TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi de la Redundància

Autor: Sergi Marqués Paniagua

Director: Francesc de Asis Carmona Pontaque

Departament: Microbiologia, Genètica i Estadística

Convocatòria: Juny 2021

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

AGRAÏMENTS

En concret al meu tutor Francesc Carmona per orientar-me i guiar-me en els moments més decisius del treball. La seva ajuda i la rapidesa en resoldre els diferents dubtes que han sorgit ha estat d'agrair.

Finalment, als meus pares i a la meva àvia per tots els consells i el recolzament mostrat en aquesta aventura.

RESUM

En aquest treball es desenvolupa una tècnica de visualització i modelització de dades ecològiques, en concret, s'estudia una tècnica d'anàlisi multivariant anomenada anàlisi de la redundància des d'un punt de vista teòric, observant la metodologia que segueix i, posteriorment s'aplica a unes dades ambientals. A més, per a dur a terme aquest estudi, es fa l'anàlisi usant models lineals múltiples tradicionals analitzant totes les característiques que presenten contra el model multivariant lineal múltiple (*RDA*). Estudiant així, si amb un únic conjunt de variables explicatives es poden predir diferents objectius, en concret, els nivells de CO_2 i NO_x.

Paraules clau: Regressió Múltiple, Regressió Multivariant, Anàlisi canònic, Anàlisi de la Redundància, Biplots, Triplots, R.

Classificació AMS: *62Hxx Multivariate analysis, 62J05 Linear regression, 62J07 Ridge regression; shrinkage estimators, 62J20 Diagnostics, 62H25 Factor analysis and principal components; correspondence analysis*

ABSTRACT

In this work, a visualization and modeling technique of ecological data is developed, specifically, a multivariate analysis technique called redundancy analysis is studied from a theoretical point of view, observing the methodology that follows and, subsequently, it is applied to some data environmental. In addition, to carry out this study, the analysis is done using traditional multiple linear models analyzing all the characteristics that they present against the multiple linear multivariate model (*RDA*). Thus, studying whether a single set of explanatory variables can predict different objectives, specifically, the levels of CO_2 and NO_x.

Keywords: Multiple Regression, Multivariate Regression, Canonical Analysis, Redundancy Analysis, Biplots, Triplots, R.

Clasification AMS: *62Hxx Multivariate analysis, 62J05 Linear regression, 62J07 Ridge regression; shrinkage estimators, 62J20 Diagnostics, 62H25 Factor analysis and principal components; correspondence analysis*

Índex

LLISTAT D'IL·LUSTRACIONS

LLISTAT DE TAULES

I Introducció

Avui dia es viu en una societat on el gran volum de dades que es generen ja siguin poblacionals, biològiques, etc, són molt importants per tal de conèixer possibles factors que poden alterar, millorar o perjudicar qualsevol objecte d'estudi. En concret, el món del *Big Data* ha generat molta expectació donat que hi ha tècniques que s'estudiaven només des d'un punt de vista teòric, i ara s'han pogut implementar, fer previsions i interpretacions de manera molt més acurada.

El treball consisteix en estudiar a fons la regressió tradicional ajustant cada variable resposta de forma independent i, veure si per cada model es compleixen les hipòtesis que avalen aquests mètodes de predicció. I alhora comprovar si és imprescindible usar totes les explicatives com a regressores o bé a partir de diferents mètodes analitzar quin seria el conjunt òptim a utilitzar.

En concret, s'estudia una tècnica d'anàlisi multivariant no vista durant el grau que permet estudiar dos conjunts de dades en el mateix moment. Aquest mètode és més eficient que no pas altres donat que a partir d'un sol model, el multivariant múltiple, ja és possible analitzar dos o més variables objectiu. Per tant, fent una combinació entre la regressió multivariant i l'anàlisi de components principals es pot dur a terme l'Anàlisi de la Redundància *(RDA),* mètode que permet estudiar dos conjunts de dades i veure en un mateix gràfic la correlació entre les diferents variables i per quina d'elles es troben millor explicades les diferents observacions o llocs.

Aquest és un mètode utilitzat principalment en investigacions en l'àmbit de l'Ecologia i la Biologia, però recentment s'ha començat a investigar com pot donar resposta a diferents qüestions socioeconòmiques com ara l'estudi que es va fer a Polònia l'any 2017 per analitzar les relacions entre alguns factors socioeconòmics i la intensitat del delicte contra la propietat en aquell país (Lodz University, 2017).

Aquest treball es basa principalment en reduir la dimensionalitat, és a dir, el nombre de variables explicatives, derivant o obtenint un altre conjunt de variables que contenen gran part de la informació de les dades originals. Es detallen els passos realitzats i es compararan amb els obtinguts amb el programari usat. A més de la representació gràfica en espais bidimensionals en lloc de multidimensionals.

Com a metodologia emprada, s'ha recopilat diverses fonts bibliogràfiques i webgrafies pel desenvolupament dels continguts teòrics i pràctics. Respecte al tractament de les dades, s'ha utilitzat el programari R i diferents paquets per desenvolupar els gràfics i els conceptes que es mostren.

II Anàlisi Canònic

Una ordenació simple analitza una matriu de dades i en revela la major estructura en un gràfic construït a partir d'un conjunt reduït d'eixos ortogonals. Per tant, hi ha una forma d'anàlisi passiva, és a dir, l'usuari interpreta els resultats de l'ordenació a posteriori.

L'ordenació canònica, al contrari, associa dos o més conjunts de dades en el propi procés d'ordenació. En conseqüència, es vol extreure estructures d'un conjunt de dades relacionades amb d'altres bases, i comprovar formalment hipòtesis estadístiques sobre la importància d'aquestes relacions.

Els mètodes d'ordenació canònica es poden classificar en dos grups: simètrics i asimètrics.

Els diferents mètodes d'anàlisi canònic permeten realitzar comparacions directes (també anomenat *Anàlisi del gradient directe)* entre els mateixos objectes, en aquest cas, el mateix territori. Tot i que, també hi ha altres mètodes com ara, *l'Anàlisi del gradient Indirecte*, el qual la matriu de predictores X no intervé en l'ordenació de Y.

El directe explora explícitament les relacions entre les dues matrius: una matriu de resposta i una matriu explicativa fent ús de mètodes asimètrics o bé amb dues matrius amb rols simètrics. Totes dues són usades en l'ordenació.

Els mètodes simètrics són l'Anàlisi Discriminant Lineal *(LDA)* que busca una combinació de variables quantitatives per explicar una agrupació predefinida dels objectes. L'Anàlisi de Correlació Canònica *(CCorA)*, l'Anàlisi de Co-Inèrcia *(CoIA)* i l'Anàlisi de factors Múltiples *(MFA),* calculen els vectors propis descrivint l'estructura comuna de diversos conjunts de dades.

Els dos mètodes simètrics que s'utilitzen principalment en ecologia actualment són, l'Anàlisi de Correspondència Canònica *(CCA)* i l'Anàlisi de la Redundància *(RDA).* Tots dos són una extensió de la regressió lineal múltiple fent ús de l'Anàlisi de les Components principals *(PCA)* o l'Anàlisi de Correspondències *(CA).*

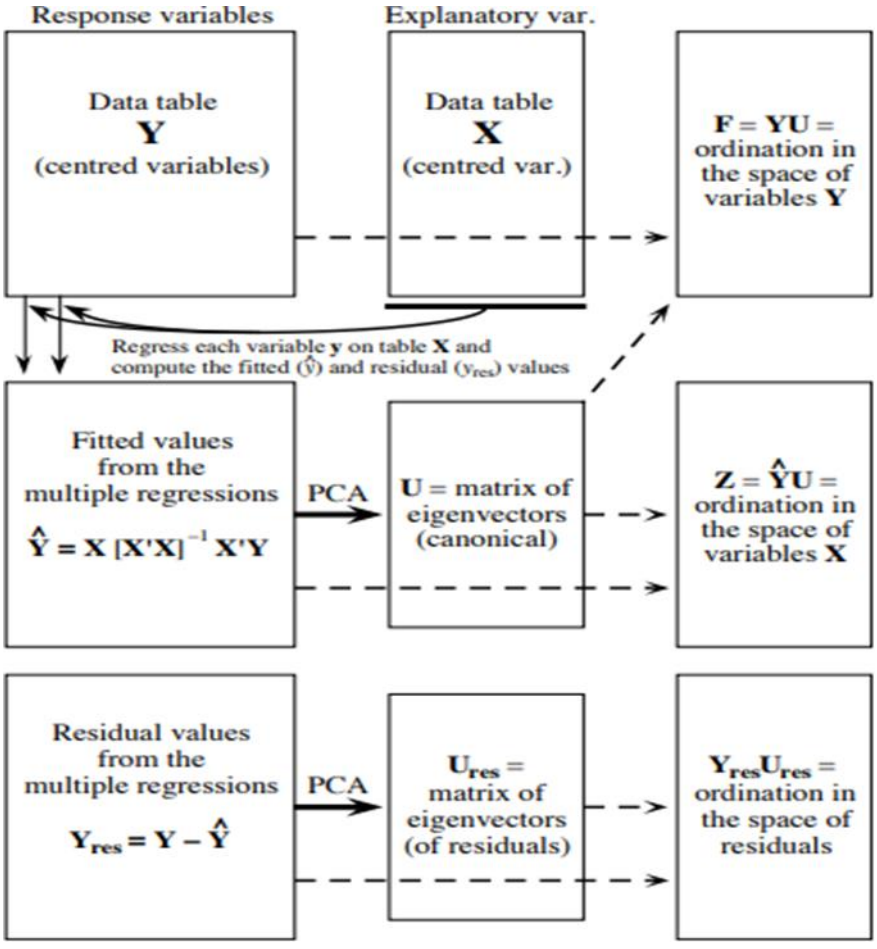
III Anàlisi de la Redundància

L'anàlisi de la redundància és un mètode d'anàlisi canònic asimètric que conté una matriu Y de variables resposta quantitatives i, una matriu de variables explicatives X. Un dels objectius d'aquest anàlisi pot ser l'estudi entre la matriu Y que descriu els nivells de $CO\ i\ NOx$, i una matriu X que conté informació ambiental, observat a la mateixa localització, en aquest cas, Turquia.

El *RDA* és un mètode que combina la regressió lineal múltiple amb l'anàlisi de components principals *(PCA).* És més, és una extensió directa de la modelització estadística/matemàtica per poder modelar matrius de resposta multivariants.

En *RDA*, els eixos d'ordenació s'obtenen mitjançant el *PCA* d'una matriu, calculada ajustant les variables Y a X fent ús de la regressió multivariant. Per tant, *RDA* conserva la distància euclidiana entre objectes.

Aquest tècnica funciona de la manera següent:



Il·lustració III-1: Passos a seguir per a realitzar l'Anàlisi de la Redundància (Legendre, 2012)

- Sigui $Y\ (n\ x\ s)$ una matriu de variables resposta i $X\ (n\ x\ m)$ una matriu de variables explicatives ambdues centrades.

- Es calcula una regressió per cada variable y ajustada per totes les variables de x, i es calculen els vectors ajustats $\hat{y}_i$ i els residuals $\hat{y}_{res}$. Finalment s'adjunten tots aquests valors en una matriu $\hat{Y}$ de valors ajustats.

- Seguidament es calcula el *PCA* de la matriu $\hat{Y}$. Aquest anàlisi produeix un vector de valors propis canònics i una matriu U de vectors propis canònics.

- S'obtenen dues ordenacions, una ($F = YU$) a l'espai de les variables resposta i l'altre ($Z = \hat{Y}U$) a l'espai de les variables explicatives.

- Finalment també es pot calcular un *PCA* dels valors residuals.

3.1 Estadístics del RDA

Un cop calculada la matriu de valors ajustats seguidament es pot passar a calcular els diferents estadístics informatius.

3.1.1 L'estadístic de la Redundància

A partir de les matrius Y i $\hat{Y}$ es pot calcular el R^2 canònic o també conegut com l'estadístic de la redundància bivariant, el qual mesura la relació entre $Y\ i\ X$:

$$R^2_{Y|X} = \frac{RSS}{TSS}$$

On TSS és la suma de quadrats total i RSS és la suma de quadrats residuals. Aquest estadístic aporta la mateixa informació que el R^2 de la regressió lineal múltiple, és a dir, informa sobre la proporció de variància de Y explicada per les diferents variables de X.

Aquesta mesura és equivalent a calcular el coeficient de correlació múltiple al quadrat entre el conjunt predictor total i cadascuna de les variables objectiu, i després fer la mitjana ponderada dels diferents coeficients de determinació al quadrat per obtenir un R^2 conjunt.

També cal esmentar que si no existeix cap relació entre aquestes dues matrius, el valor d'aquest estadístic no és zero, sinó $\frac{m}{n-1}$. Això es deu a que l'esperança de l'R^2 subjecte a un model amb una o m explicatives és $\frac{1}{n-1}$ o bé $\frac{m}{n-1}$, respectivament. Per tant en el cas, on $m = n - 1$ aquest presentarà un R^2 de 1 (si no hi ha cap relació entre les dues matrius) (Peres-Neto, 2006).

3.1.2 L'estadístic Ajustat de la Redundància

Aquest R^2 ajustat es calcula com:

$$R_a^2 = 1 - (1 - R_{Y|X}^2)\frac{(n-1)}{(n-m-1)}$$

On m és el nombre de variables explicatives de X, o de forma més compacte, és el rang de la matriu de variàncies-covariàncies de X.

3.1.3 L'estadístic F conjunt

L'estadístic F de la prova global de significació és:

$$F = \frac{\frac{R_{Y_{stand}|X}^2}{mp}}{\frac{(1 - R_{Y_{stand}|X}^2)}{(n-m-1)p}}$$

La hipòtesi nul·la que es vol contrastar és si la relació, mesurada pel R^2 canònic, no és més gran que el valor que s'obtindria per a matrius Y i X no relacionades de les mateixes mides.

Quan la matriu Y està estandarditzada i la distribució d'errors és normal, es pot provar la significació de l'estadístic mitjançant la distribució *F de Fisher* amb graus de llibertat $\eta_1 = mp$ i $\eta_2 = (n - m - 1)p$. On p és el nombre de variables resposta de Y. A més, com que s'estimen m paràmetres per a cadascuna de les p regressions múltiples utilitzades per calcular els vectors de valors ajustats, finalment es prediuen un total de mp paràmetres. Per això hi ha η_1 graus de llibertat al numerador de F.

