UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Alguns mètodes d'anàlisi multivariant

Autor: Pol Vallès Closa

Directors: Dr. Josep Vives i Dra. Carme Florit
Realitzat a: Dep. de Matemàtiques i Informàtica
Barcelona, 20 de juny de 2021

Abstract

In this project we will study some methods of multivariate analysis and give some interesting examples. In particular, we will treat the following methods:

- Multiple linear regression model: which will allow us to relate a random variable to a random vector and analyze its most important applications in Econometrics.
- Canonical correlation analysis: which will allow us to relate two random vectors.
- Principal component analysis, which allows us to transform a random vector of dimension n to dimension 2 or 3 so that it can be represented graphically.
- Discriminant analysis, which given two populations, tells us where and individual belongs to.

All of this will allow us to have a very strong idea about the different methods of multivariate analysis.

Resum

Al llarg d'aquest treball estudiarem diversos mètodes d'anàlisi multivariant i exposarem alguns exemples interessants. En concret, estudiarem els següents mètodes:

- Model de regressió lineal múltiple: ens permetrà relacionar una variable aleatòria amb un vector aleatori i analitzarem les seves aplicacions més importants en l'Econometria.
- Anàlisi de correlació canònica: ens permetrà relacionar dos vectors aleatoris.
- Anàlisi de components principals: ens permet transformar un vector aleatori de dimensió n a dimensió 2 o 3 per poder-lo representar gràficament.
- Anàlisi discriminant: donades dues poblacions, ens diu a quina pertany un individu.

Tot això ens permetrà tenir una idea molt sòlida sobre els diferents mètodes d'anàlisi multivariant.

2010 Mathematics Subject Classification. 60E05, 62H20, 62H25, 62P20, 62J05

Agraïments

Vull agrair al Dr. Josep Vives i Santa-Eulàlia i a la Dra. Carme Florit Selma per ajudar-me i guiar-me en el desenvolupament del treball.

També voldria agrair als meus pares i a la meva àvia, per fer-me sempre costat, així com al Sixte, al Víctor i al Roger, per tants moments divertits durant aquests últims quatre anys.

Índex

1 El model de regressió lineal múltiple

L'anàlisi de correlació múltiple és una tècnica d'anàlisi multivariant que ens permet relacionar una variable aleatòria y amb un vector aleatori $X = (x_1, \ldots, x_k)$ a partir del model de regressió lineal múltiple.

Definició 1.1. Un **model de regressió lineal múltiple** (MRLM), és aquell en el qual una variable y, anomenada **endògena**, ve determinada per un conjunt de variables $x_1, \ldots, x_k$, anomenades explicatives o **exògenes** de la forma:

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k + u$$

on u és el terme de pertorbació (part no determinista) i $\beta_0, \ldots, \beta_k$ són constants.

El nostre objectiu serà determinar els coeficients $\beta_0, \ldots, \beta_k$ per tal d'obtenir les millors prediccions a partir d'una certa mostra.

Exemple 1.2. Per exemple, ens podríem preguntar si una major densitat d'habitants en una comarca fa augmentar el preu de venda mitjà de l'habitatge d'obra nova per metre quadrat. Tot i que el sentit comú ens pot fer pensar que a major densitat poblacional, major serà el preu de l'habitatge, el sentit comú no ens proporciona cap resposta quantitativa a la pregunta de quin és exactament l'efecte de la densitat poblacional sobre el preu de l'habitatge. Per obtenir aquesta resposta haurem d'examinar l'evidència empírica, és a dir, l'evidència basada en les dades que relacionen la densitat poblacional amb el preu de l'habitatge.

Ara, podríem construir el següent MRLM:

$$PREU = \beta_0 + \beta_1 \cdot DENS + \beta_2 \cdot DIST + \beta_3 \cdot PIB + \beta_4 \cdot RENDA + \beta_5 \cdot EDAT + u$$

on el preu de venda mitjà de l'habitatge d'obra nova per metre quadrat (PREU) és la variable endògena i la densitat poblacional (DENS), la distància entre la capital de comarca i Barcelona (DIST), el PIB per habitant (PIB), la renda familiar per habitant (RENDA) i la mitjana de l'edat (EDAT) són les variables exògenes. Com dèiem, el nostre objectiu serà determinar les constants $\beta_0, \ldots, \beta_5$, per poder predir de la millor manera el preu de l'habitatge.

Definició 1.3. Direm que un conjunt de dades són de **tall transversal** si recullen dades d'un conjunt d'individus en un mateix moment temporal, per exemple, un cens electoral.

Observació 1.4. Hi ha altres tipus de dades, com les de sèrie temporal, on es recullen dades d'una mateixa variable en diferents moments temporals, per exemple, la temperatura a les 8:00 a Barcelona recollides cada dia del 2020. En aquest treball, però, ens centrarem únicament en dades de tall transversal.

Agafant com a exemple un conjunt de n dades de tall transversal es pot especificar una equació per cada individu:

$$\begin{cases} y_1 = \beta_0 + \beta_1 \cdot x_{1,1} + \cdots + \beta_k \cdot x_{k,1} + u_1 \\ \quad \vdots \\ y_n = \beta_0 + \beta_1 \cdot x_{1,n} + \cdots + \beta_k \cdot x_{k,n} + u_n \end{cases}$$

on y_i és la variable endògena per a l'observació i-èssima i $x_{j,i}$ és la variable explicativa x_j per l'observació i-èssima. Aquest sistema es pot escriure matricialment com:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & \dots & x_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \dots & x_{k,n} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \quad Y_{n\times 1} = X_{n\times(k+1)} \cdot \beta_{(k+1)\times 1} + U_{n\times 1}.$$

1.1 Definició d'hipòtesis bàsiques

Per poder determinar les propietats dels estimadors i el tipus de contrastos a realitzar, és necessari enunciar un conjunt d'hipòtesis bàsiques:

- **Hipòtesis bàsiques generals sobre el model:**

  1. És un model estocàstic, és a dir, és un sistema amb un comportament no determinista (hi ha un terme de pertorbació, u).
  2. És un model lineal o linealitzable. Un model és linealitzable quan una relació no lineal es pot transformar en alguna lineal, per exemple, $y_i = \beta_0 \cdot x_{1,i}^{\beta_1} \cdots x_{k,i}^{\beta_k}$ és no lineal, però aplicant logaritmes obtenim:

     $$ln(y_i) = ln(\beta_0) + \beta_1 \cdot ln(x_{1,i}) + \cdots + \beta_k \cdot ln(x_{k,i}), \text{ que és lineal.}$$

  3. Existeix informació estadística suficient sobre les variables, és a dir, ha de ser $k+1 \leq n$, ja que en cas contrari tindríem més paràmetres a estimar que nombre d'observacions. Tot i que $n = k+1$ és el mínim, a major nombre d'observacions (major n), s'obtindrà una estimació més precisa. Anomenem **graus de llibertat** del model a la diferència

     $$gl = n - (k+1) \geq 0.$$

- **Hipòtesis bàsiques sobre les variables explicatives:**

  1. Les variables explicatives estan incorrelacionades amb el terme d'error, per tant, $E(x_{j,i} \cdot u_i) = 0, \ \forall j = 1, \dots, k, \ \forall i = 1, \dots, n.$
  2. No existeix cap relació lineal exacta entre les variables explicatives, i per tant, $rang(X_{n\times(k+1)}) = k+1$ (rang màxim), on X és la matriu definida anteriorment.
  3. Les variables explicatives estan mesurades sense error.
  4. En el model no s'ha inclòs cap variable irrellevant ni s'ha omès cap variable rellevant.

- **Hipòtesis bàsiques sobre els paràmetres:**

  1. Els coeficients β_i són constants per a tota la mostra. Aquest fet s'anomena hipòtesi de permanència estructural.

- **Hipòtesis bàsiques sobre el terme de pertorbació:**

  1. L'esperança del terme d'error és 0: $E(u_i) = 0, \ \forall i = 1, \dots n.$
  2. La variància del terme d'error és constant: $Var(u_i) = \sigma^2, \ \forall i = 1, \dots, n.$ Aquesta propietat s'anomena hipòtesis d'**homoscedasticitat**. D'altra banda, si $\exists i, j \in \{1, \dots, n\}$ tals que $Var(u_i) = \sigma_i^2 \neq \sigma_j^2 = Var(u_j)$ es diu que hi ha **heteroscedasticitat**.
  3. Els termes d'error són incorrelacionats entre si, és a dir, no existeix autocorrelació:

     $$Cov(u_i, u_j) = E[(u_i - E(u_i)) \cdot (u_j - E(u_j))] \overset{E(u_i)=0}{=} E(u_i \cdot u_j) = 0, \ \forall i \neq j.$$

     Definició 1.5. Si el terme d'error és homoscedàstic i no existeix autocorrelació direm que el terme d'error és **esfèric** i, en aquest cas, la matriu de variàncies i covariàncies de Σ és escalar, com veurem a continuació:

     $$\Sigma = Var(U) = E(U \cdot U^T) = E\left[\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \cdot \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix}\right] =$$

     $$= E\left[\begin{pmatrix} u_1 \cdot u_1 & \cdots & u_1 \cdot u_n \\ \vdots & \ddots & \vdots \\ u_n \cdot u_1 & \cdots & u_n \cdot u_n \end{pmatrix}\right] = \begin{pmatrix} E(u_1 \cdot u_1) & \cdots & E(u_1 \cdot u_n) \\ \vdots & \ddots & \vdots \\ E(u_n \cdot u_1) & \cdots & E(u_n \cdot u_n) \end{pmatrix}$$

     i ara, utilitzant que $E(u_i \cdot u_i) = \sigma^2$ i $E(u_i \cdot u_j) = 0 \ \forall i \neq j$, obtenim que $\Sigma = \sigma^2 \cdot I_{n \times n}$, com volíem veure.
  4. El terme d'error segueix una distribució Normal.

  Observació 1.6. Les quatre hipòtesis relatives al terme de pertorbació es poden resumir mitjançant les expressions:

  $$u_i \sim N(0, \sigma^2), \quad U \sim N(0, \sigma^2 \cdot I_{n \times n}).$$

  Proposició 1.7. *La variable endògena es distribueix com una llei Normal amb esperança $X\beta$ i variància σ^2.*

  Demostració. $E(Y) = E(X\beta + U) = X\beta + E(U) = X\beta + 0 = X\beta$ i
  $Var(Y) = E[(Y - E(Y))^2] = [E(\cancel{X\beta} + U - \cancel{X\beta})^2] = E(U^2) = \sigma^2 \cdot I_{n \times n}.$ $\square$

1.2 Estimació del MRLM

Un cop especificat el MRLM i definides les hipòtesis bàsiques del model, el següent pas és calibrar-lo, i per tant, obtenir estimacions dels paràmetres del MRLM a partir de la informació mostral disponible. Denotarem com $\hat{\beta}_0, \dots, \hat{\beta}_k$ els paràmetres estimats.

Definició 1.8. Donat un MRLM: $y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k + u$, definim els **errors d'estimació** (e_i) com:

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{1,i} + \cdots + \hat{\beta}_k \cdot x_{k,i}) = y_i - \hat{y}_i, \ \forall i = 1, \dots, n.$$

1.2.1 Estimació per mínims quadrats ordinaris (MQO)

L'objectiu d'aquest apartat serà trobar els $\hat{\beta}_0, \ldots, \hat{\beta}_k$ que minimitzin la suma dels quadrats dels errors, i estudiar les seves propietats.

Definició 1.9. Definim la suma dels quadrats dels errors (**SQE**) com:

$$SQE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = e_1^2 + \cdots + e_n^2.$$

Podem reescriure l'expressió del SQE matricialment com

$$SQE = (Y - \hat{Y})^T \cdot (Y - \hat{Y}) = (Y - X\hat{\beta})^T \cdot (Y - X\hat{\beta})$$

$$= (Y^T - \hat{\beta}^T X^T) \cdot (Y - X\hat{\beta}) = Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}.$$

Ara bé, com tots els termes de la darrera igualtat són reals (matrius 1×1) i $(Y^T X\hat{\beta})^T = \hat{\beta}^T X^T Y$, podem afirmar que

$$SQE = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}.$$

Ara ja podem procedir a trobar el mínim del SQE, derivant i igualant a 0,

$$\frac{\partial SQE}{\partial \hat{\beta}} = -2X^T Y + 2X^T X\hat{\beta}$$

$$\frac{\partial SQE}{\partial \hat{\beta}} = 0 \iff X^T Y = X^T X\hat{\beta} \iff \hat{\beta} = (X^T X)^{-1} \cdot X^T \cdot Y.$$

Notem que aquesta expressió no es pot simplificar fent

$$(X^T X)^{-1} X^T Y = X^{-1} (X^T)^{-1} X^T Y = X^{-1} Y,$$

ja que la matriu X no és quadrada (és de dimensió $n \times (k+1)$) i per tant, X^{-1} no existeix.
Quedaria, per tant, provar que és realment un mínim,

$$\frac{\partial^2 SQE}{\partial^2 \hat{\beta}} = 2X^T X = 2 \cdot \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_{1,1} & x_{1,2} & \ldots & x_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k,1} & x_{k,2} & \ldots & x_{k,n} \end{pmatrix} \cdot \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{k,1} \\ 1 & x_{1,2} & \ldots & x_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \ldots & x_{k,n} \end{pmatrix}$$

$$= 2 \cdot \begin{pmatrix} n & \sum_{i=1}^n x_{1,i} & \ldots & \sum_{i=1}^n x_{k,i} \\ \sum_{i=1}^n x_{1,i} & \sum_{i=1}^n x_{1,i}^2 & \ldots & \sum_{i=1}^n x_{1,i} \cdot x_{k,i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{k,i} & \sum_{i=1}^n x_{k,i} \cdot x_{1,i} & \ldots & \sum_{i=1}^n x_{k,i}^2 \end{pmatrix}$$

i aquesta expressió és semidefinida positiva, ja que és el producte d'una matriu per la seva transposada ($\sqrt{2}X^T \cdot \sqrt{2}X$).

Per tant, ara ja podem afirmar que $\hat{\beta}_{MQO}$

$$\boxed{\hat{\beta}_{MQO} = (X^T X)^{-1} \cdot X^T \cdot Y}$$

minimitza el SQE.

Vegem ara algunes propietats dels estimadors obtinguts, $\hat{\beta}_{MQO}$.

Proposició 1.10. *Els estimadors són una combinació lineal dels veritables paràmetres poblacionals β, les variables explicatives i el terme de pertorbació.*

Demostració.

$$\hat{\beta}_{MQO} = (X^T X)^{-1} \cdot X^T Y = (X^T X)^{-1} \cdot X^T \cdot (X\beta + U)$$

$$= (X^T X)^{-1}(X^T X)\beta + (X^T X)^{-1} \cdot X^T U = \beta + (X^T X)^{-1} X^T U.$$

□

Observació 1.11. A més, com que el terme de pertorbació és una variable aleatòria que es distribueix seguint la llei Normal, els estimadors $\hat{\beta}_{MQO}$ també seguiran una llei Normal.

Proposició 1.12. *Els estimadors són no esbiaixats.*

Demostració.

$$E(\hat{\beta}_{MQO}) = E(\beta + (X^T X)^{-1} X^T U) = \beta + (X^T X)^{-1} X^T \cdot E(U) \overset{E(U)=0}{=} \beta$$

i per tant, Biaix$(\hat{\beta}_{MQO}) = E(\hat{\beta}_{MQO}) - \beta = \beta - \beta = 0$. □

Proposició 1.13. *(Teorema de Gauss-Markov): De tots els estimadors lineals i no esbiaixats, els estimadors MQO són els que tenen variància mínima.*

Demostració. La demostració d'aquest teorema es pot trobar a les pàgines 65-68 del llibre *Econometric Theory*, publicat per Arthur S. Goldberg el 1964. □

Observació 1.14. Tot i no veure en aquest treball la demostració del teorema de Gauss-Markov, és essencial trobar la variància dels estimadors $\hat{\beta}_{MQO}$:

$$Var(\hat{\beta}_{MQO}) = E((\hat{\beta}_{MQO} - \beta)(\hat{\beta}_{MQO} - \beta)^T) \overset{Prop.8}{=} E((X^T X)^{-1} X^T U \cdot ((X^T X)^{-1} X^T U)^T)$$

$$= (X^T X)^{-1} X^T \cdot E(U \cdot U^T) \cdot X((X^T X)^{-1})^T = (X^T X)^{-1} X^T \cdot \sigma^2 \cdot X(X^T X)^{-1}$$

on he utilitzat que $E(UU^T) = \sigma^2 \cdot I_{n \times n}$ i que $X^T X$ és simètrica, d'on $(X^T X)^{-1}$ és simètrica,

$$Var(\hat{\beta}_{MQO}) = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Observació 1.15. L'absència de biaix no garanteix que el valor numèric d'un estimador estigui molt a prop del vertader valor, sinó que només diu que en mitjana coincidirà amb aquest.

Definició 1.16. Definim l'error quadràtic mitjà (**EQM**) com

$$EQM(\hat{\beta}_{MQO}) = E((\hat{\beta}_{MQO} - \beta)^2).$$

Proposició 1.17. *L'error quadràtic mitjà es pot escriure com*

$$EQM(\hat{\beta}_{MQO}) = Var(\hat{\beta}_{MQO}) + (Biaix(\hat{\beta}_{MQO}))^2.$$

Demostració.

$$E((\hat{\beta}_{MQO} - \beta)^2) = E\left[(\hat{\beta}_{MQO} - E(\hat{\beta}_{MQO}) + E(\hat{\beta}_{MQO}) - \beta)^2\right]$$

$$= E\left[(\hat{\beta}_{MQO} - E(\hat{\beta}_{MQO}))^2 + 2(\hat{\beta}_{MQO} - E(\hat{\beta}_{MQO}))(E(\hat{\beta}_{MQO}) - \beta) + (E(\hat{\beta}_{MQO}) - \beta)^2\right]$$

$$= E[(\hat{\beta}_{MQO} - E(\hat{\beta}_{MQO}))^2] + 2E[(\hat{\beta}_{MQO} - E(\hat{\beta}_{MQO}))(E(\hat{\beta}_{MQO}) - \beta)] + E[(E(\hat{\beta}_{MQO}) - \beta)^2]$$

$$= Var(\hat{\beta}_{MQO}) + 2(E(\hat{\beta}_{MQO}) - \beta)E[\hat{\beta}_{MQO} - E(\hat{\beta}_{MQO})] + Biaix(\hat{\beta}_{MQO}))^2$$

$$= Var(\hat{\beta}_{MQO}) + Biaix(\hat{\beta}_{MQO}))^2$$

ja que $E(\hat{\beta}_{MQO} - E(\hat{\beta}_{MQO})) = 0$, com volíem veure. □

Definició 1.18. Diem que un estimador és **consistent** si $\lim_{n\to\infty} EQM(\hat{\beta}_{MQO}) = 0$. En altres paraules, un estimador és consistent si en incrementar la mida de la mostra, l'estimador s'aproxima al seu valor poblacional.

Proposició 1.19. *Sota el supòsit que la matriu* $\lim_{n\to\infty} \left(\frac{X^T X}{n}\right)^{-1}$ *existeix i és finit, l'estimador* $\hat{\beta}_{MQO}$ *és consistent.*

Demostració.

$$\lim_{n\to\infty} EQM(\hat{\beta}_{MQO}) = \lim_{n\to\infty} (Var(\hat{\beta}_{MQO}) + (Biaix(\hat{\beta}_{MQO}))^2)$$

$$\overset{\lim_{n\to\infty}(Biaix(\hat{\beta}_{MQO}))=0}{=} \lim_{n\to\infty} Var(\hat{\beta}_{MQO}) = \lim_{n\to\infty} \sigma^2 (X^T X)^{-1} = \lim_{n\to\infty} \frac{\sigma^2}{n} \left(\frac{X^T X}{n}\right)^{-1} = 0$$

on hem suposat que $\left(\frac{X^T X}{n}\right)^{-1}$ existeix i és un nombre finit. □

Corol·lari 1.20. *A partir dels resultats anteriors, podem afirmar que*

$$\hat{\beta}_{MQO} \sim N(\beta, \sigma^2 (X^T X)^{-1}).$$

Definició 1.21. Definim els **errors** o **residus de l'estimació** com

$$e_{MQO} = Y - \hat{Y} = Y - X\hat{\beta}_{MQO}.$$

Proposició 1.22. *El vector e_{MQO} és una combinació lineal de la variable endògena.*

Demostració.

