
coef(summary(model))[, "Std. Error"]
coef(summary(model))[, "t value"]
coef(summary(model))[, "Pr(>|t|)"]
```

Для того, щоб отримати більш детальну інформацію одночасно, використаємо

```
summary(model)
```

В результаті отримаємо

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
Min      1Q        Median      3Q        Max
-1.3780  -0.6855   -0.0448     0.5484    1.5407

Coefficients:
             Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)  0.9563     1.5197       0.629       0.54670
x1           0.4800     0.1062       4.521       0.00195 **
x2           0.8345     0.4647       1.796       0.11026
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.107 on 8 degrees of freedom
Multiple R-squared:  0.7623,Adjusted R-squared:  0.7029
F-statistic: 12.83 on 2 and 8 DF,  p-value: 0.00319
```

В розділі Residuals ми отримали мінімальне значення (-1.3780), перший квартіль, медіану та третій квартіль (-0.6855, -0.0448, 0.5484), та максимальне значення (1.5407) залишків.

Друга колонка - це оцінки на коефіцієнти, які ми отримали в est_b. Третя колонка- стандартне відхилення від теоретичного значення, тобто діагональні елементи матриці

$$\frac{SSE}{n-k}(X'X)^{-1}.$$

(ми підставили замість σ^2 її незміщену оцінку $\frac{SSE}{n-k}$, див. (49)).

Отримати третю колонку "вручну" можна наступним чином:

```
k=length(model$coefficients)          # кількість змінних +1, k=3, q=k-1
n=length(model$residuals)             # n=11
s2<-sum(model$residuals**2)/(n-k)     # s2 = SSE/(n-k) n=11,
Sigma<- s2*solve(t(X)%*%X)            # матриця коваріацій оцінок коефіцієнтів
sqrt(diag(Sigma))                     # стандартні відхилення похибок
```

За нульової гіпотези H_0: $b_i = 0$, $i = 0, 1, 2$, випадкові величини $t_i = \hat{b}_i/\sigma_i$ мають розподіл Стьюдента. Тому при перевірці гіпотези H_0 для кожного коефіцієнта ми обчислюємо значення t_i та відповідні $p - value$. Наприклад, перший елемент четвертого стовбчика- це

```
t_0<-est_b[1]/sqrt(diag(Sigma))[1]
t_0
```

```
0.6292939
```

Відповідно, $p - value$- це ймовірність не потрапити в критичну область, і обчислюється наступним чином:

```
2*(1-pt(t1, n-k))
```

```
0.5467026
```

```
# Залишкова стандартна похибка (Residual Standard Error)
R_Err<- sqrt(SSE/(n-k))
R_Err
```

```
1.10729
```

```
SST<-sum((y-mean(y))**2)
R2<-(SST-SSE)/SST          # коефіцієнт детермінації (Multiple R Squared)
R2
```

```
0.7623434
```

```
AdjR2<-1-(SSE/SST)*(n-1)/(n-k)
AdjR2                    # скоригований коефіцієнт детермінації (Ajusted R Squared)
```

```
0.7029292
```

Нарешті, обчислимо F-статистику:

```
F<-((SST-SSE)/(k-1))/ (SSE/(n-k))
F
```

```
12.83101
```

та порівняємо її із теоретичним квантілем:

```
pf(0.05, k-1, n-k)
```

```
0.04847572
```

Отже, ми відхиляємо нульову гіпотезу про те, що коефіцієнти дорівнюють 0, оскільки отримане значення статистики значно більше за теоретичне

Або, обчислимо $p-value$:

```
1-pf(F, k-1, n-k)
```

```
0.003190067
```

(що співпадає з $p-value$, яке обчислено у вбудованій функції).

Нарешті, "Signif. codes:" означає, наскільки впливає коефіцієнт на залежну змінну. Наприклад, "***" означає, що $p-value$ знаходиться в межах $[0, 0.001]$, "**"- в межах $(0.001, 0.01]$, "*"- в межах $(0.01, 0.05]$, "·"- в межах $(0.05, 0.1]$, та порожнє значення означає, що $p-value$ знаходиться в межах $(0.1, 1.0]$. В останніх двох випадках ми приймаємо гіпотезу про те, що коефіцієнт дорівнює нулю, або що відповідна змінна не є значущою.

Розглянемо ще одну задачу.

```
chem<-read.table('c:/your_root/T10_1_CHEM.DAT',
                 col.names = c('V1', 'Y1','Y2','Y3','X1','X2','X3') )
head(chem)
```

```
 V1   Y1   Y2   Y3  X1  X2  X3
1  1 41.5 45.9 11.2 162 23  3
2  2 33.8 53.3 11.2 162 23  8
3  3 27.7 57.5 12.7 162 30  5
4  4 21.7 58.8 16.0 162 30  8
5  5 19.9 60.6 16.2 172 25  5
6  6 15.0 58.0 22.6 172 25  8
```

Розіб'ємо на вектори на застосуємо функцію lm() для того, щоб побудувати модель лінійної регресії:

```
y<- cbind(chem_y$Y1, chem_y$Y2, chem_y$Y3)
x<- cbind(chem_x$X1, chem_x$X2, chem_x$X3)

chem.lm<-lm(y~x)
chem.lm
```

```
Call:
lm(formula = y ~ x)

Coefficients:
                 [,1]       [,2]       [,3]
(Intercept)   332.1110   -26.0353  -164.0789
x1             -1.5460     0.4046     0.9139
x2             -1.4246     0.2930     0.8995
x3             -2.2374     1.0338     1.1535
```

Ми отримали оцінку матриці B (див. (56)):

$$\hat{B} = \begin{pmatrix} -1.5460 & 0.4046 & 0.9139 \\ -1.4246 & 0.2930 & 0.8995 \\ -2.2374 & 1.0338 & 1.1535 \end{pmatrix} \tag{59}$$

Для перевірки гіпотези H_0 використаємо MANOVA.

```
chem.manova<- summary(manova(y~x), test="Wilks")
chem.manova
```

```
          Df    Wilks      approx F    num Df    den Df    Pr(>F)
x          3    0.033158   10.787      9         31.789    1.884e-07 ***
Residuals 15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Тут num Df = df1, den Df = df2, та approx F- це статистика (16). Ми відхиляємо H_0, оскільки p-value (тобто Pr(>F)) <0.05.

9 Дискримінантний аналіз

DA

Розглянемо наступну задачу. Потрібно знайти лінійну функцію від змінних, за допомогою якої можна було краще розбити вибірку на 2 або більше груп. Дискримінантна функція– це лінійна комбінація змінних, яка найкращим чином розділяє вибірку на групи.

Припустимо, що ми маємо 2 популяції $y_{11}, y_{12}, \ldots, y_{1n_1}$ та $y_{21}, y_{22}, \ldots, y_{2n_2}$ в $\mathbb{R}^p$, з однаковою коваріаційною матрицею Σ, але різними середніми μ_1 та μ_2.

Дискримінантна функція є лінійною комбінацією координат, яка максимізує відстань між двома (перетвореними) групами векторів. Позначимо $z = a'y$:

$$\begin{aligned} z_{1i} &= a'y_{1i} = a_1y_{1i1} + a_2y_{1i2} + \ldots a_py_{1ip}, \quad i = 1, \ldots, n_1, \\ z_{2i} &= a'y_{2i} = a_1y_{2i1} + a_2y_{2i2} + \ldots a_py_{2ip}, \quad i = 1, \ldots, n_2, \end{aligned} \tag{60}$$

або

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \ldots \\ y_{1n_1} \end{pmatrix} \curvearrowright \begin{pmatrix} z_{11} \\ z_{12} \\ \ldots \\ z_{1n_1} \end{pmatrix}, \quad \begin{pmatrix} y_{21} \\ y_{22} \\ \ldots \\ y_{2n_1} \end{pmatrix} \curvearrowright \begin{pmatrix} z_{21} \\ z_{22} \\ \ldots \\ z_{2n_1} \end{pmatrix},$$

Тут y_{1i}, $i = 1, \ldots, n_1$ та y_{2i}, $i = 1, \ldots, n_2$, є векторами, а z_{1i} $i = 1, \ldots, n_1$ та z_{2i}, $i = 1, \ldots, n_2$, скалярами.

Позначимо

$$\bar{z}_{1\cdot} = \sum_{i=1}^{n_1} \frac{z_{1i}}{n_1} = a'\bar{y}_{1\cdot}, \quad \bar{z}_{2\cdot} = \sum_{i=1}^{n_1} \frac{z_{2i}}{n_2} = a'\bar{y}_{2\cdot},$$

де

$$\bar{y}_{1\cdot} = \sum_{i=1}^{n_1} \frac{y_{1i}}{n_1}, \quad \bar{y}_{2\cdot} = \sum_{i=1}^{n_2} \frac{y_{2i}}{n_2}.$$

Значення $\bar{z}_{1\cdot}$ і $\bar{z}_{2\cdot}$ є "центрами" груп 1 та 2. Наша задача - обрати вектор a таким чином, щоб максимізувати відстань між цими двома побудованими центрами. Оберемо вектор a наступним чином:

$$a = \arg\max_a \left(\frac{\bar{z}_{1\cdot} - \bar{z}_{2\cdot}}{s_z} \right)^2 = \arg\max_a \left(\frac{[a'(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})]^2}{a'S_{pl}a} \right).$$

де $s_z^2 = a'S_{pl}a$. А саме, ми будемо максимізувати не саму відстань $\bar{z}_{1\cdot} - \bar{z}_{2\cdot}$, а нормовану. Зауважимо, що для того, щоб існувала S_{pl}^{-1}, необхідно виконання нерівності $n_1 + n_2 - 2 > p$.

Будемо шукати a у вигляді $a = A(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})$. Підставимо цей вектор a в $\left(\frac{\bar{z}_{1\cdot} - \bar{z}_{2\cdot}}{s_z}\right)^2$:

$$\left(\frac{\bar{z}_{1\cdot} - \bar{z}_{2\cdot}}{s_z} \right)^2 = \frac{(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})'A'(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})}{(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})'A'S_{pl}A(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})}. \tag{61}$$

max1

Зауважимо, що скалярний добуток векторів максимальний за модулем, коли вектори є колінеарними. Тому максимум у наведеному вище виразі досягається, якщо вектори $A'(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})$ та $A'S_{pl}A(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})$ колінеарні. Тоді для A має виконуватись рівність $\lambda I_p = S_{pl}A$, а отже, $A = \lambda S_{pl}^{-1}$ (будемо вважати, що $\lambda = 1$, оскільки параметр змінює лише масштаб). Оскільки матриця S_{pl} симетрична, $S_{pl} = S_{pl}'$, то при такому виборі вектору a отримаємо

$$\max \left(\frac{\bar{z}_{1\cdot} - \bar{z}_{2\cdot}}{s_z} \right)^2 = (\bar{y}_{1\cdot} - \bar{y}_{2\cdot})'S_{pl}^{-1}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}).$$

Отже, з точністю до множника, максимум є статистикою Хотеллінга T^2. А отже, можна перевірити гіпотезу про рівність середніх: H_0: $\mu_1 = \mu_2$.

Розглянемо багатовимірний випадок.

1. Якщо груп більше, ніж дві, то потрібно більше ніж дві дискримінантні функції, які б описували розбіття на групи. Якщо точки $\mathbb{R}^p$-мірного простору спроектувати на двовимірну площину, задану першими двома дискримінантними функціями, то ми отримаємо найкраще можливе наближення розбиття на групи.

2. Знайдемо множину вихідних змінних, використання яких дозволяє розбити на групи максимально якісно.

3. Впорядкуємо змінню по мірі ваги внеску до процедури розбиття.

4. Інтерпретуємо нові змінні, які задаються за допомогою дискримінантих функцій.

5. Виконуємо аналіз MANOVA.

Припустимо, що є k груп спостережень $z_{ij} = a'y_{ij}$, $i = 1, \dots, k$, $j = 1, \dots, n_j$. Як у цьому випадку обирати вектор(и) a? Потрібно знайти аналог виразу $\frac{\bar{z}_1 - \bar{z}_2}{s_z}$ у багатовимірному випадку.

Зауважимо, як виглядає матриця H у випадку двох груп (див. (33)):

$$H = \sum_{i=1}^{2} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})' = \frac{n_1 n_2}{n_1 + n_2} (\bar{y}_{1\cdot} - \bar{y}_{2\cdot})(\bar{y}_{1\cdot} - \bar{y}_{2\cdot})'. \tag{62}$$ H2k

Обчислимо $a'Ha$:

$$a'Ha = \frac{n_1 n_2}{n_1 + n_2} (\bar{z}_{1\cdot} - \bar{z}_{2\cdot})^2.$$

(оскільки $(\bar{z}_{1\cdot} - \bar{z}_{2\cdot})' = (\bar{y}_{1\cdot} - \bar{y}_{2\cdot})'a \in \mathbb{R}$, тобто є числом!).

Аналогічно, при $k = 2$ можна трасформувати E (див. (34)) наступним чином:

$$E = \sum_{i=1}^{2} \sum_{j=1}^{n_1} (y_{ij} - \bar{y}_{i\cdot})(y_{ij} - \bar{y}_{i\cdot})' = (n_1 + n_2 - 2) S_{pl}. \tag{63}$$ E2k

Домножаючи на вектори a' і a відповідно, отримаємо

$$a'Ea = (n_1 + n_2 - 2) a' S_{pl} a.$$

Отже,

$$\frac{(\bar{z}_{1\cdot} - \bar{z}_{2\cdot})^2}{a' S_{pla} a} = \frac{(n_1 + n_2 - 2)(n_1 + n_2)}{n_1 n_1} \cdot \frac{a'Ha}{a'Ea}.$$

З іншого боку, $a'Ha$ та $a'Ea$ можна записати через $SSH(z)$ та $SSE(z)$, які застосовані до змінних z, а не до y:

$$a'Ha = \sum_{i=1}^{2} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})(\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})' = SSH(z),$$

$$a'Ea = \sum_{i=1}^{2} \sum_{j=1}^{n_1} (z_{ij} - \bar{z}_{i\cdot})^2 = SSE(z).$$

Ці формули мають узагальнення на випадок k груп. Отже, у випадку $k > 2$ природньо максимізувати не (61), а співвідношення

$$\lambda = \frac{SSH(z)}{SSE(z)} = \frac{a'Ha}{a'Ea}. \tag{64}$$ max2

Знайдемо, для яких рівняння (64) має розв'язок. З рівняння $a'Ha = \lambda a'Ea$ знайдемо

$$a'(Ha - \lambda Ea) = 0. \tag{65}$$

Розв'язок $a = 0$ не підходить, оскільки ми маємо тоді $\lambda = 0/0$. Іншими розв'язками є числа $\lambda_1, \ldots, \lambda_s$, де $s = \text{rank}(E^{-1}H)$. Впорядкуємо ці λ_i. Тоді

$$\lambda_1 = \max_a \frac{a'Ha}{a'Ea}.$$

Нехай $a_1, \ldots, a_s$ - власні вектори, що відповідають власним числам $\lambda_1, \ldots, \lambda_s$. Дискримінантними функціями тоді є $z_1 = a_1'y$, $z_2 = a_2'y$, $\ldots$, $z_s = a_s'y$. Функції z_i, $i = 1, \ldots, s$, описують різницю між середніми $\bar{y}_1$, $\ldots$, $\bar{y}_k$. Власні вектори a_i є некорельованими, але не ортогональними, оскільки матриця $E^{-1}H$ не є симетричною. Зауважимо також, що побудоване розбиття відповідає власним числам, а не початковому розбиттю на k груп. Функція, яка максимально розбиває на групи– це дискримінантна функція, побудована за першим власним вектором $z_1 := a_1'y$. "Відносну важливість" дискримінантних функій можна порівняти, порівнюючи внесок λ_i, а саме, відношення

$$\frac{\lambda_i}{\sum_{i=1}^s \lambda_i}, \quad i = 1, \ldots, s.$$

Як правило, двох-трьох дискримінантних фунцій досить для того, щоб розділити виборку на підгрупи.

10 Приклади

Розглянемо Приклад 8.2 з [R02] [5]. Зчитаємо наступні дані, див. Таблиця 8.1 з [R02]. Дані являють собою температури плавління 2х типів металу. Наша задача- вдало розділити дані на 2 групи.