I com que els graus de llibertat residuals de cada regressió lineal múltiple són $(n - m - 1)$ i es realitzen p regressions, el denominador presenta η_2 graus de llibertat.

3.1.4 Criteri d'Informació d'Akaike (AIC)

És una mesura que permet numerar la qualitat relativa d'un model estadístic, a més, també es fa servir per a la selecció de models.

$$AIC = n \ln\left(\frac{\hat{\sigma}}{n}\right) + 2k = n \ln\left(\frac{1 - R^2}{n}\right) + n\, ln(TSS) + 2k$$

On:

- R^2 és l'estadístic de la redundància.
- K és el nombre de paràmetres en el model.

- n és el nombre d'observacions.

- $\hat{\sigma} = RSS$ és la suma de quadrats residuals.

- TSS és la suma de quadrats residuals totals.

3.2 Eixos Canònics

A partir dels passos explicats anteriorment, és a dir, de l'obtenció dels valors ajustats a partir de la regressió lineal múltiple multivariant i seguit del *PCA* fet a la matriu de valors ajustats, la naturalesa asimètrica del *RDA* prové del fet que la regressió multivariant és un anàlisi asimètric, igual que la seva versió univariant, on y és el vector de resposta i X és la matriu explicativa.

La matriu que conté els vectors propis canònics normalitzats u_k s'anomena U. Aquests vectors donen les aportacions dels descriptors als diversos eixos canònics. U, de mida $(p\ x\ p)$, conté només $min\{p, m, n\ -\ 1\}$ vectors propis amb valors propis diferents de zero, ja que el nombre d'aquests vectors no poden superar aquest mínim.

	Canonical eigenvalues and eigenvectors	Non-canonical eigenvalues and eigenvectors
RDA	$\min[p, m, n-1]$	$\min[p, n-1]$

Il·lustració III-2: Nombre màxim de valors propis diferents de zero i vectors propis que es poden obtenir a partir de l'anàlisi canònic de la matriu Y i una matriu X mitjançant el RDA (Legendre, 2012)

Per tant el nombre d'eixos canònics que s'obtindran sempre serà més petit o igual al min {p, m, (n − 1)}. Dit d'una altra manera, no pot excedir:

- P que és la dimensió de l'espai de referència de la matriu Y.

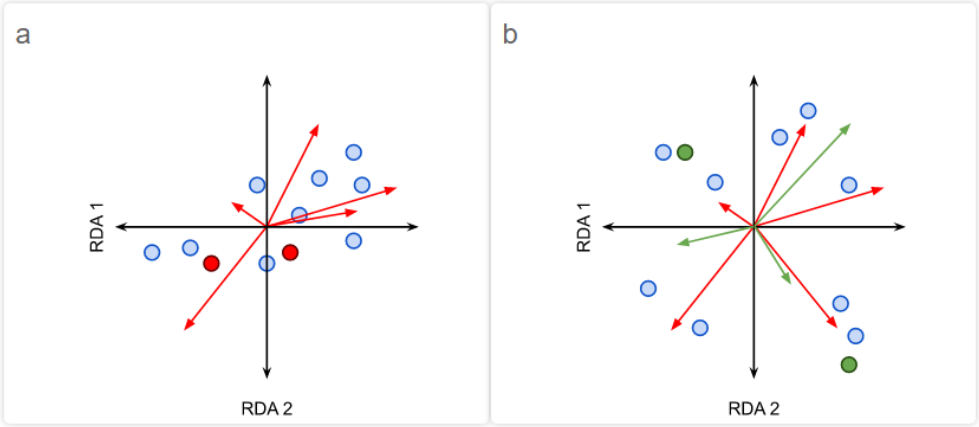
- m que és el nombre de variables de la matriu X que alhora també coincideix amb el rang d'aquesta.

- $(n-1)$ que es el nombre màxim de dimensions necessàries per representar n punts en l'espai euclidià.

3.3 Lectura de Biplots i Triplots

Per visualitzar les ordenacions d'aquesta tècnica es poden dibuixar *biplots* o *triplots* depenent si es tenen dos o tres conjunts: les puntuacions del lloc/observacions (matrius F o Z, de la il·lustració III-1), les variables resposta de Y i les variables explicatives de X. Cada parell de conjunts de punts pot ser dibuixat en un biplot, el qual ajuda a la interpretació de l'ordenació en termes de Y o X.

Per a fer una correcte interpretació cal triar anteriorment quina escala usar. En general, si les variables explicatives són dicotòmiques o nominals s'utilitza l'escalat tipus I, altrament tipus II.



Il·lustració III-3: Representació esquemàtica d'un Biplot (a) i un Triplot (b) (GUSTA ME)

El *biplot (a):* Ordena objectes com a punts i variables resposta o explicatives com a vectors (fletxes vermelles). Els nivells de variables nominals es representen en punts (vermells).

El triplot (b): Ordena els objectes com a punts (blaus) mentre que les variables explicatives i resposta es representen com a vectors (fletxes vermelles i verdes, respectivament).

3.4 Tipus d'Escala

Com s'ha esmentat anteriorment hi ha dos tipus d'escala.

3.4.1 Escala tipus I

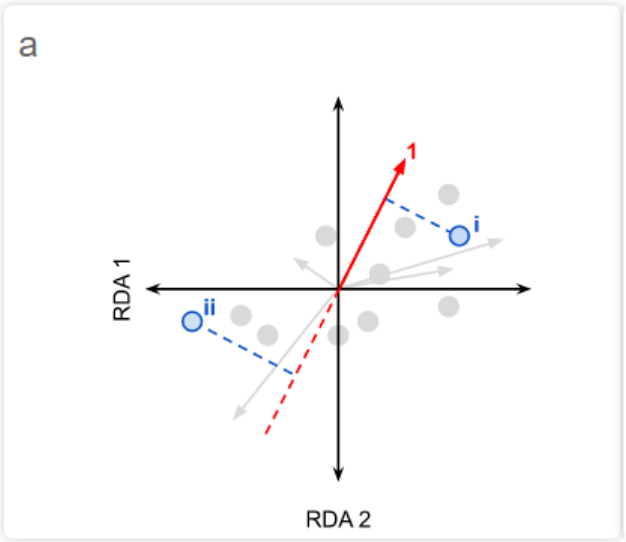
Aquesta té una gran utilitat per variables nominals o binàries:

- Les distàncies entre punts d'objectes preserva la distància euclidiana. Per tant, es pot esperar que els objectes ordenats més propers tinguin valors similars.

- Els angles entre vectors que representen variables resposta no tenen sentit.

- Els angles entre vectors que representen variables resposta i explicatives reflecteixen la seva correlació (lineal).

Des d'un punt de vista gràfic



Il·lustració III-4: Projecció d'un punt ordenat sobre un vector (GUSTA ME)

La projecció d'un punt ordenat sobre un vector variable, tal com es mostra per al punt *i* al panell *a*, la inspecció visual suggereix que tingui valors més alts de la variable 1 en relació amb la majoria dels altres objectes. Tanmateix, es pot esperar que l'objecte *ii* tingui valors més baixos de la variable 1 en relació amb els altres.

3.4.2 Escala Tipus II

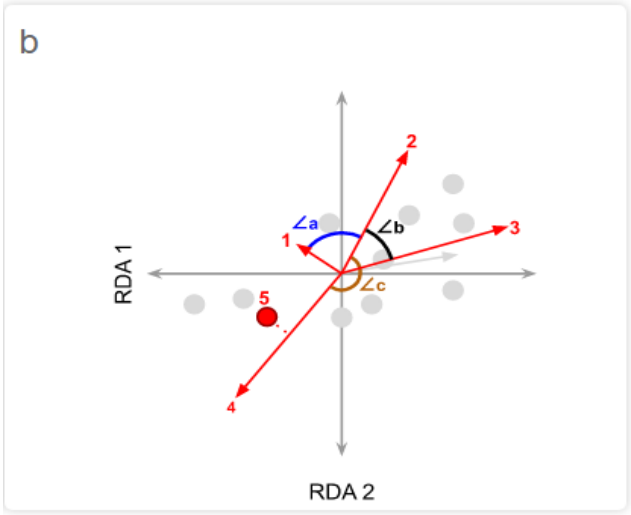
Aquesta s'utilitza per a la resta de variables:

- Les distàncies entre punts d'objecte en un *biplot* no són aproximacions de les seves distàncies euclidianes.

- Les projeccions en angle recte de punts d'objecte sobre vectors resposta aproximen els valors d'un objecte determinat.

- Els angles entre tots els vectors reflecteixen la seva correlació (lineal). I aquesta, és igual al cosinus de l'angle entre vectors.

Des d'un punt de vista gràfic



Il·lustració III-5: Angles entre vectors (GUSTA ME)

Els cosinus d'angles entre vectors s'aproximen a la correlació entre les variables que representen. L'angle *∠a* s'aproxima a 90º, cosa que suggereix que les variables *"1"* i *"2"* mostren una correlació molt reduïda donat que *cos(90º) = 0* (és a dir, són gairebé ortogonals). Si *∠b* és inferior a 90°, suggereix una correlació positiva entre les variables *"2"* i *"3"* mentre que *∠c* s'aproxima a 180° , suggerint una forta correlació negativa entre les variables *"2"* i *"4".* La 5 no és quantitativa i està representada per un centroide.

IV Altres maneres de calcular el RDA

4.1 Partial RDA

RDA parcial és l'anàlisi explicatiu de la matriu de resposta Y en presència de variables explicatives addicionals, W, anomenades covariables.

Els efectes lineals de la matriu explicativa X sobre la matriu de resposta Y s'ajusta als efectes de les covariables W, com en la regressió lineal parcial. És un mètode que permet controlar els efectes lineals coneguts, és a dir, mesura l'efecte de X sobre Y mentre es controla la influència lineal de la matriu covariable W.

A més, permet aïllar l'efecte d'una sola variable o factor explicatiu. És a dir, mesura i comprova l'efecte únic a Y d'una sola variable explicativa, i la resta posant-la a la matriu W. Finalment això es repeteix regressora a regressora.

Per a dur a terme aquest mètode hi ha dues maneres:

0. Centrar les matrius Y, X i W tal que les mitjanes de les columnes sigui 0.
1. Calcular els residus de Y a W: $Y_{res|W} = Y - W[W'W]^{-1}W'Y$.
2. Calcular els residus de X a W: $X_{res|W} = X - W[W'W]^{-1}W'X$.

On el *RDA* de $Y_{res|W}$ a $X_{res|W}$ produeix el *partial* R^2. I el *RDA* de Y a $X_{res|W}$ produeix el *semipartial* R^2.

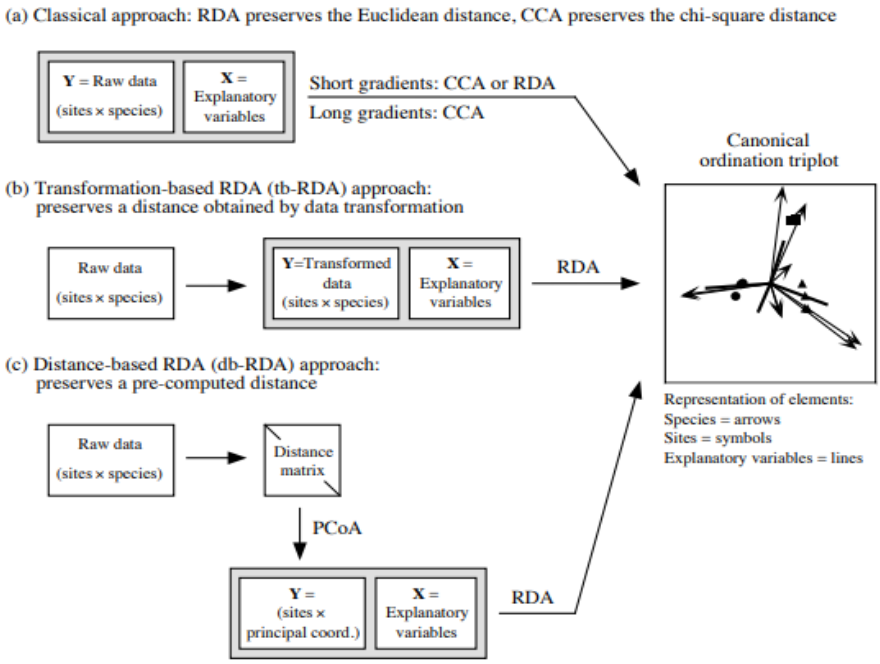
4.2 Distanced-Based RDA i Transformation-Based RDA

L'anàlisi de la redundància basada en la distància (*db-RDA*) és un tècnica que a partir d'una transformació duu a terme l'*RDA*. Aquest mètode està destinat a detectar relacions lineals sobre disimilaritats generades per mesures que poden ser no lineals. S'utilitza una matriu de disimilaritat calculada mitjançant una mesura adequada a les dades de resposta, com a entrada a una anàlisi de coordenades principals *(PCoA)* .

El resultat d'això és un conjunt de coordenades que representen les disimilaritats en un espai euclidià, que és apropiat per a l'anàlisi mitjançant l'*RDA* estàndard.

L'anàlisi de la redundància basada en la modificació *(tb-RDA)* consisteix en aplicar alguna transformació a la matriu Y. Per tant, l'*RDA* calculat en dades transformades mitjançant aquestes equacions, realment es conservarà l'acord, el perfil, la ingerència, la distància chi-quadrat o mètrica chi-quadrat entre els llocs o observacions, en funció de la transformació utilitzada.