$e_{MQO} = Y - \hat{Y} = Y - X\hat{\beta}_{MQO} = Y - X(X^TX)^{-1}X^TY = [I_{n\times n} - X(X^TX)^{-1}X^T]Y$

i fent $M = I_{n\times n} - X(X^TX)^{-1}X^T$ obtenim que $e_{MQO} = MY$, com volíem veure. □

Proposició 1.23. *La matriu M és simètrica de dimensió $n\times n$, idempotent ($MM = M$), ortogonal respecte a X, és a dir, $MX = 0$ i amb traça $n - (k+1)$.*

Demostració. Clarament la matriu M és quadrada $n \times n$, ja que la matriu Y té dimensió $n \times 1$, per tant, al treure factor comú Y, apareix la matriu identitat de dimensió $n\times n$. Observem també que la matriu $X(X^TX)^{-1}X^T$ és de dimensió $n\times n$, ja que la matriu X és de dimensió $n \times (k+1)$.

D'altra banda, com $I_{n\times n}$ és simètrica, només queda veure que $X(X^TX)^{-1}X^T$ és simètrica. Per demostrar-ho utilitzarem la propietat que diu que si A és una matriu simètrica $p \times p$ i B és una matriu $p \times q$, aleshores B^TAB és simètrica $q \times q$, en el nostre cas prenem $B = X^T$ de dimensió $(k+1) \times n$ i $A = (X^TX)^{-1}$ (ja hem vist que és simètrica) de dimensió $(k+1) \times (k+1)$. Per tant, al ser M la resta de dues matrius simètriques, és simètrica.

Vegem ara que M és idempotent, és a dir, $MM = M$:

$$MM = (I_{n\times n} - X(X^TX)^{-1}X^T)(I_{n\times n} - X(X^TX)^{-1}X^T) = I_{n\times n} - 2X(X^TX)^{-1}X^T +$$
$$+ X\cancel{(X^TX)^{-1}X^T \cdot X}(X^TX)^{-1}X^T = I_{n\times n} - X(X^TX)^{-1}X^T = M \Longrightarrow MM = M$$

També cal veure que $MX = 0$

$$MX = (I_{n\times n} - X(X^TX)^{-1}X^T)X = X - X\cancel{(X^TX)^{-1}X^TX} = X - X = 0,$$

on 0 representa la matriu de dimensió $n \times (k+1)$ amb tots els seus elements 0.

Finalment cal provar que $tr(M) = n - (k+1)$:

$$tr(M) = tr(I_{n\times n}) - tr(X(X^TX)^{-1}X^T) = n - rang(X(X^TX)^{-1}X^T)$$

perquè clarament, $X(X^TX)^{-1}X^T$ és idempotent i la traça d'una matriu idempotent és el seu rang, d'aquí podem deduir que $tr(M) = n - (k+1)$, com volíem veure. □

Proposició 1.24. *El vector e_{MQO} és una combinació lineal de U.*

Demostració.

$$e_{MQO} = MY = M(X\hat{\beta}_{MQO} + U) = (MX)\hat{\beta}_{MQO} + MU \overset{MX=0}{=} MU.$$

□

Proposició 1.25. *El vector e_{MQO} és ortogonal a la matriu X.*

Demostració.

$$e_{MQO}^TX = (MU)^TX = U^TM^TX \overset{M \text{ simètrica}}{=} U^T(MX) \overset{MX=0}{=} 0$$
$$X^Te_{MQO} = X^T(MU) \overset{M \text{ simètrica}}{=} X^TM^TU = (MX)^TU \overset{MX=0}{=} 0.$$

□

Proposició 1.26. *La mitjana mostral dels residus és 0 quan el model té terme independent.*

Demostració. De la proposició anterior sabem que $X^T e_{MQO} = 0$, per tant,

$$X^T e_{MQO} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k,1} & x_{k,2} & \dots & x_{k,n} \end{pmatrix} \cdot \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n e_i \\ \sum_{i=1}^n x_{1,i} e_i \\ \vdots \\ \sum_{i=1}^n x_{k,i} e_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

d'on deduïm que, igualant la primera component del vector

$$\sum_{i=1}^n e_i = 0 \Longrightarrow \bar{e}_{MQO} = \frac{1}{n} \sum_{i=1}^n e_i = 0.$$

□

Proposició 1.27. *El vector e_{MQO} es distribueix segons una llei Normal.*

Demostració. Com que $e_{MQO} = MU$ i $U \sim N(0, \sigma^2 I_{n\times n})$, tindrem que $e_{MQO} \sim N(0, \sigma^2 M)$. □

Un cop vistos aquests resultats, ja estem preparats per estimar la variància del terme de pertorbació. Aquest resultat ens serà molt útil, ja que és necessari conèixer la variància per poder fer contrastos d'hipòtesis.

$$E(SQE) = E(e^T e) = E[(MU)^T (MU)] = E(U^T M^T MU) \overset{M \text{ simètrica}}{=} E(U^T MMU)$$

$$\overset{MM=M}{=} E(U^T MU) \overset{U^T MU \text{ matriu } 1\times 1}{=} E(tr(U^T MU)) \overset{tr(AB)=tr(BA)}{=} E(tr(MUU^T))$$

$$= tr(E(MUU^T)) = tr(ME(UU^T)) = tr(M\sigma^2 Id_{n\times n}) = \sigma^2 tr(M) = \sigma^2(n-(k+1))$$

i per tant, deduïm que

$$\boxed{\hat{\sigma}^2_{MQO} = \frac{e^T e}{n-(k+1)} = \frac{\sum_{i=1}^n e_i^2}{n-(k+1)}}.$$

1.2.2 Estimació per màxima versemblança (MV)

El mètode de la màxima versemblança proposa com a estimador del paràmetre aquell valor que maximitza la probabilitat d'obtenir les observacions mostrals disponibles. Sabem que $U \sim N(0, \sigma^2 I_{n\times n})$, per tant, utilitzant que la funció de densitat d'una distribució Normal $N(\mu, \sigma^2)$ és

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

obtenim la funció de densitat de cadascun dels termes de pertorbació

$$f_i(u_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma^2}}, \ \ \forall i = 1, \ldots, n.$$

Ara, podem calcular la funció de densitat conjunta del vector U:

$$f(U) = \prod_{i=1}^{n} f_i(u_i) = (2\pi\sigma^2)^{-n/2} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} u_i^2\right) = (2\pi\sigma^2)^{-n/2} exp\left(-\frac{1}{2\sigma^2}U^T U\right),$$

fent el canvi de variable $Y = g(U) = X\beta + U \implies U = g^{-1}(Y) = Y - X\beta$, $J_{g^{-1}}(Y) = 1$, obtenim

$$f(Y) = (2\pi\sigma^2)^{-n/2} exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\right).$$

Els estimadors de màxima versemblança s'obtindran a partir de maximitzar la funció de versemblança, per tant caldrà calcular les derivades parcials de la funció de versemblança respecte β i σ^2, i igualar-les a 0.
Per facilitar els càlculs, faré la derivada a partir del logaritme de la funció de versemblança,

$$ln(L(y; \beta, \sigma^2)) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$$

$$\frac{\partial lnL}{\partial \beta} = -\frac{1}{2\sigma^2}[-2X^T(Y - X\hat{\beta})] = \frac{1}{\sigma^2}[X^T Y - X^T X\hat{\beta}] = 0 \tag{1}$$

$$\frac{\partial lnL}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2\sigma^4}[(Y - X\hat{\beta})^T(Y - X\hat{\beta})] = 0 \tag{2}$$

de l'equació (1) deduïm que

$$X^T Y - X^T X\hat{\beta} = 0 \iff \boxed{\hat{\beta}_{MV} = (X^T X)^{-1}(X^T Y)}$$

i de l'equació (2) deduïm que

$$\frac{1}{2\hat{\sigma}^2}\left(-n + \frac{1}{\hat{\sigma}^2}[(Y - X\hat{\beta}_{MV})^T(Y - X\hat{\beta}_{MV})]\right) = 0 \iff$$

$$\iff \boxed{\hat{\sigma}^2_{MV} = \frac{(Y - X\hat{\beta}_{MV})^T \cdot (Y - X\hat{\beta}_{MV})}{n} = \frac{e^T e}{n}}$$

Podem observar que $\hat{\beta}_{MV} = \hat{\beta}_{MQO}$, de manera que l'estimador MV té les mateixes propietats que l'estimador MQO.
No succeeix el mateix, però amb l'estimació MV de σ^2, de fet, aquest estimador és esbiaixat.

$$E(\hat{\sigma}^2_{MV}) = E\left(\frac{e^T e}{n}\right) = \frac{1}{n}E(e^T e) = \frac{1}{n}\sigma^2(n - (k+1)) = \frac{n - (k+1)}{n}\sigma^2 \neq \sigma^2.$$

Tot i ser esbiaixat, és asimptòticament no esbiaixat, ja que, a major mida mostral, el biaix s'aproxima a 0,

$$Biaix(\hat{\sigma}^2_{MV}) = -\frac{k+1}{n}\sigma^2 \overset{n\to\infty}{\longrightarrow} 0.$$

1.3 Validació del model

1.3.1 Bondat de l'ajust

Un cop especificat el model, caldrà validar-lo, és a dir, veure si és realment un bon model.

Definició 1.28. Definim la suma dels quadrats totals (**SQT**) com:

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

on $\bar{y}$ és la mitjana de la variable endògena. Notem que podem prendre la SQT com una mesura de la dispersió de la variable endògena.

Definició 1.29. Definim la suma dels quadrats de la regressió (**SQR**) com:

$$SQR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

on $\hat{y}_i$ és la variable endògena estimada per l'observació i-èssima. Notem que la SQR es pot entendre com la part de la variabilitat total de la variable endògena que es veu explicada pel model estimat.

Observació 1.30. Recordem que havíem definit la suma dels quadrats dels errors (SQE) com:

$$SQE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2,$$

Notem que la SQE s'entén com la part de la variabilitat total de la variable endògena que no es veu explicada pel model estimat.

Proposició 1.31. *Sempre que el model tingui terme independent es complirà que*

$$SQT = SQR + SQE.$$

Demostració.

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 = \sum_{i=1}^{n} (e_i + (\hat{y}_i - \bar{y}))^2$$

$$= \sum_{i=1}^{n} [e_i^2 + (\hat{y}_i - \bar{y})^2 + 2 \cdot e_i(\hat{y}_i - \bar{y})] = SQE + SQR + 2 \sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}),$$

per tant, tindrem

$$SQT = SQR + SQE \iff \sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i \hat{y}_i + \bar{y} \sum_{i=1}^{n} e_i = 0,$$

i com que el model té terme independent, per la proposició 1.26 se satisfà que

$$\sum_{i=1}^{n} e_i = 0, \text{ tindrem que } SQT = SQR + SQE \Longleftrightarrow \sum_{i=1}^{n} e_i \hat{y}_i = 0 \text{ on,}$$

$$\sum_{i=1}^{n} e_i \hat{y}_i = \sum_{i=1}^{n} e_i(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{1,i} + \cdots + \hat{\beta}_k \cdot x_{k,i}) = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} e_i x_{1,i} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} e_i x_{k,i}.$$

Ara, a partir de la demostració de la proposició 1.26 també es pot veure que

$$\sum_{i=1}^{n} e_i \cdot x_{j,i} = 0, \ \ \forall j = 2, \ldots, k \Longrightarrow \sum_{i=1}^{n} e_i \hat{y}_i = 0 \Longrightarrow SQT = SQR + SQE,$$

com volíem veure. □

Observació 1.32. Si el model no tingués terme independent, es podria complir que $SQT < SQR + SQE$, per exemple, considerem el model $y_i = \beta_1 \cdot x_{1,i}$ amb les $n = 3$ observacions següents: $y_1 = 4, x_{1,1} = 0.25$, $y_2 = 4, x_{1,2} = 0.05$, $y_3 = 4, x_{1,3} = -0.20$, en aquest cas, com que $\bar{y} = 4$, tindrem

$$SQT = \sum_{i=1}^{3} (y_i - \bar{y})^2 = \sum_{i=1}^{3} (4 - 4)^2 = 0$$

i, d'altra banda, el SQE seria positiu, ja que fixat $\hat{\beta}_1$, mai es complirà que $SQE = 0$, ja que això portaria a

$$4 = \hat{\beta}_1 \cdot 0.25 = 0, \ \ 4 = \hat{\beta}_1 \cdot 0.05 = 0, \ \ 4 = \hat{\beta}_1 \cdot (-0.20) = 0, \text{ alhora,}$$

la qual cosa no es pot complir. És a dir, hem trobat un exemple que ens permet entendre perquè no es compleix $SQT = SQR + SQE$ quan el model no té terme independent.

Com a reflexió final, notem que si introduïm el terme independent definint un nou model $y_i = \beta_0 + \beta_1 \cdot x_{1,i}$, podríem definir $\hat{\beta}_0 = 4$ i $\hat{\beta}_1 = 0$, amb el que obtindríem $SQT = SQR + SQE$, ja que tots els termes són 0.

A partir d'ara suposarem sempre que el model té terme independent.

Definició 1.33. Definim el **coeficient de determinació ($\mathbf{R^2}$)**, com

$$\boxed{R^2 = 1 - \frac{SQE}{SQT}}$$

Aquest coeficient ens dona una mesura sobre la bondat de l'ajust. De fet, el coeficient R^2 mesura quina proporció de la variabilitat total de la variable endògena es veu explicada pel model. Així, com major sigui el valor del coeficient de determinació, millor serà l'ajust del model.

Proposició 1.34. *Es compleix que*

$$R^2 = \frac{SQR}{SQT} \in [0,1].$$

Demostració. Com estem tractant models amb terme independent, tenim que

$$R^2 = 1 - \frac{SQE}{SQT} = \frac{SQT - SQE}{SQT} \overset{SQT=SQR+SQE}{=} \frac{SQR}{SQT}.$$

A més, com $SQE \geq 0$ podem afirmar que $SQT \geq SQR$ i, com SQT i SQR també són positius, podem afirmar que

$$R^2 = \frac{SQR}{SQT} \in [0,1].$$

□

Exemple 1.35. Si el model $y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_K \cdot x_k + u$ té un coeficient de determinació $R^2 = 0.83$, diem que les variables $x_1, \ldots, x_k$ expliquen un 83% de la variabilitat total de la variable endògena, i per tant, el model deixaria d'explicar un 17% de la variabilitat de y.

Definició 1.36. Diem que dos models són **ennierats** si tenen la mateixa variable endògena i les variables exògenes d'un dels dos models són també variables exògenes de l'altre model.

Exemple 1.37. Els models

$$\begin{cases} y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_K \cdot x_k + u \\ y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_K \cdot x_k + \beta_{k+1} \cdot x_{k+1} + \cdots + \beta_l \cdot x_l + u' \end{cases}$$

són ennierats per $l > (k+1)$, ja que tenen la mateixa variable endògena (y) i les variables exògenes del primer model ($x_1, \ldots, x_k$) són també variables exògenes del segon.

Malgrat la utilitat del coeficient de determinació, aquest presenta un problema quan el volem utilitzar per a seleccionar el millor model entre un conjunt de models ennierats, ja que sempre acabarà seleccionant el model més general de tots, perquè a mesura que incorporem variables explicatives, la SQE disminueix. Per tant, el coeficient de determinació no ha de ser utilitzat com a criteri per seleccionar entre models ennierats.

Definició 1.38. Definim el **coeficient de determinació corregit** ($\bar{R}^2$) com

$$\boxed{\bar{R}^2 = 1 - \frac{n-1}{n-(k+1)}(1-R^2) \in [0,1]}$$

on $k+1$ és el nombre de variables que té el model.

Amb aquest nou coeficient, en augmentar el nombre de variables, augmenta $n-(k+1)$ i per tant, el coeficient pot disminuir en comparar dos models ennierats.

Resumint: El coeficient R^2 ens indica quin percentatge de la variable endògena està explicat per les variables exògenes, i per tant, ens permet comparar models no ennierats per deduir quin dels dos s'ajusta més a la realitat. En canvi, si volem comparar dos models ennierats, caldrà utilitzar el $\bar{R}^2$ per saber quin dels dos s'ajusta més a la realitat.

1.3.2 Significació de paràmetres

Començarem aquest apartat estudiant la **significació econòmica** dels paràmetres.

Si tenim el model $y = \beta_0+\beta_1\cdot x_1+\cdots+\beta_k\cdot x_k+u$, observem que $\beta_j = \frac{\partial y}{\partial x_j}$ i per tant, els paràmetres reflecteixen l'efecte que tenen, en mitjana, variacions unitàries de les variables explicatives sobre la variable endògena. És important que les estimacions obtingudes siguin coherents, per exemple, si esperem que un paràmetre estimat sigui positiu i surt negatiu, hauríem de revisar el que ha passat.

Un altre pregunta que ens podríem fer és: Quina variable explicativa té més influència sobre la variable endògena? Com que les variables exògenes, poden estar en unitats diferents, no seria correcte mirar el valor del paràmetre estimat. Hi ha dues possibilitats:

1. Contribució mitjana d'una variable: es calcula $\hat{\beta}_j \cdot \bar{x}_j,\ \forall j = 1,\ldots,k$ i es compara quin és més gran.

2. Coeficient beta estandarditzat: Es pot calcular de dues maneres. La primera és a partir de l'expressió

   $\hat{\beta}_j^* = \hat{\beta}_j \dfrac{s_{x_j}}{s_y}$, on s_{x_j} i s_y són les desviacions típiques de x_j i y respectivament

   i es comparen aquests paràmetres. La segona opció és transformar el model on cada variable estigui dividida per la seva desviació típica i estimar aquest nou model:
   $$\frac{y}{s_y} = \beta_0 + \beta_1 \cdot \frac{x_1}{s_{x_1}} + \cdots + \beta_k \cdot \frac{x_k}{s_{x_k}} + u$$
   i comparar els paràmetres d'aquesta nova estimació.

Un cop fets aquests apunts sobre la significació econòmica ens centrarem en la **significació estadística**, on analitzarem, mitjançant contrastos, la significació estadística individual i conjunta de les variables del model.

- El **contrast de significació individual** de la variable x_j contrasta si aquesta variable és estadísticament significativa, és a dir, si realment explica a la variable endògena. Aquest contrast té la següent hipòtesi nul·la:
  $$\begin{cases} H_0 :\ \beta_j = 0 \\ H_A :\ \beta_j \neq 0 \end{cases}$$

Pel corol·lari 1.20 sabem que $\hat{\beta}_{MQO} \sim N(\beta, \sigma^2(X^TX)^{-1})$. Estandarditzant aquesta distribució Normal i utilitzant que sota H_0 tenim $\beta_j = 0$, obtenim que

$$\frac{\hat{\beta}_{MQO,\,j} - \beta_j}{\sqrt{\sigma^2((X^TX)^{-1})_{j,j}}} \sim N(0,1) \overset{H_0:\,\beta=0}{\Longrightarrow} \frac{\hat{\beta}_{MQO,\,j}}{\sqrt{\sigma^2((X^TX)^{-1})_{j,j}}} \sim N(0,1),$$

a més, tenim que

$$\frac{n-k}{\sigma^2} \cdot \hat{\sigma}^2_{MQO,\,j} = \frac{\cancel{n-k}}{\sigma^2} \cdot \frac{SQE}{\cancel{n-k}} = \frac{SQE}{\sigma^2} \sim \mathcal{X}^2_{n-(k+1)}.$$

Ara, utilitzant la definició de la llei t de Student: si $X \sim N(0,1)$ és independent de $Y \sim \mathcal{X}^2_n$, aleshores

$$\frac{X}{\sqrt{Y/n}} \text{ es coneix com la t de Student amb } n \text{ graus de llibertat}$$

i, aplicant aquesta definició al nostre cas, obtenim que

$$\frac{\hat{\beta}_{MQO,\,j}/\sqrt{\cancel{\sigma^2}((X^TX)^{-1})_{j,j}}}{\sqrt{\cancel{(n-k)}\hat{\sigma}^2_{MQO,\,j}/[\cancel{\sigma^2}\cancel{(n-k)}]}} = \frac{\hat{\beta}_{MQO,\,j}}{\sqrt{\hat{\sigma}^2_{MQO,\,j}((X^TX)^{-1})_{j,j}}} \sim t_{n-(k+1)},$$

en conclusió, hem deduït que l'estadístic de prova per al contrast de significació individual t_0 es comporta com una t-Student amb $n-(k+1)$ graus de llibertat

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}} \sim t_{n-(k+1)}.$$

Denotarem per $t_{n;\alpha/2}$ a l'antiimatge de la funció de distribució t de Student amb n graus de llibertat per a un nivell de significació α. El criteri de decisió que s'utilitza és el següent:

- Si $|t_0| \geq t_{n-(k+1);\alpha/2}$ rebutgem la hipòtesi nul·la i per tant, considerem que el paràmetre β_j és estadísticament diferent de $0 \Longrightarrow$ la variable x_j és rellevant.
- Si $|t_0| < t_{n-(k+1);\alpha/2}$ acceptem la hipòtesi nul·la i per tant, considerem que el paràmetre β_j és estadísticament igual a $0 \Longrightarrow$ la variable x_j no és rellevant.