```
Temp<-read.table("c:/your_path/T8_1_STEEL.DAT")
Temp
```

```
      V1 V2 V3
[1,]   1 33 60
[2,]   1 36 61
[3,]   1 35 64
[4,]   1 38 63
[5,]   1 40 65
[6,]   2 35 57
[7,]   2 36 59
[8,]   2 38 59
[9,]   2 39 61
[10,]  2 41 63
[11,]  2 43 65
[12,]  2 41 59
```

[5] https://rpubs.com/aaronsc32/classification-linear-discriminant-analysis

```
Temp<-as.matrix(Temp)          # зчитаємо у вигляді матриці
Temp1<-Temp[1:5,2:3]           # Температури 1
Temp2<-Temp[6:12,2:3]          # Температури 2

plot(Temp2, col = "red", xlim=c(32,48), ylim=c(55,68))   # точки Temp2
points(Temp1, col= "blue")                               # точки Temp1
```

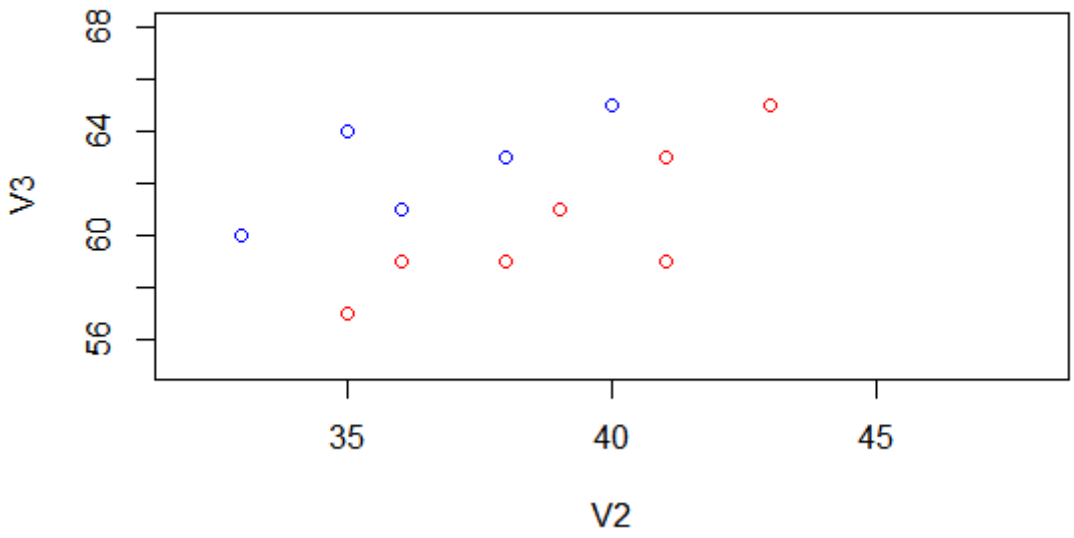


fig6-1

Рис. 3: Температури T1 та T2

Схоже на те, що вибірку можна розділити на 2 групи. Для цього побудуємо дискримінантну функцію.

```
y1<-c(mean(Temp1[,1]),mean(Temp1[,2]))   # вектор середніх Temp1
y2<-c(mean(Temp2[,1]),mean(Temp2[,2]))   # вектор середніх Temp2

# обчислимо зважену коваріаційну матрицю

S1<-cov(Temp1)
S2<-cov(Temp2)
d1<-length(Temp1[,1])   # кількість елементів в Temp1
d2<-length(Temp2[,1])   # кількість елементів в Temp2

Spl<-((d1-1)*S1+ (d2-1)*S2)/(d1+d2-2)    # зважена коваріаційна матриця
```

```
     V2       V3
V2 7.92   5.680000
V3 5.68   6.291429
```

```
a<-solve(Spl)%*%(y1-y2)   # дискримінантна функція
z1<- Temp1 %*% a
z2<- Temp2 %*% a
```

Отримаємо:

```
> a
    [,1]
V2 -1.633377
V3  1.819779

> z1
     [,1]
[1,] 55.28530
[2,] 52.20494
[3,] 59.29766
[4,] 52.57775
[5,] 52.95055

> z2
     [,1]
[1,] 46.55921
[2,] 48.56539
[3,] 45.29863
[4,] 47.30481
[5,] 47.67762
[6,] 48.05042
[7,] 40.39850
```

Розділимо на 2 групи виходячи з того, де знаходяться елементи по відношенню до середнього знячення

```
zmean<-0.5*(mean(z1)+mean(z2))
t<- as.matrix(Temp[,2:3])
z<-t%*%a
z
```

```
     [,1]
[1,] 55.28530
[2,] 52.20494
[3,] 59.29766
[4,] 52.57775
[5,] 52.95055
[6,] 46.55921
[7,] 48.56539
[8,] 45.29863
```

```
[9,] 47.30481
[10,] 47.67762
[11,] 48.05042
[12,] 40.39850
```

Розділимо тепер елементи t на групи 1 та 2:

```
group <- ifelse(z[,1] > zmean, 1, 2)
group
```

Отже, ми можемо віднести перші 5 елементів до першої групи, а останні 7- до другої.

```
1 1 1 1 1 2 2 2 2 2 2 2
```

Розглянемо ще один приклад, а саме, Приклад 8.4.1 з [R02]. Для цього завантажимо наступні бібліотеки:

```
library(car)
library(MASS)
library(dplyr)
```

та завантажимо Таблицю 8.3 (див. Глава 8). В таблиці знаходяться вимірювання параметрів шоломів в залежності від того, до якої групи відноситься людина (гравець шкільної команди, гравець команди коледжа, або людина взагалі не грає в футбол). Для простоти, ми перейменуємо колонки цієї таблиці.

```
Foot<-read.table('c:/your_path/T8_3_FOOTBALL.DAT',
      col.names = c('Group', 'V2', 'V3', 'V4', 'V5', 'V6','V7'))
head(Foot)
```

Визначимо залежні змінні (у вигляді матриці) та визначимо змінну, за якою робимо групування:

```
dependent.vars3<-cbind(Foot$V2,Foot$V3,Foot$V4,Foot$V5,Foot$V6,Foot$V7)

DepVar<-as.matrix(dependent.vars3)

Foot$Group <- as.factor(Foot$Group)
```

Застосуємо функцію lda:

```
lda1<-lda(Foot$Group~., data=Foot)
```

```
Call:
lda(Foot$Group ~ ., data = Foot)

Prior probabilities of groups:
1         2         3
0.3333333 0.3333333 0.3333333

Group means:
  V2    V3       V4       V5       V6       V7
1 15.20 58.93700 20.10833 13.08333 14.73333 12.26667
2 15.42 57.37967 19.80333 10.08000 13.45333 11.94333
3 15.58 57.77000 19.81000 10.94667 13.69667 11.80333

Coefficients of linear discriminants:
LD1           LD2
V2  0.948423100  1.4067750094
V3 -0.003639865 -0.0005126312
V4 -0.006439599 -0.0286176430
V5 -0.647483088  0.5402700415
V6 -0.504360916 -0.3839132257
V7 -0.828535064 -1.5288556226

Proportion of trace:
LD1   LD2
0.943 0.057
```

Вектори $LD1$ та $LD2$- це перші 2 власні вектори матриці . Відповідно. 0.943 та 0.057- це відношення $\frac{\lambda_1}{\sum_{i=1}^{5}\lambda_i}$ та $\frac{\lambda_2}{\sum_{i=1}^{5}\lambda_i}$.

Цей самий аналіз можна зробити вручну наступним чином. Застосуємо функцію manova для того, щоб знайти матриці E та H, та знайти власні числа та власні вектори матриці $E^{-1}H$.

```
Foot.manova<-summary(manova(dependent.vars3~ Foot$Group), test="Wilks")
Foot.manova
H<-Foot.manova$SS[1]
E<-Foot.manova$SS[2]
E
```

```
$Residuals
      [,1]      [,2]      [,3]       [,4]       [,5]       [,6]
[1,] 37.2560  50.28120 13.739000   7.290000 10.8360000 19.8360000
[2,] 50.2812 275.01513 88.737783  56.785300 29.5855333 43.9434333
[3,] 13.7390  88.73778 47.497083   6.682167 11.1873333 13.8430000
[4,]  7.2900  56.78530  6.682167 107.224333 27.3833333  3.6946667
```

```
[5,] 10.8360  29.58553 11.187333  27.383333 53.7710000  0.7943333
[6,] 19.8360  43.94343 13.843000   3.694667  0.7943333 32.6700000
```

Як ми бачимо, можна записати у вигляді матриці наступним чином (спробуйте напряму застосувати операцію solve(E) чи solve(H)):

```
E1<-as.matrix(E$Residuals)
H1<-as.matrix(H$`Foot$Group`)
```

Тепер обчислимо власні вектори та власні числа матриці $E^{-1}H$:

```
H1<-as.matrix(H$`Foot$Group`)
E1<-as.matrix(E$Residuals)
eigen(solve(E1) %*% H1)$values              # власні числа
lam1<-eigen(solve(E1) %*% H1)$values[1]     # перше власне число
lam2<-eigen(solve(E1) %*% H1)$values[2]     # друге власне число
```

Перші 2 власних числа є дійсними, тобто 1.9177 та 0.1159, інші є комплексними.

```
eigen(solve(E1) %*% H1)$vectors             # нормовані власні вектори
a1<-eigen(solve(E1) %*% H1)$vectors[,1]     # перший власний вектор
a2<-eigen(solve(E1) %*% H1)$vectors[,2]      # другий власний вектор
```

Зауважимо, що ці власні вектори є нормованими! Ці вектори відрізняються від тих, що ми вже отримали за допомогою функції lda. Зробимо наступне:

```
lda1$scaling                          # власні вектори
lda1$scaling[,1]                      # перший власний вектор
b1<- as.vector(lda1$scaling[,1])
b2<- as.vector(lda1$scaling[,2])
b1/(sqrt(sum(b1^2)))                  # це як раз a1
```

Поубудуємо дискримінантні функції і побачимо, як виглядає картинка після розділення на групи за допомогою чих функцій:

```
z11<- DepVar[1:30,]%*%b1
z12<- DepVar[1:30,]%*%b2
z21<- DepVar[31:60,]%*%b1
```

```
z22<- DepVar[31:60,]%*%b2
z22<- DepVar[31:60,]%*%b2
z31<- DepVar[61:90,]%*%b1
z32<- DepVar[61:90,]%*%b2

plot(z11,z12, col = "red", xlim= c(-15,-5), ylim = c(0,8))
points(z21,z22, col= "blue")
points(z31,z32,col="green")
```

11 Метод головних компонент

При великій кількості спостережень матриця дисперсій стає дуже великою. Потрібно зменшити кількість змінних та знайти змінні, які є базисними. Зауважимо, що метод головних компонент може бути застосований до будь-якого розподілу, не обов'язково нормального. Зауважимо також, що тести в MANOVA використовують $E^{-1}H$, а отже, ця матриця не має бути сингулярною. Перед тим, як робити аналіз MANOVA, потрібно зменшити розмірність методом головних компонент.

11.1 Геометричний підхід

Якщо змінні є корельованими, то еліпсоїд, в якщому графічно можна зобразити значення y_1, y_2, ..., y_p, не є орієнтованим вздовж жодної з осей. Знайдемо ці головні осі. Для цього розглянемо матрицю повороту A, яка повертає еліпс так, щоб в нових координатах (головних компонентах, principal components) змінні були некорельованими.

Припустимо, що y_i єцентрованими, інакше розглянемо $y_i - \bar{y}$. Нехай A є ортогональною матрицею, та розглянемо

$$z_i = Ay_i.$$

Оскільки $AA' = I$, то

$$z_i' z_i = y_i' A' A y_i = y_i' y_i.$$

Таким чином, перетворення A зберігає відстані від початку координат.

Знаходження головних осей еліпсоїда еквівалентно знаходженню такої ортогональної матриці A, яка повертає осі таким чином, щоб нові змінні z_i були некорельованими, тобто матриця вибіркових дисперсій після перетворення має вигляд

$$S_z = ASA' = \begin{pmatrix} s_{z_1}^2 & 0 & \cdot & 0 \\ 0 & s_{z_2}^2 & \cdot & 0 \\ 0 & & \cdot & s_{z_p}^2 \end{pmatrix}.$$

Тут S -вибіркова коваріаційна матриця y_1, ..., y_n. Нехай C -матриця, утворена з нормованих (тобто $a'a = 1$) власних векторів За ??? , $C'SC = D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$.

$$A = C' = \begin{pmatrix} a_1' \\ a_2' \\ \cdots \\ a_p' \end{pmatrix}$$

Головними компонентами є змінні $z_1 = a_1'y$, $z_2 = a_2'y$, $\ldots$, $z_p = a_p'y$. Наприклад,$z_1 = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1p}y_p$. При цьому вибіркові дисперсії дорівнюють власним числам: $s_{z_i}^2 = \lambda_i$. Тому "частка поясненої дисперсії"- це і є відношення перших k власних чисел до всіх власних чисел:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\mathrm{trace}(S)}.$$

Таким чином, ми замінемо весь вектор $(y_{i1}, y_{i2}, \ldots, y_{ip})$ головними компонентами $(z_{i1}, z_{i2}, \ldots, z_{ik})$, які дають максимальний внесок в дисперсію. Якщо кілька змінних сильно корельовані, фактична розмірність значно менше, ніж p.

11.2 Алгебраїчний підхід

Ми шукаємо лінійну комбінацію векторів, яка максимізує дисперсію. Вибіркова дисперсія вектору $z = a'y$ дорівнює $a'Sa$. Оскільки $a'Sa$ не досягає максимуму якщо норма $\|a\|$ не обмежена, тому ми будемо шукати максимум

$$\lambda = \frac{a'Sa}{a'a}.$$

Задача зводиться до задачі Лагранжа

$$L(a) := a'Sa - \lambda(a'a - 1) \mapsto \max.$$

Використовуючи (2) та враховуючи симетрію S отримаємо

$$\frac{\partial L}{\partial a} = 2Sa - 2\lambda a = 0,$$
$$\frac{\partial L}{\partial \lambda} = a'a - 1 = 0.$$

Отже, a є розв'язком рівняння

$$Sa = \lambda a, \quad a'a = 1,$$

що співпадає з результатом, отриманим у попередньому підрозділі.

Власний вектор a_1, який відповідає власному числу λ_1. При цьому дисперсія $z_1 = a_1'y$ є максимальною.

На відміну від дискримінантного аналізу, ми не обчислюємо тут обернену матрию, а отже, S може бути сингулярною, а отже, деякі власні числа (а отже, і власні вектори) можуть бути нульовими. Наприклад, матриця сингулярна, якщо $n < p$, тобто розмір вибірки менше, ніж розмірність простору. Далі так само, як і в геометричному підході, ми відбираємо власні вектори з найвищим рівнем впливу на величину дисперсії.

11.3 Приклади

Розглянемо наступний приклад (див. Приклад 12.2.1 з [R02]). В наступній таблиці наведені вимірюван ня параметрів голови (довжина та ширина щелепа) для першого (колонки $V1$ та $V2$) та другого (відповідно, колонки $V3$ та $V4$) синів в родині.