Finalment per sintetitzar tota aquesta informació, les diferents vies es poden representar de la manera següent:



Il·lustració IV-1: Comparació del RDA clàssic respecte les dues alternatives (Legendre, 2012)

V Descripció i Resum de la Base de Dades

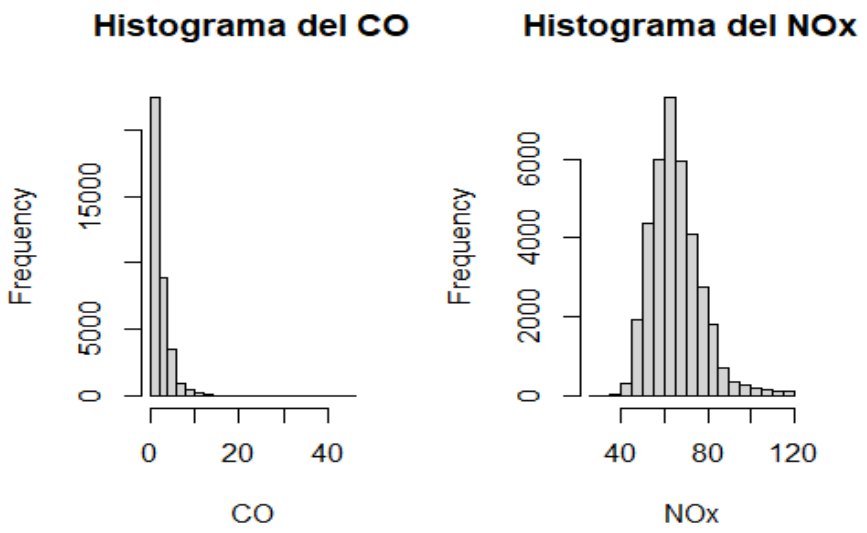
El conjunt de dades que s'ha considerat per a dur a terme la part pràctica està formada per 36.733 observacions recollides durant 5 anys, especialment, de 2011 al 2015. Hi ha 11 mesures diferents, les quals es classifiquen en nou variables d'entrada, és a dir, explicatives i dues variables objectiu. El que es busca és estudiar, modelar i utilitzar les diferents tècniques explicades anteriorment a les emissions de gasos de combustió, en concret, al monòxid de carboni (CO) i a la suma entre el diòxid de nitrogen i el monòxid de nitrogen $(NO_x = NO_2 + NO)$.

A continuació s'adjunta un resum analític de les diferents variables:

Variables	Abre.	Unitats	Min	Max	Mean	Des.Tip
Ambient temperature	AT	ºC	-6.23	37.10	17.71	7.45
Ambient pressure	AP	mbar	985.85	1036.60	1013.07	6.46
Ambient humidity	AH	(%)	24.09	100.20	77.87	14.46
Air filter difference pressure	AFDP	mbar	2.09	7.61	3.93	0.77
Gas turbine exhaust pressure	GTEP	mbar	17.70	40.72	25.56	4.20
Turbine inlet temperature	TIT	ºC	1000.80	1100.90	1081.43	17.54
Turbine after temperature	TAT	ºC	511.04	550.61	546.16	6.84
Turbine energy yield	TEY	MWH	100.02	179.50	133.51	15.62
Compressor discharge pressure	CDP	mbar	9.85	15.16	12.06	1.09
Carbon monoxide	CO	mg/m^3	0.00	44.10	2.37	2.26
Nitrogen oxides	NOX	mg/m^3	25.91	119.91	65.29	11.68

Taula V-1: Resum estadístic de totes les variables de la base de dades

Es grafiquen les dues variables resposta:



Il·lustració V-1: Histogrames de les dues variables objectiu

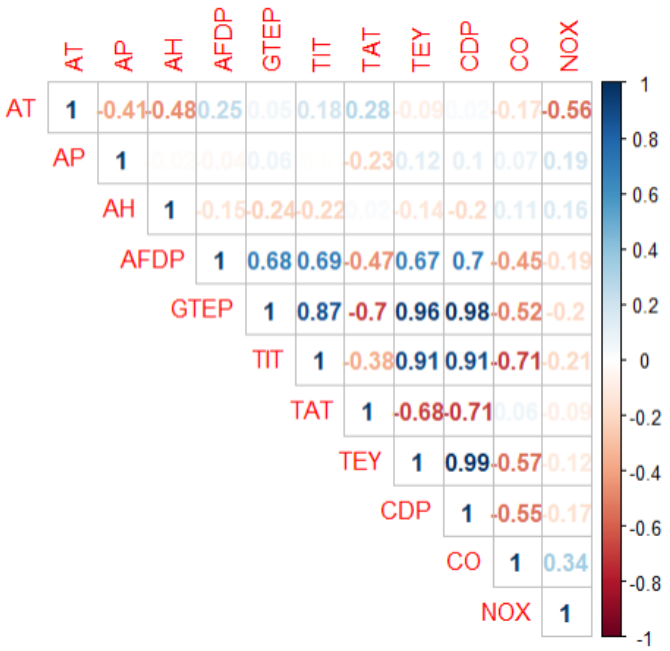
Cap de les dues variables presenten normalitat, donat que mostren un apuntament molt elevat, és a dir, una forma leptocúrtica substancial, i a més, en cap dels dos casos les funcions plasmades són simètriques.

Tot seguit es calcula la matriu de correlacions per veure quines variables estan més correlacionades o bé, si són independents entre elles.

A partir d'aquest estadístic

$$\rho_{XY} = \rho(X,Y) = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

S'ha construït el següent gràfic:



Il·lustració V-2: Matriu triangular superior de les correlacions

S'observa l'existència d'una dependència lineal positiva forta entre les variables explicatives rendiment energètic de la turbina *(TEY)* i pressió de descàrrega del compressor *(CDP) (0.99)*. La pressió del gas del destapament de la turbina *(GTEP)* i *CDP* (0.98) i entre *GTEP i TEY* (0.96). Aquests valors exposen que hi ha informació redundant, és a dir, que hi ha variables que ja es troben explicades per altres. I en conseqüència, en un futur model podrien ser eliminades.

Si s'observen les correlacions de les variables objectiu respecte les variables de la turbina, es pot veure que aquests valors són més elevats per *CO* que no pas per NOx.

A més, si s'analitzen les característiques individuals i les variables objectiu es pot observar que a una major temperatura d'entrada de la turbina *(TIT)* i de la temperatura ambiental *(AT)* es redueixen els nivells de *CO* i *NOx*, donat que hi ha una correlació lineal negativa forta.

VI Creació models

Inicialment es creen els dos models lineals amb totes les variables explicatives tret de les dues que són resposta.

```
Call:
lm(formula = CO ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY +
    CDP, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-11.967  -0.686  -0.084   0.536  34.669

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.287e+02  2.209e+00  58.293   <2e-16 ***
AT          -4.993e-02  3.011e-03 -16.580   <2e-16 ***
AP          -3.675e-03  1.484e-03  -2.477   0.0133 *
AH          -7.983e-03  6.723e-04 -11.873   <2e-16 ***
AFDP        -1.388e-01  1.590e-02  -8.726   <2e-16 ***
GTEP         9.675e-02  1.016e-02   9.520   <2e-16 ***
TIT         -7.110e-02  2.758e-03 -25.783   <2e-16 ***
TAT         -8.071e-02  3.648e-03 -22.127   <2e-16 ***
TEY         -1.921e-01  7.980e-03 -24.068   <2e-16 ***
CDP          1.952e+00  1.131e-01  17.260   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.495 on 36723 degrees of freedom
Multiple R-squared:  0.5634,    Adjusted R-squared:  0.5633
F-statistic:  5266 on 9 and 36723 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = NOX ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY +
    CDP, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-36.364  -4.812  -0.036   3.764  52.649

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -54.987575  11.979648   -4.590 4.45e-06 ***
AT           -1.762792   0.016334 -107.921  < 2e-16 ***
AP           -0.235852   0.008048  -29.305  < 2e-16 ***
AH           -0.222782   0.003647  -61.091  < 2e-16 ***
AFDP          0.698462   0.086246    8.098 5.74e-16 ***
GTEP         -0.111960   0.055123   -2.031  0.04225 *
TIT           1.409075   0.014959   94.198  < 2e-16 ***
TAT          -1.527747   0.019786  -77.214  < 2e-16 ***
TEY          -1.950592   0.043284  -45.065  < 2e-16 ***
CDP          -1.749285   0.613482   -2.851  0.00436 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.11 on 36723 degrees of freedom
Multiple R-squared:  0.5178,    Adjusted R-squared:  0.5177
F-statistic:  4382 on 9 and 36723 DF,  p-value: < 2.2e-16
```

Taula VI-1: Resum models

En conjunt els models són significatius al llindar escollit (0.05) i, a més, la bondat de l'ajust és positiu i superior a 0.5. És a dir, existeix una associació lineal lleu entre les dues variables objectiu i les diferents predictores.

A més, és important tenir en compte la variància de l'estimador perquè si aquest augmenta hi ha una major tendència a quedar-se per sota del valor crític, i en conseqüència acceptar la hipòtesi nul·la. Com es pot arribar a concloure que una variable no és rellevant quan en realitat sí que ho és, serà necessari fer una diagnosi i estudiar les millores oportunes en el cas que es pugui.

VII Hipòtesis Bàsiques i Transformació Variables

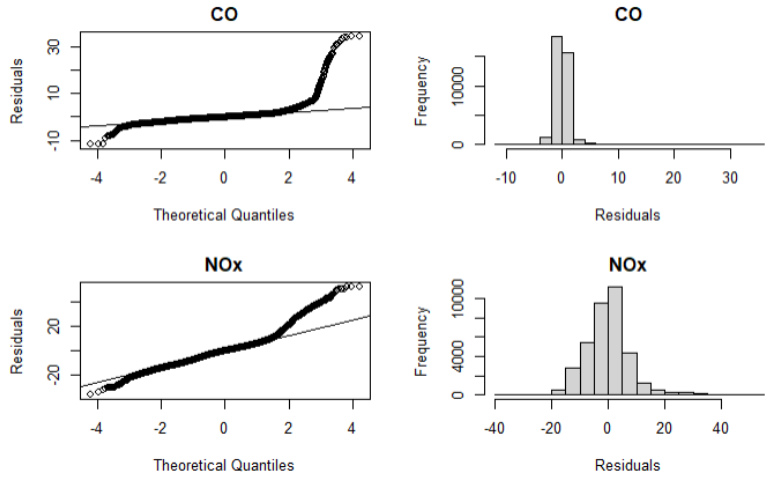
7.1 Normalitat

Per estudiar la possible normalitat o no normalitat dels residus del model des d'un punt de vista gràfic, s'ha considerat fer ús de diferents funcions de l'R, com ara, *QQ-plot* i histogrames.

El contrast és:

$$\begin{cases} H_0: \epsilon_i \sim N(E(\epsilon_i), Var(\epsilon_i)) \\ H_1: \epsilon_i \nsim N(E(\epsilon_i), Var(\epsilon_i)) \end{cases}$$

I les imatges corresponents:



Il·lustració VII-1: Normalitat

Analitzant l'extrem inferior i superior esquerre es pot veure com els diferents residus no segueixen una normal ja que la línia contínua mostra una *Normal(0,1)* teòrica, i els cercles com es distribueixen els diferents valors residuals. A més, a la cua dreta és on presenta una major diferenciació.

Tot seguit en els histogrames s'observa el mateix, és a dir, la distribució no s'aproxima a una normal. Tot i que, els residus de NOx poden portar a confusió, però si es té en compte l'esquema leptocúrtic que presenta també es pot descartar.

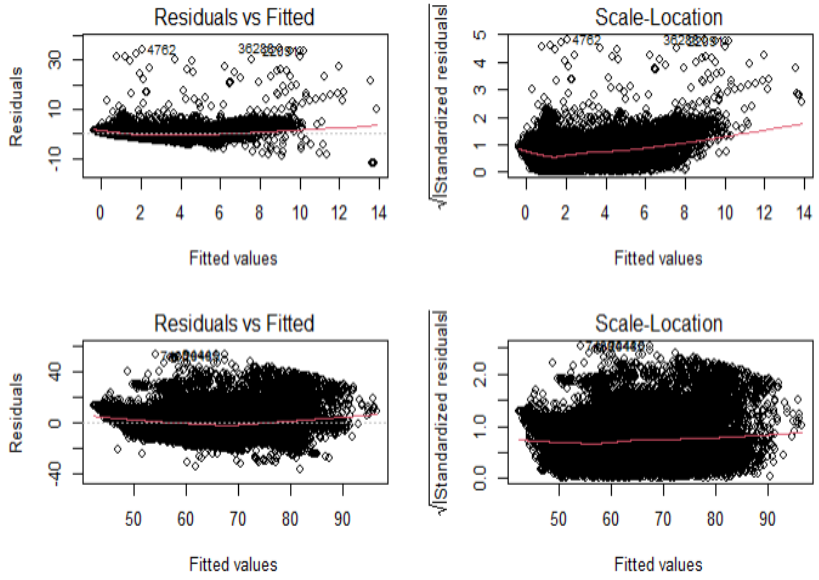
7.2 Homocedasticitat

El contrast que es vol dur a terme per tal de corroborar o desmentir la igualtat de variàncies és:

$$\begin{cases} H_0: var(\epsilon_i) = var(\epsilon_j) = \sigma_\epsilon^2 \ \forall i \neq j \\ H_1: var(\epsilon_i) \neq var(\epsilon_j) \neq \sigma_\epsilon^2 \ \forall i \neq j \end{cases}$$

On la hipòtesi nul·la representa les variàncies constants per tota la mostra (homoscedasticitat) contra l'alternativa, és a dir, variàncies no constants (heteroscedasticitat).

Un cop analitzada la normalitat dels residus, ara és important centrar-se amb la variància.



Il·lustració VII-2: Variància dels residus

Els gràfics situats a l'extrem superior i inferior esquerre permeten veure si es pot ajustar a un model lineal, el qual sembla que sí ja que les dades es veuen prou centrades al voltant del zero.

A més, es veu que independent del model utilitzat, la línia vermella no és establa, sinó que presenta una petita tendència positiva.

Finalment, per concloure es pot dir que la variància del terme pertorbador no és constant per tota la mostra, és a dir, que presenta heteroscedasticitat.

7.3 Transformació de Variables

Hi ha diferents mètodes que es poden usar per a dur a terme l'escull de la funció que millor s'ajusti a les dades, com ara, el mètode de *Box-Cox* o l'escala de Poders de *Tuckey*. Les transformacions més clàssiques són el logaritme i l'arrel quadrada, tot i que pot ser qualsevol altra funció.