- El **contrast de significació conjunta** del model contrasta si les variables explicatives són estadísticament significatives, en conjunt. Aquest contrast té la següent hipòtesi i estadístic de prova:

$$\begin{cases} H_0:\ \beta_1 = \ldots = \beta_k = 0 \\ H_A:\ \exists j \in \{1,\ldots,k\} \text{ tal que } \beta_j \neq 0 \end{cases} \qquad i \ \ F_0 = \frac{\dfrac{SQR}{k}}{\dfrac{SQE}{n-(k+1)}} \sim F_{k;n-(k+1)}$$

on $F_{k;n-(k+1)}$ representa la distribució F de Fisher (o F de Snedecor) amb k graus de llibertat al numerador i $n-(k+1)$ graus de llibertat al denominador.

Observació 1.39. L'estadístic de prova F_0 es pot reescriure, si el model té terme independent, com:

$$F_0 = \frac{\frac{SQR}{k}}{\frac{SQE}{n-(k+1)}} = \frac{\frac{SQR/SQT}{k}}{\frac{SQE/SQT}{n-(k+1)}} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-(k+1)}} \sim F_{k;n-(k-1)}$$

Denotarem per $F_{k;n;\alpha/2}$ a l'antiimatge de la funció de distribució F de Fisher amb k graus de llibertat al numerador i n graus de llibertat al denominador per a un nivell de significació α. El criteri de decisió que s'utilitza és el següent:

- Si $F_0 \geq F_{k-1;n-(k+1);\alpha}$ rebutgem la hipòtesi nul·la i per tant, considerem que el model és globalment significatiu.
- Si $F_0 < F_{k-1;n-(k+1);\alpha}$ acceptem la hipòtesi nul·la i per tant, considerem que el model no és globalment significatiu.

Observació 1.40. De l'expressió de F_0 es pot deduir que, si R^2 és proper a 1, aleshores F_0 serà major i per tant, es conduirà a rebutjar la hipòtesi nul·la.

1.4 Predicció

L'objectiu d'aquesta secció és mostrar com fer prediccions, que és un dels grans objectius de l'Econometria.

Quan es tenen dades de tall transversal, pot ser de gran interès predir el comportament d'un individu que no hagi estat inclòs en la mostra, en canvi, si tenim dades de tall temporal, pot ser d'interès predir el comportament futur de la variable endògena. Cal tenir present que les prediccions poden ser puntuals (es prediu un valor) o per interval (es determina un interval de possibles valors entre els quals estigui el valor de la variable endògena).

Notem que per fer les prediccions s'ha de suposar que les hipòtesis bàsiques es mantindran per a les observacions de fora de la mostra, en concret, és important que es compleixi la hipòtesi de permanència estructural, ja que en cas contrari no tindria sentit fer la predicció.

1.4.1 Predicció puntual

Suposem que a partir d'una mostra de n observacions s'ha estimat el següent model:

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k + u$$

obtenint uns coeficients beta estimats

$$y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \cdots + \hat{\beta}_k \cdot x_k.$$

Ara, si volem predir el valor de la variable endògena per una observació $n+1$, de la qual coneixem el valor de les variables explicatives $x_{1,n+1}, \ldots, x_{k,n+1}$, utilitzarem:

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{1,n+1}, \ldots, \hat{\beta}_k \cdot x_{k,n+1}.$$

Aquesta expressió es pot reescriure matricialment com

$$\hat{y}_{n+1} = X_{n+1}^T \cdot \hat{\beta} = \begin{pmatrix} 1, & x_{1,n+1}, \ldots, x_{k,n+1} \end{pmatrix} \cdot \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}.$$

1.4.2 Predicció per interval

Donat un nivell de significació α, podem obtenir la predicció per interval sobre la variable endògena per a l'observació $n+1$, X_{n+1}, a partir de la següent expressió:

$$P\left(|y_{n+1} - \hat{y}_{n+1}| \leq t_{n-k;\alpha/2} \sqrt{\hat{\sigma}^2 (1 + X_{n+1}^T (X^T X)^{-1} X_{n+1})} \right) = 1 - \alpha$$

és a dir, la variable endògena per l'observació $n+1$, estarà a l'interval

$$\left[\hat{y}_{n+1} - t_{n-k;\alpha/2} \cdot C,\ \hat{y}_{n+1} + t_{n-k;\alpha/2} \cdot C \right]$$

on

$$C = \sqrt{\hat{\sigma}^2 (1 + X_{n+1}^T (X^T X)^{-1} X_{n+1})}$$

amb una probabilitat de $1 - \alpha$. Notem que $X_{n+1} = (1, x_{1,n+1} + \cdots + x_{k,n+1})$.

1.4.3 Valoració de les prediccions

Per tal de valorar les prediccions que es fan en un MRLM existeixen diverses mesures que permeten avaluar la capacitat predictiva del model. Sigui H el nombre d'observacions que s'han predit. Definim:

Definició 1.41. Error absolut mitjà (**EAM**):

$$EAM(H) = \frac{1}{H} \cdot \sum_{h=1}^{H} |e(h)| = \frac{1}{H} \cdot \sum_{h=1}^{H} |y_{n+h} - \hat{y}_{n+h}|.$$

Definició 1.42. Error quadràtic mitjà (**EQM**):

$$EQM(H) = \frac{1}{H} \cdot \sum_{h=1}^{H} e(h)^2 = \frac{1}{H} \cdot \sum_{h=1}^{H} (y_{n+h} - \hat{y}_{n+h})^2.$$

Observació 1.43. Notem que perquè un conjunt de prediccions siguin bones, tant l'EAM com l'EQM han de ser tan petits com sigui possible. El problema principal d'aquests errors és que estan influïts per les unitats de mesura de la variable endògena, és a dir, són mesures no adimensionals. És a dir, en cap cas sabrem, mitjançant l'EAM o l'EQM si un conjunt de prediccions són bones o no.

Per resoldre aquest problema podem definir:

Definició 1.44. Error percentual absolut mitjà (**EPAM**):

$$EPAM(H) = \frac{100}{H} \cdot \sum_{h=1}^{H} \frac{e(h)}{|y_{n+h}|}$$

i utilitzarem el següent criteri per saber si una predicció és bona o no:

- Si $EPAM \leq 1\% \implies$ Molt bona capacitat predictiva.
- Si $1\% \leq EPAM \leq 3\% \implies$ Bona capacitat predictiva.
- Si $3\% \leq EPAM \leq 5\% \implies$ Capacitat predictiva regular.
- Si $5\% \leq EPAM \implies$ Baixa capacitat predictiva.

1.5 Contrastació de restriccions lineals

Fins ara hem estudiat dos tipus de contrastos (els de significació individual i els de significació conjunta dels paràmetres d'un MRLM), però podríem estar interessats a contrastar altres supòsits, com per exemple $\beta_1 + \beta_3 = 0$.

Aquests contrastos on s'analitzen combinacions lineals de coeficients s'anomenen contrastos de restriccions lineals.

Per fer una contrastació de q restriccions lineals s'hauran de seguir dos passos:

1. Formulació matricial de les restriccions lineals: escrivim de la forma $R_{q\times k} \cdot \beta_{k\times 1} = r_{q\times 1}$, on R és una matriu construïda pels coeficients associats a cada una de les q restriccions i r és un vector que conté els termes independents.

   Exemple 1.45. Si tenim el model $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + u$ i volem contrastar $\beta_2 = 5$ i $\beta_1 = 3\beta_3$ construiríem les matrius

   $$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -3 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$$

2. Construcció de l'estadístic de prova: Primer cal definir les hipòtesis nul·la i alternativa corresponents, a partir de les matrius R i r calculades anteriorment.

   $$\begin{cases} H_0 : R\beta = r \quad (\text{les } q \text{ restriccions són certes}) \\ H_A : R\beta \neq r \end{cases}$$

Definició 1.46. Donat un MRLM i un contrast d'hipòtesis, anomenem **model ampliat** al model en el qual es rebutja la hipòtesi nul·la i anomenem **model restringit** al model en el qual s'accepta la hipòtesi nul·la.

Exemple 1.47. Seguint amb l'exemple anterior, si tenim el MRLM $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + u$ i volem contrastar $\beta_2 = 5$ i $\beta_1 = 3\beta_3$,

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + u \quad \text{és el model ampliat, i}$$

$$y = \beta_0 + 3 \cdot \beta_3 \cdot x_1 + 5 \cdot x_2 + \beta_3 \cdot x_3 + u \Longrightarrow y = \beta_0 + \beta_3(3x_1 + x_3) + 5x_2 + u$$

és el model restringit.

Per al contrast anterior ($H_0 : R\beta = r$) cal utilitzar l'estadístic de prova

$$F_0 = \frac{(SQE_R - SQE_A)/q}{SQE_A/(n-(k+1))} \sim F_{q;n-(k+1)},$$

on SQE_A i SQE_R són la suma dels quadrats dels errors del model ampliat i del model restringit, respectivament.

Ara, utilitzant que $R^2 = 1 - SQE/SQT \Longrightarrow SQE = SQT(1-R^2)$ i que $SQT_A = SQT_R$, ja que

$$SQT = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{no depèn dels coeficients estimats,}$$

podem deduir que

$$F_0 = \frac{(SQT_R(1-R_R^2) - SQT_A(1-R_A^2))/q}{SQT_A(1-R_A^2)/(n-k)} = \frac{\cancel{SQT_A}(1-R_R^2-1+R_A^2)/q}{\cancel{SQT_A}(q-R_A^2)/(n-k)} \Rightarrow$$

$$\boxed{F_0 = \frac{(R_A^2 - R_R^2)/q}{(1-R_A^2)/(n-(k+1))} \sim F_{q;n-(k+1)}}$$

on R_A^2 i R_R^2 són els coeficients de determinació del model ampliat i del model restringit, respectivament.

Per un nivell de significació α utilitzarem el següent criteri de decisió:

- Si $F_0 \geq F_{q;n-(k+1);\alpha} \Longrightarrow$ Es rebutja la hipòtesi nul·la, és a dir, les restriccions no són certes.
- $F_0 < F_{q;n-(k+1);\alpha} \Longrightarrow$ No es rebutja la hipòtesi nul·la, és a dir, les restriccions són certes.

A tall d'exemple provaré que aquest estadístic de prova coincideix amb l'estadístic del contrast de significació conjunta descrit a l'apartat *1.3.2. Significació de paràmetres.*

Observació 1.48. Contrast de significació conjunta: Si tenim el MRLM $y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k + u$ i volem contrastar

$$\begin{cases} H_0 : \beta_1 = \cdots = \beta_k = 0 \\ H_A : \text{ no } H_0 \end{cases} \implies q = k$$

tindríem l'estadístic de prova:

$$F_0 = \frac{(SQE_R - SQE_A)/k}{SQE_A/(n-(k+1))} \sim F_{k;n-(k+1)},$$

on el model restringit satisfà $\beta_1 = \cdots = \beta_k = 0$ i per tant, $y = \beta_0 + u$.

Notem que en estimar aquest model restringit, com es vol minimitzar el SQE, tindrem que

$$SQE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0)^2,$$

i aquest es minimitza quan

$$\frac{\partial SQE}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0)\cdot(-1) = 2\sum_{i=1}^{n}(\hat{\beta}_0 - y_i) = 2\left(\sum_{i=1}^{n} y_i - n\hat{\beta}_0\right) = 0 \iff$$

$$\iff \sum_{i=1}^{n} y_i/n = \hat{\beta}_0 \iff \hat{\beta}_0 = \bar{y}.$$

Per tant, com el model restringit té terme independent, es compleix que $SQT_R = SQE_R + SQR_R$, on

$$SQR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \overset{\hat{y}_i=\bar{y}}{=} \sum_{i=1}^{n}(\bar{y} - \bar{y})^2 = 0 \implies SQT_R = SQE_R,$$

a més, sabem que

$$R_R^2 = 1 - \frac{SQE_R}{SQT_R} = 1 - 1 = 0 \implies R_R^2 = 0,$$

i per tant, podem reescriure l'estadístic de prova com:

$$F_0 = \frac{(R_A^2 - R_R^2)/k}{(1-R_A^2)/(n-(k+1))} \overset{R_R^2=0}{=} \frac{R_A^2/k}{(1-R_A^2)/(n-(k+1))} \sim F_{k;n-(k+1)},$$

que efectivament coincideix amb l'estadístic de prova que hem vist a la secció *1.3.2.*

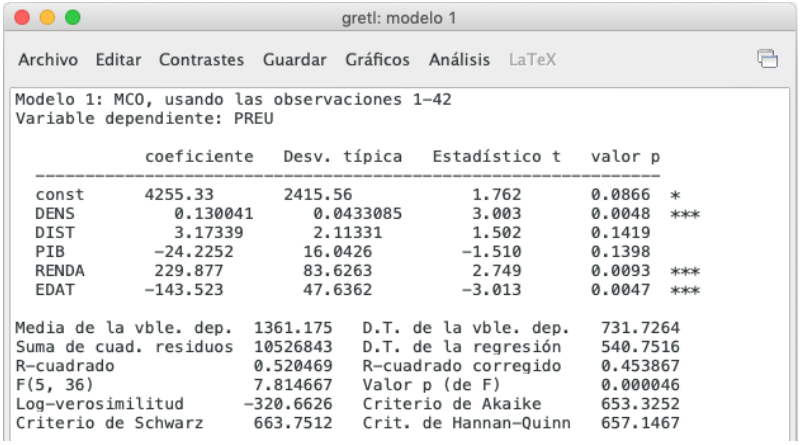
1.6 Exemple amb Gretl

En aquesta secció utilitzaré el programa *Gretl* que, entre altres utilitats, serveix per obtenir l'estimació per MQO donada una mostra.

Primer de tot he creat un fitxer Excel amb les dades del preu de venda mitjà de l'habitatge d'obra nova per metre quadrat, densitat poblacional, distància a Barcelona, PIB per habitant, renda per habitant i mitjana d'edat (a la primera secció de l'annex es pot veure amb detall d'on les he obtingut). En segon lloc, he estimat el següent model per MQO:

$$PREU = \beta_0 + \beta_1 \cdot DENS + \beta_2 \cdot DIST + \beta_3 \cdot PIB + \beta_4 \cdot RENDA + \beta_5 \cdot EDAT + u$$

un cop el *Gretl* ha estimat el model, apareix la següent finestra:



Notem que a la part superior de la finestra, s'indica que s'ha estimat el model per MQO utilitzat 42 observacions (una observació per cada comarca de Catalunya).

Al mig de la finestra apareixen diferents columnes: En la primera hi apareixen la constant i les variables exògenes; en la segona columna (coeficient) apareixen les estimacions per MQO i, per tant, el model estimat és el següent:

$$PREU = 4255.33 + 0.130041 \cdot DENS + 3.17339 \cdot DIST - 24.2252 \cdot PIB + \\ + 229.877 \cdot RENDA - 143.523 \cdot EDAT$$

en la tercera columna (desv. típica) s'indica la desviació típica de cada variable; en la quarta columna (estadístic t) s'indica el valor de l'estadístic de significació individual de cada variable, és a dir,

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{S(\hat{\beta}_j)} \sim t_{42-6;\alpha/2} = t_{36;\alpha/2},$$

per tant, a la quarta columna apareix la divisió entre el coeficient (segona columna) i la desviació típica (tercera columna); finalment, a la cinquena columna (valor p) apareix el $p - valor$ que té la següent utilitat: Per un nivell de significació α, $|t_0| \geq t_{n-k;\alpha/2} \iff p - valor \leq \alpha$.

Utilitzant un nivell de significació del 5% ($\alpha = 0.05$) podem veure com el contrast de significació individual ens porta a la conclusió que:

1. Les variables "DENS", "RENDA" i "EDAT" són estadísticament significatives, ja que tenen un $p - valor$ inferior a 0.05.
2. Les variables "DIST" i "PIB" no són estadísticament significatives, ja que tenen un $p - valor$ superior a 0.05.

A la part inferior de la finestra trobem diversos paràmetres, entre els quals destaquem:

- La mitjana i desviació típica de la variable dependent (en aquest cas, $PREU$).
- La suma dels quadrats dels residus, és a dir, la SQE.
- El coeficient de determinació ($R^2 = 0.520469$), que ens indica que el model explica un 52.0469 % de la variabilitat de $PREU$.
- El coeficient de determinació corregit ($\bar{R}^2 = 0.453867$), que ens permet poder comparar entre diferents models ennierats.
- L'estadístic de significació conjunta del model, que podem comprovar que està ben calculat a partir de la seva fórmula, ja que coneixem el valor de n, k i R^2 :

$$F_0 = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-(k+1)}} = \frac{\frac{0.520469}{5}}{\frac{1-0.520469}{42-(5+1)}} = 7.81467 \sim F_{k;n-(k+1)} = F_{5,36}.$$

- El $p - valor$ del contrast de significació conjunta, que té el següent significat: per un nivell de significació α, $|F_0| \geq F_{k;n-(k+1);\alpha} \iff p - valor \leq \alpha$. En el nostre cas, com el $p - valor$ és $0.000046 \leq 0.05$, podem deduir que el model és globalment significatiu per un nivell de significació del 5% ($\alpha = 0.05$).