```
sons <- read.table('c://your_path/T3_7_SONS.DAT')
sons
```

```
  V1  V2  V3  V4
1 191 155 179 145
2 195 149 201 152
3 181 148 185 149
4 183 153 188 149
5 176 144 171 142
6 208 157 192 152
```

Щоб проілюструвати, що знаходження головних компонент - це фактично ротація осей координат.

```
son1<-matrix(cbind(sons$V1,sons$V2), nrow = 25, byrow = FALSE)
m1<-mean(sons$V1)
m2<-mean(sons$V2)
ybar<-c(m1,m2)
plot(sons$V1-m1,sons$V2-m2, col="green")  # малюємо центровані координати
Sson1<-cov(son1)                          # коваріація
eigen(Sson1)$values                       # власні числа
eigen(Sson1)$vectors                      # власні вектори
a1<- -eigen(Sson1)$vectors[,1]            # перша дискримінантна функція
a2<- -eigen(Sson1)$vectors[,2]            # друга дискримінантна функція
t(a1)%*%a2                                # перевіряємо ортонрнальність
```

Центруємо всю вибірку і побудуємо графік в нових координатах:

```
son2<-matrix(c(son1[,1]-m1,son1[,2]-m2), nrow = 25, byrow = FALSE)
z1<- son2%*% a1
z2<-son2%*% a2
plot(sons$V1-m1,sons$V2-m2, col="green")
points(z1,z2, col="red")
```

Ми бачимо, що в нових осях координат наші дані розташовані вздовж осі OX.

В наступній таблиці наведено результати 6 тестів 20 інженерів-студентів та 20 пілотів[6]. Ми наведемо перші 6 значень таблиці; тут 1 означає, що мова йде про студента, а -про пілота. Тому ми зробимо групування данних за принципом, чи є людина, що бере участь в експерименті, студентом чи пілотом.

```
pilots <- read.table('c:/your_path/T5_6_PILOT.DAT',
col.names =  c('Group', 'V1', 'V2','V3', 'V4', 'V5','V6'))
head(pilots)
pilots$Group<- ifelse(pilots$Group == 1, 'Apprentice', 'Pilot')
```

[6]https://rpubs.com/aaronsc32/principal-component-analysis

```
   Group  V1  V2 V3 V4  V5 V6
1      1  121 22 74 223 54 254
2      1  108 30 80 175 40 300
3      1  122 49 87 266 41 223
4      1  77  37 66 178 80 209
5      1  140 35 71 175 38 261
6      1  108 37 57 241 59 245
```

За допомогою функції отримаємо унфомацію про власті числа та власні вектори матриці коваріацій:

```
S <- cov(pilots[,2:7])
sum(diag(S))  # сумарна дисперсія, тобто сума власних значень

s.eigen <- eigen(S)
s.eigen
```

```
eigen() decomposition
$values
[1] 1722.0424  878.3578  401.4386  261.0769  128.9051   50.3785

$vectors
       [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
[1,] -0.21165160 -0.38949336  0.88819049  0.03082062 -0.04760343  0.10677164
[2,]  0.03883125 -0.06379320  0.09571590 -0.19128493 -0.14793191 -0.96269790
[3,] -0.08012946  0.06602004  0.08145863 -0.12854488  0.97505667 -0.12379748
[4,] -0.77552673  0.60795970  0.08071120  0.08125631 -0.10891968 -0.06295166
[5,]  0.09593926 -0.01046493  0.01494473  0.96813856  0.10919120 -0.20309559
[6,] -0.58019734 -0.68566916 -0.43426141  0.04518327  0.03644629 -0.03572141
```

Для того, щоб отримати власні вектори або власні числа, можна застосувати функції, відповідно, s.eigen$vectors та s.eigen$values.

Обчислимо пропорцію власних чисел до сумарної дисперсії $\lambda_i/(sum\lambda_i)$:

```
x<-c()
for (s in s.eigen$values) {
        x<-c(x,s / sum(s.eigen$values))
}
print(x)
```

```
0.50027387 0.25517343 0.11662269 0.07584597 0.03744848 0.01463556
```

```
plot(s.eigen$values, xlab = 'Номер власного числа', ylab = 'Власні числа')
lines(s.eigen$values)
```

З іншого боку, щоб отримати стандартні відхилення (тобто корені з власних чисел), можна застосувати функції princomp та prcomp до pilots[,2:7]. Отримаємо:

```
princomp(pilots[,2:7])
```

```
princomp(x = pilots[, 2:7])

Standard deviations:
Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
40.975497 29.264294 19.783897 15.954623 11.210822  7.008498

6  variables and  40 observations.
```

```
prcomp(pilots[,2:7])
```

```
Standard deviations (1, .., p=6):
[1] 41.497499 29.637102 20.035932 16.157875 11.353640  7.097781

Rotation (n x k) = (6 x 6):
    PC1         PC2         PC3         PC4         PC5         PC6
V1  0.21165160 -0.38949336  0.88819049 -0.03082062 -0.04760343 -0.10677164
V2 -0.03883125 -0.06379320  0.09571590  0.19128493 -0.14793191  0.96269790
V3  0.08012946  0.06602004  0.08145863  0.12854488  0.97505667  0.12379748
V4  0.77552673  0.60795970  0.08071120 -0.08125631 -0.10891968  0.06295166
V5 -0.09593926 -0.01046493  0.01494473 -0.96813856  0.10919120  0.20309559
V6  0.58019734 -0.68566916 -0.43426141 -0.04518327  0.03644629  0.03572141
```

Зауважимо, що наведені PC відрізняються від власних векторів знаком!

Для того, щоб отримати повну інформацію і про стандартні відхтлення, і про власні числа, і про власні вектори, можна використати функцію

```
summary(pilots.pca)
```

Запишемо тепер дані в координатах, коли осі- це перешкальовані власні вектори $PC1$ та $PC2$. В якості шкалюючаго множника використаємо $sqrt(nrow(pilots))$ (корінь квадратний з кількості випробування).

```
pc1 <- rowSums(t(t(sweep(pilots[,2:7], 2, colMeans(pilots[,2:7])))
                 * s.eigen$vectors[,1] * -1) / scaling[1])
pc2 <- rowSums(t(t(sweep(pilots[,2:7], 2,  colMeans(pilots[,2:7])))
                 * s.eigen$vectors[,2]) / scaling[2])
```

Тут використовується функція sweep для того, щоб від компонент (pilots[,2:7] віднімається середнє (індекс 2 тут використовується для того, щоб зазначити, що ми проводимо цю операцію зі стовбчиками; якби було 1, це означало б, що операція проводиться із рядками).

Тепер ми задамо дані в цих нових координатах, і зобразимо графічно за допомогою ggplot:

```
df <- data.frame(pc1, pc2, c(rep('Студент', 20), rep('Пілот', 20)))
colnames(df) <- c('PC1', 'PC2', 'Група')

ggplot(df, aes(x=PC1, y=PC2, color=Group)) +
geom_point()
```

Результат зображено на малюнку 5 Таке розділення на групи можна також зробити

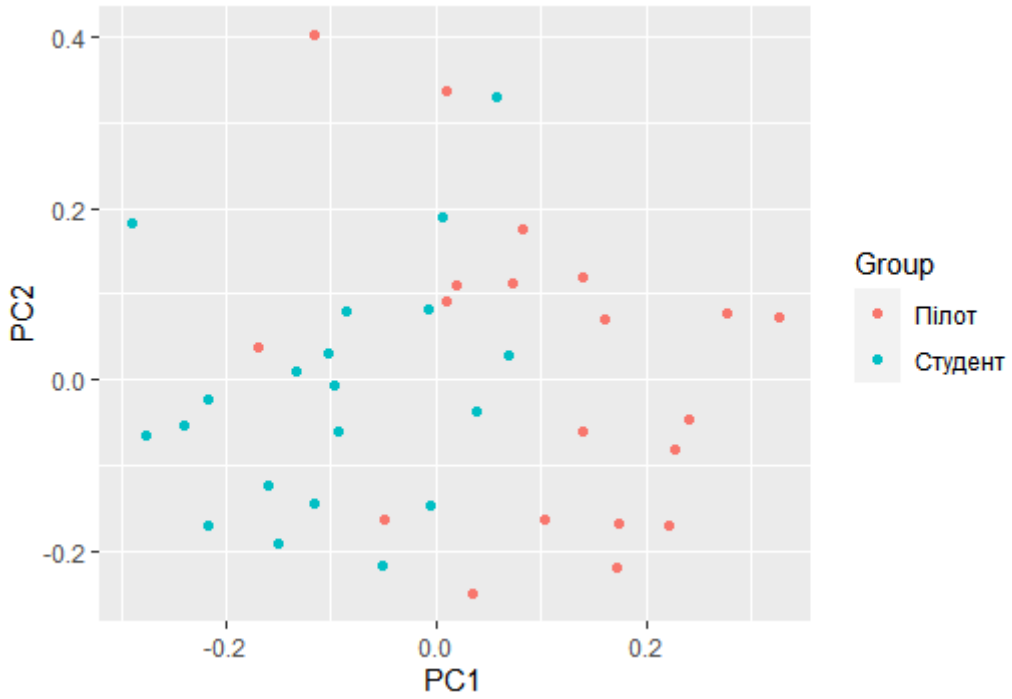


fig6-2

Рис. 4: Розділення на групи методом головних компонент

вбудованою функцією autoplot (при цьому нам знадобиться бібліотека ggfortify)

```
library(ggfortify)
pca.plot <- autoplot(pilots.pca, data = pilots, colour = 'Group')
pca.plot
```

Альтернативно, такий самий аналіз можна зробити, використовуючи не коваріаційну матрицю, а кореляційну.

За допомоги вектора головних компонент можна, наприклад, стискати фотографії. Розглянемо наступний приклад[7].

[7]https://rpubs.com/aaronsc32/image-compression-principal-component-analysis

Ми застосуємо метод головних компонент для того, щоб стиснути фотографію котика: Для цього завантажимо пакет jpeg та бібліотеку library(jpeg). Функція readJPEG перетво-

Рис. 5: Кошеня Патрик

fig6-2

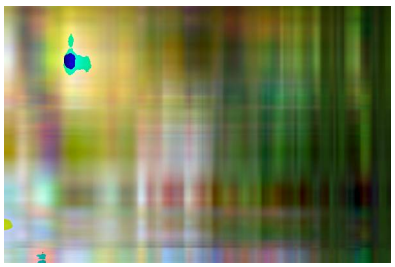
рює фотографію в матрицю.

```
cat <- readJPEG('cat.jpg')
dim(cat)
ncol(cat)
nrow(cat)
```

Запустивши ці функції ми побачимо, що масив має розмірність 398x300x3, тобто що фотографію можна зобразити за допомогою трьох матриць розміру 600x398, кожна матриця є представленням кольорів в схемі RGB. Наша задача - зменшити розмірності, вибравши "ключові кольори".

Спочатку виділимо кольори в окремі матриці, і виконаємо аналіз методом головних компонент для кожної з матриць.

```
r <- cat[,,1]
g <- cat[,,2]
b <- cat[,,3]
cat.r.pca <- prcomp(r, center = FALSE)
cat.g.pca <- prcomp(g, center = FALSE)
cat.b.pca <- prcomp(b, center = FALSE)
rgb.pca <- list(cat.r.pca, cat.g.pca, cat.b.pca)
```



(а) m=3

(б)

Рис. 6: Main caption

(а)

(б)

Рис. 7: Main caption

(а)

(б)

Рис. 8: Main caption

Ми обрали center = FALSE для того, щоб зберегти зображення кольорів, сентрування змістило б кольоровий спектр. Далі ми зібрали головні компоненти в один список.

Тепер ми будемо обирати, відповідно, 3, 10, 20, 50, 100, 300 головних компонент, та перетворимо відповідні матриці на фотографії (функція writeJPEG).