7.3.1 Transformació de Box-Cox i Tuckey

Box-Cox (Faraway, 2015) determina quina transformació s'ajusta millor a la variable resposta. S'usa per dades estrictament positives i escull la millor funció. El mètode consisteix en transformar la resposta $y \rightarrow g_\lambda(y)$ on la família escollida ve determinada per λ, és a dir:

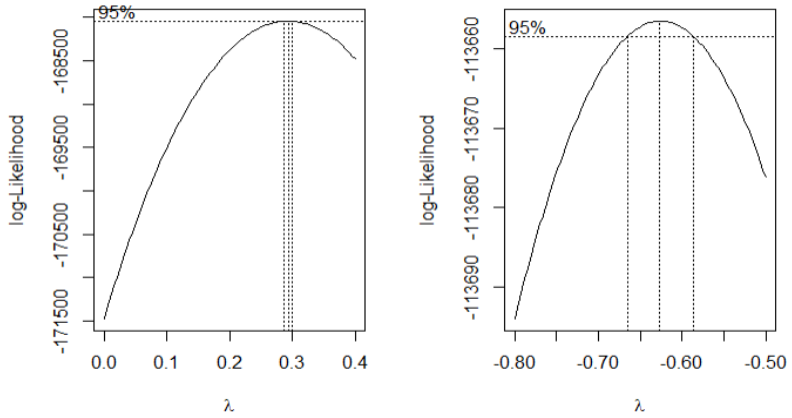
$$g_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

Si $y \geq 0$ i $g_\lambda(y)$ és continua en λ. La tria del valor d'aquesta (λ) es fa a partir de maximitzar la versemblança. Assumint normalitat dels residus el perfil de la log-versemblança seria:

$$L(\lambda) = -\frac{n}{2}\log\left(\frac{RSS_\lambda}{n}\right) + (\lambda - 1)\sum_{i=1}^{n} log(y_i)$$

On RSS_λ és la suma de quadrats residuals quan $g_\lambda(y)$ és la resposta. Com que el propòsit és fer bones prediccions es pot usar directament y^λ en lloc d'aquell quocient.

Per visualitzar aquesta informació es pot executar la següent funció:



Il·lustració VII-3: Funció log-versemblant per diferents valors λ

Tot i així, és adient esbrinar quin és el valor de λ que maximitza la funció de log-versemblança.

```
  Lmax CO   Lmax NOx
0.2949495 -0.6272727
```

Taula VII-1: λ òptimes

Donat aquest valor (*Lmax CO)* una aproximació que es podria fer servir és el logaritme, ja que és proper a zero. Tot i que, l'arrel cúbica o l'arrel quadrada també podria ser una opció. En canvi, com l'altre λ és negativa *(LmaxNOx),* les transformacions que es poden aplicar serien la inversa o l'arrel quadrada inversa.

L'enfocament de *Tukey's Ladder of Powers* (Lane) utilitza una transformació de potència en un conjunt de dades. Per exemple, augmentar les dades a una potència de 0.5 equival a aplicar una transformació basada en l'arrel quadrada; augmentar les dades a una potència de 0.33 equival a aplicar l'arrel cúbica.

A partir del paquet *rcompanion* de R, aquest realitza proves iteratives de *Shapiro-Wilk* i troba el valor λ que maximitza l'estadístic W. En essència, es troba la transformació potencial que fa que les dades s'ajustin o s'aproximin a la distribució normal.

Els valors inclinats a l'esquerra, és a dir, inferiors a 0 s'han d'ajustar mitjançant una constant, per convertir la inclinació d'esquerra a dreta (positiu). En alguns casos de dades esbiaixades, pot ser beneficiós afegir una constant per fer que tots els valors de les dades siguin positius abans de la transformació.

Aquest mètode consisteix en transformar la resposta $y \rightarrow h_\lambda(y)$ on la família escollida ve determinada per λ, és a dir, l'estructura que segueix:

$$h_\lambda(y) = \begin{cases} x^\lambda & si\ \lambda > 0 \\ \log(x) & si\ \lambda = 0 \\ -x^\lambda & si\ \lambda < 0 \end{cases}$$

Per visualitzar la informació explicada anteriorment s'utilitzarà la funció `transformTuckey()` obtenint així la següent taula:

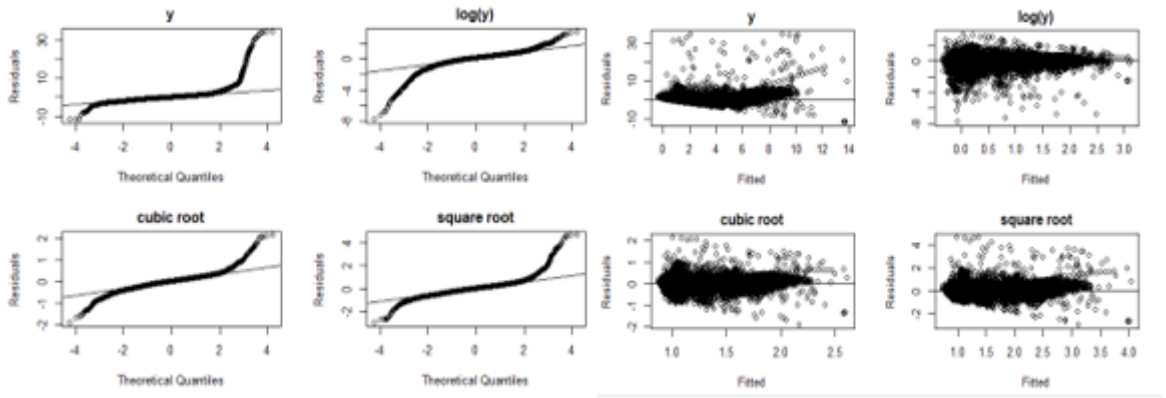
	lambda <dbl>	W <dbl>	Shapiro.p.value <dbl>
406	0.125	0.9856	3.387e-22
	lambda <dbl>	**W** <dbl>	**Shapiro.p.value** <dbl>
373	-0.7	0.9965	1.473e-09

Taula VII-2: λ que fa màxim l'estadístic W

Aquests valors de λ s'aproximen als obtinguts anteriorment, per tant, posa de relleu que el logaritme i la inversa o arrel inversa en els respectius models siguin les millors propostes.

Per decidir quina pot ser l'òptima atenent a tots els resultats mostrats, s'estudiarà des d'un punt de vista gràfic quina seria la funció que millor s'ajusta a les dades en termes de normalitat i homocesdascitat.

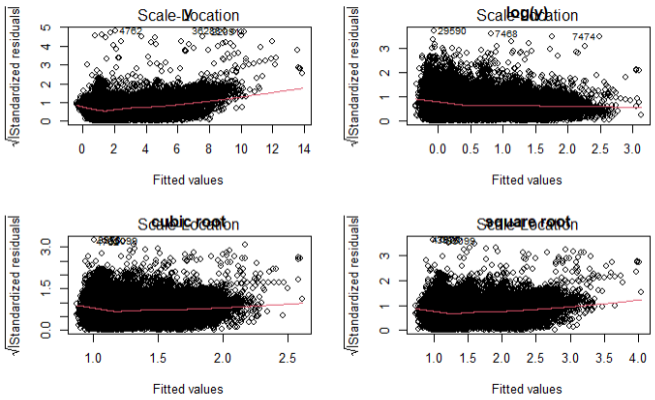
Començant pel model on la variable depenent és el *monòxid de carboni (CO)*:



Il·lustració VII-4: Normalitat i Esperança

La normalitat continua sense complir-se, però els residus es troben molt més estabilitzats. A més, en terme mig els residus es troben al voltant del zero independentment de la transformació aplicada, tot i presentar una forma triangular la qual fa referència a la no igualtat de variàncies (heteroscedasticitat).

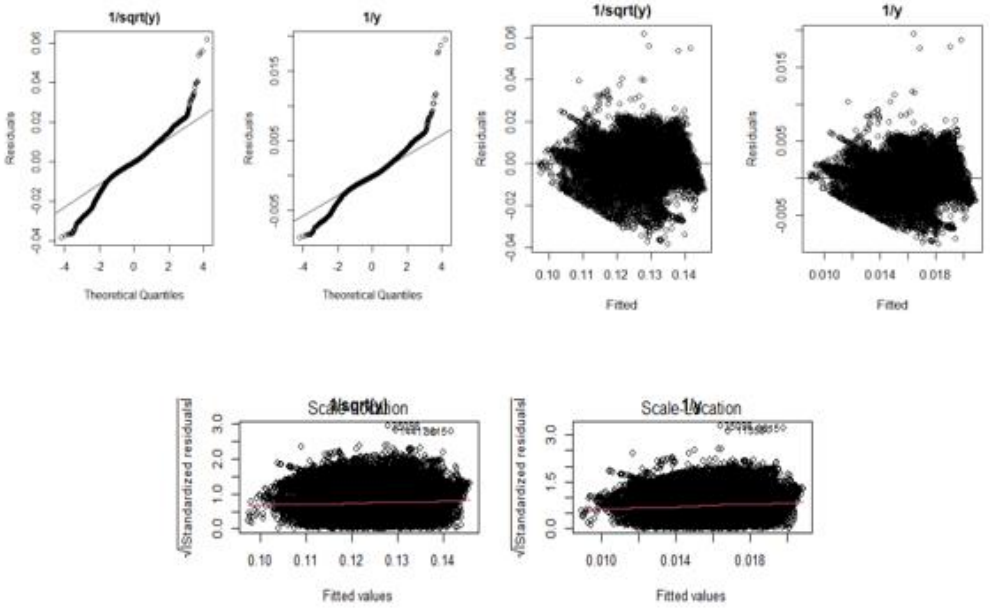
I seguint aquesta última observació pot ser precís considerar el model en termes de variància.



Il·lustració VII-5: Dispersió dels residus

L'arrel cúbica i la quadrada (gràfics inferiors) no acaben d'estabilitzar-la donat que a mida que augmenten els valors ajustats, la línia vermella presenta una tendència ascendent. Però observant la variabilitat associada al logaritme aquesta és més estable que no pas les altres. És per això, que finalment s'usarà aquesta funció.

Tot seguit, per esbrinar la millor funció pel model on la variable objectiu és el *NOx* es representaran els mateixos gràfics.



Il·lustració VII-6: Normalitat, Esperança i Variància

A partir d'aquesta col·lecció d'imatges es pot observar que ambdues transformacions presenten les mateixes millores, ja que els residus estan al voltant del zero i en termes de variància semblen estables.

És per això, que s'ha decidit usar la transformació inversa en lloc de l'arrel inversa ja que la interpretació que se'n pot derivar posteriorment pot ser més senzilla.

VIII Multicol·linealitat

La multicol·linealitat es defineix com el grau de correlació existent entre les variables explicatives o regressores del model lineal múltiple.

Es poden distingir tres situacions:

- Multicol·linealitat perfecta es produeix quan existeix una variable que és combinació lineal de la resta de predictores del model.

- Absència de multicol·linealitat es troba quan les variables explicatives són ortogonals entre sí, és a dir, totalment independents entre elles:

$$Cov(X_i, X_j) = 0 \; \forall i \neq j$$

- Presència de multicol·linealitat és la situació més present a la pràctica i suposa que existeix un cert grau de correlació entre les variables explicatives.

8.1 Detecció a partir de la Matriu de Correlacions

Per a la detecció i valoració d'aquest fenomen hi ha diversos instruments que es poden fer servir, en aquest cas, s'utilitza:

- El càlcul del determinant de la matriu de correlacions el qual proporciona la següent informació:

$$|R_x| = \begin{vmatrix} 1 & r_{2,3} & \cdots & r_{2,k} \\ r_{3,2} & 1 & \cdots & r_{3,k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k,2} & r_{k,3} & \cdots & 1 \end{vmatrix}$$

  Sabent que $|\mathrm{R_x}| \in [0,1]$, si $|\mathrm{R_x}| = 0$ s'estarà davant de multicol·linealitat perfecta, però si $|\mathrm{R_x}| = 1$ s'estarà davant d'absència total, en canvi, si el determinant es proper a zero, llavors es pot dir que sí que existeix un cert grau elevat de multicol·linealitat.

8.2 Factor d'Increment de la Variància

El càlcul del Factor d'Increment de la Variància *(FIV)*, consisteix en comparar la variància real d'un paràmetre β_j amb la que s'obtindria en cas que hi hagués absència total de multicol·linealitat.

Variància en cas de regressors perpendiculars: $Var(\hat{\beta}_j^*) = \frac{\sigma_\epsilon^2}{\sum_{j=1}^{p} x_j^2}$

Variància en cas d'alguna correlació lineal: $Var(\hat{\beta}_j) = \frac{\sigma_\epsilon^2}{(1-R_j^2)\sum_{j=1}^{p} x_j^2}$

on R_j^2 és la correlació de la regressió auxiliar de la variable x_j i la resta d'explicatives.

Per tant, el *FIV* per aquesta variable seria:

$$FIV_j = \frac{Var(\hat{\beta}_j)}{Var(\hat{\beta}_j^*)} = \frac{\frac{\sigma_\epsilon^2}{(1-R_j^2)\sum_{j=1}^{p} x_j^2}}{\frac{\sigma_\epsilon^2}{\sum_{j=1}^{p} x_j^2}} = \frac{1}{(1-R_j^2)}$$

Sabent que $R_j^2 \in [0,1] \Rightarrow FIV_j \in [0,\infty)$.

8.3 Detecció en els models

Com que ambdós models presenten les mateixes variables explicatives també presentaran els mateixos valors de *FIV's* i el mateix determinant. És per això, que s'ha decidit agafar un dels models independentment de la variable resposta per estudiar aquest fenomen. Tot i així, alhora de buscar solucions, en cas que hi hagi, sí que es tornarà a diferenciar cada model.

Inicialment, es calcula el determinant de la matriu de correlacions.

```{r}
explicatives <- d[c("AT","AP","AH","AFDP","GTEP","TIT","TAT","TEY","CDP")]
det(cor(explicatives))
```

```
[1] 4.928008e-07
```

Taula VIII-1: Determinant matriu correlacions

Seguidament els *FIV's*.

```{r}
vif(mod)
```

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP
8.263717	1.511004	1.553099	2.488056	29.873876	38.426405	10.235054	255.219703	249.151945

Taula VIII-2: Factor d'Increment de la Variància

Com que la gran majoria de valors obtinguts són superiors a 10, això implica que s'està davant d'un problema greu de multicol·linealitat.

A més, també és interessant comprovar si usant les fórmules clàssiques els valors coincideixen amb els obtinguts amb R.