A més a més, ens podríem preguntar quina variable exògena té més influència sobre la variable endògena ($PREU$), per això, calcularem el coeficient beta estandarditzat, que s'obté a partir de la fórmula

$$\hat{\beta}_j^* = \hat{\beta}_j \frac{s_{x_j}}{s_y}, \text{ on } s_{x_j},\ s_y \text{ són les desviacions típiques:}$$

$$\hat{\beta}_1^* = 0.130041 \cdot \frac{0.0433085}{731.7264} = 7.6967 \cdot 10^{-6};\ \ \hat{\beta}_2^* = 3.17339 \cdot \frac{2.11331}{731.7264} = 9.1638 \cdot 10^{-3}$$

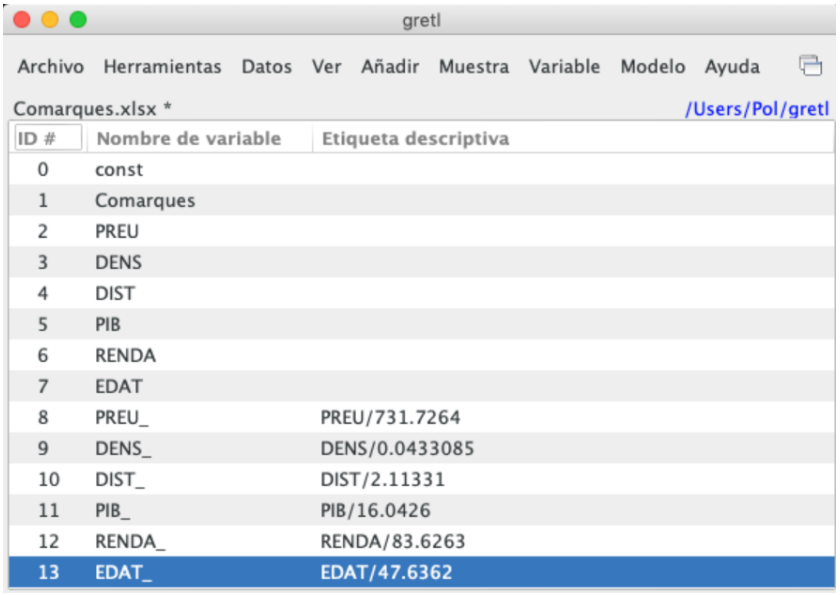
$$\hat{\beta}_3^* = -24.2252 \cdot \frac{16.0426}{731.7264} = -0.5311;\ \ \hat{\beta}_4^* = 229.877 \cdot \frac{83.6263}{731.7264} = 26.2718$$

$$\hat{\beta}_5^* = -143.523 \cdot \frac{47.6362}{731.7264} = -9.3435.$$

Observem que $\hat{\beta}_5^*$ és el coeficient beta estandarditzat més gran en valor absolut, i per tant, la variable $RENDA$ és la que té més influència sobre la variable endògena $PREU$. Una altra manera de calcular aquests coeficients és estimant per MQO el model:

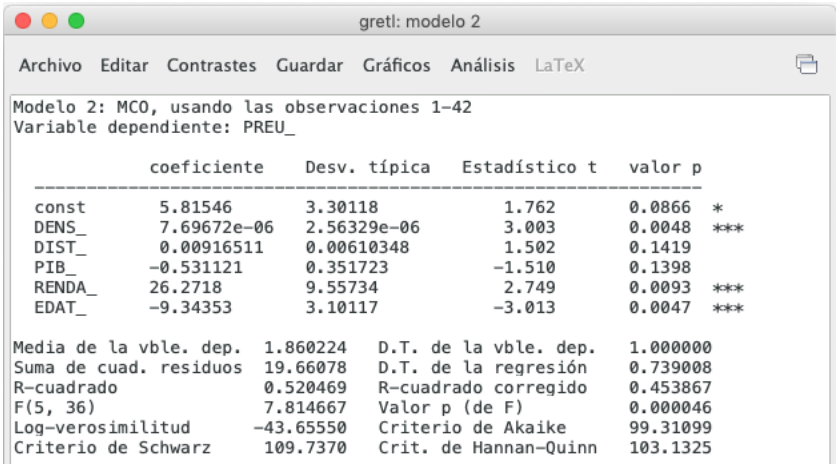
$$\frac{y}{s_y} = \beta_1 + \beta_2 \cdot \frac{x_2}{s_{x_2}} + \cdots + \beta_K \cdot \frac{x_k}{s_{x_k}} + u$$

per tant, si definim les noves variables al *Gretl* x_j/s_{x_j} com veiem a continuació



ID #	Nombre de variable	Etiqueta descriptiva
0	const	
1	Comarques	
2	PREU	
3	DENS	
4	DIST	
5	PIB	
6	RENDA	
7	EDAT	
8	PREU_	PREU/731.7264
9	DENS_	DENS/0.0433085
10	DIST_	DIST/2.11331
11	PIB_	PIB/16.0426
12	RENDA_	RENDA/83.6263
13	EDAT_	EDAT/47.6362

i estimant el model de les noves variables per MQO, obtenim els resultats:




```
Modelo 2: MCO, usando las observaciones 1-42
Variable dependiente: PREU_

            coeficiente    Desv. típica   Estadístico t   valor p
  ------------------------------------------------------------------
  const       5.81546       3.30118          1.762       0.0866  *
  DENS_       7.69672e-06   2.56329e-06      3.003       0.0048  ***
  DIST_       0.00916511    0.00610348       1.502       0.1419
  PIB_       -0.531121      0.351723        -1.510       0.1398
  RENDA_     26.2718        9.55734          2.749       0.0093  ***
  EDAT_      -9.34353       3.10117         -3.013       0.0047  ***

Media de la vble. dep.  1.860224   D.T. de la vble. dep.   1.000000
Suma de cuad. residuos  19.66078   D.T. de la regresión    0.739008
R-cuadrado              0.520469   R-cuadrado corregido    0.453867
F(5, 36)                7.814667   Valor p (de F)          0.000046
Log-verosimilitud      -43.65550   Criterio de Akaike      99.31099
Criterio de Schwarz     109.7370   Crit. de Hannan-Quinn   103.1325
```

on podem comprovar, que els coeficients estimats $\hat{\beta}$ coincideixen amb els coeficients beta estandarditzats calculats anteriorment. D'altra banda, també observem com els estadístics de significació individual t, així com els seus p-valors coincideixen en els dos models. El mateix passa amb l'estadístic de significació global F, el seu p-valor, amb el coeficient de determinació (R^2) i amb el coeficient de determinació corregit ($\bar{R}^2$).

2 Anàlisi de correlació canònica

L'objectiu d'aquest capítol serà estudiar la relació multivariant que hi pugui haver entre dos grups vectors aleatoris $X = (x_1, \ldots, x_p)^T$ i $Y = (y_1, \ldots, y_q)^T$.

Notem que aquest anàlisi és una generalització de l'anàlisi de correlació múltiple, ja que, en el capítol anterior hem vist que l'anàlisi de correlació múltiple relaciona una variable aleatòria y amb un vector aleatori X.

2.1 La correlació canònica

Sigui X i Y dos vectors aleatoris de dimensions p i q respectivament, volem trobar dues variables

$$U = X^T a = a_1 x_1 + \cdots + a_p x_p \quad i \quad V = Y^T b = b_1 y_1 + \cdots + b_q y_q,$$

$$on \quad a = (a_1, \ldots, a_p)^T \quad i \quad b = (b_1, \ldots, b_q)^T$$

tals que la correlació $Corr(U, V)$ sigui màxima.

Denotem per S_{11} i S_{22} les matrius de covariàncies de les variables X i Y, respectivament, i denotem per S_{12} la matriu de covariàncies de les variables X amb les variables Y (anàlogament, $S_{21} = S_{12}^T$ la matriu de covariàncies de les variables Y amb les variables X).

Notem que S_{11} és de dimensió $p \times p$, S_{22} és de dimensió $q \times q$, S_{12} és de dimensió $p \times q$ i S_{21} és de dimensió $q \times p$.

Ara, suposant que les variàncies de les variables U i V són 1:

$$Var(U) = a^T S_{11} a = 1, \quad Var(V) = b^T S_{22} b = 1.$$

El problema de maximitzar la correlació entre U i V es redueix a:

$$\text{maximitzar} \quad a^T S_{12} b \quad \text{restringit a} \quad Var(U) = Var(V) = 1.$$

Definició 2.1. Els vectors a i b que satisfan les condicions anteriors s'anomenen **primer vectors canònics** i anomenarem **primera correlació canònica** (r_1) a la màxima correlació entre U i V.

Teorema 2.2. *Els primers vectors canònics satisfan les equacions:*

$$S_{12} S_{22}^{-1} S_{21} a = \lambda S_{11} a \quad i \quad S_{21} S_{11}^{-1} S_{12} b = \lambda S_{22} b.$$

Demostració. Per fer aquesta demostració utilitzarem el mètode de maximització dels multiplicadors de Lagrange, que consisteix a definir una funció que tingui una primera part que correspon a la funció que volem maximitzar i una segona part, on hi hagi tants termes com restriccions multiplicats per una constant. En el nostre cas tenim dues restriccions: $a^T S_{11} a - 1 = 0$ i $b^T S_{22} b - 1 = 0$. Dit això, definim la funció

$$\phi(a, b) = a^T S_{12} b - \frac{\alpha}{2}(a^T S_{11} a - 1) - \frac{\beta}{2}(b^T S_{22} b - 1),$$

on α i β són multiplicadors de Lagrange.

Ara, fent les derivades parcials respecte a i b obtenim

$$\begin{cases} \dfrac{\partial \phi}{\partial a} = 0 \iff S_{12}b - \alpha S_{11}a \overset{(1)}{=} 0 \iff a^T S_{12} b = \alpha a^T S_{11} a \\ \dfrac{\partial \phi}{\partial b} = 0 \iff S_{21}a - \beta S_{22}b \overset{(2)}{=} 0 \iff b^T S_{21} a = \beta b^T S_{22} b \end{cases}$$

i, utilitzant que $a^T S_{11} a = 1 = b^T S_{22} b$ i que $a^T S_{12} b = (a^T S_{12} b)^T = b^T S_{21} a$, ja que és un nombre real i $S_{21} = S_{12}^T$, veiem que

$$\alpha = a^T S_{12} b = b^T S_{21} a = \beta \Longrightarrow \alpha = \beta.$$

Finalment, a partir de les equacions (1) i (2) veiem que

$$\begin{cases} (1):\ S_{12}b = \alpha S_{11}a \iff a \overset{(3)}{=} \dfrac{1}{\alpha} S_{11}^{-1} S_{12} b \\ (2):\ S_{21}a = \beta S_{22}b \iff b \overset{(4)}{=} \dfrac{1}{\beta} S_{22}^{-1} S_{21} a \end{cases}$$

i substituint (4) a l'equació (1) i definint $\lambda = \alpha \cdot \beta$ obtenim

$$S_{12}\frac{1}{\beta} S_{22}^{-1} S_{21} a = \alpha S_{11} a \Longrightarrow S_{12} S_{22}^{-1} S_{21} a = \lambda S_{11} a.$$

Anàlogament, substituint (3) a l'equació (2) obtenim

$$S_{21}\frac{1}{\alpha} S_{11}^{-1} S_{12} b = \beta S_{22} b \Longrightarrow S_{21} S_{11}^{-1} S_{12} b = \lambda S_{22} b,$$

com volíem veure. □

Teorema 2.3. *Els vectors canònics normalitzats estan relacionats per*

$$a = \lambda^{-1/2} S_{11}^{-1} S_{12} b \quad \text{i} \quad b = \lambda^{-1/2} S_{22}^{-1} S_{21} a$$

i la primera correlació canònica és $r_1 = \sqrt{\lambda_1}$*, on* λ_1 *és el major VAP de* $S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$.

Demostració. A partir de les equacions (1) i (2) de la demostració anterior veiem que $a = \alpha^{-1} S_{11}^{-1} S_{12} b$ i $b = \beta^{-1} S_{22}^{-1} S_{21} a$ i ara, utilitzant que $\alpha = \beta$ i $\lambda = \alpha \cdot \beta$ tenim que $\alpha = \beta = \lambda^{1/2} \Longrightarrow \alpha^{-1} = \beta^{-1} = \lambda^{-1/2}$, d'on deduïm que

$$a = \lambda^{-1/2} S_{11}^{-1} S_{12} b \quad \text{i} \quad b = \lambda^{-1/2} S_{22}^{-1} S_{21} a.$$

D'altra banda, la correlació entre a i b és $r_1 = a^T S_{12} b$ i sabem que, com tractem vectors canònics normalitzats,

$$1 = a^T S_{11} a = a^T S_{11} \cdot \lambda^{-1/2} S_{11}^{-1} S_{12} b = \lambda^{-1/2} a^T S_{12} b \Longrightarrow \lambda^{1/2} = a^T S_{12} b,$$

d'on deduïm que $r^2 = \lambda$, on pel Teorema 2.2 sabem que

$$S_{11}^{-1} S_{12} S_{22}^{-1} S_{21} a = \lambda a \Longrightarrow \ \lambda \text{ representa els VAPs de } S_{11}^{-1} S_{12} S_{22}^{-1} S_{21},$$

i com que volem maximitzar la correlació r_1, prendrem $r_1^2 = \lambda_1$, el major VAP de $S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$. □

Observació 2.4. Notem que, les equacions en valors i vectors propis tenen altres solucions, concretament hi ha $m = min\{p, q\}$ parelles de vectors canònics $(a_1, b_1), \ldots, (a_m, b_m)$ que defineixen variables $U_i = X^T a_i$, $V_i = Y^T b_i$ amb correlacions canòniques $r_1, \ldots, r_m$.

Teorema 2.5. *Suposant que $r_1 > \cdots > r_m$, aleshores es compleix que:*

1. *Les variables $U_1, \ldots, U_m$ estan incorrelacionades (ídem amb $V_1, \ldots, V_m$).*
2. *La primera correlació canònica $r_1 = Corr(U_1, V_1)$ és la màxima correlació entre una combinació lineal de X i una combinació lineal de Y.*
3. *La segona correlació canònica $r_2 = Corr(U_2, V_2)$ és la màxima correlació entre les combinacions lineals de X incorrelacionades amb U_1 i les combinacions lineals de Y incorrelacionades amb V_1.*
4. *$Corr(U_i, V_i) = 0$ si $i \neq j$.*

Demostració. Sigui $i \neq j$, multiplicant per a_i^T i per a_j^T les equacions del Teorema 2.2, tenim que

$$\begin{cases} a_i^T S_{12} S_{22}^{-1} S_{21} a_j = \lambda_j a_i^T S_{11} a_j \\ a_j^T S_{12} S_{22}^{-1} S_{21} a_i = \lambda_i a_j^T S_{11} a_i \end{cases}$$

ara, si restem les dues equacions, obtenim que

$$a_j^T S_{12} S_{22}^{-1} S_{21} a_i - a_i^T S_{12} S_{22}^{-1} S_{21} a_j = \lambda_i a_j^T S_{11} a_i - \lambda_j a_i^T S_{11} a_j \iff$$

$$\iff 0 = (\lambda_i - \lambda_j) a_i^T S_{11} a_j \overset{\lambda_i \neq \lambda_j}{\iff} a_i^T S_{11} a_j = 0 \iff Corr(U_i, U_j) = 0.$$

Fent el mateix raonament amb b_i, b_j deduïm que $cor(V_i, V_j) = 0$, com volíem veure.

D'altra banda, també pel Teorema 2.2 sabem que

$$\begin{cases} S_{11}^{-1} S_{12} S_{22}^{-1} S_{21} a_i = \lambda_i a_i \iff b_j^T S_{21} S_{11}^{-1} S_{12} S_{22}^{-1} S_{21} a_i = \lambda_i b_j^T S_{21} a_i \\ S_{11}^{-1} S_{12} S_{22}^{-1} S_{21} b_j = \lambda_j b_j \iff a_i^T S_{21} S_{11}^{-1} S_{12} S_{22}^{-1} S_{21} b_j = \lambda_j a_i^T S_{21} b_j \end{cases}$$

ara, si restem les dues equacions, obtenim que

$$0 = (\lambda_i - \lambda_j) a_i^T S_{12} b_j = 0 \overset{\lambda_1 \neq \lambda_2}{\iff} a_i^T S_{12} b_j = 0 \iff Corr(U_i, V_j) = 0.$$

Amb això queden provats els punts (1) i (4) del teorema; els altres dos punts expliquen com construir les components canòniques. □

2.2 Tests de significació de les correlacions canòniques i d'independència

Fins ara hem trobat les variables i correlacions canòniques a partir de les matrius de covariàncies. Ara ens podríem preguntar quines correlacions canòniques

$$r_1 \geq r_2 \geq \cdots \geq r_m$$

on $m = min\{p, q\}$, són significatives.

Denotarem $r_0 = 1$ i definim la següent hipòtesi nul·la:

$$H_0^k : \ \ r_k > r_{k+1} = \cdots = r_m = 0, \ \ k \in \{0, 1, \ldots, m\}$$

que, matricialment, equival a:

$$H_0^k : \ \ rang(S_{22}^{-1} S_{21}) = k, \ \ k \in \{0, 1, \ldots, m\}.$$

Teorema 2.6. ***Test de Bartlett-Lawley****: Si la hipòtesi nul·la anterior H_0^k és certa, aleshores l'estimador*

$$L_k = -[n - 1 - k - \frac{1}{2}(p + q + 1) + \sum_{i=1}^{k} r_i^{-2}] log \left[\prod_{i=k+1}^{m} (1 - r_i^2) \right]$$

es comporta asimptòticament com la distribució χ^2 amb $(m - k) \cdot (p - k)$ graus de llibertat. Si L_i resulta ser significatiu per $i = 0, 1, \ldots, k-1$ però L_k no és significatiu, llavors acceptarem H_0^k.

També ens podríem preguntar si els vectors aleatoris X i Y són independents. Suposant normalitat, aquest fet equival a plantejar el test

$$\begin{cases} H_0 : \ S_{12} = 0 \\ H_A : \ S_{12} \neq 0 \end{cases}$$

Resoldrem aquest test pel principi d'unió intersecció: Considerem les variables

$$U = a^T X = a_1 X_1 + \cdots + a_p X_p, \ \ i \ \ V = b^T Y = b_1 Y_1 + \cdots + b_q Y_q$$

que tenen correlació

$$r(U, V) = \frac{a^T S_{12} b}{\sqrt{a_T S_{11} a} \sqrt{b^T S_{22} b}},$$

ara, aplicant el principi d'unió intersecció, acceptarem H_0 si $r(U, V)$ no és significativa per tot U, V, en canvi, rebutjarem H_0 si és significativa per algun parell U, V. Aquest criteri ens porta a estudiar la significació de $r_1 = \underset{U,V}{max}\, r(U, V)$.

Per tant, el test anterior que té per hipòtesi nul·la $H_0 : \ S_{12} = 0$ equival a

$$\begin{cases} H_0 : \ r_1 = 0 \\ H_A : \ r_1 > 0 \end{cases}$$

i per veure si r_1 és significativa podem aplicar l'estadístic L_0 de Bartlett-Lawley que hem utilitzat per al test de significació de les correlacions canòniques.

2.3 Exemple amb RStudio

En aquesta secció utilitzaré el programa *RStudio* que servirà per obtenir les variables i correlacions canòniques.

Per fer aquest exemple he creat el fitxer *Comarques_Nom_Eleccions_LOG* en format Excel que conté les dades (a l'annex es pot veure en detall d'on les obtinc), per a cadascuna de les 42 comarques de Catalunya, dels següents vectors aleatoris:

- X conté el logaritme del percentatge de vot de cadascun dels 5 partits polítics que van obtenir una major representació parlamentària a les eleccions autonòmiques de Catalunya, celebrades el passat 14 de febrer de 2021 (és a dir: ERC, PSC, JxCat, VOX i la CUP), així, he definit el vector aleatori

$$X^T = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{pmatrix} = \begin{pmatrix} \log(\text{percentatge vots ERC}) \\ \log(\text{percentatge vots PSC}) \\ \log(\text{percentatge vots JxCat}) \\ \log(\text{percentatge vots VOX}) \\ \log(\text{percentatge vots CUP}) \end{pmatrix}$$

- Y pretén analitzar si la població de cada comarca és més o menys catalanoparlant, així, he definit el següent vector aleatori que recull el logaritme del quocient entre el nombre de persones que tenen el mateix nom on, al numerador surt la versió catalana del nom i al denominador, la versió castellana del nom, per tant, a major quocient s'entén que la població d'aquella comarca té més tendència a parlar català, en canvi, si el quocient és més proper a 0, predominarà el castellà

$$Y^T = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \log(\text{quocient Josep/José}) \\ \log(\text{quocient Anna/Ana}) \end{pmatrix}$$

L'objectiu d'aquest exemple serà veure si hi ha relació entre les comarques on predominen els partits independentistes (ERC, JxCat i la CUP) i les comarques on predominen els noms catalans (és a dir, on major és el logaritme dels quocients Josep/José i Anna/Ana).

Primer de tot cal obrir el fitxer *Comarques_Nom_Eleccions_LOG.xlsx* i un cop obert definim una matriu X que contingui les dades de les variables $X_1, \ldots, X_5$ i una matriu Y que contingui les dades de les variables Y_1, Y_2.

```
> # Primer obrim l'arxiu amb les dades i l'anomenem "Mostra_ACC"
> Mostra_ACC <- read.xlsx("/Users/Pol/Desktop/Comarques_Nom_Eleccions_LOG.xlsx")
> # Ara definim les matrius X=(X1,...,X5) i Y=(Y1,Y5)
> X <- Mostra_ACC[,(2:6)]
> Y <- Mostra_ACC[,(7:8)]
> # A continuació, podem estudiar les diferents matrius de covariàncies:
> cov(X)
                    X1=log(.Vots.ERC) X2=log(.Vots.PSC) X3=log(.Vots.JxCat) X4=log(.Vots.VOX) X5=log(.Vots.CUP)
X1=log(.Vots.ERC)         0.006531943      -0.008819681         0.005658387       -0.01233091       0.005273884
X2=log(.Vots.PSC)        -0.008819681       0.027237983        -0.021706997        0.03362993      -0.017432929
X3=log(.Vots.JxCat)       0.005658387      -0.021706997         0.022385100       -0.02710158       0.016029132
X4=log(.Vots.VOX)        -0.012330912       0.033629931        -0.027101576        0.05673058      -0.026578755
X5=log(.Vots.CUP)         0.005273884      -0.017432929         0.016029132       -0.02657876       0.017515449
> cov(Y)
                    Y1=log(.Josep/José) Y2=log(.Anna/Ana)
Y1=log(.Josep/José)           0.1092337        0.07561450
Y2=log(.Anna/Ana)             0.0756145        0.06073859
> cov(X,Y)
                    Y1=log(.Josep/José) Y2=log(.Anna/Ana)
X1=log(.Vots.ERC)            0.01561084       0.008944457
X2=log(.Vots.PSC)           -0.04606641      -0.034166750
X3=log(.Vots.JxCat)          0.04185804       0.031778373
X4=log(.Vots.VOX)           -0.06212583      -0.042158374
X5=log(.Vots.CUP)            0.03263851       0.021259642
```

Com a primera conclusió podem destacar que la covariància entre X_1, X_3 i X_5 és positiva, és a dir, quan una variable creix, l'altre també i per tant, els partits independentistes es comporten de manera similar i el mateix amb les variables X_2 i X_4, que tenen covariància positiva i per tant, els dos partits no independentistes es comporten de manera similar. En canvi, si mirem les covariàncies entre partits independentistes i partits no independentistes, és a dir:

$Cov(X_1, X_2)$, $Cov(X_1,\ X_4)$, $Cov(X_3, X_2)$, $Cov(X_3,\ X_4)$, $Cov(X_5, X_2)$ i $Cov(X_5,\ X_4)$

notem que són negatives, per tant, quan un creix, l'altre decreix, fet que és bastant lògic seguint amb el raonament anterior.