```
for (i in c(3,10,20,50,100,300)) {
        pca.img <- sapply(rgb.pca, function(j) {
                compressed.img <- j$x[,1:i] %*% t(j$rotation[,1:i])
        }, simplify = 'array')
        writeJPEG(pca.img, paste('c:/your_path/cat_compressed_',
        round(i,0), '_components.jpg', sep = ''))
}
```

Як ми бачимо, зі збільшенням кількості головних компонент ми отримаємо більш якісну картинку.

12 Факторний аналіз

Нехай є змінні y_1, y_2, ..., y_p. Якщо ці змінні залежні, то фактична розмірність системи менша за p.

Задача факторного аналізу полягає в тому, щоб зменшити розмірність шляхом "зменшення повторень".

Наприклад, розглянемо кореляційну матрицю

$$\begin{pmatrix} 1 & 0.9 & 0.05 & 0.05 & 0.05 \\ 0.9 & 1 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 1 & 0.9 & 0.9 \\ 0.05 & 0.05 & 0.9 & 1 & 0.9 \\ 0.05 & 0.05 & 0.9 & 0.9 & 1 \end{pmatrix}.$$

Виходячі з значень корелляцій, доцільно розділити систему на два фактори: y_1, y_2- фактор 1, y_3, y_4, y_5- фактор 2.

Порівняємо методи факторного аналізу з методом головних компонент.

Метод головних компонент	Факторний аналіз
Головні компоненти є комбінаціями вихідних змінних	Вихідні компоненти є комбінаціями факторів
Пояснюємо найбільшу частку дисперсії	Обираємо фактори згідно з тим, як кореллюють змінні.

Розглянемо наступну модель. У факторному аналізі у нас одна вибірка з популяції, з середнім значенням μ та коваріаційною матрицею Σ. Зобразимо змінні як лінійні комбінації факторів f_1, f_2, …, f_m, та похибки ϵ_k, $1 \le k \le m$. Наша модель має вигляд:

$$\begin{aligned} y_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \dots \lambda_{1m} f_m + \epsilon_1 \\ y_2 - \mu_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \dots \lambda_{2m} f_m + \epsilon_2 \\ &\dots \\ y_p - \mu_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \dots \lambda_{pm} f_m + \epsilon_p. \end{aligned} \tag{66}$$

В ідеалі m значно менше за p, інакше описання y за допомоги факторів f не є якісним. Тут f- випадкові змінні, які описують y. Коефіцієнти λ_{ij} називаються навантаженнями (loadings), наприклад вплив фактору j на змінну y_i. Принципова відмінність моделі факторного аналізу від моделі лінійної регресії полягає в тому, що 1) насправді ми не спостерігаємо фактори, а їх треба знайти; у нас не n спостережень, а одне.

Нам потрібно оцінити λ_{ij} та адекватно підібрати фактори.

Зробимо наступні припущення.

1. $\mathbb{E} f_i = 0$, $\operatorname{Var} f_1 = 1$;
2. $\mathbb{E}\epsilon_i = 0$, $\operatorname{Var}\epsilon_i = \psi_i$, $\operatorname{cov}(\epsilon_i, \epsilon_j) = 0$.
3. $\operatorname{cov}(f_i, f_j) = 0$, $i \ne j$, $\operatorname{cov}(\epsilon_i, f_j) = 0$;

$$\operatorname{Var} y_i = \lambda_{i1}^2 + \lambda_{i2}^2 + \dots + \lambda_{im}^2 + \psi_i. \tag{67}$$

$$y - \mu = \Lambda f + \mathcal{E}. \tag{68}$$

$y = (y_1, y_2, \ldots, y_p)'$, $f = (f_1, f_2, \ldots, f_m)'$, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_m)'$

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2m} \\ & & \ldots & \\ \lambda_{p1} & \lambda_{p2} & \ldots & \lambda_{pm} \end{pmatrix}, \quad \operatorname{cov}(\epsilon) = \Psi = \begin{pmatrix} \Psi_1 & 0 & \ldots & 0 \\ 0 & \Psi_2 & \ldots & 0 \\ & & \ldots & \\ 0 & 0 & \ldots & \Psi_p \end{pmatrix}, \quad \operatorname{cov}(f, \mathcal{E}) = 0.$$

Нехай $(\sigma_{ij}) = \operatorname{cov}(y) = \operatorname{cov}(\Lambda f + \mathcal{E})$. Оскільки $\operatorname{cov}(f, \mathcal{E}) = 0$,

$$\operatorname{cov}(y) = \operatorname{cov}(\Lambda f) = \Lambda I \Lambda' + \Psi = \Lambda I \Lambda' + \Psi.$$

Розглянемо випадок $m = 2$.Тоді $\operatorname{cov}(y_1, y_2) = \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22}$. Якщо y_1 та y_2 мають "багато спільного", то вони мають схожі навантаження на f_1 та f_2 У цьому випадку або $\lambda_{11}\lambda_{21}$ або $\lambda_{12}\lambda_{22}$ великі. Якщо y_1 та y_2 мають мало спільного, то навантаження λ_{11} та λ_{21} на f_1 та навантаження λ_{12}, λ_{22} на f_2 різні. У цьому випадку $\lambda_{11}\lambda_{21}$ та $\lambda_{12}\lambda_{22}$ малі.

Розглянемо $\operatorname{cov}(y, f)$. Наприклад,

$$\begin{aligned} \operatorname{cov}(y_1, f_2) &= \mathbb{E}(y_1 - \mu_1)(f_2 - \mu_2) \\ &= \mathbb{E}((\lambda_{11}f_1 + \cdots + \lambda_{1m}f_m)f_2) - \mathbb{E}(\lambda_{11}f_1 + \cdots + \lambda_{1m}f_m)\mathbb{E}f_2 \\ &= \lambda_{11}\operatorname{cov}(f_1, f_2) + \lambda_{12}\operatorname{cov}(f_2, f_2) + \cdots + \lambda_{1m}\operatorname{cov}(f_m, f_2) \\ &= \lambda_{12}. \end{aligned}$$

Отже,

$$\operatorname{cov}(y_i, f_j) = \lambda_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, m. \tag{69}$$

У матричному вигляді цей вираз має вигляд

$$\operatorname{cov}(y, f) = \Lambda.$$

Також,

$$\sigma_{ii} = \operatorname{Var}(y) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \Psi_i = h_i^2 + \Psi_i. \tag{70}$$

Величина h_i^2 називається спільною дисперсією (communality), а величина Ψ_i -залишковою дисперсією.

Навантаження на фактори обираються неоднозначно. Нехай T ортогональна матриця, тобто $TT' = I$. Тоді

$$y - \mu = \underbrace{\Lambda T}_{\Lambda^*}\underbrace{T'f}_{f^*} + \epsilon = \Lambda^* f^*,$$

При цьому $\Lambda^*(\Lambda^*) = \Lambda T(\Lambda T)' = \Lambda\Lambda'$.

Фактори f^* задовольняють ті самі умови, що і f: $\mathbb{E}f^* = 0$, $\operatorname{cov}(f^*) = I$, $\operatorname{cov}(f^*, \epsilon) = 0$. Величини h_i^2 теж не змінилися. Дійсно, запишемо h_i^2 як $h_i^2 = \lambda_i'\lambda$. Зауважимо, що iй рядок матриці Λ^*- це $\lambda_i'T$. Тоді

$$(h_i^*)^2 = (\lambda_i^*)'(\lambda_i^*) = \lambda_i'T(\lambda_i'T) = \lambda_i'\lambda_i = h_i^2.$$

У наступних двох підрозділах ми розглянемо методи оцінки навантажень та спільної дисперсії.

12.1 Метод головних компонент у факторному аналізі

Нехай S вибіркова коваріація. Наша задача– знайти оцінку $\hat{\Lambda}$ матриці навантажень Λ. При цьому має виконуватись співвідношення

$$S = \hat{\Lambda}\hat{\Lambda}' + \Psi.$$

Запишемо спектральний розклад $S = CDC'$, де C- ортогональна матриця, побудована з власних векторів S, $D = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_p)$, де $\theta_i > 0$, $1 \le i \le p$ -власні числа S.

Тоді

$$S = CD^{1/2}D^{1/2}C' = CD^{1/2}(CD^{1/2})' =: \Lambda^*(\Lambda^*)'.$$

Але $CD^{1/2}$ має розмірність $p \times p$, а нам хотілося б зменшити кількість факторів, тобто було б добре задати як матрицю розмірністю

Нехай $D_1 = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_m)$, $C_1 = (c_1, c_2, \ldots, c_m)$. Покладемо $\hat{\Lambda} = C_1 D_1^{1/2}$. Наприклад, при $p = 3$, $m = 2$

$$\begin{pmatrix} \hat{\lambda}_{11} & \hat{\lambda}_{12} \\ \hat{\lambda}_{21} & \hat{\lambda}_{22} \\ \hat{\lambda}_{31} & \hat{\lambda}_{32} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix} \begin{pmatrix} \sqrt{\theta_1} & 0 \\ 0 & \sqrt{\theta_2} \end{pmatrix} = \begin{pmatrix} \sqrt{\theta_1}c_{11} & \sqrt{\theta_2}c_{12} \\ \sqrt{\theta_1}c_{21} & \sqrt{\theta_2}c_{22} \\ \sqrt{\theta_1}c_{31} & \sqrt{\theta_2}c_{32} \end{pmatrix} \quad (71) \quad \boxed{\texttt{com-f}}$$

Наведена вище формула пояснює назву "метод головних компонент". Стовбчики в (71) є прапорційними власним векторам S, а отже навантаження на j-й фактор є прапорційним до j-ї головної компоненти. Фактори пов'язані з m першими головними компонентами. Але після повороту навантажень інтерпретація як правило змінюється, тому не зовсім якно, як інтерпретувати ці нові фактори.

З попереднього, i-й діагональний елемент $\hat{\Lambda}$ дорівнює $\sum_{i=1}^m \hat{\lambda}_{ij}^2$. Отже, щоб довести до кінця апроксимацію S, покладемо

$$\hat{\psi}_i = s_{ii} - \sum_{i=1}^m \hat{\lambda}_{ij}^2.$$

та

$$S = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}, \quad (72) \quad \boxed{\texttt{Sest}}$$

де $\Psi = \mathrm{diag}(\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_p)$. Дисперсії в (72) задаються точно, в той час як ми лише апроксимуємо недіагональні елеметни матриці S.

Сума квадратів елементів $\hat{\Lambda}$ по колонкам і по рядкам дорівнює спільній дисперсії $\hat{h}_i^2$ та власному числу θ_j, відповідно.

$$\hat{h}_i^2 = \sum_{j=1}^m \hat{\lambda}_{ij}^2,$$

$$\sum_{i=1}^p \hat{\lambda}_{ij}^2 = \sum_{i=1}^p (\sqrt{\theta_j}c_{ij})^2 = \theta_j \sum_{i=1}^p c_{ij}^2 = \theta_j,$$

де ми використали те, що власні вектори є нормованими. Отже, дисперсія змінної розкладається в частину, яка пояснюється факторами і частину, яка є притаманною змінній:

$$s_{ii} = \hat{h}_i^2 + \hat{\psi}_i.$$

Таким чином, jй фактор має внесок $\hat{\lambda}_{ij}^2$ до дисперсії s_{ii}, а внесок в сумарну дисперсію $\mathrm{trace}(S) = s_{11} + s_{22} + \cdots + s_{pp}$ дорівнює, відповідно, $\sum_{i=1}^p \hat{\lambda}_{ij}^2$. Отже, пропорція j-го фактору дорівнює

$$\frac{\sum_{i=1}^p \hat{\lambda}_{ij}^2}{\mathrm{trace}(S)} = \frac{\theta_j}{\mathrm{trace}(S)}.$$

Наскільки хорошою є підгонка можна побачити, розглянувши матрицю похибок

$$E = S - \hat{\Lambda}\hat{\Lambda}' - \hat{\Psi}.$$

Діагональні елементи цієї матриці дорівнюють 0. Якщо підгонка є хорошою, то елементи близькі до 0.

Якщо елемени не є пропорційними, то можна проробити аналогічні обчислення з матрицею кореляцій R замість матриці S.

12.2 Метод головних факторів

(principle factor method, principle axis method) Ψ S (R)

В методі головних компонент факторному аналізі ми працювали з , але не з . У методі головних факторів ми спочатку будуємо оцінку $\hat{\Psi}$ та задаємо $\hat{\Lambda}$ за допомоги $S - \hat{\Psi} \approx \hat{\Lambda}\hat{\Lambda}'$ (або $R - \hat{\Psi} \approx \hat{\Lambda}\hat{\Lambda}'$).

Діагональним елементом $S - \hat{\Psi}$ є $\hat{h}_i^2 = \hat{s}_{ii}^2 - \hat{\psi}_i$ (або $\hat{h}_i^2 = 1 - \hat{\psi}_i$, якщо виходити з матриці R). В якості оцінки $\hat{h}_i^2$ можна обрати де множинний коефіцієнт кореляції R_i^2 між y_i на іншими $p-1$ змінніми. Нагадаємо, що

$$R_i^2 =$$

Тоді

$$\hat{h}_i^2 = s_{ii} - \frac{1}{s^{ii}} = s_{ii}R_i^2, \tag{73}$$ hi

де s_{ii} є iм елементом S на діагоналі, та where s^{ii} є iм елементом S^{-1} на діагоналі.

Остання рівність випливає з :::::

Для того, щоб використати оцінку (73) матриця не має бути сингулярною.

Після того, як ми отримали оцінки на середні дисперсії $\hat{h}_i^2$, обчислимо власні значення і влсні числа матриць $S-\hat{\Psi}$ ($S-\hat{\Psi}$), та використаємо рівняння $\hat{\Lambda} = C_1 D_1^{1/2}$ для оцінювання навантажень на фактори. Тоді колонки і рядки можна використати для отримання нових власних значень та середніх дисперсій. Сума квадратів j-го стовбчика $\hat{\Lambda}$ є j-м власним значенням $S-\hat{\Psi}$, та сума квадратів j-го рядка є середньою дісперсією y_i. Частка поясненої j- фактором дисперсієї дорівнює

$$\frac{\theta_j}{\text{trace}(S - \hat{\Psi})} = \frac{\theta_j}{\sum_{i=1}^{p} \theta_i}.$$

Матриця $S-\hat{\Psi}$ (відповідно, $R-\hat{\Psi}$) не обов'язково додатньо визначена, а отже, може мати від'ємні власні числа. У цьому випадку відносна частка поясненої дисперсії може бути більше 1, а потім спадає до 1 по мірі додавання власних чисел.

12.3 Вибір кількості факторів та інтерпретація

Існує декілька критеріїв, щоб обрати кількість факторів m.

1. Оберати кількість факторів m, якої достатньо, щоб пояснити 80% дисперсії.

2. Оберати кількість факторів m, яка дорівнює кількості власних чисел, які більше серенього з них. Для матриці S це число дорівнює $\sum_{j=1}^{p} \theta_j/p$. Якщо використовувати матрицю R, то це середнє число дорівнює 1.

3. Використаємо графічне зображення для того, щоб оцінити, скільки власних значень S (або R) досить. Якщо графік ламаної має критий злам, то для того, то доцідьно вибрати кількість до цього крутого зламу.

4. Перевірити гіпотезу про те, що вірне значення в розмірності матриці , тобто H_0: $\Sigma = \Lambda\Lambda' + \Psi$, де Λ має розмірність $p \times m$.

В останньому випадку використовується статистика

$$\left(n - \frac{2p + 4m + 11}{11}\right) \ln\left(\frac{|\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}|}{|S|}\right),$$

яка за виконання гіпотези H_0 має розподіл χ^2_ν, де $\nu = \frac{1}{2}[(p-m)^2 - m - p]$, а $\hat{\Lambda}$ та $\hat{\Psi}$ є оцінками максимальної вирогідності.

В залежності від потреби можна проводити факторний аналіз, використовуючи коваріаційну або кореляційну матриці. Якщо дисперсії дуже розрізняються, має сенс розглядати матрицю кореляцій R замість S. Те, що дисперсія однієї з компонент значно більша за інші означає, що ця компонента домінує. З іншого боку, якщо фактори некорельовані, їх можна обрати у якості головних компонент.

Головні компоненти отримані за допомоги повороту осей. Якщо після повороту результат не має прозорої інтерпретації, можа здійснити ще один поворот, шукаючи підпростір певної розмірності, щоб інтерпретувати фактори. Тобто, при проектуванні на цей підпостір більшість коефіцієнтів перед факторами має дорівнювати 0.

13 Приклади

Розглянемо приклади[8]. За допомогою факторного аналізу проаналізуємо дані (див. розділ „,) (не забуваємо завантажити library(ggplot2) та library(car))

Обчислимо коваріаційну матрицю та її власні числа.