Sabent que

$$R_x^{-1} = \begin{pmatrix} 1 & r_{2,3} & \cdots & r_{2,k} \\ r_{3,2} & 1 & \cdots & r_{3,k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k,2} & r_{k,3} & \cdots & 1 \end{pmatrix}^{-1} = \begin{pmatrix} FIV_2 & \cdots & \cdots & \cdots \\ \cdots & FIV_3 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & FIV_k \end{pmatrix}^{-1}$$

I substituint

```{r}
X <- cbind(d$AT,d$AP,d$AH,d$AFDP,d$GTEP,d$TIT,d$TAT,d$TEY,d$CDP)# matriu explicatives
R <- cor(X) # matriu correlacions
DR <- det(R) # determinant
IR <- solve(R) # calcula la inversa
FIV <- diag(IR) # afaga la diagonal
FIV
```

```
[1]   8.263717   1.511004   1.553099   2.488056  29.873876  38.426405  10.235054 255.219703
[9] 249.151945
```

Taula VIII-3: FIV's calculats manualment

Es comprova que s'obté el mateix valor que amb el programari.

Finalment atenent a tots els resultats mostrats abans, és a dir, un determinant molt proper a zero i factors d'inflació superiors a 10, són clars indicadors de que existeix una elevada multicol·linealitat.

IX Possibles solucions a la Multicol·linealitat elevada

Hi ha diverses solucions per intentar dissoldre aquest fenomen, ja que si no es corregeix els intervals de predicció seran més grans donat que els estimadors són poc estables i imprecisos.

9.1 Reespecificació del model

Consisteix en transformar les diferents variables per les seves diferències o bé dividint alguna variable en concret.

Cal esmentar que té sentit diferenciar el model exclusivament quan les dades són temporals, cosa que en aquest estudi no es pot aplicar ja que es tenen dades de tall transversal. Finalment dividir per una variable és quelcom complexa ja que en algun cas pot reduir la multicol·linealitat, però afecta negativament al terme de pertorbació.

9.2 Anàlisi de les Components Principals

Des d'un punt de vista teòric és el millor donat que permet obtenir una combinació lineal de variables ortogonals entre elles. Per tant, si aquestes noves variables es fan servir com a explicatives del model, s'estarà davant d'absència total de multicol·linealitat (apartat VIII).

Com a la pràctica presenta una difícil interpretació, per estudiar l'anàlisi estructural i/o la predicció d'aquest s'utilitzaran altres mètodes.

9.3 Selecció de Variables

Els diferents mètodes anteriors a vegades no es poden usar o bé no presenten les condicions idònies. És per això que s'analitzarà si és imperatiu utilitzar-les totes o no.

Tenint en compte les explicatives en el model una pregunta podria ser si totes han d'estar-hi, i en cas negatiu, saber quines han de ser-hi i quines no.

En general, si s'inclouen cada vegades més, l'ajust a les dades millora, donat que augmenta la quantitat de paràmetres a estimar però disminueix la precisió individual, és a dir, major variància. Per tant, en aquesta funció de regressió es pot produir un sobre ajust. Tot i que, si s'inclouen menys variables de les necessàries, les variàncies es redueixen però el baix augmentarà.

Finalment l'objectiu d'aquests mètodes (Martínez, 2018) és buscar un model que s'ajusti bé a les dades, i alhora sigui possible trobar un equilibri entre ajust i facilitat.

El mètode de mínims quadrats ordinaris *(MQO)* és una forma comú d'estimar el vector de paràmetres desconeguts (β). I a més, presenta un bon comportament quan les variables del model són rellevants. Tot i que hi ha dues raons essencials pel qual aquest no seria un mètode adequat:

- **Precisió de la predicció**: les estimacions per MQO tenen poc biaix però molta variància.

- Algunes d'aquestes estimacions poden proporcionar valors molt propers a zero i fer que la variable associada contribueixi poc al model, però no l'eliminaria. És a dir, que hi ha situacions on no es redueix el conjunt de regressores.

És per això que serà necessari utilitzar altres algorismes, com ara:

- **Mètode *Forward*:** (selecció cap endavant) parteix d'un model molt senzill i es van afegint termes sota un criteri, fins que no procedeix a afegir cap terme més, és a dir, a cada etapa s'introdueix la variable més significativa fins una certa regla de parada.

- **Mètode *Backward*:** (Eliminació enrere) Es parteix d'un model complexa, que incorpora tots els efectes que poden influir en la resposta, i en cada etapa s'elimina la variable menys influent, fins que no procedeixi suprimir-ne cap més.

- **Mètode *Stepwise***: Aquest procediment és una combinació dels dos anteriors. Comença com el d'introducció progressiva, però en cada etapa es planteja si totes les variables introduïdes han d'estar.

Quan s'apliquen algun d'aquests mètodes explicats s'ha de tenir en compte quina serà la condició per afegir o suprimir variables. Per això es poden considerar dos criteris: criteris de significació del terme (1) i criteris d'ajust global (2).

1. **Criteris de significació:** En un mètode *backward* o *forward* es suprimirà el terme que resulti menys o més significatiu, respectivament. El criteri usat pel *forward selection* és considerar aquella que presenti una *F* més gran o bé el p-valor més petit, i al revés pel *backward*.

2. **Criteris globals:** En lloc d'utilitzar el criteri anterior, es pot usar un generalitzat de forma que tingui en compte l'ajust i l'excés de paràmetres. S'escollirà aquell model que presenti un Criteri d'Informació d'Akaike *(AIC)* o Criteri d'Informació de Bayes *(BIC)* més petit, ja que en aquell cas hauria una versemblança molt gran i pocs paràmetres.

Però com aquests mètodes s'han usat durant el grau s'ha considerat fer servir una alternativa:

- ***Best subset*:** Donades p variables explicatives, aquest mètode consisteix en formar tots els possibles subconjunts de variables i efectuar totes les possibles regressions, retenint aquella que, d'acord amb el criteri de bondad d'ajust que s'hagi escollit, sembli millor (coeficient de determinació, coeficient de determinació ajustat i el coeficient de C_p *de* $Mallows$). L'inconvenient és el gran volum de càlcul que és precís realitzar.

  Els passos que realitza l'algorisme són els següents:

  A. Es parteix d'un model nul Ψ_0, és a dir, sense regressores on la predicció correspon a la mitjana global per cada observació.

  B. Per $k = 1, \dots, p$:

1. S'ajusten tots els possibles models, és a dir, els $\binom{p}{k}$ que contenen exactament k predictors.

2. Seleccionar el millor depenent de tots els anteriors Ψ_k

C. Seleccionar un únic model entre els $\Psi_1, \dots, \Psi_p$ utilitzant l'error de predicció per validació creuada, $C_p, AIC, BIC \ o \ R^2_{adj}$.

A partir dels models creats inicialment *(VI)*, s'ha cregut convenient usar aquesta opció aprofitant que el nombre de variables explicatives és petit.

- Partint del model amb el monòxid de carboni com a variable independent subjecte a totes les explicatives i fent ús de la funció `ols_step_all_possible()` s'obté el següent:

	Index <int>	N <int>	Predictors <chr>	R-Square <dbl>	Adj. R-Square <dbl>	Mallow's Cp <dbl>
6	1	1	TIT	0.498824448	0.498810804	5428.64852
8	2	1	TEY	0.324686352	0.324667967	20076.71471
9	3	1	CDP	0.303630422	0.303611463	21847.88751
5	4	1	GTEP	0.269267069	0.269247175	24738.44782
4	5	1	AFDP	0.201084630	0.201062879	30473.78606
1	6	1	AT	0.030389478	0.030363080	44832.24028
3	7	1	AH	0.011360569	0.011333653	46432.90504
2	8	1	AP	0.004495706	0.004468604	47010.36031
7	9	1	TAT	0.003405121	0.003377989	47102.09766
40	10	2	TIT TAT	0.550721764	0.550697301	1065.17457

1-10 of 511 rows Previous 1 2 3 4 5 6 ... 52 Next

Taula IX-1: Models pels diferents nombres de predictors

Aquesta sortida és una extracció de tots els subconjunts possibles i, en concret, es veuen només els models amb una variable explicativa, quin és el seu coeficient de determinació, a més de l'ajustat i el $C_p \ de \ Mallows$. Hi ha 511 models diferents.
Per tant, el primer que tendeix un a pensar és escollir aquell que presenti un coeficient de determinació més alt. Tot i així, si es tingués en compte el millor model seria novament l'explicat abans. Però a partir de la propera funció es podrà fer ús de mètodes gràfics i analítics per tal d'ajudar a decidir quin és el nombre òptim de variables a escollir.

```
                Best Subsets Regression
-------------------------------------------------
Model Index    Predictors
-------------------------------------------------
      1        TIT
      2        TIT TAT
      3        AH TIT TAT
      4        TIT TAT TEY CDP
      5        AT TIT TAT TEY CDP
      6        AT AH TIT TAT TEY CDP
      7        AT AH GTEP TIT TAT TEY CDP
      8        AT AH AFDP GTEP TIT TAT TEY CDP
      9        AT AP AH AFDP GTEP TIT TAT TEY CDP
-------------------------------------------------
```

Taula IX-2: Millor model per cada variable inclosa

S'aprecien els diferents models atenent en cada cas a les variables introduïdes.

Subsets Regression Summary

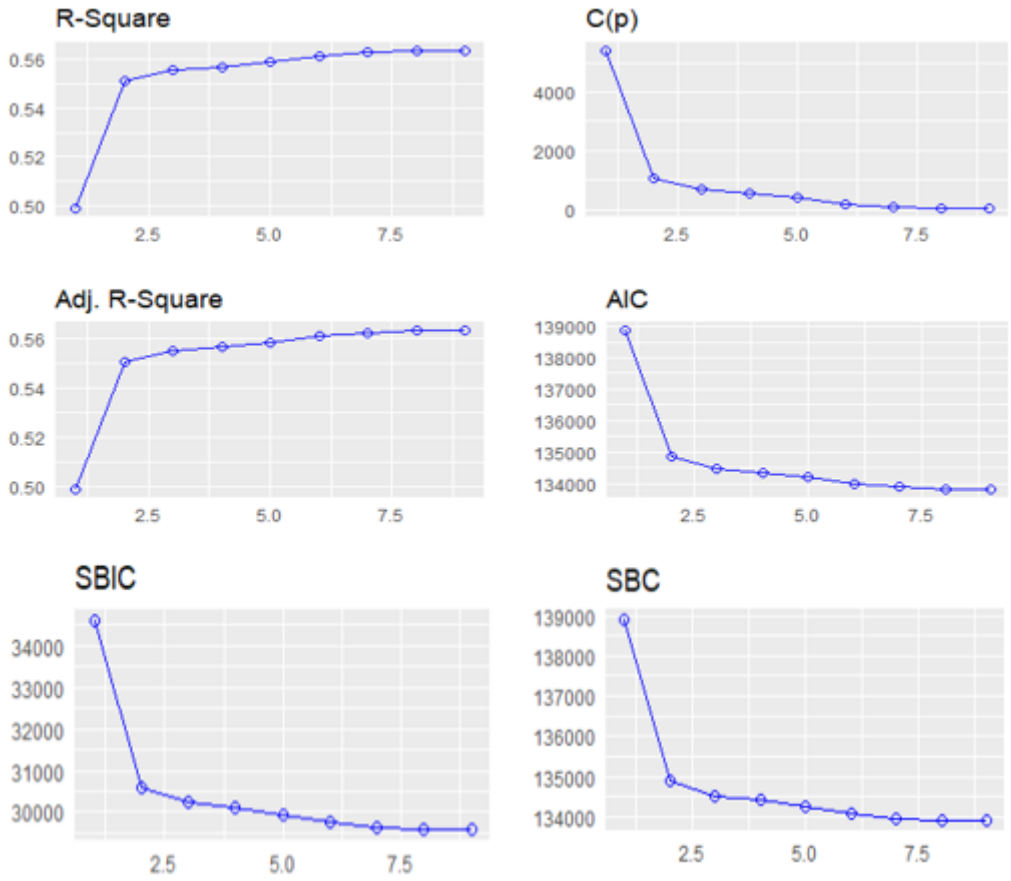
Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.4988	0.4988	0.4987	5428.6485	138862.0119	34617.7258	138887.5462	94254.2986	2.5661	1e-04	0.5012
2	0.5507	0.5507	0.5505	1065.1746	134848.5782	30604.6702	134882.6239	84496.4559	2.3005	1e-04	0.4494
3	0.5552	0.5552	0.5549	689.0047	134481.1556	30237.2710	134523.7127	83653.2141	2.2776	1e-04	0.4449
4	0.5567	0.5567	0.5564	565.0139	134359.2493	30115.3618	134410.3179	83373.7840	2.2700	1e-04	0.4434
5	0.5587	0.5586	0.5584	400.0565	134196.4077	29952.5437	134255.9877	83002.7371	2.2600	1e-04	0.4414
6	0.5612	0.5611	0.5608	191.2901	133989.2502	29745.4447	134057.3417	82533.7104	2.2473	1e-04	0.4390
7	0.5625	0.5624	0.5621	87.7834	133886.0999	29642.3304	133962.7028	82300.0317	2.2410	1e-04	0.4377
8	0.5634	0.5633	0.5629	14.1334	133812.5198	29568.7834	133897.6341	82133.1056	2.2365	1e-04	0.4369
9	0.5634	0.5633	0.563	10.0000	133808.3853	29564.6524	133902.0110	82121.6264	2.2362	1e-04	0.4368

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error
HSP: Hocking's Sp
APC: Amemiya Prediction Criteria

Taula IX-3: Justificació analítica dels models

Atenent a l'*AIC* i al coeficient de determinació ajustat es pot observar que el millor model seria el de 9 variables. Tot i que, a partir de 6 la millora que es produeix és pràcticament nul·la, és a dir, que un es pot preguntar si és necessari introduir-les totes quan amb 6 pràcticament s'explica el mateix.

Per resoldre aquesta qüestió es farà ús de diferents mètodes gràfics:



Il·lustració IX-1: Nombre òptim d'explicatives a utilitzar

Qualsevol dels gràfics mostrats es pot observar com a partir de 6 variables aquesta tendència que mostra la línia blava va desapareixent. Finalment es confirma allò que s'havia dit, és a dir, que per la millora que es produeix no és computacionalment viable considerar-les totes.