Finalment, a partir de la matriu $Cov(X, Y)$ notem que la covariància entre els partits independentistes (X_1, X_3, X_5) i els quocients (Y_1, Y_2) són positives i per tant, es reforça la idea que en les comarques on predominen els noms catalans sobre els castellans es tendeix a votar a partits independentistes, mentre que la covariància entre els partits no independentistes (X_2, X_4) i els quocients (Y_1, Y_2) són negatives i per tant, a menor quocient (quan predominen els noms castellans sobre els catalans) major és el percentatge de vot dels partits no independentistes.

Un cop fetes aquestes primeres consideracions podem calcular les components principals, que és el que ens interessa a partir de les fórmules:

$$\begin{cases} \lambda a = S_{11}^{-1} S_{12} S_{22}^{-1} S_{21} a \Longrightarrow \lambda \cdot a = A \cdot a \\ \lambda b = S_{22}^{-1} S_{21} S_{11}^{-1} S_{12} b \Longrightarrow \lambda \cdot b = B \cdot b \end{cases}$$

Podem traduir aquestes expressions a llenguatge R utilitzant la funció solve() que calcula la inversa d'una matriu de la següent manera:

```
> A <-solve(cov(X,X))%*%cov(X,Y)%*%solve(cov(Y,Y))%*%cov(Y,X)
> B <-solve(cov(Y,Y))%*%cov(Y,X)%*%solve(cov(X,X))%*%cov(X,Y)
```

Un cop calculades les matrius A i B ja podem trobar els VAPs i VEPs per trobar les correlacions canòniques

```
> eigen(A)$values
[1]  8.091221e-01+0.000000e+00i  4.210176e-01+0.000000e+00i -2.681898e-17+2.781635e
-16i -2.681898e-17-2.781635e-16i -2.271249e-17+0.000000e+00i
> eigen(B)$values
[1] 0.8091221 0.4210176
> a <-eigen(A)$vector[,1]
> b <-eigen(B)$vector[,1]
> a
[1]  0.08814199+0i  0.40933567+0i -0.80481331+0i  0.18313165+0i  0.37869993+0i
> b
[1] -0.3300072 -0.9439784
```

Notem que el major VAP és $r_1 = 0'8091221$. D'altra banda, notem que per obtenir les components principals s'ha de complir que

$$a^T S_{11} a = 1 \ \ i \ \ b^T S_{22} b = 1,$$

per tant, calcularem el valor de $a^T S_{11} a$ i ho normalitzarem dividint el vector a per l'arrel quadrada d'aquest valor (ídem per b).

```
> t(a)%*%cov(X,X)%*%a
                [,1]
[1,] 0.03051424+0i
> a=a/sqrt(0.03051424)
> a
[1]  0.5045818+0i  2.3433023+0i -4.6072722+0i  1.0483641+0i  2.1679234+0i
> t(a)%*%cov(X,X)%*%a
     [,1]
[1,] 1+0i
> t(b)%*%cov(Y,Y)%*%b
          [,1]
[1,] 0.1131307
> b=b/sqrt(0.1131307)
> b
[1] -0.9811449 -2.8065438
> t(b)%*%cov(Y,Y)%*%b
     [,1]
[1,]    1
```

Per tant, ja hem obtingut els primers vectors canònics:

$$a_1 = (0.5046, 2.3433, -4.6072, 1.0484, 2.1679) \ \ i \ \ b_1 = (-0.9811, -2.8065)$$

i, en conseqüència, les primeres variables canòniques (amb variància 1) són:

$$U_1 = 0.5046 \cdot X_1 + 2.3433 \cdot X_2 - 4.6072 \cdot X_3 + 1.0484 \cdot X_4 + 2.1679 \cdot X_5$$

$$V_1 = -0.9811 \cdot Y_1 - 2.8065 \cdot Y_2$$

Finalment, podríem fer el test de Bartlett-Lawley per veure quines correlacions canòniques són significatives. En el nostre cas tenim

$$r_1 = 0.8091221, \ \ r_2 = 0.4210176,$$

i sabem que la hipòtesi nul·la

$$H_0^0 : 1 \geq r_1 = r_2 = 0$$

té l'estadístic de prova

$$L_0 = -[42 - 1 - 0 - \frac{1}{2}(5 + 2 + 1)]log[(1 - r_1^2)(1 - r_2)^2] = 46.56$$

i aquest estadístic es comporta com la distribució $\mathcal{X}^2$ amb $(2-0)(5-0) = 10$ graus de llibertat. Si fem aquest contrast per un nivell de significació $\alpha = 5\%$, obtenim que $\mathcal{X}^2_{10;0.05} = 18.307$ a partir de les taules estadístiques i per tant,

$$L_0 = 46.56 > 18.3075 \Longrightarrow \text{Rebutgem la hipòtesi nul·la}$$

Com hem rebutjat la hipòtesi nul·la H_0^0 sabem que, com a mínim, r_1 és significativa. Considerem ara la hipòtesi nul·la

$$H_0^1 : r_1 > r_2 = 0$$

que té l'estadístic de prova

$$L_1 = -[42 - 1 - 1 - \frac{1}{2}(5 + 2 + 1) + r_1^{-2}]log[(1 - r_2)^2] = 7.32$$

i aquest estadístic es comporta com la distribució $\mathcal{X}^2$ amb $(2-1)(5-1) = 4$ graus de llibertat. Si fem aquest contrast per un nivell de significació $\alpha = 5\%$, obtenim que $\mathcal{X}^2_{4;0.05} = 9.49$ a partir de les taules estadístiques i per tant,

$$L_0 = 7.32 < 9.49 \Longrightarrow \text{Acceptem la hipòtesi nul·la}$$

Com hem acceptat la hipòtesi nul·la $H_0^1 : r_1 > r_2 = 0$, tenim que la primera correlació canònica és significativa, però la segona no ho és.

3 Anàlisi de components principals

Suposem que tenim una mostra de n observacions amb p variables per cada observació, si ho volguéssim representar gràficament, necessitaríem un espai de dimensió p i, si p és elevat, és impossible de visualitzar, per això, durant aquesta secció veurem un mètode que ens permet, mitjançant combinacions lineals de les p variables inicials reduir el sistema inicial de p variables a $m = 2$ o $m = 3$ variables.

3.1 Obtenció de les components principals

Definició 3.1. Suposem que hi ha n individus $w_1, \ldots, w_n$ i p variables $x_1, \ldots, x_p$. Sigui $x_{i,j} = x_j(w_i)$ l'observació de la variable x_j sobre l'individu w_i, definim la **matriu de dades multivariants** com

$$X^T = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

ara, denotant per $\bar{x}_j$ a la mitjana de la variable x_j per les observacions dels n individus, és a dir,

$$\bar{x}_j = \frac{1}{n} \sum_{j=1}^{n} x_{j,i}$$

podem definir la **matriu de covariàncies mostrals** com

$$S = \begin{pmatrix} s_{1,1} & \cdots & s_{1,p} \\ \vdots & \ddots & \vdots \\ s_{p,1} & \cdots & s_{p,p} \end{pmatrix} \quad on \;\; s_{i,j} = \frac{1}{n} \sum_{t=1}^{n} (x_{t,i} - \bar{x}_i)(x_{t,j} - \bar{x}_j).$$

Definició 3.2. Una variable y es diu **variable composta** si és una combinació lineal de les variables observables amb coeficients $a = (a_1, \ldots, a_p)^T$, és a dir,

$$y = a_1 x_1 + \ldots + a_p x_p.$$

Si X és la matriu de dades multivariants, podem escriure $Y = X^T a$.

Definició 3.3. Les **components principals** són les variables compostes

$$Y_1 = X^T t_1, \cdots, Y_p = X^T t_p \text{ construïdes de la següent manera:}$$

1. $Var(Y_1)$ és màxima condicionada a $t_1^T t_1 = 1$.
2. Entre totes les variables compostes Y, tals que $Cov(Y_1, Y) = 0$, la variable Y_2 és la que maximitza $Var(Y_2)$ condicionada a $t_2^T t_2 = 0$.
3. Y_3 és una variable incorrelacionada amb Y_1 i Y_2, és a dir, $Cov(Y_1, Y_3) = Cov(Y_2, Y_3) = 0$, amb $Var(Y_3)$ màxima condicionada a $t_3^T t_3 = 1$.
4. Anàlogament es defineixen les altres components principals.

Definició 3.4. Si $T = (t_1, \ldots, t_p)$ és la matriu $p \times p$ que té per columnes els vectors que defineixen les diferents components principals, aleshores la transformació lineal $Y = XT$ s'anomena **transformació per components principals.**

Teorema 3.5. *Siguin $t_1, \ldots, t_p$ els VEPs normalitzats de la matriu de covariàncies S, és a dir,*

$$St_i = \lambda_i t_i, \;\; t_i^T t_i = 1 \;\; \forall i = 1, \ldots, p.$$

Aleshores se satisfà que:

1. *Les variables compostes $Y_i = X^T t_i, \;\; i = 1, \ldots, p$ són les components principals.*
2. *Les variàncies són els VAPs de S, és a dir, $Var(Y_i) = \lambda_i \;\; \forall i = 1, \ldots, p$.*
3. *Les components principals són variables incorrelacionades, és a dir, $Cov(Y_i, Y_j) = 0 \;\; \forall i, j = 1, \ldots, p$ tals que $i \neq j$.*

Demostració. Suposant que $\lambda_1 > \cdots \lambda_p > 0$, tenim que

$$\begin{cases} Cov(Y_i, Y_j) = t_i^T S t_j \overset{St_j = \lambda_j t_j}{=} t_i^T \lambda_j t_j = \lambda_j t_i^T t_j \\ Cov(Y_j, Y_i) = t_j^T S t_i \overset{St_i = \lambda_i t_i}{=} t_j^T \lambda_i t_i = \lambda_i t_j^T t_i \end{cases}$$

ara, utilitzant que $Cov(Y_i, Y_j) = Cov(Y_j, Y_i)$ i que, com $t_i^T t_j$ és un real, es compleix la igualtat $t_i^T t_j = (t_i^T t_j)^T = t_j^T t_i$, obtenim

$$\lambda_j t_i^T t_j = \lambda_i t_j^T t_i \Longleftrightarrow (\lambda_j - \lambda_i) t_i^T t_j = 0.$$

Ara, cal diferenciar dos casos:

- Si $i \neq j$: En aquest cas, com $\lambda_j \neq \lambda_i$, la igualtat anterior només es compleix si $t_i^T t_j = 0$ i, en conseqüència, $Cov(Y_i, Y_j) = Cov(Y_j, Y_i) = 0$, com volíem veure.
- Si $i = j$: En aquest cas, com $\lambda_i = \lambda_j$, la igualtat anterior sempre es compleix i utilitzant que $t_i^T t_j \overset{i=j}{=} t_i^T t_i = 1$, podem deduir que $Cov(Y_i, Y_i) = Var(Y_i) = \lambda_i$.

Per tant, ja tenim demostrats els punts 2 i 3 del teorema.

Sigui $Y = \sum_{i=1}^p \alpha_i Y_i$ una variable composta tal que $\sum_{i=1}^p \alpha_i = 1$,

$$Var(Y) = Var\left(\sum_{i=1}^p \alpha_i Y_i\right) = \sum_{i=1}^p \alpha_i^2 Var(Y_i) = \sum_{i=1}^p \alpha_i^2 \lambda_i \overset{\lambda_1 > \cdots \lambda_p > 0}{\leq} \lambda_1 \sum_{i=1}^p \alpha_i^2 = \lambda_1$$

és a dir, hem vist que $Var(Y) \leq Var(Y_1)$, fet que prova que Y_1 té variància màxima.

Sigui ara Z una variable composta incorrelacionada amb Y_1, tal que $\sum_{i=1}^p \beta_i = 1$. Notem que el fet d'estar incorrelacionada amb Y_1 implica que es pot escriure de la forma $Z = \sum_{i=1}^p \beta_i X_i \overset{Cov(Z, Y_1) = 0}{=} \sum_{i=2}^p \beta_i Y_i$, aleshores

$$Var(Z) = Var\left(\sum_{i=2}^p \beta_i Y_i\right) = \sum_{i=2}^p \beta_i^2 Var(Y_i) = \sum_{i=2}^p \beta_i^2 \lambda_i \overset{\lambda_2 > \cdots \lambda_p > 0}{\leq} \lambda_2 \sum_{i=2}^p \beta_i^2 = \lambda_2$$

és a dir, hem vist que $Var(Z) \leq Var(Y_2)$, fet que prova que Y_2 té variància màxima.

Anàlogament es demostra aquest resultat per $Y_3, \ldots, Y_n$. □

Definició 3.6. Definim el **percentatge de variabilitat** explicat per les primeres $m \leq p$ components principals com

$$P = 100 \cdot \frac{\lambda_1 + \cdots \lambda_m}{\lambda_1 + \cdots \lambda_p}$$

3.2 Representació d'una matriu de dades

Definició 3.7. Definim la **distància euclidiana** entre dues files de X com

$$\delta_{i,j} = \sqrt{\sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2}$$

i denotem per $\Delta = (\delta)_{ij}$ la matriu $n \times n$ de distàncies entre les files.

Ara podríem representar les n files de X com n punts a l'espai $\mathbb{R}^p$, però si p és gran és impossible de visualitzar i, per tant, necessitem reduir la dimensió.

Definició 3.8. Definim la **variabilitat geomètrica** de la matriu de distàncies Δ com la mitjana dels seus elements al quadrat, és a dir,

$$V_\delta(X) = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{i,j}^2.$$

Observació 3.9. Si $Y = XT$ és una transformació lineal de X, on T és una matriu $p \times q$ de constants, tindrem que la distància euclidiana entre files de Y és

$$\delta_{i,j}(q) = \sqrt{\sum_{k=1}^{q} (y_{i,k} - y_{j,k})^2}$$

i la variabilitat geomètrica en dimensió $q \leq p$ és

$$V_\delta(Y)_q = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{i,j}^2(q).$$

Teorema 3.10. *La variabilitat geomètrica de la distància euclidiana és la traça de la matriu de covariàncies,*

$$V_\delta(X) = tr(S) = \sum_{i=1}^{p} \lambda_i.$$

Demostració. Sigui $x_1, \ldots, x_n$ una mostra univariant amb variància σ^2 i sigui $\bar{x}$ la seva mitjana, se satisfà que

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x} + (\bar{x} - x_j))^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})(x_j - \bar{x})$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(x_j - \bar{x})^2 + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i x_j - x_j\bar{x} - x_i\bar{x} + \bar{x}^2)$$

$$= \sigma^2 + \sigma^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i\bar{x} - \bar{x}\bar{x} - x_i\bar{x} + \bar{x}^2) = \sigma^2 + \sigma^2 + 0 = 2\sigma^2$$

i per tant, tenim que

$$\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = \sigma^2.$$

Ara bé, tot això es compleix per una mostra univariant $x_1, \ldots, x_n$, per tant, aplicant aquest procediment a cada columna de X i tenint en compte que a la diagonal principal de la matriu de covariàncies s'hi troben les variàncies, obtenim que

$$V_\delta(X) = \sum_{i=1}^{p} s_{i,i} = tr(S).$$

□

Observació 3.11. Una bona representació en dimensió q petita serà aquella que maximitzi la variabilitat geomètrica, ja que d'aquesta manera els punts estaran tan separats com sigui possible.

Teorema 3.12. *La transformació lineal T que maximitza la variabilitat geomètrica en dimensió q és la transformació per components principals.*

Demostració. Sigui T una matriu qualsevol, la variabilitat geomètrica de $Y = XT$ és

$$V_\delta(Y)_q = \sum_{i=1}^{p}\sigma^2(Y_i),$$

on $\sigma^2(Y_i) = Var(Y_i)$, per tant, assolim el màxim de la variabilitat geomètrica quan es maximitzen les variàncies, és a dir, quan Y_i és una component principal. □

Observació 3.13. El percentatge de variabilitat geomètrica explicat per Y és

$$P_q = 100 \cdot \frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_p}$$

3.3 Nombre de components principals

En aquest apartat exposarem alguns criteris per determinar el nombre de components principals $m < p$.

- **Criteri del percentatge:** Prenem el nombre de components principals tal que el percentatge P_m sigui major a un valor concret (se sol utilitzar 80% o 90%).

- **Criteri de Kaiser:** Notem que obtenir les components principals a partir de la matriu de correlacions equival a suposar que les variables observables tinguin variància 1, per tant, una component principal amb variància inferior a 1 explica menys variabilitat que una variable observada.

  Un cop fet aquest apunt el criteri de Kaiser indica que s'han de prendre les m primeres components principals tals que tinguin variància major o igual a 1, és a dir, $\lambda_1 \geq \ldots \geq \lambda_m \geq 1$.

- **Test d'esfericitat:** Es tracta de contrastar la hipòtesi

  $$H_0^{(m)} : \lambda_1 > \ldots > \lambda_m > \lambda_{m+1} = \ldots = \lambda_p.$$

  Si acceptem H_0, no tindrà sentit considerar més de m components principals.

  Es pot comprovar que el test per decidir sobre $H_0^{(m)}$ està basat en l'estadístic $\mathcal{X}^2$ i s'aplica seqüencialment: Si acceptem $H_0^{(0)} : \lambda_1 = \ldots = \lambda_p$, no hi haurà direccions principals, però si el rebutgem, haurem de repetir el test amb $H_0^{(1)} : \lambda_1 > \lambda_2 = \ldots = \lambda_p$; si acceptem $H_0^{(1)}$, aleshores serà $m = 1$, però si el rebutgem, haurem de repetir el test amb $H_0^{(2)}$ i així successivament.

3.4 Exemple amb RStudio

En aquesta secció utilitzaré el programa *RStudio* que ens servirà per obtenir les components principals i representar els resultats obtinguts.

Primer de tot s'ha d'obrir el fitxer Excel (*Comarques_ Temperatura_ Pluja_ 2019*) amb les dades d'altitud, mitjana de temperatures màximes, mitjana de temperatures mínimes, precipitació anual i humitat relativa, (a la segona secció de l'annex es poden veure les dades de la taula i d'on les he extret), i a continuació, eliminem la columna "Comarques", ja que no vull que la tracti com a una variable.