```
S<- cov(root[,2:5])
S.eigen <- eigen(S)
S.eigen
```

```
eigen() decomposition
$values
[1] 0.495986813 0.162680761 0.006924035 0.001565068

$vectors
           [,1]        [,2]        [,3]       [,4]
[1,] -0.1011191  0.09661363 -0.21551730  0.9664332
[2,] -0.7516463  0.64386366  0.06099466 -0.1294103
[3,] -0.5600279 -0.62651631 -0.52992316 -0.1141384
[4,] -0.3334239 -0.42846553  0.81793239  0.1903481
```

Останні 2 власних числа близьці до 0. Тому досить обрати кількість факторів $m = 2$. Можна також зобразити власні числа графічно:

```
plot(S.eigen$values, xlab = 'Eigenvalue Number', ylab =
        'Eigenvalue Size', main = 'Eigenvalues Size',
                                  type = 'b', xaxt = 'n')
axis(1, at = seq(1, 4, by = 1))
```

[8] https://rpubs.com/aaronsc32/factor-analysis-introduction

Тут type = b означає, що ми малюємо і лінії і відповідні точки; xaxt= n означає, що ми самі задаємо, що саме буде відображатися по осі OX, а саме, ми обираємо axis(1, at = seq(1, 4, by = 1), тобто кількість власних чисел.

Запишимо матрицю коваріацій S у вигляді $S = CD^(1/2)D^(1/2)C'$, та знайдемо C та D. Для того, щоб зобразити матрицю C, візьмемо перші 2 власних вектори. Для того, щоб отримати матрицю D, ми спочатку створюємо порожню матрицю (але відповідного розміру!), а потім додаємо в неї власні числа.

```
C <- as.matrix(S.eigen$vectors[,1:2])
D <- matrix(0, dim(C)[2], dim(C)[2])
diag(D) <- S.eigen$values[1:2]
```

Зипишемо тепер $\hat(\Lambda) = CD^(1/2)$

```
S.loadings <- C %*% sqrt(D)
S.loadings
```

```
           [,1]        [,2]
[1,] -0.07121445  0.03896785
[2,] -0.52935694  0.25969406
[3,] -0.39440707 -0.25269723
[4,] -0.23481824 -0.17281602
```

Все, що ми проробили вище, можна зробити за допомоги вбудованої функції prcomp:

```
root.pca <- prcomp(root[,2:5])$rotation[,1:2]
root.pca
```

Операція $rotation повертає значення власних векторів матриці S (порівняйте з застосуванням S.eigen$vectors[,1:2]). Операція $sdev повертає стандартне відхилення компонент:

```
root.pca2 <- prcomp(root[,2:5])$sdev
```

Обчислимо спільні дисперсії: $\hat{h}_i^2 = \sum_{j=1}^{m} \hat(\lambda)_{ij}^2$:

```
S.h2 <- rowSums(S.loadings^2)
S.h2
```

```
0.006589992 0.347659774 0.219412829 0.085004979
```

Тепер знайдемо $\psi_i = s_{ii} - \hat{h}_i^2$.

```
S.u2 <- diag(S) - S.h2
S.u2
```

```
  V1           V2           V3           V4
1.783368e-03 5.197004e-05 1.964786e-03 4.688978e-03
```

Занйдемо пропорції власних чисел:

```
prop.loadings <- colSums(S.loadings^2)
prop.var <- cbind(prop.loadings[1] / sum(S.eigen$values),
 prop.loadings[2] / sum(S.eigen$values))
prop.var
```

Тобто, перше власне значення має внесок 0.7434338 в загальну суму власних значень, а друге 0.2438419. Зауважимо, що сума цих двох власних чисел (sum(prop.var)) менша за 1 (sum(prop.var)).

Якщо виключити інші власні числа, то отримаємо наступні внески:

```
prop.exp <- cbind(prop.loadings[1] / sum(prop.loadings),
 prop.loadings[2] / sum(prop.loadings))
prop.exp
```

```
          [,1]      [,2]
[1,] 0.7530154 0.2469846
```

З іншого боку, можна застосувати вбудовану функцію fa() з пакету psych:

```
install.packages("psych")
library(psych)
```

Функція principal() виконує факторний аналіз за допомоги методу головних компонент. Поки що ми не використовуємо ротацію факторів, тому поставили rotate = 'none'. Єдине що- для аналізу використовується матриці кореляцій, а не матриця коваріацій[9]

```
root.fa.covar <- principal(root[,2:5], nfactors = 2, rotate = 'none',
                  covar = TRUE)
root.fa.covar
```

[9]https://m-clark.github.io/posts/2020-04-10-psych-explained/

```
Principal Components Analysis
Call: principal(r = root[, 2:5], nfactors = 2, rotate = "none",
covar = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
   PC1   PC2   h2    u2  com
V1 0.79  0.57 0.95 0.051 1.8
V2 0.85  0.47 0.94 0.061 1.6
V3 0.87 -0.45 0.97 0.027 1.5
V4 0.82 -0.55 0.98 0.022 1.7

PC1  PC2
SS loadings           2.78 1.05
Proportion Var        0.70 0.26
Cumulative Var        0.70 0.96
Proportion Explained  0.73 0.27
Cumulative Proportion 0.73 1.00

Mean item complexity =  1.7
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.03
with the empirical chi square  0.39  with prob <  NA

Fit based upon off diagonal values = 1
```

Перевіримо, що саме ці результати ми отримаємо, якщо використаємо кореляційну матрицю у попередніх обчисленнях.

```
R<- cor(root[,2:5])
R.eigen <- eigen(R)
plot(R.eigen$values, xlab = 'Eigenvalue Number', ylab =
                'Eigenvalue Size', main = 'Eigenvalues Size', type = 'b', xaxt =
axis(1, at = seq(1, 4, by = 1))
```

Зобразимо R у вигляді $R = CR(DR)^(1/2)(DR)^(1/2)CR'$ та знайдемо CR та DR.

```
CR <- as.matrix(R.eigen$vectors[,1:2])
DR <- matrix(0, dim(CR)[2], dim(CR)[2])
diag(DR) <- R.eigen$values[1:2]
R.loadings <- CR %*% sqrt(DR)
R.loadings
```

```
          [,1]       [,2]
[1,] 0.7865453  0.5749668
[2,] 0.8493229  0.4666140
[3,] 0.8749282 -0.4549787
[4,] 0.8240901 -0.5466267
```

Тобто, R.loadings - це як раз вектори $PC1$, $PC2$, які знайдено функцією principal(). Обчислимо тепер $\hat{h}_i^2 = \sum_{j=1}^{m} \hat{\lambda}_{ij}^2$ та $\psi_i = r_{ii} - \hat{h}_i^2$

```
R.h2 <- rowSums(R.loadings^2)
R.h2
```

```
0.9492403 0.9390781 0.9725050 0.9779253
```

(порівняйте з колонкою h2 вище). Значення $colSums(R.loadings\hat{}2)$ (або, що те саме, R.eigen$values[1:2]), дорівнюють, відповідно, 2.784627 та 1.054122 (порівняйте з рядком SS loadings вище).

Далі,

```
R.u2 <- diag(R) - R.h2
R.u2
```

```
V1         V2         V3         V4
0.05075965 0.06092192 0.02749496 0.02207470
```

(порівняйте з колонкою u2). Середню дисперсію можна обчислити наступним чином:

```
comR <- rowSums(R.loadings^2)^2 / rowSums(R.loadings^4)
comR
```

(порівняйте з колонкою com). Чим ближчі до 1 значення com, тим карще поясненою є наша модель. Знайдемо пропорцію власних чисел:

```
prop.loadingsR <- colSums(R.loadings^2)
prop.varR <- cbind(prop.loadingsR[1] / sum(R.eigen$values),
        prop.loadingsR[2] / sum(R.eigen$values))
prop.varR
```

(отримаємо, відповідно, рядок Proportion Var). Якщо розглядати лише ці два перші власні числа, то отримаємо

```
prop.expR <- cbind(prop.loadingsR[1] / sum(prop.loadingsR),
        prop.loadingsR[2] / sum(prop.loadingsR))
prop.expR
```

(Отримаємо, відповідно, рядок Proportion explained).

Функція varimax() дозволяє знайти найкращу комбінацію факторів (тобто здійснити "поворот" осей) для того, щоб знайти найкращі навантаження. Найкращим розв'язком був би такий, в якому складність була б близькою до 1, що в свою чергу означає, що одна змінна найкраще пояснюється одним фактором.

В результати ми отримали б $\Lambda^* = \Lambda T$, де матриця T є ортогональною та такою, що максимізує дисперсії навантажень в кожному стовбчику матриці Λ^*. Зауважимо, що при цьому середні дисперсії не змінюються: ортогональне перетворення не змінює власні числа.

```
factors<-R.loadings
varimax(factors)
factors.v <- varimax(factors)$loadings
round(factors.v, 2)
```

```
Loadings:
     [,1] [,2]
[1,] 0.16 0.96
[2,] 0.28 0.93
[3,] 0.94 0.29
[4,] 0.97 0.19

               [,1]  [,2]
SS loadings    1.928 1.907
Proportion Var 0.482 0.477
Cumulative Var 0.482 0.959
```

(ми округлили до другого знаку після коми).

```
h2.v <- rowSums(factors.v^2)
h2.v
```

```
0.9492403 0.9390781 0.9725050 0.9779253
```

(що те саме, що R.h2) Перевіримо, що середні дисперсії не змінилися:

```
u2.v <- 1 - h2.v
u2.v
```

```
0.05075965 0.06092192 0.02749496 0.02207470
```

(що те саме, що R.u2)

Порахуємо складність:

```
com.v <- rowSums(factors.v^2)^2 / rowSums(factors.v^4)
com.v
```

```
1.054355 1.179631 1.185165 1.074226
```

Весь цей аналіз можна проробити за допомоги наступної вбудованої функції:

```
root.fa2 <- principal(root[,2:5], nfactors = 2, rotate = 'varimax')
root.fa2
```

```
Principal Components Analysis
Call: principal(r = root[, 2:5], nfactors = 2, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
   RC1  RC2  h2     u2  com
V1 0.16 0.96 0.95 0.051 1.1
V2 0.28 0.93 0.94 0.061 1.2
V3 0.94 0.29 0.97 0.027 1.2
V4 0.97 0.19 0.98 0.022 1.1

RC1  RC2
SS loadings           1.94 1.90
Proportion Var        0.48 0.48
Cumulative Var        0.48 0.96
Proportion Explained  0.50 0.50
Cumulative Proportion 0.50 1.00

Mean item complexity =  1.1
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.03
with the empirical chi square  0.39  with prob <  NA

Fit based upon off diagonal values = 1
```

Розглянемо аналіз методом головних факторів.

Замінимо діагональні елементи на оцінки (73) (у випадку, коли ми розглядаємо матрицю кореляцій, а не матрицю коваріацій, то $s_{ii} = 1$ в (73), тобто $\hat{h}_i^2 = 1 - \frac{1}{r_{ii}}$)

```
R.smc <- (1 - 1 / diag(solve(R)))
R.smc
diag(R) <- R.smc
round(R, 2)
```

(*.smc squared multiple correlation). Тепер знайдемо власні числа та власні вектори матриці $R - \hat{\Psi}$.

```
r.eigen <- eigen(R)
r.eigen$values
```

Така матриця R вже не є додатньо визначеною, оскільки ми вже замінили елементи на діагоналі їх оцінками. Отже, можуть бути декільки від'ємних власних чисел. Оскільки від'ємні власні числа не використовуються для оцінки λ, оберемо $m = 2$.

```
r.lambda <- as.matrix(r.eigen$vectors[,1:2]) %*%
                          diag(sqrt(r.eigen$values[1:2]))
r.lambda
```

Обчислимо середні дисперсії (communalities), (specific variances) та складність навантажень:

```
r.h2 <- rowSums(r.lambda^2)
r.u2 <- 1 - r.h2
com <- rowSums(r.lambda^2)^2 / rowSums(r.lambda^4)
```

Зберемо результати в наступну data.frame.

```
cor.pa <- data.frame(cbind(round(r.lambda, 2),
                  round(r.h2, 2), round(r.u2, 3), round(com, 1)))
colnames(cor.pa) <- c('PA1', 'PA2', 'h2', 'u2', 'com')
cor.pa
```

```
    PA1   PA2   h2    u2 com
1 -0.74  0.54 0.84 0.158 1.8
2 -0.80  0.45 0.85 0.150 1.6
3 -0.88 -0.41 0.93 0.067 1.4
4 -0.83 -0.50 0.93 0.072 1.6
```

Функція fa() може застосовувати ітеративно метод головних факторів до тих пір, поки ми не отримуємо задовільний результат.

```
root.cor.fa <- fa(root[,2:5], nfactors = 2, rotate = 'none',
                  fm = 'pa', max.iter = 1)
root.cor.fa
```

Застосуємо тепер вбудовану функцію fa(). Параметри fm = 'pa' означає, що ми використовуємо метод головних факторів. SMC= TRUE означає, що ми використаємо квадратичну множинну кореляцію як першу апроксимацію h^2 (інакше $h^2 = 1$). Насправді, цей параметр використовується за замовчуванням, тобто його можна опустити. max.iter = 1 означає, що ми робимо лише одну ітерацію.