Però tornant a la taula *(IX-1),* en la qual es podia veure quins són tots els possibles submodels amb n_i variables. Ara es vol estudiar els models amb 6 explicatives i escollir el que no presenti multicol·linealitat elevada:

```r
models_6_variables <- models[models$n == 6,]
models_6_variables
```

Description: df[,6] [84 × 6]

	Index <int>	N <int>	Predictors <chr>	R-Square <dbl>	Adj. R-Square <dbl>	Mallow's Cp <dbl>
431	382	6	AT AH TIT TAT TEY CDP	0.5612057	0.5611340	191.2901
437	383	6	AT GTEP TIT TAT TEY CDP	0.5604214	0.5603495	257.2689
436	384	6	AT AFDP TIT TAT TEY CDP	0.5601696	0.5600977	278.4479
463	385	6	AH AFDP TIT TAT TEY CDP	0.5595242	0.5594522	332.7352
465	386	6	AFDP GTEP TIT TAT TEY CDP	0.5590127	0.5589406	375.7656
427	387	6	AT AH GTEP TIT TAT TEY	0.5589834	0.5589113	378.2303
459	388	6	AH AFDP GTEP TIT TAT TEY	0.5588644	0.5587923	388.2373
457	389	6	AP AFDP TIT TAT TEY CDP	0.5587361	0.5586640	399.0297
416	390	6	AT AP TIT TAT TEY CDP	0.5587007	0.5586286	402.0046
464	391	6	AH GTEP TIT TAT TEY CDP	0.5583997	0.5583276	427.3257

1-10 of 84 rows

Previous 1 2 3 4 5 6 … 9 Next

Taula IX-4: Combinacions amb 6 regressores

Hi ha 84 models candidats, però aquesta taula no informa per cada model si presenta multicol·linealitat o no. Per tant s'ha considerat construir de manera independent el següent algorisme .

```r
vif_efectiu <- c()
for(i in 1:nrow(models_6_variables)){
   var_exp <- unlist(strsplit(models_6_variables$predictors[i], " "))
   b <- vif(lm(d[,"CO"] ~ d[,var_exp[1]]+d[,var_exp[2]]+
               d[,var_exp[3]]+d[,var_exp[4]]+d[,var_exp[5]]+d[,var_exp[6]]))
  if((b[1]<=5) & (b[2]<=5) & (b[3]<=5) & (b[4]<=5) & (b[5]<=5) & (b[6]<=5)){
    vif_efectiu <- c(vif_efectiu,i)
  }
}
models_6_variables[vif_efectiu,]
```

Description: df[,6] [4 × 6]

	Index <int>	N <int>	Predictors <chr>	R-Square <dbl>	Adj. R-Square <dbl>	Mallow's Cp <dbl>
386	404	6	AT AP AH AFDP TIT TAT	0.5562920	0.5562195	604.617
389	441	6	AT AP AH AFDP TAT TEY	0.5464132	0.5463391	1435.600
390	456	6	AT AP AH AFDP TAT CDP	0.5329103	0.5328340	2571.427
383	462	6	AT AP AH AFDP GTEP TAT	0.4794802	0.4793952	7065.838

4 rows

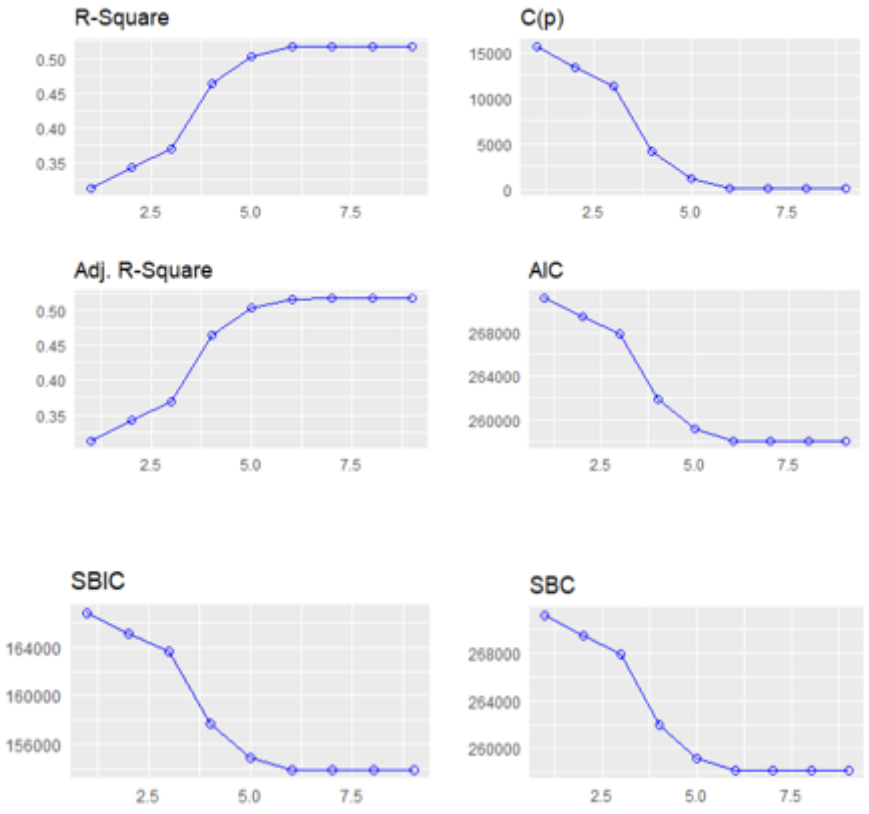
Taula IX-5: Algorisme per seleccionar un model sense multicol·linealitat

Aquest algorisme a partir de la taula *(IX-4)* agafa la columna *"predictors"* es calculen tots els *FIV's* i només es guarden aquells que presenten una multicol·linealitat inferior o igual a 5.

Finalment queden 4 models candidats, el qual s'escull el d'índex *404* ja que és aquell que presenta un R^2 i R^2_{adj} major, i a més, mostra un C_p de Mallows menor.

Per concloure aquesta selecció cal remarcar que a partir d'aquests mètodes s'ha reduït la dimensionalitat del model i alhora la multicol·linealitat elevada.

- Partint del model que vol explicar els diferents valors de *NOx*, s'ha usat la mateixa funció que pel model anterior obtenint així una taula on es mostra la importància a l'hora de crear els submodels (*Taula XIV2, Apèndix I*).
  A més, un cop s'han obtingut els *best subsets* (*Taula XIV 3, Apèndix I*) s'ha pogut veure quin és el millor en cada cas atenent al criteri *AIC,* R^2_{adj}, etc... Però tot i tenint aquesta informació també s'ha considerat convenient fer una interpretació gràfica:



Il·lustració IX-2: Nombre d'explicatives a utilitzar

En aquest cas, també s'observa que amb 6 variables la tendència que mostra la línia blava desapareix dràsticament.

Finalment un cop construïts els diferents models s'escull el d'índex *406* ja que és aquell que presenta un R^2 i R^2_{adj} major, i a més, mostra un C_p de *Mallows* menor. (*Taula XIV 4, Apèndix I*)

9.4 Altres mètodes de selecció de variables

Molt sovint per resoldre la multicol·linealitat es fan servir altres mètodes de selecció de variables com ara: el mètode Bridge, Lasso, etc... Els quals introdueixen una penalització a l'estimació dels paràmetres.

Aquests mètodes (Martínez, 2018) es poden diferenciar com:

- **Mètode *Lasso*:** mètode que combina la contracció d'alguns paràmetres cap a zero imposant una restricció i penalització sobre els coeficients de regressió. *Lasso* estima el model minimitzant el problema dels mínims quadrats ponderats (MQP):

$$\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Essent λ el paràmetre de penalització o regulació. Per valors grans de λ, els coeficients β_j es contrauen cap a zero i fins i tot alguns s'eliminen.

- **Mètode *Bridge*:** mètode que combina la contracció d'alguns paràmetres cap a zero. Els estimadors Bridge es troben resolent el següent problema:

$$\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^{\gamma}$$

Essent λ el paràmetre de penalització. Quan $\gamma = 1$ es tracta del mètode *Lasso* explicat anteriorment i quan $\gamma = 2$ es coneix com a Ridge.

- **Mètode SCAD:** és una extensió de les tècniques presentades anteriorment, es fa servir sobretot quan el mètode *Lasso* no és prou satisfactori. L'estimador SCAD es defineix com el que minimitza la següent expressió:

$$\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda\sum_{j=1}^{p}p_{\lambda}^{SCAD}(\beta_j)$$

On

$$p_{\lambda}^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j| & si\ 0 \le |\beta_j| \le \lambda \\ -\dfrac{(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2)}{(2(a-1))} & si\ \lambda \le |\beta_j| \le a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & si\ |\beta_j| \ge a\lambda \end{cases}$$

Finalment per concloure aquests mètodes penalitzats, remarcar la importància que té *λ* en el procés d'optimització. Veient que una *λ* amb valors grans fa que els coeficients es contraguin cap a zero.

Hi ha diferents mètodes per seleccionar la millor *λ*, com ara:

- **CV (validació creuada):** un dels mètodes més utilitzats per estimar el paràmetre *λ* és el mètode *k-fold-validation*.

Sigui $Q = \{(x_i, y_i), i = 1, \dots, n\}$ el conjunt de totes les dades, i els conjunts d'entrenament i prova per $Q - Q^v$ *i* Q^v respectivament, per $v = 1, \dots, p$.

Cada Q^v està format per una proporció fixe d'elements de Q seleccionats a l'atzar.

Sigui $\widehat{\beta}_{J_n}^{(v)}$ l'estimador de β utilitzant el conjunt de dades d'entrament $Q - Q^v$.

L'estadístic de validació creuada es defineix com

$$CV(\lambda) = \sum_{v=1}^{k} \sum_{(y_k, x_k) \in Q^v} \left\{ y_k - x_k' \hat{\beta}_n^{(v)}(\lambda) \right\}^2$$

Per tant, l'estimador del paràmetre de penalització λ s'escriurà com $\hat{\lambda}$ i serà el valor que minimitza l'expressió.

9.4.1 Mètode Bridge

S'usarà aquest mètode per $\gamma = 2$, donant lloc a la regressió *Ridge* ja esmentada abans. La qual consisteix en introduir un biaix a l'estimació dels paràmetres per *MQO* però reduint la seva variància de forma que tingui un menor Error Quadràtic Mig *(EQM)* i siguin més estables. Amb aquesta regressió es treballa amb les variables estandarditzades.

L'estimador resolent el problema d'optimització matemàtica (apartat 9.4) és:

$$\hat{\beta}_{RIDGE} = (X'X + \lambda I)^{-1}(X'Y) = (R + \lambda I)^{-1}(X'Y)$$

La determinació de λ es farà per mètodes gràfics o numèrics *(CV)* seleccionant el valor més petit que estabilitza les estimacions dels paràmetres.

Per dur a terme aquesta regressió és important segmentar la base de dades en dos subconjunts. 2/3 de la base original serà per entrenar *(training)* i l'altre terç per testejar *(testing)*.

```
```{r}
n <- nrow(d)
learn <- 1:round(0.67*n)
nlearn <- length(learn)
ntesting <- n - nlearn
training <- d[1:nlearn,] # Q-Q^v
testing <- d[24612:nrow(d),]# Q^v
```
```

Il·lustració IX-3: Subconjunts d'entrenament i prova

En aquesta imatge s'observa que $Q^v = testing$ *i* $Q - Q^v = training$.

Per seleccionar la λ òptima.

```{r}
x_vars <- as.matrix(training[,-c(10,11)])
y_var <- training$CO

set.seed(123)

cv_output <- cv.glmnet(x_vars, y_var, alpha = 0)
cv_output$lambda.min
```

[1] 0.1649405

```{r}
x_vars <- as.matrix(training[,-c(10,11)])
y_var <- training$NOX

set.seed(123)

cv_output <- cv.glmnet(x_vars, y_var, alpha = 0)
cv_output$lambda.min
```

[1] 0.6979598

Taula IX-6: λ que minimitza el biaix introduït

S'especifica l'ús de la regressió usada introduint l'argument $\alpha = 0$ dins de la funció `cv.glmnet()`.

Si es posa 1 o bé un valor comprés entre aquests dos anteriors, fan referència a *LASSO* i *Regressió Elàstica*, respectivament.

Seguidament es calculen les prediccions d'aquest model tenint en compte la λ trobada anteriorment. I es procedeix a calcular el *root mean square error (rmse)*.

CO	NOx
1.516409	8.131984

Taula IX-7: RMSE

El *rmse* associat, és a dir, la mesura que quantifica quant ens equivoquem és prou baix. Tot i així, com l'objectiu és dur a terme el *RDA*, fent aquesta regressió no es podrien comparar directament ambdós mètodes. Tot i que, si es fes un *RDA* penalitzat es podria comparar, però no és l'objectiu del treball.

X Detecció d'Observacions Atípiques i Influents

10.1.1 Observacions Potencialment Influents

Una observació presenta influència potencial (Faraway, 2015) si és prou diferent a la resta d'observacions en termes de les variables explicatives.

El *leverage* (h_{ii}) serveix per conèixer el grau de palanquejament d'una observació i-èssima. Aquest s'obté com l'element i-èssim de la diagonal principal de la matriu *H (hat matrix)*:

$$H = X(X'X)^{-1}X'$$

Altrament també es possible obtenir aquest valor a partir de la següent expressió:

$$h_{ii} = \frac{1}{n}[1 + d_i] = \frac{1}{n}[1 + (x_{ji} - \bar{x}_j)S^{-1}(x_{ji} - \bar{x}_j)']$$

On:

- d_i és la distància de Mahalanobis que proporciona la distància d'una observació en relació a un conjunt d'observacions.

- $(x_{ji} - \bar{x}_j)$ és un vector fila de dimensió $(1\ x\ k)$ que conté els valors centrats de l'observació i-èssima respecte de cadascuna de les K variables.