Ara, aplicant la funció *prcomp*, aplicada a la nostra mostra estandarditzada (*center=TRUE* i *scale=TRUE*) obtenim les 5 components principals:

```
> #Primer de tot obrim l'arxiu i li posem el nom "Mostra_ACP
> Mostra_ACP <- read.xlsx("/Users/Pol/Desktop/Comarques_Temperatura_Pluja_2019.xlsx")
> #Ara eliminem la columna "Comarques" de l'arxiu Mostra_ACP
> Mostra_ACP$Comarques<-NULL
> #A continuació, apliquem la funció "prcomp" a l'arxiu Mostra_ACP estandaritzat
> ACP<-prcomp(Mostra_ACP, center=TRUE, scale=TRUE)
> #Aquesta funció ens permet obtenir les 5 components principals
> print(ACP)
Standard deviations (1, .., p=5):
[1] 1.7468457 1.0578425 0.6754136 0.5762847 0.2030069

Rotation (n x k) = (5 x 5):
                                PC1        PC2         PC3        PC4        PC5
Altitud                   0.5333249 -0.2599852  0.25200751  0.1074396 0.75691558
Mitjana.temperatures.màximes -0.5047565  0.1769587  0.22617909 -0.6859116 0.43849154
Mitjana.temperatures.mínimes -0.4949366 -0.1334876 -0.60539689  0.4149782 0.44554069
Precipitació.anual.(mm)       0.4587146  0.2060123 -0.71333239 -0.4851702 0.05391331
Humitat.relativa              0.0735126  0.9169696  0.09993364  0.3322424 0.18273156
```

Amb la funció *print(ACP)*, poem veure les desviacions estàndards de cada component així com, el pes de cada variable en cada component principal, per exemple, podem veure que la primera component principal està molt influïda per l'altitud, la mitjana de temperatures màximes, la mitjana de temperatures mínimes i la precipitació anual (magnituds de 0.45 a 0.53 en valor absolut), mentre que la variable humitat relativa pràcticament no afecta a la primera component principal, ja que té un coeficient de 0.07. Respecte a la segona component principal, veiem com la gran part del seu pes recau sobre la variable humitat relativa, ja que té un coeficient de 0.91, mentre que les altres 4 variables tenen molt menys pes. En concret, tenim

$$CP1 = 0.533 \cdot Alt - 0.505 \cdot T\_max - 0.495 \cdot T\_min + 0.459 \cdot Prec + 0.074 \cdot Hum$$

$$CP2 = -0.260 \cdot Alt + 0.177 \cdot T\_max - 0.133 \cdot T\_min + 0.206 \cdot Prec + 0.917 \cdot Hum$$

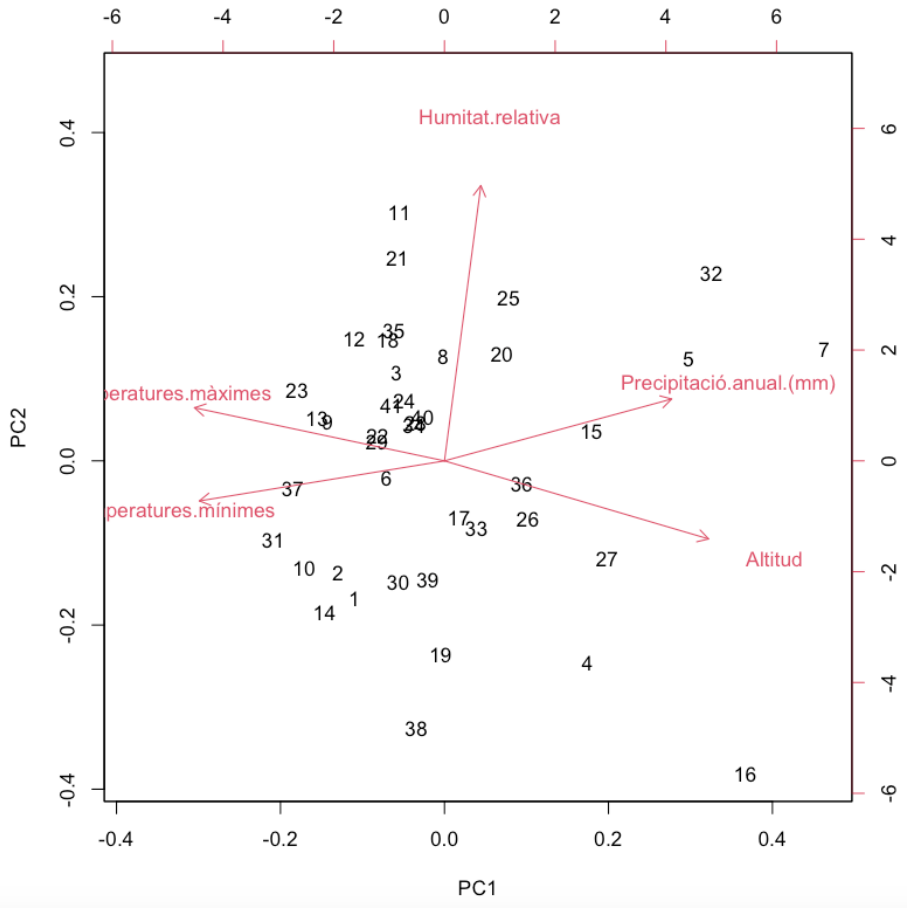
A més a més, la funció *print(ACP)* també ens proporciona la informació necessària per poder aplicar el criteri de Kaiser, que ens portaria a utilitzar les dues primeres components principals, ja que són les úniques que tenen una desviació estàndard major que 1.

Aplicant ara la funció *summary*, podem obtenir el percentatge acumulat per cada component:

```
> summary(ACP)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     1.7468 1.0578 0.67541 0.57628 0.20301
Proportion of Variance 0.6103 0.2238 0.09124 0.06642 0.00824
Cumulative Proportion  0.6103 0.8341 0.92534 0.99176 1.00000
>
```

A partir d'aquest output, podem aplicar el criteri del percentatge: Si fixem $P = 80\%$, aquest criteri ens portaria a escollir les dues primeres components principals, ja que tenen un percentatge acumulat del 83.41%, mentre que si fixem $P = 90\%$, hauríem de seleccionar les 3 primeres components principals, que tenen un percentatge acumulat del 92.53%.

Per tant, ara ja podem pensar a representar les dades en dues dimensions, on les variables siguin $PC1$ i $PC2$. Per representar-la, podem utilitzar la funció *biplot(ACP)* i obtenim la següent representació:



on cada número representa la comarca informada en la fila i-èssima. Notem primer que a l'eix de les abscisses hi apareix la component principal 1 i a l'eix de les ordenades, la component principal 2. Ja sabem que la CP1, es veu explicada positivament per les variables altitud i la precipitació anual i es veu explicada negativament per la mitjana de temperatures màximes i mínimes, mentre que la humitat relativa pràcticament no té influència, per tant, trobarem a la part dreta de la imatge les comarques que tinguin una gran altitud i molta precipitació, per exemple, destaquen les observacions 7, 16, 32 i 5, que corresponen a les comarques Aran (alt: 1002, prec:1063), Cerdanya (alt:1213, prec: 678), Ripollès (alt: 852, prec: 757) i l'Alta Ribagorça (alt:823, prec:843) mentre que la mitjana d'altituds és 357 i la mitjana de precipitacions és 514, per tant, realment té sentit que aquestes observacions estiguin a la part dreta de la representació.

Anàlogament a aquest raonament, com la component principal 2, ve explicada pràcticament en la seva totalitat per la humitat relativa, per tant, trobarem a la part superior del gràfic, les observacions amb major humitat relativa. Per exemple, a la part superior del gràfic podem destacar les observacions 11, 21, 32 i 25, que corresponen a les comarques Baix Empordà, Gironès, Ripollès i Osona, que de fet,

són les 4 observacions amb major humitat relativa, en concret, tenen els valors 75, 73, 76 i 74 respectivament mentre que la mitjana se situa en 64.52.

Notem que les fletxes indiquen cap a on es tendeix a representar cada observació si predomina aquella variable respecte a les altres.

Podem concloure, per tant, que la primera component principal s'explica, pràcticament en la seva totalitat com l'altitud, ja que, en general, a més alçada, menors són les temperatures màximes i mínimes i major és la precipitació anual (és a dir, aquestes 4 variables estan molt correlacionades), mentre que la humitat relativa té un comportament totalment diferent, ja tant les comarques de costa com d'interior, poden tenir una humitat elevada.

4 Anàlisi discriminant

4.1 Classificació en dues poblacions

Siguin Ω_1, Ω_2 dues poblacions i $X_1, \ldots, X_p$ variables observables. El nostre objectiu serà determinar a quina població pertany un individu w.

Notem que aquest problema és bastant freqüent, per exemple, en decidir si una persona és apte o no per un treball (en funció dels seus estudis, experiència, etc.) o per decidir si comprar o no una marca de cereals (en funció de les calories, el preu, etc.).

Definició 4.1. Una **regla discriminant** és un criteri que ens permet determinar a quina població pertany un individu w. Per això, caldrà definir una **funció discriminant** $D(x_1, \ldots, x_p)$ i utilitzarem la següent regla de classificació:

$$\begin{cases} Si \;\; D(x_1, \ldots, x_p) \geq 0 \Longrightarrow \; w \in \Omega_1 \\ Si \;\; D(x_1, \ldots, x_p) < 0 \Longrightarrow \; w \in \Omega_2 \end{cases}$$

Observació 4.2. Notem que aquesta classificació divideix $\mathbb{R}^n$ en dues regions

$$R_1 = \{x : \; D(x) \geq 0\} \;\; i \;\; R_2 = \{x : \; D(x) < 0\}.$$

Definició 4.3. Utilitzant la notació anterior, definim la **probabilitat de classificació errònia** (pce) com

$$pce = P(R_2|\Omega_1) \cdot P(\Omega_1) + P(R_1|\Omega_2) \cdot P(\Omega_2),$$

és a dir, la pce és la probabilitat d'assignar w a la població que no li pertoca.

- **Discriminador lineal de Fisher:** Siguin μ_1, μ_2 els vectors de les mitjanes de les variables $X_1, \ldots, X_p$ en Ω_1 i Ω_2 respectivament, i suposem que les dues poblacions tenen la mateixa matriu de covariàncies Σ.

  Definició 4.4. Donat un individu w amb observacions $x = (x_1, \ldots, x_p)^T$, definim la **distància de Mahalanobis** a la població Ω_i com

  $$M^2(x, \mu_i) = (x - \mu_i)^T \Sigma^{-1} (x - \mu_i), \;\; i = 1, 2.$$

  Ara, intuïtivament podem expressar aquesta regla com una funció discriminant, de la següent manera

  $$\begin{aligned} M^2(x, \mu_2) - M^2(x, \mu_1) &= (x - \mu_2)^T \Sigma^{-1} (x - \mu_2) - (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) \\ &= \cancel{x\Sigma^{-1}x} + \mu_2^T \Sigma^{-1} \mu_2 - 2x^T \Sigma^{-1} \mu_2 - \cancel{x\Sigma^{-1}x} - \mu_1^T \Sigma^{-1} \mu_1 + 2x^T \Sigma^{-1} \mu_1 \\ &= (\mu_2 + \mu_1)^T \Sigma^{-1} (\mu_2 - \mu_1) + 2x^T \Sigma^{-1} (\mu_1 - \mu_2). \end{aligned}$$

  Podem definir, per tant, la funció discriminant següent

  $$L(x) = \left[x - \frac{1}{2}(\mu_1 + \mu_2) \right] \Sigma^{-1} (\mu_1 - \mu_2).$$

Definició 4.5. La funció discriminant anterior rep el nom de **discriminador lineal de Fisher** i s'utilitzarà el següent criteri:

$$\begin{cases} Si\ \ L(x) \geq 0 \Longrightarrow\ w \in \Omega_1 \\ Si\ \ L(x) < 0 \Longrightarrow\ w \in \Omega_2 \end{cases}$$

- **Criteri de la màxima versemblança:** Aquest criteri consisteix a assignar w a la població on la versemblança de les observacions x és major, és a dir, si $f_1(x)$, $f_2(x)$ són les densitats de x en Ω_1 i Ω_2 respectivament, definim la funció discriminant

$$V(x) = log(f_1(x)) - log(f_2(x)),$$

i utilitzarem el criteri

$$\begin{cases} Si\ \ V(x) \geq 0 \Longrightarrow\ w \in \Omega_1 \\ Si\ \ V(x) < 0 \Longrightarrow\ w \in \Omega_2 \end{cases}$$

Notem que $V(x) \geq 0 \Longleftrightarrow log(f_1(x)) \geq log(f_2(x)) \Longleftrightarrow f_1(x) \geq f_2(x)$.

- **Criteri de Bayes:** Aquest criteri té com a supòsit que coneixem les probabilitats que w pertanyi a cadascuna de les poblacions, és a dir, coneixem

$$q_1 = P(\Omega_1),\ \ q_2 = P(\Omega_2),\ \ q_1 + q_2 = 1.$$

Aleshores, si tenim les observacions $x = (x_1, \ldots, x_p)$ per a l'individu w, utilitzant el teorema de Bayes de la probabilitat condicionada sabem que

$$P(\Omega_i|x) = \frac{q_i f_i(x)}{q_1 f_1(x) + q_2 f_2(x)},\ \ i = 1,\, 2.$$

Ara, definim el discriminador de Bayes com

$$B(x) = log(f_1(x)) - log(f_2(x)) + log(q_1/q_2).$$

Notem que si $q_1 = q_2 = 1/2$, tindrem que $log(q_1/q_2) = log(1) = 0 \Longrightarrow B(x) = V(x)$.

Teorema 4.6. *El criteri de Bayes minimitza la probabilitat de classificació errònia (pce).*

Demostració. Farem aquesta demostració per contradicció, per tant, suposarem que existeix un altre criteri que classifica de la següent manera:

$$\begin{cases} Si\ \ x \in R_1^* \Longrightarrow\ w \in \Omega_1 \\ Si\ \ x \in R_2^* \Longrightarrow\ w \in \Omega_2 \end{cases}$$

on R_1^* i R_2^* són regions complementàries de l'espai mostral.

Aleshores, la probabilitat de classificació errònia d'aquest criteri és

$$pce^* = q_1 \int_{R_2^*} f_1(x)dx + q_2 \int_{R_1^*} f_2(x)dx$$

$$= \int_{R_2^*} (q_1 f_1(x) - q_2 f_2(x))dx + q_2 \left(\int_{R_2^*} f_2(x)dx + \int_{R_1^*} f_2(x)dx \right)$$

$$= \int_{R_2^*} (q_1 f_1(x) - q_2 f_2(x))dx + q_2$$

on a la primera igualtat hem sumat i restat el terme $q_2 \int_{R_2^*} f_2(x)$ i a la segona igualtat hem utilitzat que $\int_{R_2^*} f_2(x)dx + \int_{R_1^*} f_2(x)dx = 1$, ja que R_1^* i R_2^* són complementaris.

Per tant, minimitzar pce^* equival a minimitzar la integral

$$\int_{R_2^*} (q_1 f_1(x) - q_2 f_2(x))dx$$

i aquesta integral és mínima si R_2^* inclou totes les x tal que $q_1 f_1(x) - q_2 f_2(x) < 0$ i exclou totes les x tals que $q_1 f_1(x) - q_2 f_2(x) > 0$, per tant, pce^* és mínima si $R_2^* = \{x : B(x) < 0\}$, com volíem veure. □

4.2 Classificació en poblacions normals

En aquesta secció, estudiarem diferents criteris de classificació suposant que les variables $X_1, \ldots, X_p$ és distribueixen seguint una llei Normal en cadascuna de les dues poblacions, és a dir, la distribució de $X_1, \ldots, X_p$ en Ω_1 és $N_p(\mu_1, \Sigma_1)$ i en Ω_2 és $N_p(\mu_2, \Sigma_2)$, per tant, tindrem

$$f_{i,X}(X(x)) = (2\pi)^{-p/2} |\Sigma_i^{-1}|^{1/2} exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\}, \;\; i = 1, 2.$$

- **Discriminador lineal:** Suposant que $\mu_1 \neq \mu_2$ i que $\Sigma_1 = \Sigma_2 = \Sigma$, aleshores

$$V(x) = log(f_1(x)) - log(f_2(x)) = \cancel{log\left[(2\pi)^{-p/2}|\Sigma^{-1}|^{1/2}\right]} - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)$$

$$\cancel{-log\left[(2\pi)^{-p/2}|\Sigma^{-1}|^{1/2}\right]} + \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)$$

$$= -\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) = L(x)$$

per tant, els discriminadors de màxima versemblança (V) i el lineal (L), basat en el criteri de la mínima distància, coincideixen.

Sigui α la distància de Mahalanobis entre les dues poblacions, és a dir,

$$\alpha = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2).$$

Si suposem que $x \in \Omega_2 \sim N_p(\mu_2, \Sigma)$, utilitzant que $(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) \sim \mathcal{X}_p^2$ i que l'esperança d'una distribució chi-quadrat amb p graus de llibertat és p, tindríem que:

$$E\left[(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right] = E\left[(x-\mu_2+\mu_2-\mu_1)\Sigma^{-1}(x-\mu_2+\mu_2-\mu_1)\right]$$
$$= E\left[(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)+\alpha+2(x-\mu_2)^T\Sigma^{-1}(\mu_2-\mu_1)\right]$$
$$= p+\alpha+2\cdot E\left[x-\mu_2\right]\cdot\Sigma^{-1}(\mu_2-\mu_1) = p+\alpha,$$

on a l'última igualtat he utilitzat que $E(x-\mu_2) = E(x)-\mu_2 = 0$, al ser $x \in \Omega_2$. Amb això, ja podem calcular l'esperança de L :

$$E(L(x)) = -\frac{1}{2}E\left[(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right] + \frac{1}{2}E\left[(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right]$$
$$= -\frac{1}{2}(p+\alpha)+\frac{1}{2}p = -\frac{1}{2}\alpha.$$

Observació 4.7. Notem que aquest mateix raonament també es pot fer pel cas en el qual suposem que $x \in \Omega_1$. Tots els passos són anàlegs i acabaríem obtenint $E(L(x)) = \frac{1}{2}\alpha$.

Un cop trobades les esperances es pot comprovar com $Var(L(x)) = \alpha$ i per tant, podem utilitzar el següent criteri:

$$\begin{cases} Si\ \ x \in N_p(\mu_1,\Sigma) \Longrightarrow L(x) \sim N(\frac{1}{2}\alpha) \\ Si\ \ x \in N_p(\mu_2,\Sigma) \Longrightarrow L(x) \sim N(-\frac{1}{2}\alpha) \end{cases}$$

- **Criteri de Bayes:** Suposant com abans que $\mu_1 \neq \mu_2$, $\Sigma_1 = \Sigma_2 = \Sigma$ i coneixem les probabilitats

$$q_1 = P(\Omega_1),\ \ q_2 = P(\Omega_2),\ \ q_1+q_2 = 1,$$

tenim que

$$B(x) = log(f_1(x)) - log(f_2(x)) + log(q_1/q_2) = L(x) + log(q_1/q_2),$$

és a dir, la funció discriminant de Bayes en el cas de poblacions normals és el discriminador lineal (L) més la constant $log(q_1/q_2)$.

- **Probabilitat de classificació errònia:** La probabilitat d'assignar x a Ω_2 quan prové de Ω_1, és a dir, quan $L(x) \sim N(\frac{1}{2}\alpha,\alpha)$, és

$$P(L(x)<0) \overset{Normalitzem}{=} P\Big((L(x)-\frac{1}{2}\alpha)/\sqrt{\alpha} < -\frac{1}{2}\frac{\alpha}{\sqrt{\alpha}}\Big) = P\Big(N(0,1) < -\frac{1}{2}\sqrt{\alpha}\Big),$$

per tant, $P(L(x)<0) = \Phi\Big(-\frac{1}{2}\sqrt{\alpha}\Big)$, on $\Phi(z)$ denota la funció de distribució de la Normal $N(0,1)$.

Vist això, la probabilitat de classificació errònia és

$$pce = q_1P(L(x)<0|\Omega_1) + q_2P(L(x)>0|\Omega_2) = \Phi\Big(-\frac{1}{2}\sqrt{\alpha}\Big),$$

per tant, la *pce* és una funció decreixent de la distància de Mahalanobis α entre les dues poblacions. Aquest resultat és intuïtiu, ja que a major distància entre les dues poblacions, més diferents seran entre elles, fet que disminueix la *pce*.

- **Discriminador quadràtic:** Suposem ara que $\mu_1 \neq \mu_2$ i $\Sigma_1 \neq \Sigma_2$, aleshores, el criteri de la màxima versemblança ens proporciona el discriminador

$$Q(x) = log(f_1(x)) - log(f_2(x)) = \cancel{log[(2\pi)^{-p/2}]} + log|\Sigma_1^{-1}|^{1/2}$$

$$-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) \cancel{-log[(2\pi)^{-p/2}]} - log|\Sigma_2^{-1}|^{1/2} + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)$$

$$= \frac{1}{2}x^T(\Sigma_2^{-1} - \Sigma_1^{-1})x + x^T(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) + \frac{1}{2}\mu_2^T\Sigma_2^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1$$

$$- \frac{1}{2}log|\Sigma_1| + \frac{1}{2}log|\Sigma_2|$$

  $Q(x)$ s'anomena discriminador quadràtic i, anàlogament, podríem obtenir el discriminador quadràtic de Bayes $BQ(x) = Q(x) + log(q_1/q_2)$.

Notem, però que en les aplicacions pràctiques no coneixem μ_1, μ_2, Σ_1, Σ_2 i per tant, s'hauran d'estimar a partir de mostres de mida n_1 i n_2 de les dues poblacions substituint μ_1, μ_2 pels vectors de mitjanes $\bar{x}_1$, $\bar{x}_2$, Σ_1, Σ_2 per les matrius de covariàncies S_1, S_2 i Σ per S on

$$S = (n_1S_1 + n_2S_2)/(n_1 + n_2).$$

D'altra banda no coneixem la distribució de $\hat{L}(x)$ però podem utilitzar que asimptòticament es comporta com una distribució Normal.