```
root.cor.fa2 <- fa(root[,2:5], nfactors = 2, rotate = 'none',
                   fm = 'pa', SMC= TRUE, max.iter = 1)
root.cor.fa2
```

```
Factor Analysis using method =  pa
Call: fa(r = root[, 2:5], nfactors = 2, rotate = "none", max.iter = 1,
fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
PA1   PA2   h2    u2 com
V1 0.74  0.54 0.84 0.158 1.8
V2 0.80  0.45 0.85 0.150 1.6
V3 0.88 -0.41 0.93 0.067 1.4
V4 0.83 -0.50 0.93 0.072 1.6

PA1  PA2
SS loadings           2.65 0.91
Proportion Var        0.66 0.23
Cumulative Var        0.66 0.89
Proportion Explained  0.74 0.26
Cumulative Proportion 0.74 1.00

Mean item complexity =  1.6
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  6  and the objective function was
4.19 with Chi Square of  187.92
The degrees of freedom for the model are -1  and the objective function was  0.17

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  NA

The harmonic number of observations is  48 with the empirical chi square  0.24
with prob <  NA
The total number of observations was  48  with Likelihood Chi Square =  7.26  with prob <

Tucker Lewis Index of factoring reliability =  1.281
Fit based upon off diagonal values = 1
Measures of factor score adequacy
PA1  PA2
Correlation of (regression) scores with factors   0.98 0.93
Multiple R square of scores with factors          0.95 0.86
Minimum correlation of possible factor scores     0.90 0.72
```

Порівняємо результати, якщо встановити поворот осей (тобто параметр rotate = 'varimax')

```
root.cor.fa <- fa(root[,2:5], nfactors = 2, rotate = 'varimax',
                  fm = 'pa', max.iter = 1)
root.cor.fa
```

```
Factor Analysis using method =  pa
Call: fa(r = root[, 2:5], nfactors = 2, rotate = "varimax", max.iter = 1,
fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
PA1  PA2   h2    u2 com
V1 0.19 0.90 0.84 0.158 1.1
V2 0.30 0.87 0.85 0.150 1.2
V3 0.92 0.29 0.93 0.067 1.2
V4 0.95 0.18 0.93 0.072 1.1

PA1  PA2
SS loadings           1.87 1.68
Proportion Var        0.47 0.42
Cumulative Var        0.47 0.89
Proportion Explained  0.53 0.47
Cumulative Proportion 0.53 1.00

Mean item complexity =  1.1
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  6  and the objective function was
4.19 with Chi Square of  187.92
The degrees of freedom for the model are -1  and the objective function was  0.17

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  NA

The harmonic number of observations is  48 with the empirical chi square  0.24
with prob <  NA
The total number of observations was  48  with Likelihood Chi Square =  7.26  with prob <

Tucker Lewis Index of factoring reliability =  1.281
Fit based upon off diagonal values = 1
Measures of factor score adequacy
PA1  PA2
Correlation of (regression) scores with factors   0.97 0.94
Multiple R square of scores with factors          0.94 0.87
Minimum correlation of possible factor scores     0.88 0.75
```

14 Задачі класифікації

Нехай є вибірка спостережень, про яку попередньо відомо те, що ця вибірка складається з k груп. Наша задача- класифікувати кожне спостереження, тобто віднести його до однієї з груп.

Один із методів – це зробити класифікацію, беручи до уваги вже відому інформацію про групи. А саме, на основі результатів попередньої класифікації на "групі для тренування". Наприклад, можна порівнювати спостередення з векторами середніх, отриманих на попередньому кроці, та обрати групу відповідно тому, до якого середнього ближче наше спостереження.

Задача класифікації може виникнути, наприклад, коли треба класифікувати пацієнтів,

чи мають вони певну хворобу чи ні на основі тестів.

Розглянемо деякі методи класифікації.

14.1 Метод Фішера (лінійний класифікатор), 2 групи

Нехай є дві популяції, які мають однакові коваріаційні матриці $\Sigma_1 = \Sigma_2$ (тут ми не припускаємо, що популяції нормально розподілені). Побудуємо дискримінантну функцію

$$z = a'y = (\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1} y.$$

Тоді ми будемо говорити, що спостереження y належить до G_1, якщо z знаходиться ближче до $\bar{z}_1 = a'\bar{y}_1$ ніж до $\bar{z}_2 = a'\bar{y}_2$. В іншому випадку відносимо z до G_2.

Нехай

$$z > \frac{\bar{z}_1 + \bar{z}_2}{2}. \tag{74}$$ z1

З того, як обране a, маємо

$$\bar{z}_1 - \bar{z}_2 = a'(\bar{y}_1 - \bar{y}_2) = (\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1} (\bar{y}_1 - \bar{y}_2) > 0.$$

Оскільки $\frac{\bar{z}_1 + \bar{z}_2}{2}$ є середньою точкою, то якщо виконано (74), то z знаходиться ближче до $\bar{z}_1 = a'\bar{y}_1$ ніж до $\bar{z}_2 = a'\bar{y}_2$. Отже, ми класифікуємо z до групи G_1.

Це правило класифікації, яке має назку метод Фішера, є лінійним. Більш того, якщо розподіл виборок є нормальним, то таке розбиття є асимптотично оптимальним (див. 9).

14.2 Байєсів класифікатор

Розглянемо спочатку випадок двох груп. Нехай $p_i = \mathbb{P}(\text{спостереження належить } G_i)$, $i = 1, 2$, $p_1 + p_2 = 1$, є апріорними ймовірностями того, що дане спостереження належить до групи G_i. Припустимо також, що відомі щільності розподілів $f(y|G_i)$, $i = 1, 2$, за умови того, що спостереження знаходиться в групі G_i. Тоді відносимо спостереження y до G_1, якщо

$$p_1 f(y|G_1) > p_2 f(y|G_2), \tag{75}$$ welch1

інакше відносимо y до G_2.

Якщо $p_1 = p_2$, то $y \in G_1$, якщо $f(y|G_1) > f(y|G_2)$, тобто відношення вирогідностей > 1.

Запишемо (75), $i = 1, 2$, $p_1 + p_2 = 1$, у випадку, коли у нас є номально розподілена вибірка.

Приклад 2. Нехай $f(y|G_i)$ є щільностями нормального розподілу $\mathrm{N}_p(\mu_i, \Sigma)$, $i = 1, 2$. Тоді (75) має місце тоді і тільки тоді, коли

$$(\mu_1 - \mu_2)' \Sigma^{-1} y > \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2) + \ln \frac{p_2}{p_1}. \tag{76}$$ welch2

Дійсно, використовуючи те, що матриця Σ симетрична, а отже $a'\Sigma^{-1}b = b'\Sigma^{-1}a$, $a, b \in \mathbb{R}^p$, отримаємо

$$\frac{p_2}{p_1} < \frac{f(y|G_1)}{f(y|G_2)} = \exp\left\{(\mu_1 - \mu_2)'\Sigma^{-1} y - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)\right\}.$$

Взявши логарифм і перегрупувавгши, і отримаємо (76).

На практиці, в нерівності (76) ми не знаємо ні середні μ_i, $i = 1, 2$, ні коваріаційну матрицю Σ. Тому для класифікації нам треба спочатку оцінити ці середні і коваріаційну матрицю (через $\bar{y}_1$, $\bar{y}_2$ і S_{pl}, відповідно). Отже, нам потрібно замінити нерівність (76) на

$$(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1} y > \frac{1}{2}(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1} (\bar{y}_1 + \bar{y}_2) + \ln \frac{p_2}{p_1}. \tag{77}$$ welch22

Таким чином, якщо для спостереження виконується (77), то відносимо спостереження y до групи G_1, якщо не виконується– то до групи G_2.

За теоремою Байєса ми можемо переоцінити апріорні ймовірності p_i:

$$\begin{aligned}\mathbb{P}(G_i|y) &= \frac{\mathbb{P}(G_i)f(y|G_i)}{\mathbb{P}(G_1)f(y|G_1)+\mathbb{P}(G_2)f(y|G_2)} \\ &= \frac{p_2 f(y|G_2)}{p_1 f(y|G_1)+p_2 f(y|G_2)}, \quad i=1,2.\end{aligned} \tag{78}$$

bayes1

У випадку двох груп така переоцінка нічого не змінює: $\mathbb{P}(G_1|y) > \mathbb{P}(G_2|y)$ тоді і тільки тоді, коли має місце (75).

14.3 Випадкок декількох груп, спільна коваріаційна матриця

Нехай у нас є k груп нормально розподілених спостережень, та відомі апріорні ймовірності

$$p_i = \mathbb{P}(\text{спостереження належить } G_i), \quad i=1,\dots,k.$$

Ми будемо відносити спостереження y до групи G_i, якщо $p_i f(y|G_i) > p_j f(y|G_j)$, $j = 1,\dots,k$, $j \neq i$.

Розглянемо

$$\begin{aligned}\ln(p_i f(y|G_i)) &= \ln p_i - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(y-\mu_i)'\Sigma^{-1}(y-\mu_i) \\ &= \ln p_i + \mu_i'\Sigma^{-1}y - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \frac{1}{2}y'\Sigma^{-1}y.\end{aligned} \tag{79}$$

lndens

Розглянемо функцію (замінимо теоретичні середні μ_i на вибіркові $\bar{y}_i$)

$$L_i(y) = \ln p_i + \bar{y}_i'\Sigma^{-1}y - \frac{1}{2}\bar{y}_i'S_{pl}^{-1}\bar{y}_i.$$

Тут $L_i(y)$ є лінійною фунцією від y; нагадаємо, що у випадку k груп зважена коваріація визначається наступним чином:

$$S_{pl} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)S_i = \frac{E}{N-k}, \quad N = \sum_{i=1}^{k} n_i.$$

Розглянемо перший рядок (79), в якому теоретичні середні і коваріація замінені на $\bar{y}_i$, $i=1,2$, та S_{pl}. Зауважимо, що вираз

$$\ln p_i - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(y-\bar{y}_i)'S_{pl}^{-1}(y-\bar{y}_i)$$

досягає максимума тоді і тільки тоді, коли $L_i(y)$, оскільк частина $\frac{1}{2}y'\Sigma^{-1}y$ є спільною для всіх k груп спостережень.

Нехай

$$D_i^2(y) = (y-\bar{y}_i)'S_{pl}^{-1}(y-\bar{y}_i).$$

Тоді

$$D_i^2(y) \quad \text{мінімальне} \quad \Longleftrightarrow \quad L_i(y) \quad \text{максимальне}.$$

Отже, ми відносимо спостереження y до групи G_i, якщо функція $L_i(y)$ є максимельною для серед всіх k функцій.

Розглянемо апостеріорні ймовірності

$$\mathbb{P}(G_i|y) = \frac{p_i f(y|G_i)}{\sum_{j=1}^k p_j f(y|G_j)}.$$

Ми відносимо y до групи G_i, якщо ймовірність $\mathbb{P}(G_i|y)$ максимальна; зауважимо, що це відбувається тоді і тільки тоді, коли максимальною є $p_i f(y|G_i)$.

Приклад 3. Нехай $f(y|G_i)$ є щільностями нормального розподілу $\mathrm{N}_p(\mu_i, \Sigma)$, $i = 1, 2$, але вектор середніх μ_i і коваріація Σ - невідомі. Тоді розглянемо апостеріорні ймовірності, в яких теоретичні значення замінені оцінками:

$$\tilde{\mathbb{P}}(G_i|y) := \frac{p_i e^{-D_i^2/2}}{\sum_{j=1}^k p_j e^{-D_j^2/2}}, \quad i = 1, \ldots, k.$$

Відповідно, робимо класифікацію на основі ймовірностей $\tilde{\mathbb{P}}(G_i|y)$, $i = 1, \ldots, k$.

14.4 Похибка класифікації

Розглянемо похибку класифікації у випадку двох груп (див. **??**с.240]R98):

$$\begin{aligned}\mathcal{E} &= \mathbb{P}\{\text{випробування } y \text{ класифіковано до іншої групи }\} \\ &= p_1 \mathbb{P}(\ y \text{ класифіковано до } G_2|G_1) + p_2 \mathbb{P}(y \text{ класифіковано до } G_1|G_2),\end{aligned}$$

де p_1, p_2- апріорні ймовірності. $\mathcal{E}$- це ймовірність класифікувати y до неправильної групи. Знайдемо $\mathbb{P}$(класифіковано до $G_i|G_j$), $i, j = 1, 2$. Якщо ми класифікували y до G_1, то виконується (76). Позначимо

$$\alpha' y := (\mu_1 - \mu_2)'\Sigma^{-1} y, \quad \Delta^2 := (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2).$$

Тоді

$$\begin{aligned}\mathbb{P}(\text{класифіковано до } G_1|G_2) &= \mathbb{P}\left(\alpha' y > \frac{1}{2}(\mu_1 - \mu_2)'\Sigma(\mu_1 + \mu_2) + \ln\frac{p_2}{p_1}\right) \\ &= \mathbb{P}\left(\frac{\alpha' y - \alpha'\mu_2}{\Delta} > \frac{\frac{1}{2}(\mu_1 - \mu_2)'\Sigma(\mu_1 + \mu_2) + \ln\frac{p_2}{p_1}\alpha'\mu_2}{\Delta}\right),\end{aligned}$$

де ми використали те, що

$$\alpha'\Sigma\alpha = (\mu_1 - \mu_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1 - \mu_2) = \Delta^2,$$

тобто є відстанню Махаланобіса. Покладемо $w = \frac{\alpha' y - \alpha'\mu_2}{\Delta}$. Оскільки насправді $y \sim \mathrm{N}(\mu_2, \Sigma)$, то після нормування $w \sim \mathrm{N}_p(0, 1)$. Тоді

$$\begin{aligned}\mathbb{P}(\text{класифіковано до } G_1|G_2) &= \mathbb{P}\left(\frac{\alpha' y - \alpha'\mu_2}{\Delta} > \frac{\frac{1}{2}(\mu_1 - \mu_2)'\Sigma(\mu_1 - \mu_2) + \ln\frac{p_2}{p_1} - \alpha'\mu_2}{\Delta}\right) \\ &= \mathbb{P}\left(w > \frac{\frac{1}{2}\Delta^2 + \ln\frac{p_2}{p_1}}{\Delta}\right) \\ &= \mathbb{P}\left(w < -\frac{\frac{1}{2}\Delta^2 + \ln\frac{p_2}{p_1}}{\Delta}\right) \\ &= \Phi\left(\frac{-\frac{1}{2}\Delta^2 - \ln\frac{p_2}{p_1}}{\Delta}\right)\end{aligned}$$

де Φ- функція розподілу $w \sim \mathrm{N}(0,1)$. Аналогічно,

$$\mathbb{P}(\text{класифіковано до } G_2|G_1) = \Phi\left(\frac{-\frac{1}{2}\Delta^2 - \ln\frac{p_1}{p_2}}{\Delta}\right).$$

Отже,

$$\mathcal{E} = p_1\Phi\left(\frac{-\frac{1}{2}\Delta^2 - \ln\frac{p_2}{p_1}}{\Delta}\right) + p_2\Phi\left(\frac{-\frac{1}{2}\Delta^2 - \ln\frac{p_1}{p_2}}{\Delta}\right).$$

Підставивши замість Δ^2 оцінку $D^2 = (\bar{y}_1 - \bar{y}_2)'S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2)$, отримаємо оцінку похибки класифікації

$$\hat{\mathcal{E}} = p_1\Phi\left(\frac{-\frac{1}{2}D^2 - \ln\frac{p_2}{p_1}}{D}\right) + p_2\Phi\left(\frac{-\frac{1}{2}D^2 - \ln\frac{p_1}{p_2}}{D}\right).$$

Якщо спостереження не є нормальними, можна застосувати ядерні оцінки щільності для оцінювання $f(y|G)$, див. Додаток 1.

14.5 Метод найближчих сусідів

Одним з простих найбільш поширених методів є метод k найближчих сусідів. Розглянемо відстань між y_i та y_j:

$$d(y_i, y_j) = (y_i - y_j)'S_{pl}^{-1}(y_i - y_j).$$

Для того, щоб віднести спостереження до тієї чи іншої групи, дивимося на k найближчих сусідів цього спостереження. Якщо більшість з них належить до G_1, то відносимо y до G_1, і навпаки. Якщо вибірки різного розміру, то відносимо y_i до G_1, якщо $\frac{k_1}{n_1} > \frac{k_2}{n_2}$. Якщо відомі апріорні ймовірності p_1, p_2, то ми відносимо y_i до G_1, якщо

$$p_1\frac{k_1}{n_1} > p_2\frac{k_2}{n_2}.$$

15 Приклади

Розглянемо спочатку класифікацію за допомоги дискримінантної функції [10].

Завантажимо таблицю, в якій наведені вимірювання 2х типів жуків (для простоти позначимо вимірювання через колонки V3–V6).