- S és la matriu de variàncies i covariàncies mostrals entre les variables explicatives. $S = \frac{1}{n}(X'X)$

A partir de l'anterior expressió, es pot observar que a mesura que augmenta, d_i, augmenta també el *leverage* h_{ii}, i que:

$$\frac{1}{n} \leq h_{ii} \leq 1$$

Per tant, el que es vol dur a terme és:

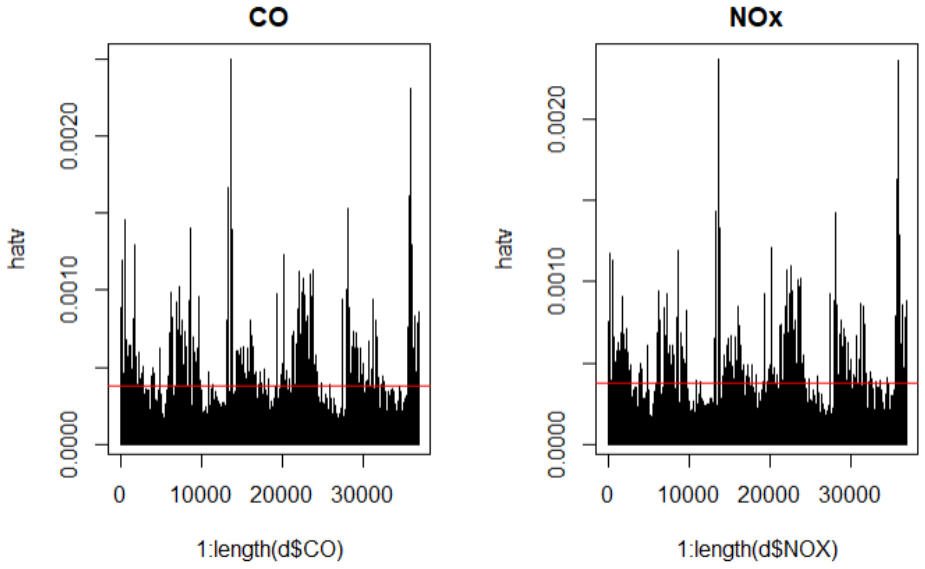
$$\begin{cases} h_{ii} \leq 2\bar{h} \\ h_{ii} > 2\bar{h} \end{cases}$$

On:

- $\bar{h} = \frac{k+1}{n}$

On si es compleix que és més gran implica que presenta influència potencial, altrament no.

Primerament es graficaran les observacions sota la condició anterior.



Il·lustració X-1: Observacions que presenten influència potencial

Les observacions que es troben per sobre de la línia vermella, corresponen a aquelles que presenten un *leverage* superior al llindar establert i, en conseqüència, mostren influència potencial.

10.1.2 Observacions atípiques

Els *outliers* o observacions atípiques (Faraway, 2015) són aquelles observacions que es comporten de manera diferent a la resta.

Els residus estudentitzats són una metodologia que permet estandarditzar el residu per la seva desviació estàndard en lloc de per la del terme de pertorbació:

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}_\epsilon(1-h_{ii})}} = \frac{e_i}{\sqrt{\frac{e_i{}'e_i}{n-k}(1-h_{ii})}}$$

On en el numerador es troba el residu (e_i) de l'estimació associat a l'observació i-èssima i, al denominador la desviació estàndard del terme de pertorbació estimat.

Per decidir si els valors són *outliers* o no, es segueix el següent criteri:

$$\begin{cases} t_i \leq t_{n-k;\frac{\alpha}{2}} \\ t_i > t_{n-k;\frac{\alpha}{2}} \end{cases}$$

Si el residu estudentitzats es troba per sobre d'aquesta *t-Student* comporta que l'observació i-èssima és atípica.

Description: df[,3] [10 x 3]

	rstudent <dbl>	unadjusted p-value <dbl>	Bonferroni p <dbl>
29590	-13.864568	1.3378e-43	4.9142e-39
7468	-12.773124	2.7815e-37	1.0217e-32
7465	-12.063293	1.9095e-33	7.0141e-29
7474	-11.965105	6.2331e-33	2.2896e-28
33730	-11.448841	2.6828e-30	9.8547e-26
7476	-10.833586	2.6245e-27	9.6404e-23
33688	-10.029481	1.2133e-23	4.4567e-19
14180	-9.841205	7.9838e-23	2.9327e-18
7473	-9.422586	4.6489e-21	1.7077e-16
7786	-9.357321	8.6261e-21	3.1686e-16

1-10 of 10 rows

Taula X-1: Observacions presenten residu alt, p-valor i p-valor corregit per Bonferroni

A partir de la funció `outlierTest()` retorna l'observació, el valor del residu estudentitzat, el *p-valor*, i el *p-valor corregit* per *bonferroni*, d'aquelles que mostren un residu en valor absolut major a 2.

Finalment fent el mateix pel segon model (*Taula XV, Apèndix II*), les observacions que es mostren són dades que han de tenir una atenció especial perquè difereixen notablement de la resta. Sí es conegués els motius exactes d'aquestes diferències es podria prescindir d'elles, però sinó se sap amb una certesa absoluta és millor no eliminar-les.

10.1.3 Observacions amb Influència Real

Una observació té influència real (Faraway, 2015) quan té un major efecte en l'ajust que la resta d'observacions.

La distància de *Cook* (DC_i) i el *DFFITS* serveixen per detectar si l'observació i-èssima presenta influència real.

La distància es calcula fent:

$$DC_i = \frac{\frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})}{K}}{\frac{e'e}{(N-K)}} = \frac{\frac{(\hat{\beta} - \hat{\beta}_{(i)})'X'X(\hat{\beta} - \hat{\beta}_{(i)})}{K}}{\frac{e'e}{(N-K)}}$$

On:

- $\hat{Y}_{(i)}$ és l'estimació de la variable resposta sense incloure l'observació i-èssima.
- $\hat{\beta}_{(i)}$ és l'estimació dels paràmetres sense incloure l'observació i-èssima.
- $e'e$ és la suma de quadrats dels errors (SQE) del model estimat per MQO amb totes les observacions.

Tot i que, es pot reescriure de la manera següent:

$$DC_i = \frac{r_i^2}{K}\frac{h_{ii}}{1 - h_{ii}}$$

On:

- r_i és el residu estudentitzat de la i-èssima observació i h_{ii} el seu leverage.

Així doncs,

$$\begin{cases} DC_i \leq F_{k;N-K;\frac{\alpha}{2}} \\ DC_i > F_{k;N-K;\frac{\alpha}{2}} \end{cases}$$

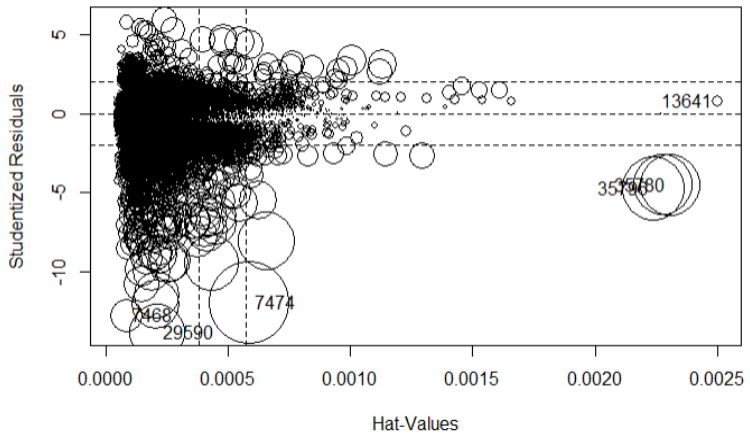
Fent ús de la funció `influencePlot()` permet obtenir la taula següent:

	StudRes <dbl>	Hat <dbl>	CookD <dbl>
7468	-12.7731239	8.510664e-05	0.0019750739
7474	-11.9651054	5.865602e-04	0.0119570643
13641	0.8140811	2.499692e-03	0.0002372546
29590	-13.8645681	2.106547e-04	0.0057560153
35780	-4.5530672	2.307378e-03	0.0068453988
35796	-4.7501635	2.241137e-03	0.0072361387

6 rows

Taula X-2: Observacions amb influència real

Les observacions que apareixen presenten un residu estudentitzat elevat, un *leverage* alt i una distància de *Cook* elevada. Per tant, les observacions 7468, 7474, 13641, 29590, 35780, 35796, i fent el mateix pel segon model (*Taula XV-2, Apèndix II*), les 3615, 11359, 13641, 14417 i 35780 presenten influència real.



Il·lustració X-2: Observacions amb influència real

Tot seguit, en l'eix de les ordenades del gràfic es troben els residus estudentitzats (t_i), i les línies discontínues horitzontals mostren quan $t = 2$ *i* $t = -2$. Aquells valors que es troben fora d'aquest interval són possibles *outliers*. A més, en l'eix de les abscisses es troben els *hat-values* i, les dues línies verticals es projecten calculant $2pn$ i $3pn$, respectivament.

El gràfic representa en forma de cercles cada observació de la variable resposta, i la distància de *Cook* influeix en el radi dels cercles, és a dir, que és proporcional. Com més gran sigui el radi, més significatiu/influència té.

Per veure el gràfic referent a la variable objectiu *NOx*, *Apèndix II*.

XI Anàlisi de la Redundància Aplicat

En aquest apartat es vol posar de relleu el mètode d'anàlisi multivariant explicat anteriorment, fent ús de les fórmules plantejades i alhora resseguint l'esquema mostrat inicialment. És a dir, es farà una comparativa dels valors obtinguts usant les funcions de l'R amb els càlculs fets manualment i veure si coincideixen.

11.1 Regressió Lineal Multivariant Múltiple

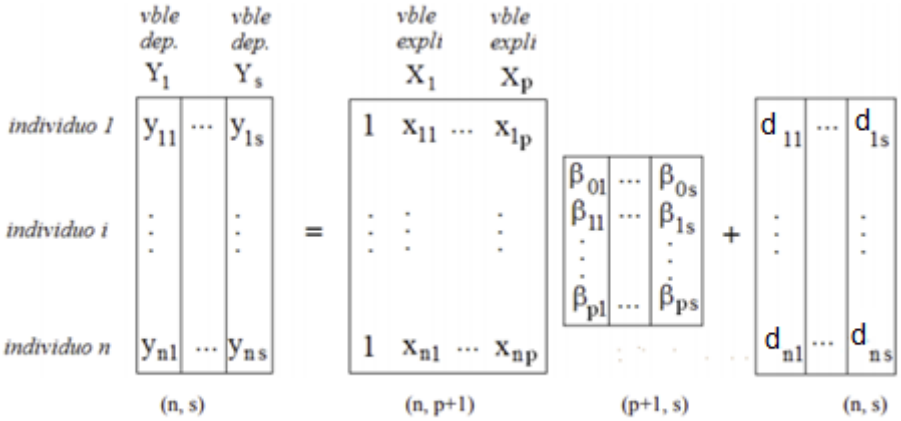
Es considera el següent problema de modelització que consisteix a estudiar les relacions entre s respostes $Y_1, \dots, Y_s$ i una col·lecció de p variables explicatives $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{p}$.

La principal diferència amb els models dels apartats anteriors és que $Y, B\ i\ \mathrm{D}$ ara són matrius i abans eren vectors.

Aquests models (Christensen, 2019) assumeixen que cada variable resposta segueix la següent expressió:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{i1} + \cdots + \beta_{pj}X_{ip} + d_{ij} \quad \forall j = 1, \dots, s$$

Matricialment,



Il·lustració XI-1: Forma matricial del model lineal multivariant múltiple

Per tant, el model multivariant és

$$Y = XB + D$$

On:

- Y és la matriu coneguda d'observacions.
- X és la matriu de disseny

- B és la matriu de paràmetres desconeguts
- D és la matriu desconeguda de termes perturbadors
- $\theta \equiv XB$ és la matriu desconeguda de mitjanes.
- $E(d_i) = 0 \; \forall i$
- $Cov(d_i, d_j) = \sigma_{ik} I \;\; \forall \, i, k = 1, \dots, s$

El model multivariant juxtaposa S models múltiples, un per cada variable de Y. El submodel j-èssim es forma amb les columnes j de $Y, B \; i \; D$:

$$y^j = X\beta^j + d^j, j = 1, \dots, s$$

Però amb una particularitat rellevant, com ara, que els s residus d_i d'un individu estan correlacionats.

El primer pas per a dur a terme l'anàlisi de la redundància és resoldre l'equació de la regressió multivariant.

L'estimació es fa sota el criteri dels mínims quadrats. Per tant, es tracta de trobar la matriu de $\hat{B}$ que minimitza la següent expressió:

$$RSS = \sum_{i=1}^{n} d_i^2 = d'd = \left(Y - \hat{Y}\right)'\left(Y - \hat{Y}\right) = (Y - XB)'(Y - XB) = (Y' - X'B')(Y - XB)$$

$$= Y'Y - Y'XB - B'X'Y + B'X'XB = Y'Y - 2B'X'Y + B'X'XB$$

Així doncs,

$$\min RSS = min \sum_{i=1}^{n} d_i^2 = \min(Y'Y - 2B'X'Y + B'X'XB)$$

Si es deriva matricialment respecte B:

$$\frac{d(RSS)}{d(B)} = \frac{d(d'd)}{d(B)} = \frac{d(Y'Y - 2B'X'Y + B'X'XB)}{d(B)} = -2X'Y + 2X'XB$$

Si $\frac{d(RSS)}{d(B)} = 0 \iff -2X'Y + 2X'XB = 0 \implies \boldsymbol{X'XB = X'Y}$.

D'aquesta manera s'obté les equacions normals del model.

A més, per comprovar que realment és un mínim és imperatiu calcular la segona derivada i veure si el resultat és positiu:

$$\frac{d^2(RSS)}{d^2(B)} = \frac{d^2(Y'Y - 2B'X'Y + B'X'XB)}{d^2(B)} = \frac{d^2(-2X'Y + 2X'XB)}{d^2(B)} = \boldsymbol{2X'X \geq 0}$$

Aquesta expressió és positiva obtenint així un mínim, ja que $X'X$ és sempre definida positiva donat que a la diagonal d'aquest producte de matrius hi ha la suma de quadrats de les observacions.

Per tant, si el rang és màxim llavors $X'X$ té inversa i l'única solució de les equacions normals és:

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

I el càlcul de la matriu de les observacions estimades pel model i la matriu d'errors associats és:

$$\hat{Y} = X\hat{B} = X(X'X)^{-1}X'Y \quad i \quad D = Y - \hat{Y} = [I - (X'X)^{-1}X']Y$$

Per comprovar si el rang de la matriu X és màxim o no és necessari definir el conjunt de variables explicatives que es faran servir. La selecció de variables feta a l'apartat 9.3 hi ha 5 variables que coincideixen i dos que no pels dos models. Per tant, s'ha considerat partir d'aquelles que tenen en comú *(AT, AH, AP, AFDP i TAT)* i afegir aquelles dues diferents *(TIT, GTEP)*.

```{r}
qr(matriu_disseny)$rank
```

```
[1] 7
```

Taula XI-1: Rang de la matriu de disseny

Com que el rang de la matriu de disseny coincideix amb el nombre de variables explicatives sí que es poden resoldre les equacions normals, obtenint així unes estimacions i prediccions úniques. Si aquest rang no fora màxim, s'hauria d'utilitzar una *g-inversa* per tal d'obtenir una solució, la qual sí que proporcionaria prediccions úniques però no pas estimacions exactes.