Amb tot això, definint

$$\alpha = (\bar{x}_1 - \bar{x}_2)^T S^{-1}(\bar{x}_1 - \bar{x}_2)$$

podem utilitzar el següent criteri:

$$\begin{cases} Si \;\; x \in N_p(\mu_1, \Sigma) \Longrightarrow \hat{L}(x) \sim N(\frac{1}{2}\alpha, \alpha) \\ Si \;\; x \in N_p(\mu_2, \Sigma) \Longrightarrow \hat{L}(x) \sim N(-\frac{1}{2}\alpha, \alpha) \end{cases}$$

4.3 Classificació en k poblacions

Suposarem en aquesta secció que l'individu w pot pertànyer a k poblacions $\Omega_1, \ldots, \Omega_k$ on $k \geq 3$. Com en els casos anteriors, el nostre objectiu serà, donades unes observacions $x = (x_1, \ldots, x_p)^T$ de l'individu w, assignar w a una de les k poblacions.

- **Discriminadors lineals:** Suposem que la mitjana de les variables a cada població Ω_i és μ_i i que les k poblacions tenen la mateixa matriu de covariàncies, és a dir, $\Sigma_1 = \cdots = \Sigma_k = \Sigma$. Considerem ara la distància de Mahalanobis de w a cada població Ω_i

$$M^2(x, \mu_i) = (x - \mu_i)^T\Sigma^{-1}(x - \mu_i), \;\; i = 1, \ldots, k$$

aleshores, assignarem w a la població més propera, és a dir, caldrà calcular les k distàncies de Mahalanobis i veure quina és la més petita,

$$si\ \ M^2(x,\mu_i) = min\{M^2(x,\mu_1),\ldots,M^2(x,\mu_2)\} \implies w \in \Omega_i.$$

Si volem expressar aquesta assignació com una funció discriminant lineal podem definir

$$\begin{aligned}2L_{ij}(x) = M^2(x,\mu_i) - M^2(x,\mu_j) &= (x-\mu_i)^T\Sigma^{-1}(x-\mu_i) - (x-\mu_j)^T\Sigma^{-1}(x-\mu_j)\\ &= (x-\mu_i-x+\mu_j)^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_i + \mu_i^T\Sigma^{-1}\mu_i + x^T\Sigma^{-1}\mu_j - \mu_j^T\Sigma^{-1}\mu_j\\ &= (\mu_j-\mu_i)\Sigma^{-1}x - \mu_i^T\Sigma^{-1}x + \mu_j^T\Sigma^{-1}x + (\mu_i-\mu_j)^T\Sigma^{-1}(\mu_i+\mu_j)\\ &= 2(\mu_j-\mu_i)\Sigma^{-1}x + (\mu_i-\mu_j)^T\Sigma^{-1}(\mu_i+\mu_j)\end{aligned}$$

és a dir, la funció discriminant lineal és,

$$L_{ij}(x) = (\mu_i-\mu_j)^T\Sigma^{-1}x + \frac{1}{2}(\mu_i-\mu_j)^T\Sigma^{-1}(\mu_i+\mu_j),$$

i utilitzarem el següent criteri:

$$si\ \ L_{ij}(x) > 0\ \ \forall j \neq i \implies w \in \Omega_i$$

per tant, caldrà calcular $k-1$ funcions discriminants.

- **Criteri de la màxima versemblança**: Sigui $f_i(x)$ la funció de densitat de x en la població Ω_i, assignarem w a la població on la versemblança sigui major, és a dir,
$$si\ \ f_i(x) = max\{f_1(x),\ldots,f_k(x)\} \implies w \in \Omega_i.$$
Si volem expressar aquesta assignació com una funció discriminant definim
$$V_{ij}(x) = log(f_i(x)) - log(f_j(x))$$
i utilitzarem el criteri
$$si\ \ V_{ij}(x) > 0\ \ \forall j \neq i \implies w \in \Omega_i.$$

- **Criteri de Bayes**: Si a més de les funcions de densitats coneixem les probabilitats
$$q_1 = P(\Omega_1),\ldots,q_k = P(\Omega_k),\ \ q_1+\cdots+q_k = 1$$
el criteri de Bayes és el següent
$$si\ \ q_if_i(x) = max\{q_1f_1(x),\ldots,q_kf_k(x)\} \implies w \in \Omega_i.$$
Aquest criteri està associat a les funcions discriminants
$$B_{ij}(x) = log(f_i(x)) - log(f_j(x)) + log(q_i/q_j).$$

4.4 Exemple amb RStudio

En aquesta secció utilitzaré el programa *RStudio* que servirà per calcular les matrius de covariàncies de les diferents poblacions.

Per fer aquest exemple he creat un fitxer Excel (*Exemple_ AD.xlsx*) que conté les dades (a l'annex es pot veure en detall d'on les obtinc), per cadascuna de les 42 comarques de Catalunya, dels següents vectors aleatoris:

- X és un vector aleatori que conté dues variables: la primera recull el percentatge de vot dels partits independentistes que van obtenir representació parlamentària a les eleccions autonòmiques de Catalunya, celebrades el passat 14 de febrer de 2021 (és a dir: ERC, JxCat i la CUP), mentre que la segona variable recull el percentatge de vot dels partits no independentistes que van obtenir representació parlamentària (és a dir: PSC, VOX, ECP, PP i Cs)

$$X^T = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \text{percentatge vots ERC+JxCat+CUP} \\ \text{percentatge vots PSC+VOX+ECP+PP+Cs} \end{pmatrix}$$

  A partir d'aquest vector aleatori es poden definir dues poblacions: Classificarem una comarca en la població A si els partits independentistes tenen major percentatge de vot que els no independentistes, és a dir, si $X_1 > X_2$. En cas contrari assignarem la comarca a la població B.

  Podem veure a partir de la taula de l'annex que hi ha 7 comarques que pertanyen a la població B (Aran, Baix Llobregat, Baix Penedès, Barcelonès, Garraf, Tarragonès i Vallès Occidental) mentre que les 35 comarques restants pertanyen a la població A.

- Y pretén analitzar si la població de cada comarca és més o menys catalanoparlant, així, he definit el següent vector aleatori que recull el quocient entre el nombre de persones que tenen el mateix nom on, al numerador surt la versió catalana del nom i al denominador, la versió castellana del nom, per tant, a major quocient s'entén que la població d'aquella comarca té més tendència a parlar català, en canvi, si el quocient és més proper a 0, predominarà el castellà

$$Y^T = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} \text{quocient Josep/José} \\ \text{quocient Joan/Juan} \\ \text{quocient Anna/Ana} \\ \text{quocient Carme/Carmen} \end{pmatrix}$$

Durant tot l'apartat, suposarem que les dues poblacions A i B es distribueixen com una llei Normal, és a dir, $A \sim N(\mu_A, \Sigma_A)$ i $B \sim N(\mu_B, \Sigma_B)$. D'altra banda, com no coneixem els valors de μ_A, μ_B, Σ_A, Σ_B ho substituirem pels vectors de mitjanes $\bar{x}_A$, $\bar{x}_B$ i per les matrius de covariàncies S_A, S_B, a més, si n_A, n_B són les mides de les dues poblacions, substituirem Σ per $S = (n_a \cdot S_A + n_B \cdot S_B)/(n_1 + n_B)$.

Primer de tot cal obrir el fitxer *Exemple_ AD.xlsx*. Per fer-nos una idea de les relacions entre les 4 variables Y_1, Y_2, Y_3, Y_4 i les dues poblacions A i B podem graficar-ho utilitzant les funcions *ggplot* i *ggarrange*.

```
> # Obrim l'arxiu amb les dades i l'anomenem Mostra_AD
> Mostra_AD <- read.xlsx("/Users/Pol/Desktop/Exemple_AD.xlsx")
> p1 <- ggplot(data = Mostra_AD, aes(x = Y1, fill = Població))+geom_histogram(position = "identity", alpha = 0.5)
> p2 <- ggplot(data = Mostra_AD, aes(x = Y2, fill = Població))+geom_histogram(position = "identity", alpha = 0.5)
> p3 <- ggplot(data = Mostra_AD, aes(x = Y3, fill = Població))+geom_histogram(position = "identity", alpha = 0.5)
> p4 <- ggplot(data = Mostra_AD, aes(x = Y4, fill = Població))+geom_histogram(position = "identity", alpha = 0.5)
> ggarrange(p1, p2, p3, p4, nrow = 4, common.legend = TRUE)
```

A partir del gràfic anterior veiem que les comarques que pertanyen a la població B (el percentatge de vots dels partits no independentistes és superior al percentatge de vots dels partits independentistes), les variables Y_1, Y_2, Y_3, Y_4 prenen valors molt petits, és a dir, predominen els noms castellans respecte als catalans mentre que les comarques que pertanyen a la població A, les variables Y_1, Y_2, Y_3, Y_4 prenen valors més grans.

Un cop feta aquesta puntualització podem trobar diferents estimadors:

- Discriminador lineal de Fisher: Com hem vist anteriorment el discriminador lineal és

$$L(x) = \left[x - \frac{1}{2}(\mu_A + \mu_B)\right] \Sigma^{-1}(\mu_A^T - \mu_B^T)$$

  i s'utilitza el següent criteri:

$$\begin{cases} Si \;\; L(x) \geq 0 \Longrightarrow w \in A \\ Si \;\; L(x) < 0 \Longrightarrow w \in B \end{cases}$$

  Aquest discriminador, però, suposa que $\Sigma_A = \Sigma_B = \Sigma$, és a dir, que les dues poblacions tenen la mateixa matriu de covariàncies. Per tant, definim la matriu de covariàncies com $\Sigma = (n_a \cdot S_A + n_B \cdot S_B)/(n_1 + n_B)$:

```
> #Definim A com les observacions que pertanyen a la població A (files 8 a 42 del fitxer)
> A <- Mostra_AD[(8:42), (3:6)]
> #Definim B com les observacions que pertanyen a la població B (files 1 a 7 del fitxer)
> B <- Mostra_AD[(1:7),(3:6)]
> # Ara, la inversa de la matriu de covariàncies és:
> S <- solve((35*cov(A)+7*cov(B))/(35+7))
> S
           Y1         Y2         Y3         Y4
Y1  39.661495 -41.911497   4.730631 -17.937797
Y2 -41.911497  70.213712 -22.848343   7.635516
Y3   4.730631 -22.848343  18.499081  -6.132067
Y4 -17.937797   7.635516  -6.132067  53.174687
```

és a dir,

$$\Sigma^{-1} = \begin{pmatrix} 39.66 & -41.91 & 4.73 & -17.94 \\ -41.91 & 70.21 & -22.85 & 7.64 \\ 4.73 & -22.85 & 18.50 & -6.13 \\ -17.94 & 7.64 & -6.13 & 53.17 \end{pmatrix}$$

Ara, només queda calcular les mitjanes $\bar{x}_A$, $\bar{x}_B$ a partir de les dades del fitxer *Exemple_AD*:

$$\bar{x}_A = (1.106,\ 1.216,\ 1.231,\ 0.326),\ \ \bar{x}_B = (0.241,\ 0.478,\ 0.425,\ 0.077)$$

Amb tot això, donada una observació x, com ja coneixem μ_A, μ_B i Σ^{-1}, ja podrem decidir si pertany a la població A o a la B.

Per exemple, si tinguéssim una quaranta-tresena comarca amb

$$x = (0.5,\ 0.8,\ 0.7,\ 0.2),$$

tindríem $L(x) = -0.5035 < 0 \implies x \in B$.

- Criteri de Bayes: El criteri de Bayes, igual que el discriminador lineal de Fisher suposa que $\Sigma_A = \Sigma_B$ i a més, es coneixen les probabilitats $q_A = P(A),\ \ q_B = P(B)$. En el nostre cas tenim

$$q_A = \frac{35}{42} = \frac{5}{6} = 0.83 \quad \text{i} \quad q_B = \frac{7}{42} = \frac{1}{6} = 0.17$$

Ara, com que hem suposat que A i B es distribueixen com una llei Normal, podem utilitzar la següent fórmula per obtenir el discriminador de Bayes:

$$B(x) = L(x) + log(q_A/q_B).$$

Ara, com coneixem els valors de q_A, q_B podem trobar el logaritme

$$\frac{q_A}{q_B} = \frac{5/6}{1/6} = 5 \implies log(q_A/q_B) = log(5) = 1.609$$

Per tant, si volguéssim classificar, segons el criteri de Bayes l'observació

$$x = (0.5,\ 0.8,\ 0.7,\ 0.2),$$

tindríem que

$$B(x) = -0.5035 + 1.609 = 1.105 > 0$$

i per tant, arribaríem a la conclusió que $x \in A$.

- Discriminador quadràtic: Aquest discriminador, a diferència dels anteriors, suposa que $\Sigma_A \neq \Sigma_B$ i es pot calcular a partir de la següent fórmula:

$$Q(x) = \frac{1}{2}x^T(\Sigma_B^{-1} - \Sigma_A^{-1})x + x^T(\Sigma_A^{-1}\mu_1 - \Sigma_B^{-1}\mu_2) + \frac{1}{2}\mu_B^T\Sigma_B^{-1}\mu_B - \frac{1}{2}\mu_A^T\Sigma_A^{-1}\mu_A -$$

$$- \frac{1}{2}log|\Sigma_A| + \frac{1}{2}log|\Sigma_B|$$

És a dir, necessitem calcular Σ_A^{-1}, Σ_B^{-1}, $|\Sigma_A|$ i $|\Sigma_B|$:

```
> # Primer calculem les inverses de les matrius de covariàncies:
> solve(cov(A))
           Y1        Y2         Y3         Y4
Y1  33.647267 -35.58452   4.067574 -15.302752
Y2 -35.584522  59.33784 -19.262933   6.741710
Y3   4.067574 -19.26293  15.529658  -5.210047
Y4 -15.302752   6.74171  -5.210047  44.622360
> solve(cov(B))
           Y1        Y2         Y3         Y4
Y1  429.65394 -441.0686  -44.20525  168.60852
Y2 -441.06859 1556.3587 -500.95595 -929.43661
Y3  -44.20525 -500.9560  500.11155   20.71278
Y4  168.60852 -929.4366   20.71278 3027.02543
> # Ara calculem els determinants de les matrius de covariàncies:
> det(cov(A))
[1] 2.238946e-05
> det(cov(B))
[1] 4.830131e-12
```

$$\Sigma_A^{-1} = \begin{pmatrix} 33.65 & -35.58 & 4.07 & -15.30 \\ -35.58 & 59.34 & -19.26 & 6.74 \\ 4.07 & -19.26 & 15.53 & -5.21 \\ -15.30 & 6.74 & -5.21 & 44.62 \end{pmatrix}, \quad |\Sigma_A| = 2.24 \cdot 10^{-5}$$

$$\Sigma_B^{-1} = \begin{pmatrix} 429.65 & -441.07 & -44.21 & 168.61 \\ -441.07 & 1556.36 & -500.96 & -929.44 \\ -44.21 & -500.96 & 500.11 & 20.71 \\ 168.61 & -929.44 & 20.71 & 3027.03 \end{pmatrix}, \quad |\Sigma_B| = 4.83 \cdot 10^{-12}$$

Per tant, si volguéssim classificar, segons el criteri del discriminador quadràtic l'observació

$$x = (0.5,\ 0.8,\ 0.7,\ 0.2),$$

tindríem que

$$Q(x) = 21.01 - \frac{1}{2}log(2.25 \cdot 10^{-5}) + \frac{1}{2}log(4.83 \cdot 10^{-12}) = 13.41 > 0$$

i per tant, arribaríem a la conclusió que $x \in A$.

- Discriminador quadràtic de Bayes: Finalment podríem calcular el discriminador quadràtic de Bayes que es pot calcular a partir de l'expressió

$$BQ(x) = Q(x) + log(q_1/q_2)$$

Per tant, si volguéssim classificar, segons el criteri del discriminador quadràtic de Bayes l'observació

$$x = (0.5,\ 0.8,\ 0.7,\ 0.2),$$

tindríem que

$$BQ(x) = 13.41 + 1.61 = 15.02 > 0$$

i per tant, arribaríem a la conclusió que $x \in A$.

Resumint, per l'observació $x = (0.5,\ 0.8,\ 0.7,\ 0.2)$ hem calculat 4 discriminadors diferents obtenint els següents resultats:

$$\begin{cases} \text{Discriminador lineal de Fisher: } x \in B \\ \text{Discriminador de Bayes: } x \in A \\ \text{Discriminador quadràtic: } x \in A \\ \text{Discriminador quadràtic de Bayes: } x \in A \end{cases}$$

Notem que tres discriminadors (Bayes, quadràtic i quadràtic de Bayes) conclouen que $x \in A$, mentre que el discriminador lineal de Fisher conclou que $x \in B$. Com a conclusió podríem afirmar que $x \in A$, ja que el discriminador quadràtic és més precís que el discriminador lineal de Fisher.

5 Conclusions

Podem concloure que s'han assolit els dos objectius principals d'aquest treball:

D'una banda, hem pogut estudiar en detall quatre dels principals mètodes d'anàlisi multivariant: el model de regressió lineal múltiple (MRLM), que ens ha permès relacionar una variable aleatòria amb un vector aleatori, l'anàlisi de correlació canònica (ACC), que és una generalització del MRLM, ja que serveix per relacionar dos vectors aleatoris, l'anàlisi de components principals (ACP), que té com a objectiu reduir la dimensió d'un vector aleatori per poder-lo representar gràficament i l'anàlisi discriminant (AD), que serveix per classificar un individu entre diverses poblacions.

D'altra banda, he trobat dades reals amb les quals he pogut construir exemples interessants d'aplicació d'aquests mètodes, això m'ha servit per entendre quina és la finalitat de cada mètode més enllà de la seva base teòrica.

6 Annex

6.1 Mostra per l'exemple del model de regressió lineal múltiple

Per construir aquesta base de dades, he extret la informació a partir de les dades que té publicades l'IDESCAT (Institut d'Estadística de Catalunya), que és l'òrgan estadístic de la Generalitat de Catalunya. Entre altres objectius, l'IDESCAT produeix dades estadístiques oficials econòmiques, demogràfiques i socials i a més, fa el seguiment d'altres activitats estadístiques i en difon els resultats oficials.

- **Preu de venda mitjà d'habitatge d'obra nova per metre quadrat**: En el següent enllaç `https://www.idescat.cat/pub/?id=aec&n=728` es pot veure com l'IDESCAT informa el preu de venda mitjà d'habitatge d'obra nova de cada comarca per anys. A la meva base de dades he copiat les xifres relatives a l'any 2018 excepte per les comarques "Conca de Barberà", "Pallars Sobirà" i "Priorat", que he agafat les dades del 2019, ja que pel 2018 no n'hi havia i per les comarques "Alta Ribagorça" i "Pallars Jussà" he buscat una font alternativa, ja que l'IDESCAT no ha informat el preu de l'habitatge en els últims 5 anys, així que he buscat com a font alternativa les dades facilitades per la *Secretaria d'Habitat Urbà i Territori* relatives al gener-juny del 2018, com es pot veure en el següent enllaç:

  `http://habitatge.gencat.cat/web/.content/home/dades/estadistiques/01_Estadistiques_de_construccio_i_mercat_immobiliari/02_Compravenda_i_preu_de_venda/02_Compravendes_d_habitatges_registrades_i_el_preu_de_venda/Estadistica_PDF/Compravendes_gen_jun18.pdf`.

- **Densitat poblacional**: En el següent enllaç `https://www.idescat.cat/pub/?id=aec&n=249&t=2018` es pot veure com l'IDESCAT informa la densitat poblacional per comarca (hab/km^2). He agafat les dades relatives al 2018.

- **Distància entre la capital de comarca i Barcelona**: Aquesta distància no la recull l'IDESCAT, sinó que he estat jo, qui mitjançant *Google Maps* ha calculat la distància en cotxe entre la capital de cada comarca i la Plaça Catalunya de Barcelona.

- **PIB per habitant**: En el següent enllaç `https://www.idescat.cat/pub/?id=aec&n=358` es pot veure com l'IDESCAT informa el PIB per habitant (en milers d'euros) de cada comarca. Les últimes dades disponibles són les de 2018, per això en els altres camps també he agafat les dades relatives al 2018.

- **Renda familiar per habitant**: En el següent enllaç `https://www.idescat.cat/pub/?id=aec&n=941&t=2018` es pot veure com l'IDESCAT informa la renda familiar per habitant (en milers d'euros) de cada comarca al 2018.

- **Mitjana d'edat**: En el següent enllaç `https://www.idescat.cat/pub/?id=pmh&n=1180&t=201800&geo=com:01` el nombre de persones de cada edat residents a la comarca "Alt Camp", m'he descarregat les dades i he calculat la

mitjana amb un Excel. He seguit el mateix procediment per les 42 comarques de Catalunya.