```
beetles <- read.table('c:/your_path/T5_5_FBEETLES.DAT',
   col.names = c('Measurement.Number','Species','V3','V4','V5','V6'))
beetles
```

Розділимо цю таблиці на 2 за принципом: якщо Species = 1, то це перша група, якщо Species = 2- то друга.

```
beetle1 <- beetles[beetles$Species == 1,][,3:6]
beetle2 <- beetles[beetles$Species == 2,][,3:6]
```

[10] https://rpubs.com/aaronsc32/classification-linear-discriminant-analysis

Знайдемо середні значення та зважену коваріацію (нагадаємо, що 2 в функції apply означає, що ми сумуємо по стовбчикам).

```
n1 <- nrow(beetle1)
n2 <- nrow(beetle2)

beetle1.means <- apply(beetle1, 2, mean)
beetle2.means <- apply(beetle2, 2, mean)

w1 <- (n1 - 1) * var(beetle1)
w2 <- (n2 - 1) * var(beetle2)

sp1 <- 1 / (n1 + n2 - 2) * (w1 + w2)
```

Позначимо через cutoff середнє значення, за допомогою якого ми будемо визначати, до якої групи належить випробування (див. також розділ Дискримінантний аналіз).

```
cutoff <- .5 * (beetle1.means - beetle2.means) %*% solve(sp1)
                %*% (beetle1.means + beetle2.means)
cutoff
```

Якщо z<cutoff= 15.81, то ми віднесемо спостереження до першої групи, якшо ні- то до другої.

```
species.prediction <- apply(beetles[,3:6], 1, function(y) {
        z <- (beetle1.means - beetle2.means) %*% solve(sp1) %*% y
ifelse(z > cutoff, 1, 2)
})
```

Розглянемо матрицю змішування (confusion matrix), яка показує, наскільки якісн ми віднесли елементи до тієї чи іншої групи.

```
table(beetles$Species, species.prediction,
                dnn = c('Actual Group','Predicted Group'))
```

```
              Predicted Group
Actual Group      1        2
      1           19       0
      2           1        19
```

Ми класифікували вірно всі спостереження, які відносяться до першої групи, але помилково класифікували одне спостереження з другої групи. Похибка класифікації- це кількість хибних класифікацій поділити на кількість елементів у вибірці.

```
n <- dim(beetles)[1]
```

Тобто $1/n = 0.02564103$. Як ми бачимо, навіть при малому об'ємі виборки похибка є досить малою.

Для того, щоб зробити класифікацію за допомогою вбудованої функції, використаємо функцію lda() з пакету MASS.

```
library(MASS)
beetle.lda <- lda(Species ~ .-Measurement.Number, data = beetles)
lda.pred <- predict(beetle.lda)$class
```

Матриця змішання демонструє співвідношення результату класифікації та фактичної належності до тієї чи іншої групи.

```
table(beetles$Species, lda.pred,
                dnn = c('Actual Group','Predicted Group'))
```

```
             Predicted Group
Actual Group       1   2
           1       19  0
           2       1   19
```

У наступному прикладі ми розглянемо байєсівську класифікацію[11]. Для цього ми розглянемо дані по квіткам ірис та спробуємо побудувати класифікатор. Нам потрібно інсталювати наступні пакети та завантажити відповідні бібліотеки.

```
install.packages("e1071")
install.packages("caTools")
install.packages("caret")

library(e1071)
library(caTools)
library(caret)
```

База даних квітів ірис містить три типи квіток: setosa, viriginica, versicolor, та їх характеристики-довжина та ширина стебла, та довжина та ширина пелюстків.

```
data(iris)
head(iris)
```

[11] https://duttashi.github.io/blog/splitting-a-data-frame-into-training-and-testing-sets-in-r/

```
 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
```

Розіб'ємо дані на дві групи у співвідношенні 0.7:0.3, тобто 70 даних- це навчальний набір, та це тестуваотний набір:

```
iris$spl=sample.split(iris,SplitRatio=0.7)
train=subset(iris, iris$spl==TRUE)
test=subset(iris, iris$spl==FALSE)
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species   spl
1          5.1         3.5          1.4         0.2  setosa  TRUE
2          4.9         3.0          1.4         0.2  setosa FALSE
3          4.7         3.2          1.3         0.2  setosa  TRUE
4          4.6         3.1          1.5         0.2  setosa  TRUE
5          5.0         3.6          1.4         0.2  setosa FALSE
6          5.4         3.9          1.7         0.4  setosa  TRUE
```

Альтернативно, це розбиття можна зробити наступним чином:

```
split <- sample.split(iris, SplitRatio = 0.7)
train_cl <- subset(iris, split == "TRUE")
test_cl <- subset(iris, split == "FALSE")
```

Застосуємо тепер байєсівську класифікацію до наачального набору:

```
set.seed(120)
classifier_cl <- naiveBayes(Species ~ ., data = train_cl)
classifier_cl
```

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
setosa versicolor  virginica
0.33       0.34       0.33
```

```
Conditional probabilities:
Sepal.Length
Y                [,1]      [,2]
setosa     5.003030 0.3901000
versicolor 5.897059 0.5578559
virginica  6.660606 0.6123415

Sepal.Width
Y                [,1]      [,2]
setosa     3.418182 0.4216445
versicolor 2.705882 0.3365972
virginica  2.948485 0.3308334

Petal.Length
Y                [,1]      [,2]
setosa     1.487879 0.1745666
versicolor 4.226471 0.4937849
virginica  5.584848 0.5489315

Petal.Width
Y                 [,1]       [,2]
setosa     0.2454545 0.09711755
versicolor 1.3205882 0.19814199
virginica  1.9939394 0.26450354

spl
Y                FALSE      TRUE
setosa     0.3636364 0.6363636
versicolor 0.4117647 0.5882353
virginica  0.4242424 0.5757576
```

Зауважимо, що у вищенаведеній табліці наведено не умовні ймовірності, а середні значення та стандартні відхилення. При цьому ми припускали, що дані є нормально розподіленими випадковими величинами. The given table of conditional probabilities is not showing the Якщо розподіл не є наперед відомим, потрібно використовувати ядерні оцінки щільності ядра.

```
classifier2_cl <- naiveBayes(Species ~ ., data = train_cl,usekernel=TRUE,
kernel="triangular" )
classifier2_cl
```

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace, usekernel = TRUE,
kernel = "triangular")

A-priori probabilities:
Y
```

```
setosa versicolor  virginica
0.33       0.34       0.33

Conditional probabilities:
Sepal.Length
Y                [,1]      [,2]
setosa     5.003030 0.3901000
versicolor 5.897059 0.5578559
virginica  6.660606 0.6123415

Sepal.Width
Y                [,1]      [,2]
setosa     3.418182 0.4216445
versicolor 2.705882 0.3365972
virginica  2.948485 0.3308334

Petal.Length
Y                [,1]      [,2]
setosa     1.487879 0.1745666
versicolor 4.226471 0.4937849
virginica  5.584848 0.5489315

Petal.Width
Y                 [,1]       [,2]
setosa     0.2454545 0.09711755
versicolor 1.3205882 0.19814199
virginica  1.9939394 0.26450354

spl
Y                FALSE      TRUE
setosa     0.3636364 0.6363636
versicolor 0.4117647 0.5882353
virginica  0.4242424 0.5757576
```

Можна вивести інформацію, яка випробування відноситься до якого класу:

```
y_pred <- predict(classifier_cl, newdata = test_cl)
y_pred
```

```
[1] setosa     setosa     setosa     setosa     setosa     setosa     setosa
[8] setosa     setosa     setosa     setosa     setosa     setosa     setosa
[15] setosa     setosa     setosa     virginica  versicolor versicolor versicolor
[22] versicolor versicolor virginica  versicolor versicolor versicolor versicolor
[29] versicolor versicolor versicolor versicolor versicolor virginica  virginica
[36] versicolor virginica  virginica  virginica  virginica  virginica  virginica
[43] virginica  virginica  versicolor virginica  virginica  virginica  virginica
[50] virginica
Levels: setosa versicolor virginica
```

Або можна вивести матрицю змішування:

```r
cm <- table(test_cl$Species, y_pred)
cm
```

```
   y_pred
              setosa versicolor virginica
setosa         17          0         0
versicolor      0         14         2
virginica       0          2        15
```

```r
confusionMatrix(cm)
```

```
Confusion Matrix and Statistics

y_pred
setosa versicolor virginica
setosa         17          0         0
versicolor      0         14         2
virginica       0          2        15

Overall Statistics

Accuracy : 0.92
95% CI : (0.8077, 0.9778)
No Information Rate : 0.34
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.88

Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa   Class: versicolor Class: virginica
Sensitivity                1.00            0.8750           0.8824
Specificity                1.00            0.9412           0.9394
Pos Pred Value             1.00            0.8750           0.8824
Neg Pred Value             1.00            0.9412           0.9394
Prevalence                 0.34            0.3200           0.3400
Detection Rate             0.34            0.2800           0.3000
Detection Prevalence       0.34            0.3200           0.3400
Balanced Accuracy          1.00            0.9081           0.9109
```

Додаток 1

Ядерні оцінки щільності

Нехай ξ має щільність розподілу $f(y)$, та нехай $N(y_0)/n$- кількість спостережень, що попали в інтервал (y_0-h, y_0+h). Оскільки

$$\int_{y_0-h}^{y_0+h} f(z)dz = \mathbb{P}(y_0-h<\xi\le y_0+h) \asymp \frac{N(y_0)}{n},$$

то

$$\hat{f}(y) \asymp \frac{N(y_0)}{2hn}.$$

Наша задача- оцінити $N(y)$ за допомогою ядра $K(y)$, тому такі оцінки щільності називаються ядерними.

Нехай

$$K(u) = \begin{cases} \frac{1}{2}, & |u|\le 1; \\ 0, & |u|>1. \end{cases}$$

Тоді

$$N(y_0) = 2\sum_{i=1}^{n} K\left(\frac{y_0-y_i}{h}\right)$$

підраховує кількість точок, які попали в інтервал $y_0-h\le y_i\le y_0+h$ (якщо $|y_i-y_0|\le h$, то це як раз і означає, що точка y_i попала в інтервал $y_0-h\le y_i\le y_0+h$). Тоді

$$\hat{f}(y_0) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y_0-y_i}{h}\right).$$

Можні обирати і інші ядра. Наведемо кілька прикладів.

- Ядро Єпанчикова:
$$K(u) = \frac{3}{4}(1-x^2)\mathbb{1}_{|x|\le 1};$$
- Трикутне ядро:
$$K(u) = (1-|u|)\mathbb{1}_{|u|\le 1}.$$
- Гаусівске ядро:
$$K(u) = \frac{e^{-u^2/2}}{\sqrt{2\pi}}.$$
- І ще одне ядро:
$$K(u) = \frac{1}{\pi}\frac{\sin^2 u}{u^2}.$$

У багатовимірному випадку оцінка щільності виглядає наступним чином:

$$\hat{f}(y_0) = \frac{1}{nh_1h_2\dots h_p}\sum_{i=1}^{n} K\left(\frac{y_{01}-y_{i1}}{h_1}, \frac{y_{02}-y_{i2}}{h_1}, \dots, \frac{y_{0p}-y_{ip}}{h_1}\right).$$

Розглянемо кілька прикладів наближення щільності.
Для довільного ядра K можна побудувати графік $f(x)$ з (**??**).
Розглянемо дані, які вже використовувались у розділі **??**.

```
library(ggplot2)
Foot<-read.table('c:/your_path/T8_3_FOOTBALL.DAT',
                 col.names = c('Group', 'V2', 'V3', 'V4', 'V5', 'V6','V7') )
```

Ми використаємо функціію density. Параметром може бути одна з координат вектору $kernel = c("gaussian"\, "epanechnikov"\, "rectangular"\, "triangular"\, "biweight"\, "cosine"\, "optcosine")$. Наприклад,

```
den1<-density(x, bw = "nrd0", adjust = 1, kernel = "rectangular",
                 weights = NULL)
plot(den1)
```

```
den2<-density(x, bw = "nrd0", adjust = 1, kernel = "rectangular", weights = NULL)
plot(den2)
```

Можна побудувати [12]

16 Кластерний аналіз

В кластерному аналізі ми групуємо спостереження в групи таким чином, щоб групи розрізнялися між собою, але при цьому в середині групи об'єкти мали схожі властивості. Наша мета- знайти оптимальне групування.

Кластерний аналіз принципово відрізняється від класифікації даних, яку буде розглянуто у наступній главі. Задача класифакації- розподілити об'єкти по попередньо визначеним групам. В кластерному аналізі ані самі групи ані число груп не визначено напаред. Для того, щоб віднести об'єкт до певного кластеру знаходять спільні риси об'єктів (наприклад, вимірюють "відстань" між ними). Інший підхід полягає в тому, щоб обрати центр кластеру та порівнювати відстані об'єктів до центру або до іншого кластеру. МОжна також відносити об'єкти до того чи іншого кластеру, порівнюючи кореляції.

Розглянемо наступні типи кластерізації. В *ієархічній кластерізації* починаємо з n об'єктів і поступово об'єднуємо їх в кластери; на виході ми отримаємо один кластер. Можна також робити кластерізацію у зворотньому боці– почати з одного кластеру і розбивати на частини.

В *кластерізації розбиттям* ми просто розбіваємо вибірку на g кластерів. Це можна зробити обравши центри і вимірюючи відстані до центрів. Інші статистичні) методи використовують матриці H і E з MANOVA.

При вимірюванні відстані до центру використовуються наступні методи.

- **Метод найближчого сусіда (nearest neighbour method).** Відстань між кластерами задається наступним чином:

$$D(A,B) = \min\{d(y_i, y_j) : \quad y_i \in A, \quad y_j \in B\},$$

[12] https://stackoverflow.com/questions/28077500/find-the-probability-density-of-a-new-data-point-using-density-function-in-r

де $d(\cdot,\cdot)$ є евклідовою (або іншою відстанню) між точками. Два кластери, відстань між якими мінімальна, об'єднують в один. В результаті отримуємо наступне "дерево", або *дендрограму:*

- **Метод найшвидшого сусіда (fastest neighbour method).**

$$D(A,B) = \max\{d(y_i, y_j) : \quad y_i \in A, \quad y_j \in B\},$$

Два кластери, у яких $D(A,B)$ мінімальне, об'єднують в один.

- **Метод середньої відстані (average linkage method).**

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j).$$

Так само: два кластери, у яких відстань $D(A,B)$ мінімальна, об'єднують в один.

- **Метод центроїда (centroid method).**

$$D(A,B) = d(\bar{y}_A, \bar{y}_B), \quad \bar{y}_A := \frac{1}{n_A} \sum_{i=1}^{n_A}.$$

Об'єднуємо в один два кластери, у яких відстань $D(A,B)$ мінімальна, та обираємо новий центр за формулою

$$\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B}.$$

- **Метод медіан (median method).** Якщо $n_A \gg n_B$, то новий центроїд $\bar{y}_{AB}$ хначно ближчий до $\bar{y}_A$ ніж до $\bar{y}_B$. Для того, щоб не нормувати середні відповідно до ваги, можна використовувати медіану

$$m_{AB} = \frac{1}{2}(\bar{y}_A + \bar{y}_B).$$

для того, доб рахувати відстані до інших кластерів.

Два кластери з найменшою відстанню між медіанами зклеюємо в один. Зауважимо, що це не "звичайна" медіана в статистичному сенсі, це медіана в геометричному сенсі.

- **Метод Варда (Ward's method).**

Метод Варда (або the incremental sum of squares method), використовує відстані внутрі кластерів та відстані між кластерів. Якщо AB є кластером, який отримано шляхом комбінування кластерів A і B,то суми внутрі кластерів дорівнюють

$$SSE_A := \sum_{i=1}^{n_A} (y_i - \bar{y}_A)'(y_i - \bar{y}_A),$$

$$SSE_B := \sum_{i=1}^{n_B} (y_i - \bar{y}_B)'(y_i - \bar{y}_B),$$

$$SSE_{AB} := \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})'(y_i - \bar{y}_{AB}).$$

Метод Варда полягає в тому, щоб з'єднувати два кластери, у яких відстань

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B) = \frac{n_A n_B}{n_A + n_B}(\bar{y}_A - \bar{y}_B)'(\bar{y}_A - \bar{y}_B).$$

є мінімальною.

Порівняємо метод Варда і метод центроїда. Якщо піднести відстань у методі центроїда до квадрату, то побачимо, що єлина відмінність - це коефіцієнт $\frac{n_A n_B}{n_A+n_B}$. Отже, розмір кластера має вплив на метож Варда, а не на метод центроїда. Оскільки

$$\frac{n_A n_B}{n_A + n_B} = \frac{1}{1/n_A + 1/n_B},$$

то $\frac{n_A n_B}{n_A+n_B}$ зростає при зростанні n_A та n_B. З іншого боку,

$$\frac{n_A n_B}{n_A + n_B} = \frac{n_A}{1 + n_A/n_B},$$

а отже, якщо n_B зростає при фіксованому n_A, то $\frac{n_A n_B}{n_A+n_B}$ зростає.Отже, порівняно з методом центроїда, метод Варда скоріше з'єднує малі кластери або кластери однакового розміру.

Розглянемо тепер неієархічні методи: метод розбіття, методи з використанням MANOVA, та методи, що базуються на оціниванні щільності розподілу.

У кластерізації методом розбиття спостереження розбівають на g груп без ієарзічного групування. Ідеально було б розглянути всі можливі розбиття, але така процедура дуже громіздка, тому розглядають інші підходи.

- **Метод k середніх (k-means method)**

  Цей метод дозволяє пересувати елементи з одного кластеру в інший, що не є можливим в ієархічних методах кластерізації. Спочатку ми обираємо g центроїдами кластерів. Ці g елементів можна обрати випадково, наприклад, обрати перші елементів у вибірці, або точок, які мають максимальні відстані, або точок, в яких щільність (розподілу) є максимальна і т.д. При цьому число кластерів g обирається наперед. Після цього інші точки відносять до кластерів за принципом мінімальної відстані. Як тількі кластер має більш ніж один елемент, центр кластеру обчилюєтося за методом центроїда. Як тільки всі елементи розподілені між кластерами, кожен елемент кластеру повторно розглядають, чи знаходиться він ближче до іншого центроїда чи ні. Якщо так, то цей елемент відносять до іншого кластеру. Цю процедуру повторюють до тих пір, поки покращення вже неможливі. Також, метод k середніх можна комбінувати з ієархічними методами.

- **MANOVA метод**

  Розглянемо методи, засновані на аналізі матриць H та E для однофакторної моделі MANOVA. Якщо кластери "добре визначені", то матриця похибок має бути "малою", а матриця гіпотези - "великою".

  Це можна зробити наступними методами

  1. мінімізувати
  $$\text{trace}(E) = \text{trace}\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)'(y_{ij} - \bar{y}_i);$$
  2. мінімізувати $|E|$, що є еквівалентом мінімізації Λ Уілкса;

3. максимізуівати $\operatorname{trace}(E-H)$, або максимізувати

$$\operatorname{trace} E^{-1}H = \sum_{i=1}^{s} \lambda_i.$$

При цьому найбільше власне значення λ_1 (яке відповідає дискримінантній функції $z_1 = a_1'y$, де a_1 є відповідним власним вектором) має найбільший вплив на $\operatorname{trace} E^{-1}H$.

- **Метод сумішей**

У цьому методі ми припускаємо існування g багатовимірних нормальних розподілів та маємо на мені розподілити кожне з спостережень по цим g класам (див. також). Визначимо щільність суміші наступним чином:

$$h(y) = \sum_{i=1}^{g} \alpha_i f(y, \mu_i, \Sigma_i),$$

де $0 \le \alpha_i \le 1$, $\sum_{i=1}^{g} \alpha_i = 1$, а $f(y, \mu_i, \Sigma_i)$ є щільністю нормального розподілу $\mathrm{N}_p(\mu_i, \Sigma_i)$.

Далі формуємо кластери настуним чином. Спостереження y відносимо до кластеру C_i з найбільшою вирогідністю

$$\hat{P}(C_i|y) = \frac{\hat{\alpha}_i f(y, \hat{\mu}_i, \hat{\Sigma}_i)}{h(y)},$$

де $\hat{\alpha}_i$, $\hat{\mu}_i$, $\hat{\Sigma}_i$ є оцінками максимальної вирогідності.

Це підхід можна використати ітеративно. Можна віднести y відносимо до кластеру C_i з найбільшим значенням

$$\ln \hat{\alpha}_i - \frac{1}{2}\ln|\hat{\Sigma}_i| - \frac{1}{2}(y-\hat{\mu}_i)'\hat{\Sigma}_i^{-1}(y-\hat{\mu}_i),$$

де

$$\hat{\alpha}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{P}(C_i|y_j), \quad i = 1, \dots, g-1,$$

$$\hat{\mu}_i = \frac{1}{n\hat{\alpha}_i}\sum_{j=1}^{n} y_j \hat{P}(C_i|y_j), \quad i = 1, \dots, g,$$

$$\hat{\Sigma}_i = \frac{1}{n\hat{\alpha}_i}\sum_{j=1}^{n} (y-\hat{\mu}_i)(y-\hat{\mu}_i)'\hat{P}(C_i|y_j), \quad i = 1, \dots, g,$$

а $\hat{P}(C_i|y_j)$ визначена вище. Цю процедуру можна повторювати ітеративно. Якщо g невідомо, можна почати з $g = 1$, а потім спробувати $g = 2$, $g = 3$ і т.д. до тих пір, поки результат не буде задовільний.

Про методи, як оптимізувати вибір кількості кластерів, можна почитати в [R02, §14.5].

17 Приклади

Завантажимо наступні бібліотеки:

```
library(dplyr)
# install.packages("factoextra")
library(factoextra)
library(cluster)
```

(якщо не встановлено, то потрібно встановити пакет "factoextra"). Завантажимо таблицю, в якій вказані дані по злочинності у містах. (див. Таблиця 15.13 [R02]).

```
city<-read.table("c:/your_path/T14_1_CITYCRIME.dat")
city
```

Для виконання кластерного аналізу використаємо функцію hclust; за замовченням, використовується евклідова відстань.

```
hc<- hclust(dist(city[,2:8]))
hc
```

Якщо потрібно змініти відстань, треба задати інший параметр в method, наприклад, method="manhattan"

```
hc1 <- hclust(dist(city[,2:8], method="manhattan"))
hc1
```

```
hc2 <- hclust(dist(city[,2:8], method="maximum"))
hc2
```

У даному випадку всі ці методи дають однакові результати. Намалюємо дентрограму:

```
hcd1 <- as.dendrogram(hc1)
plot(hcd1, main = "Manhattan distance")
```

Застосуємо тепер метод к середніх:

```
kmeans.city<-kmeans(city[,2:8], 4, iter.max = 10, nstart = 1)
kmeans.city
```

```
K-means clustering with 2 clusters of sizes 6, 10

Cluster means:
V2       V3    V4       V5       V6        V7    V8
1  8.366667 22.36667 164.5 168.8333 1030.167  783.1667 636.5
2 10.580000 31.57000 290.9 212.7000 1583.000 1135.8000 720.6

Clustering vector:
[1] 1 1 1 2 2 2 1 2 2 2 2 1 2 2 1 2

Within cluster sum of squares by cluster:
[1]  501946.6 1035952.6
(between_SS / total_SS =  52.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Виклик kmeans.city$cluster дає розподіл по кластерам:

```
1 1 1 3 2 4 1 2 3 3 4 1 4 2 1 3
```

Задачі для самостійної роботи

Якщо не оговорено окремо, то наведені задачі взято з підручника [R02]. Файли з даними містяться у файлі multivariate.zip, який можна знайти за посиланням ????. Файли згруповано по номеру розділа та ключовим словам в назві таблиці.

Тема 1: Перевірка гіпотез про рівність середніх. 5.18 a,b 5.20 a)

Тема 2: Тестування гіпотез про коваріаційну матрицю. 7.14, 7.15, (тільки H_0: $\Sigma = \sigma I$), 7.28.

Тема 3: Дискримінантний аналіз. 8.10, 8.11 (стандартизовані коефіцієнти: за бажанням, самостійне опрацювання матеріалу).

Тема 4: Лінійна регресія.

1. Таблиця 3.1. Дослідіти модель лінійної регресії y_1 на x_1.
2. Таблиця 3.4. Дослідити модель множинної регресії y_1 на (x_1, x_2, x_3)
3. Таблиця 3.4. Дослідити модель багатовимірної множинної регресії (y_1, y_2) на (x_1, x_2, x_3).

Тема 5: Метод головних компонент. 12.8, 12.10.

Тема 6: Факторний аналіз. 13.10, 13.11.

Тема 7: Кластерний аналіз. 14.8, 14.9

Тема 8: Класифікація. 9.8.

Література

[AD54] Anderson T. W., Darling D. A. A test of goodness of fit // J. Amer. Stist. Assoc., 1954. — V. 29. — P. 765—769.

[K07] М.В. Карташов. Ймовірність, процеси, статистика. Видавничо-поліграфічний центр "Київський університетє". Київ, 2007.

[Le60] Levene, H. (1960). In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin et al. eds., Stanford University Press, pp. 278-292.

[Ma40] J. W. Mauchly. "Significance Test for Sphericity of a Normal n-Variate Distribution". (1940) The Annals of Mathematical Statistics. 11 (2): 204–209.

[Ma07] Р.Є. Майборода. Регресія: лінійні моделі. ВПЦ Київський університет. 296 стор. 2007

[PP12] K. B. Petersen, M. S. Pedersen. The Matrix Cookbook [http://matrixcookbook.com] 2012

[R98] A. C. Rencher. *Multivariate statistical inference and applications.* Wiley, NY, 1998.

[R02] A. C. Rencher. *Methods of Multivariate Analysis.* Second ed., Wiley, NY, 2002.

[RS08] A.C. Rencher, G.B Schaalje. *Linear Models in Statistics.* Wiley, NY, 2002.

[R16] J. Rickert Generating and Visualizing Multivariate Data with R. `https://blog.revolutionanalytics.com/2016/02/multivariate_data_with_r.html`

[SW65] S. S. Shapiro; M. B. Wilk. An Analysis of Variance Test for Normality (Complete Samples) *Biometrika*, Vol. 52, No. 3/4, pp. 591-611.

[Sh] A. Shegel. Посилання на сторінку в RPubs: `https://rpubs.com/aaronsc32`

[VAG09] José A. Villasenor Alva Elizabeth González Estrada. A Generalization of Shapiro–Wilk's Test for Multivariate Normality. *Communications in Statistics - Theory and Methods* Volume 38, 2009 - Issue 11. 1870–1883.