Recollint totes les dades descrites anteriorment, he construït la següent taula d'Excel:

Comarques	PREU	DENS	DIST	PIB	RENDA	EDAT
Alt Camp	821,89	81,9	104	33,2	15,1	42,16
Alt Empordà	1.812,31	103,6	146	24,7	13,9	41,60
Alt Penedès	1.258,99	182,4	60,2	29,6	16	41,14
Alt Urgell	948,73	14	175	24,1	13,5	44,40
Alta Ribagorça	1177,8	8,8	237	33,5	15,3	44,56
Anoia	1.125,34	137,5	69,6	22,2	15,2	41,45
Aran	2.370,82	15,8	276	38,8	15	41,18
Bages	1.122,76	161,6	57,7	28,3	16,3	42,86
Baix Camp	1.331,84	271	109	23,1	14,9	41,08
Baix Ebre	790,57	77,7	181	22,9	13,3	43,76
Baix Empordà	2.184,86	190,2	130	23,2	14,3	42,39
Baix Llobregat	2.571,63	1.685,60	11,9	33,1	17,9	41,25
Baix Penedès	1.377,75	345,5	69	19,1	13,8	41,99
Barcelonès	3.365,24	15.469,20	0	39,5	19,4	43,50
Berguedà	1.358,65	33	102	25,1	16,6	46,15
Cerdanya	3.071,42	32,7	158	25,7	14,5	42,58
Conca de Barberà	1057,16	30,9	118	32,6	15,4	44,27
Garraf	3.043,06	805,5	48,5	19,7	17,3	41,90
Garrigues	605,47	23,5	153	22,8	13,5	46,45
Garrotxa	1.149,42	77,2	112	30,2	16,6	43,35
Gironès	1.666,07	331,5	103	33,2	16,4	39,34
Maresme	2.298,01	1.123,70	31	22,1	17,2	41,95
Moianès	1.218,76	39,9	60,6	23,1	16,2	42,75
Montsià	681,47	91,2	174	18,6	12,2	43,57
Noguera	875,92	21,7	148	24,5	13,4	44,01
Osona	1.493,49	127,2	72	29,6	17	41,55
Pallars Jussà	711,9	9,8	186	23,2	14,3	46,64
Pallars Sobirà	945,08	5	235	27,9	14,6	44,52
Pla de l'Estany	604,34	121,8	121	27,8	16,2	41,00
Pla d'Urgell	1.295,29	120,4	137	27,2	14	42,07
Priorat	666,67	18,6	136	18,5	14	46,65
Ribera d'Ebre	919,3	26,6	154	49,4	14,6	45,82
Ripollès	809,72	26,1	107	25	17,2	46,62
Segarra	334,88	31,6	101	34,4	14	41,28
Segrià	957,6	149,4	162	29,2	15,2	41,84
Selva	1.526,96	169,9	87,6	27,1	14,1	41,55
Solsonès	980,42	13,4	107	26,6	14,1	43,20
Tarragonès	1.489,78	791,5	98,8	34,8	14,9	40,73
Terra Alta	647,24	15,5	175	24	13,6	47,99
Urgell	637,11	62,5	113	26,7	14,1	42,58
Vallès Occidental	2.258,54	1.574,10	25,7	31	17,6	40,56
Vallès Oriental	1.605,10	552,9	29,4	31,7	16,9	40,87

6.2 Mostra per l'exemple d'anàlisi de correlació canònica

La mostra d'aquest exemple conté dades de les següents fonts:

- **Percentatge de vot de cada partit a les eleccions autonòmiques de Catalunya del passat 14 de febrer:** Aquestes dades s'han extret a partir dels resultats oficials publicats al diari 'Nació Digital'.
  En el següent enllaç es poden consultar les dades relatives a la comarca de l'Alt Penedès `https://www.naciodigital.cat/eleccions/parlament2021/comarca/03?elecAnt=parlament2017`.

- **Quocient Josep/José i Anna/Ana**: Per trobar el quocient del nombre de persones que es diuen Josep i el nombre de persones que es diuen José he extret

les dades de l'IDESCAT (Institut d'Estadística de Catalunya). En el següent enllaç `https://www.idescat.cat/noms/?res=d03&sexe=0` es poden veure els noms de la població de la comarca de l'Alt Penedès així que, per construir la base de dades, he sumat al numerador el nombre de persones que es diuen Josep o Josep Maria i al denominador he col·locat la suma de persones que es diuen José o José Maria. S'ha fet el mateix procediment en el cas del quocient de les persones que es diuen Anna i les que es diuen Ana.

	X					Y	
Comarques	Vots ERC	Vots PSC	Vots JxCat	Vots VOX	Vots CUP	Josep/José	Anna/Ana
Alt Camp	25,50	14,36	28,62	5,64	10,57	0,90	0,81
Alt Empordà	21,71	15,31	27,28	9,56	7,34	0,73	0,84
Alt Penedès	25,19	16,87	25,23	4,75	9,70	0,73	0,91
Alt Urgell	26,95	15,46	28,32	3,43	8,51	0,94	0,93
Alta Ribagorça	22,83	20,44	20,98	3,23	9,26	1,14	0,71
Anoia	22,38	20,47	21,08	6,97	7,28	0,56	0,67
Aran	14,43	29,89	10,54	14,03	3,61	0,10	0,36
Bages	24,75	16,50	27,76	5,59	8,44	0,67	0,84
Baix Camp	22,30	18,81	21,08	9,59	7,16	0,49	0,70
Baix Ebre	31,43	14,77	20,63	5,32	6,48	0,43	0,59
Baix Empordà	22,64	16,89	30,75	5,86	7,72	0,80	0,93
Baix Llobregat	19,54	31,41	10,62	9,20	4,16	0,17	0,32
Baix Penedès	21,26	25,14	14,88	11,12	4,93	0,29	0,40
Barcelonès	18,33	26,21	15,47	7,70	6,20	0,27	0,48
Berguedà	25,70	9,75	36,09	2,35	11,52	0,99	1,49
Cerdanya	20,82	12,42	30,64	5,19	8,87	0,87	1,17
Conca de Barberà	31,05	9,72	31,09	2,89	9,86	2,39	1,95
Garraf	21,45	24,70	17,18	8,21	6,42	0,27	0,47
Garrigues	32,16	9,10	33,28	2,72	8,39	1,60	1,74
Garrotxa	23,16	10,08	39,88	3,05	10,53	1,10	1,49
Gironès	19,24	14,79	33,54	5,62	11,08	0,92	1,37
Maresme	22,58	19,58	22,91	7,78	6,49	0,42	0,61
Moianès	29,51	9,38	31,27	2,92	11,59	1,13	1,12
Montsià	33,55	15,50	18,65	6,21	7,00	0,45	0,68
Noguera	31,23	11,25	29,50	3,79	7,22	0,99	1,18
Osona	22,54	9,50	41,44	3,21	10,32	1,57	1,92
Pallars Jussà	23,32	16,06	33,68	2,78	9,15	0,88	1,10
Pallars Sobirà	27,86	9,31	29,92	1,31	14,02	2,17	1,33
Pla d'Urgell	28,79	10,75	32,48	3,72	6,12	2,99	3,20
Pla de l'Estany	24,79	7,27	40,76	2,00	12,01	0,80	1,05
Priorat	32,61	9,66	28,06	2,10	14,35	1,31	0,97
Ribera d'Ebre	33,03	12,78	26,25	3,24	8,80	0,65	0,80
Ripollès	22,86	11,53	36,88	3,53	8,93	0,77	1,50
Segarra	26,90	11,07	31,97	4,06	8,78	1,49	1,27
Segrià	24,08	17,89	23,03	7,50	6,05	0,42	0,60
Selva	21,14	17,96	29,04	6,81	6,93	0,73	0,92
Solsonès	24,34	9,40	36,22	2,72	11,41	2,97	4,00
Tarragonès	19,12	24,14	14,03	12,73	5,10	0,31	0,41
Terra Alta	29,91	14,79	22,57	4,39	7,71	0,56	0,72
Urgell	28,12	10,26	34,72	3,56	8,10	2,78	2,40
Vallès Occidental	19,99	25,50	16,45	8,72	5,65	0,28	0,53
Vallès Oriental	22,04	23,66	19,96	7,18	6,46	0,38	0,56

Ara bé, el que les dades que s'han fet servir per fer l'exemple són el logaritme dels percentatges de vots i el logaritme dels quocients, per tant, les dades que s'han importat a R per fer l'exemple són:

	X					Y	
Comarques	X1=log(Vots ERC)	X2=log(Vots PSC)	X3=log(Vots JxCat)	X4=log(Vots VOX)	X5=log(Vots CUP)	Y1=log(Josep/José)	Y2=log(Anna/Ana)
Alt Camp	1,41	1,16	1,46	0,75	1,02	-0,05	-0,09
Alt Empordà	1,34	1,18	1,44	0,98	0,87	-0,14	-0,07
Alt Penedès	1,40	1,23	1,40	0,68	0,99	-0,14	-0,04
Alt Urgell	1,43	1,19	1,45	0,54	0,93	-0,03	-0,03
Alta Ribagorça	1,36	1,31	1,32	0,51	0,97	0,06	-0,15
Anoia	1,35	1,31	1,32	0,84	0,86	-0,25	-0,17
Aran	1,16	1,48	1,02	1,15	0,56	-0,98	-0,44
Bages	1,39	1,22	1,44	0,75	0,93	-0,17	-0,08
Baix Camp	1,35	1,27	1,32	0,98	0,85	-0,31	-0,16
Baix Ebre	1,50	1,17	1,31	0,73	0,81	-0,37	-0,23
Baix Empordà	1,35	1,23	1,49	0,77	0,89	-0,10	-0,03
Baix Llobregat	1,29	1,50	1,03	0,96	0,62	-0,78	-0,50
Baix Penedès	1,33	1,40	1,17	1,05	0,69	-0,54	-0,40
Barcelonès	1,26	1,42	1,19	0,89	0,79	-0,57	-0,32
Berguedà	1,41	0,99	1,56	0,37	1,06	0,00	0,17
Cerdanya	1,32	1,09	1,49	0,72	0,95	-0,06	0,07
Conca de Barberà	1,49	0,99	1,49	0,46	0,99	0,38	0,29
Garraf	1,33	1,39	1,24	0,91	0,81	-0,57	-0,33
Garrigues	1,51	0,96	1,52	0,43	0,92	0,20	0,24
Garrotxa	1,36	1,00	1,60	0,48	1,02	0,04	0,17
Gironès	1,28	1,17	1,53	0,75	1,04	-0,04	0,14
Maresme	1,35	1,29	1,36	0,89	0,81	-0,38	-0,21
Moianès	1,47	0,97	1,50	0,47	1,06	0,05	0,05
Montsià	1,53	1,19	1,27	0,79	0,85	-0,35	-0,17
Noguera	1,49	1,05	1,47	0,58	0,86	0,00	0,07
Osona	1,35	0,98	1,62	0,51	1,01	0,19	0,28
Pallars Jussà	1,37	1,21	1,53	0,44	0,96	-0,06	0,04
Pallars Sobirà	1,44	0,97	1,48	0,12	1,15	0,34	0,12
Pla d'Urgell	1,46	1,03	1,51	0,57	0,79	0,48	0,51
Pla de l'Estany	1,39	0,86	1,61	0,30	1,08	-0,10	0,02
Priorat	1,51	0,98	1,45	0,32	1,16	0,12	-0,01
Ribera d'Ebre	1,52	1,11	1,42	0,51	0,94	-0,18	-0,09
Ripollès	1,36	1,06	1,57	0,55	0,95	-0,11	0,18
Segarra	1,43	1,04	1,50	0,61	0,94	0,17	0,10
Segrià	1,38	1,25	1,36	0,88	0,78	-0,37	-0,22
Selva	1,33	1,25	1,46	0,83	0,84	-0,14	-0,04
Solsonès	1,39	0,97	1,56	0,43	1,06	0,47	0,60
Tarragonès	1,28	1,38	1,15	1,10	0,71	-0,51	-0,39
Terra Alta	1,48	1,17	1,35	0,64	0,89	-0,25	-0,14
Urgell	1,45	1,01	1,54	0,55	0,91	0,44	0,38
Vallès Occidental	1,30	1,41	1,22	0,94	0,75	-0,55	-0,27
Vallès Oriental	1,34	1,37	1,30	0,86	0,81	-0,42	-0,25

6.3 Mostra per l'exemple d'anàlisi de components principals

Per construir aquesta base de dades, he extret la informació a partir de les dades que té publicades l'IDESCAT (Institut d'Estadística de Catalunya).

- **Altitud, mitjana de temperatures màximes, mitjana de temperatures mínimes, precipitació anual i humitat relativa**: En el següent enllaç `https://www.idescat.cat/pub/?id=aec&n=214` es pot veure com l'IDESCAT informa l'altitud d'una certa estació meteorològica de cada comarca juntament amb la mitjana de temperatures màximes i la mitjana de temperatures mínimes que registra aquesta estació. A sota, hi trobem una segona taula on trobem les dades relatives a la precipitació anual (mm) que recull cada estació així com la mitjana d'humitat relativa al llarg de l'any.

  Recollint totes les dades relatives al 2019 excloent la comarca del Moianès, ja que l'IDESCAT no informa cap valor dels que ens interessen, he construït la següent taula d'Excel:

Comarques	Altitud	Mitjana temperatures màximes	Mitjana temperatures mínimes	Precipitació anual (mm)	Humitat relativa
Alt Camp	287	21,8	10,6	342,1	64
Alt Empordà	24	22,1	11,6	522,9	63
Alt Penedès	240	21,9	9,5	465,5	71
Alt Urgell	849	20,2	5,3	592,2	62
Alta Ribagorça	823	20	2,8	842,9	71
Anoia	316	22,5	10,4	523,9	67
Aran	1.002	16,5	4,7	1.062,90	73
Bages	349	22,3	6,7	490,3	71
Baix Camp	29	22,1	11,5	426,7	69
Baix Ebre	179	22,2	12,3	317,1	65
Baix Empordà	29	22,3	8,8	579,8	75
Baix Llobregat	8	22,3	11,2	555,8	71
Baix Penedès	59	22,7	10,5	368,8	69
Barcelonès	33	21,6	15,2	626,6	62
Berguedà	873	19,8	7	597	71
Cerdanya	1.213	17,3	3,6	677,9	60
Conca de Barberà	446	20,7	8,3	443,2	67
Garraf	148	21,7	9,8	466,6	72
Garrigues	505	20,5	8,6	343,3	63
Garrotxa	433	21,6	6,4	613,6	71
Gironès	72	23	7,2	540	73
Maresme	81	20,8	12,1	513,9	69
Montsià	3	21,8	12,9	340,8	71
Noguera	238	21,9	7,6	389,8	70
Osona	509	21,2	6,3	528,2	74
Pallars Jussà	508	21,3	6,5	699,4	65
Pallars Sobirà	679	20,3	5,2	775,7	64
Pla de l'Estany	247	21,8	7,7	437,7	69
Pla d'Urgell	176	22,2	10,2	508	68
Priorat	359	21,8	10	465,9	64
Ribera d'Ebre	53	24,1	9,8	320,8	64
Ripollès	852	18,1	4,1	756,9	76
Segarra	554	19,8	8,4	356,2	68
Segrià	286	21,2	7,7	331,4	70
Selva	150	22,9	7,7	510,6	71
Solsonès	659	21	6,4	522,1	68
Tarragonès	5	22,2	12,4	359,9	67
Terra Alta	515	21,2	9,6	388,3	60
Urgell	427	20,9	9,2	401,6	65
Vallès Occidental	258	21,7	9,3	553	69
Vallès Oriental	176	22,3	9	506,5	69

6.4 Mostra per l'exemple d'anàlisi discriminant

La mostra d'aquest exemple conté dades de les següents fonts:

- **Població:** A partir dels resultats oficials publicats al diari 'Nació Digital', he assignat com a població A les comarques en les quals el percentatge de vot obtingut pels partits independentistes (ERC, JxCat i CUP) és superior al percentatge de vot obtingut pels partits no independentistes (PSC, VOX,

ECP, PP i Cs). En el següent enllaç es poden consultar les dades relatives a la comarca de l'Alt Penedès: `https://www.naciodigital.cat/eleccions/parlament2021/comarca/03?elecAnt=parlament2017`.

- **Quocient Josep/José, Joan/Juan, Anna/Ana i Carme/Carmen**: Anàleg a l'apartat 5.2.

Comarques	Població	Y1= Josep/ José	Y2= Joan/ Juan	Y3= Anna/ Ana	Y4= Carme/ Carmen
Aran	B	0,10	0,43	0,36	0,08
Baix Llobregat	B	0,17	0,39	0,32	0,04
Baix Penedès	B	0,29	0,50	0,40	0,07
Barcelonès	B	0,27	0,54	0,48	0,07
Garraf	B	0,27	0,45	0,47	0,07
Tarragonès	B	0,31	0,50	0,41	0,11
Vallès Occidental	B	0,28	0,54	0,53	0,10
Alt Camp	A	0,90	1,01	0,81	0,30
Alt Empordà	A	0,73	0,82	0,84	0,16
Alt Penedès	A	0,73	1,08	0,91	0,17
Alt Urgell	A	0,94	1,04	0,93	0,19
Alta Ribagorça	A	1,14	1,07	0,71	0,25
Anoia	A	0,56	0,68	0,67	0,12
Bages	A	0,67	0,85	0,84	0,24
Baix Camp	A	0,49	0,66	0,70	0,13
Baix Ebre	A	0,43	0,54	0,59	0,14
Baix Empordà	A	0,80	0,91	0,93	0,17
Berguedà	A	0,99	1,06	1,49	0,39
Cerdanya	A	0,87	1,20	1,17	0,19
Conca de Barberà	A	2,39	2,36	1,95	0,39
Garrigues	A	1,60	1,46	1,74	0,70
Garrotxa	A	1,10	1,24	1,49	0,53
Gironès	A	0,92	1,11	1,37	0,25
Maresme	A	0,42	0,75	0,61	0,10
Moianès	A	1,13	1,32	1,12	0,30
Montsià	A	0,45	0,65	0,68	0,06
Noguera	A	0,99	0,96	1,18	0,27
Osona	A	1,57	1,69	1,92	0,70
Pallars Jussà	A	0,88	0,95	1,10	0,10
Pallars Sobirà	A	2,17	1,63	1,33	0,41
Pla d'Urgell	A	2,99	3,28	3,20	0,98
Pla de l'Estany	A	0,80	0,96	1,05	0,31
Priorat	A	1,31	1,14	0,97	0,32
Ribera d'Ebre	A	0,65	0,87	0,80	0,14
Ripollès	A	0,77	1,06	1,50	0,37
Segarra	A	1,49	1,51	1,27	0,40
Segrià	A	0,42	0,61	0,60	0,11
Selva	A	0,73	0,91	0,92	0,22
Solsonès	A	2,97	3,04	4,00	0,66
Terra Alta	A	0,56	1,07	0,72	0,17
Urgell	A	2,78	2,43	2,40	1,40
Vallès Oriental	A	0,38	0,63	0,56	0,09

Referències

[1] James H. Stack; Mark M. Watson: Introducción a la econometría, *Pearson*, https://danielmorochoruiz.files.wordpress.com/2018/05/0000017.pdf

[2] Mª Victoria Esteban González; MM. Paz Moral Zuazo: Econometría básica aplicada con Gretl, *Universidad del País Vasco*, https://addi.ehu.es/bitstream/handle/10810/12496/08-09est.pdf?sequence=1&isAllowed=y

[3] Carles M. Cuadras: Nuevos métodos de análisis multivariante, http://www.est.uc3m.es/esp/nueva_docencia/getafe/estadistica/analisis_multivariante/doc_generica/archivos/metodos.pdf

[4] Arthur S. Goldberg: Econometric Theory, *Wiley publications*, 1964.

[5] Juan Gabriel Gomila Salas. (30 Abril, 2018). 21 - Análisis de componentes principales en RStudio [Arxiu de video]. https://www.youtube.com/watch?v=6BeuHCo1gZQ

[6] Joaquín Amat Rodrigo (Septiembre, 2016). Análisis discriminante lineal (LDA) y análisis discriminante cuadrático (QDA). RPubs. https://rpubs.com/Joaquin_AR/233932

[7] Jairo Ayala (17 de Febrero, 2020). Mineria de datos: Análisis de Correlación Canónica. RPubs. https://rpubs.com/JairoAyala/575486

[8] José R. Berrendero y Antonio Cuevas. An´alisis discriminante: Prácticas con R. Universidad Autónoma de Madrid. http://www.eio.uva.es/~valentin/ad3d/anadat/np/discrim/craneosTibet/craneosTibet_practicas-r.pdf