Tot seguit, tal i com s'indica a l'apartat III és necessari centrar les variables, tan explicatives com les resposta.

Una manera de centrar-les és fent ús de la funció `scale()`, però aquí es vol mostrar una alternativa usant la funció `apply()`.

```{r}
center_apply <- function(k) {
  apply(k, 2, function(y) y - mean(y))
}
x <- center_apply(matriu_disseny)
head(x)
```

```
            AT       AP         AH      AFDP       TAT      GTEP        TIT
[1,] -15.75953 7.029835  7.1179845 -1.395118 -1.238517 -5.447801 -32.728084
[2,] -16.49363 7.029835  9.6559845 -1.531818  2.341483 -6.979801 -35.928084
[3,] -16.76358 9.129835  0.4679845 -1.146618  3.791483 -3.299801 -12.628084
[4,] -16.70523 8.629835 -0.9250155 -1.108518  3.471483 -2.205801  -6.228084
[5,] -16.42693 8.529835 -1.1350155 -1.087818  3.521483 -2.080801  -5.228084
[6,] -15.88083 8.629835 -1.4560155 -1.084518  3.761483 -2.068801  -5.028084
```

Taula XI-2: Codi per centrar la matriu de disseny

Aquesta taula mostra les 6 primeres observacions de la matriu de disseny centrada, la qual s'ha obtingut creant la funció `center_apply()` que per cada columna de la matriu X li resta la seva mitjana.

I el mateix per la matriu de respostes.

	CO	NOX
[1,]	5.07663175	47.95693
[2,]	4.09593175	46.72693
[3,]	1.26103175	22.85393
[4,]	0.82473175	21.78493
[5,]	0.01083175	17.22193
[6,]	-0.29126825	15.89993

Taula XI-3: Matriu resposta centrada

Per calcular les prediccions utilitzant les fórmules explicades anteriorment cal resoldre les equacions normals.

```{r}
x_t_x <- t(x)%*%x #x'x
xtxi <- solve(t(x) %*% x) #(x'x)^-1
betes <- xtxi %*% t(x) %*% y # hat{B}
pred <- x %*% betes # hat{Y}
```

Il·lustració XI-2: Càlcul de predicció del RDA

I finalment, també és necessari calcular la matriu dels residus, és a dir, D. Obtenint així els següents valors:

	CO	NOX
[1,]	1.7300875	30.509326
[2,]	0.4559727	27.627120
[3,]	0.1445416	5.759637
[4,]	0.4966913	6.044634
[5,]	-0.1105269	2.086081
[6,]	-0.3507403	1.388087

Taula XI-4: Matriu residual

Els 6 primers residus són propers a zero a excepció de les dues primeres observacions per la variable *NOx*.

11.1.1 Contrastos estadístics multivariants

Per provar la significació dels diferents paràmetres associats a les variables explicatives, és a dir, per testejar $H_0: B = 0$ s'han proposat diferents estadístics multivariants (JOHNSON, 2007).

Una taula *MANOVA* és una generalització de les taules `anova()` usades a la regressió múltiple univariant. En conseqüència, els càlculs són extensions de l'enfocament del model lineal general. La major diferència entre aquests testos és que la situació univariant només s'usa la prova estadística F, en canvi, per la multivariant es proporcionen diverses proves alternatives i/o complementàries.

Per a dur-les a terme és necessari definir prèviament dues matrius:

1. D matriu d'errors:

$$D = \left(Y - \hat{Y}\right)^t \left(Y - \hat{Y}\right).$$

   I seguidament es pot trobar l'estimador màxim versemblant de la matriu de covariàncies dels residus:

$$\hat{\Sigma} = \frac{D}{n - p - 1}$$

   On p és el rang de la matriu de disseny.

2. H s'anomena matriu d'hipòtesis donat que el seu càlcul està subjecte a aquestes.

   Sigui $H_0: B_j = 0$ permet subdividir la matriu dels paràmetres i la de disseny de la manera següent:

$$B = \begin{bmatrix} B_j \\ B_l \end{bmatrix}, X = \left[X_j \middle| X_l\right]$$

   Aquesta nomenclatura dona lloc a analitzar un model subjecte a la hipòtesi nul·la, és a dir, un model reduït.

$$G = X_l B_l + D$$

   A partir d'aquest model es pot calcular la matriu de paràmetres i estimacions:

$$\underline{\hat{B}_l} = (X_l^t X_l)^{-1} X_l^t G \quad i \quad \underline{\hat{G}} = X_l \underline{\hat{B}_l}$$

   I finalment la matriu d'hipòtesis s'obté mitjançant la diferència entre la matriu d'errors reduïda i la matriu d'errors completa, és a dir,

$$H = D_0 - D \ (> 0)$$

S'assumeix que les matrius $H\ i\ D$ tenen $h\ i\ d$ graus de llibertat, respectivament. Sigui $\theta_i, \phi_i\ i\ \lambda_i$ els valors propis de $H(D+H)^{-1}, HD^{-1}$ i $D(D+H)^{-1}$, respectivament. Aquests valors es troben relacionats entre sí de la següent manera:

$$\theta_i = 1 - \lambda_i = \frac{\phi_i}{1+\phi_i}\ ;\ \phi_i = \frac{\theta_i}{1-\theta_i} = \frac{1-\lambda_i}{\lambda_i}\ ;\ \lambda_i = 1-\theta_i = \frac{1}{1+\phi_i}$$

Obtenint així les següents definicions:

- **La traça de Pillai:** s'utilitza com a estadística de prova i el seu valor és positiu i oscil·la entre 0 i 1. L'augment de valors significa que els efectes contribueixen més al model i s'hauria de rebutjar la hipòtesi nul·la per a valors grans.

$$\mathrm{V} = tr(H(H+D)^{-1}) = \sum_{i=1}^{l} \frac{\phi_i}{1+\phi_i}\ ; on\ l = \min(p,h)$$

  I l'estadístic F en aquest cas és:

$$F_{l(2m+l+1),l(2n+l+1)} = \frac{(2n+l+1)V}{(2m+l+1)(l-V)}$$

  On

$$m = \frac{(|p-h|-1)}{2}$$
$$n = \frac{(e-p-1)}{2}$$

  La seva aplicació en el programari R es troba usant la funció `anova()`:

```
Analysis of Variance Table

               Df  Pillai approx F num Df den Df    Pr(>F)
(Intercept)     1 0.98269  1042250      2  36724 < 2.2e-16 ***
AT              1 0.35717    10202      2  36724 < 2.2e-16 ***
AP              1 0.00275       51      2  36724 < 2.2e-16 ***
AH              1 0.03928      751      2  36724 < 2.2e-16 ***
AFDP            1 0.29541     7699      2  36724 < 2.2e-16 ***
GTEP            1 0.20643     4776      2  36724 < 2.2e-16 ***
TIT             1 0.40557    12528      2  36724 < 2.2e-16 ***
TAT             1 0.07136     1411      2  36724 < 2.2e-16 ***
Residuals   36725
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taula XI-5: Contribució marginal de les variables explicatives

  S'observa com els valors de l'estadístic (columna 2) estan fitats entre 0 i 1. A més, és un clar indicador de que hi ha variables, tot i que, els seus paràmetres associats sí siguin significatius al 5% es puguin prescindir d'elles. Com ara les variables *pressió ambiental (AP), humitat ambiental (AH) i temperatura després de la turbina (TAT)* ja que tenen valors molt propers a zero.

- **Lambda de Wilks:** Estadístic que el seu rang de valors oscil·la entre 0 i 1, essent variables que sí o no contribueixen a la millora del model, respectivament.

$$\Lambda = \frac{|D|}{|H+D|} = \prod_{i=i}^{s} \frac{1}{1+\phi_i}$$

I l'estadístic F associat és:

$$F_{ph,ft-g} = \frac{(ft-g)\left(1-\Lambda^{\frac{1}{t}}\right)}{ph\Lambda^{\frac{1}{t}}}$$

On

$$f = e - \frac{1}{2}(p-h+1)$$

$$g = \frac{ph-2}{2}$$

$$t = \begin{cases} \sqrt{\dfrac{p^2h^2-4}{p^2+h^2-5}} & si\ p^2+h^2-5>0 \\ 1 & Altrament \end{cases}$$

Els seus valors per les diferents variables es troben usant la funció `manova()` i especificant el test que es vol, en aquest cas, *test = "Wilks".*

```
Analysis of Variance Table

             Df   Wilks approx F num Df den Df    Pr(>F)
(Intercept)   1 0.01731  1042250      2  36724 < 2.2e-16 ***
AT            1 0.64283    10202      2  36724 < 2.2e-16 ***
AP            1 0.99725       51      2  36724 < 2.2e-16 ***
AH            1 0.96072      751      2  36724 < 2.2e-16 ***
AFDP          1 0.70459     7699      2  36724 < 2.2e-16 ***
GTEP          1 0.79357     4776      2  36724 < 2.2e-16 ***
TIT           1 0.59443    12528      2  36724 < 2.2e-16 ***
TAT           1 0.92864     1411      2  36724 < 2.2e-16 ***
Residuals 36725
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taula XI-6: Contribució de les variables explicatives mitjançant Wilks

La interpretació serà quelcom diferent donat que ara les variables que menys afavoreixen el model, són aquelles que presenten una Λ *de Wilks* propera a 1. Per tant, la *pressió ambiental (AP), humitat ambiental (AH) i temperatura després de la turbina (TAT)* són variables les qual la variància d'aquestes s'explica a partir de la variable independent, quan l'òptim seria tot el contrari.

- **Traça de Hotelling-Lawly:** És un estadístic de valor positiu que indica a valors més grans una major contribució dins del model. Es defineix com la suma dels valors propis, és a dir,

$$T_g^2 = d\sum_{i=1}^{l}\phi_i = \sum_{i=1}^{l}\frac{1-\lambda_i}{\lambda_i};\ on\ l = \min(p,h)$$

I l'estadístic F associat és:

$$F_{a,b} = \frac{T_g^2}{cd}$$

On

$$a = ph$$

$$b = 4 + \frac{(a+2)}{(B-1)}$$

$$c = \frac{a(b-2)}{b(e-p-1)}$$

$$B = \frac{(e+h-p-1)(e-1)}{(e-p-3)(e-p)}$$

Fent ús de la mateixa funció usada anteriorment canviant només *trace = "Hotelling-Lawly"* s'observa el següent:

```
Analysis of Variance Table

             Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
(Intercept)   1           56.761  1042250      2  36724 < 2.2e-16 ***
AT            1            0.556    10202      2  36724 < 2.2e-16 ***
AP            1            0.003       51      2  36724 < 2.2e-16 ***
AH            1            0.041      751      2  36724 < 2.2e-16 ***
AFDP          1            0.419     7699      2  36724 < 2.2e-16 ***
GTEP          1            0.260     4776      2  36724 < 2.2e-16 ***
TIT           1            0.682    12528      2  36724 < 2.2e-16 ***
TAT           1            0.077     1411      2  36724 < 2.2e-16 ***
Residuals 36725
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taula XI-7: Contribució de les variables explicatives mitjançant Hotelling-Lawly

A diferència dels altres dos els valors no estan fitats dins de cap interval. A més, cal veure que aquest estadístic és pràcticament igual a la *Traça de Pillai* per les variables *AP, AH i TAT*. Això indica una vegada més que aquestes tres variables no contribueixen de manera notable al model.

Finalment es comprova que independentment del test utilitzat indica que les variables *pressió ambiental (AP), humitat ambiental (AH) i temperatura després de la turbina (TAT)* no presenten una contribució notable pel model. Es per això, que més endavant pot ser interessant fer una

selecció de variables per veure si és necessari continuar amb aquest, o bé es pot simplificar una mica més.

11.2 Aplicació dels Estadístics del RDA

11.2.1 Coeficient de determinació i el coeficient ajustat

Per calcular el coeficient de determinació s'utilitzarà la formula explicada en l'apartat 3.1.1.

```{r}
tss <- sum(y^2)
rss <- sum(e^2)
(r2 <- 1- rss/tss)
```

```
[1] 0.4434068
```

Taula XI-8: Coeficient de determinació

Un dels inconvenients que presenta aquest estadístic es que és molt susceptible a l'hora d'incloure noves variables, ja que aquest valor sempre serà quelcom més gran.

És per això que s'usa una alternativa que exclou aquest inconvenient. A partir de l'expressió de l'apartat 3.1.2:

```{r}
(r2_ajustat <- 1- (((nrow(y)-1)/(nrow(y)-ncol(matriu_disseny)-1))*(1-r2)))
```

```
[1] 0.4433007
```

Taula XI-9: Coeficient de determinació ajustat

Un cop obtinguts aquests valors es comparen amb els que torna l'R.

```{r}
RsquareAdj(RDA)
```

```
$r.squared
[1] 0.4434068

$adj.r.squared
[1] 0.4433007
```

Taula XI-10: R^2 i R^2_{adj} usant R

Es veu clarament que s'obtenen els mateixos valors. Tanmateix també cal esmentar que aquesta diferència tan petita és un indicador que informa per prescindir o afegir alguna variable. Tot i així, més endavant es farà una selecció per dur a terme el *RDA* ja que amb els models lineals múltiples s'ha vist que sí que esqueia una reducció d'explicatives.

11.2.2 Significació global

Tot seguit s'ha descrit la significació global de la regressió multivariant. La qual tracta de contrastar si la regressió global calculada és significativa respecte el model sense variables explicatives.

```{r}
num <- r2/ncol(matriu_disseny)
den <- (1-r2)/(nrow(d)-ncol(matriu_disseny)-1)
(f_conjunt <- num/den)
```

```
[1] 4179.538
```

Taula XI-11: L'estadístic F conjunt

Tot i que, la distribució d'aquest estadístic no és coneguda, el càlcul del p-valor s'ha de fer utilitzant testos de permutacions.

Però mitjançant la funció `anova()` de l'objecte anomenat *RDA* ja conté aquest test.

```
Permutation test for rda under reduced model
Permutation: free
Number of permutations: 999

Model: rda(formula = y ~ x)
          Df Variance      F Pr(>F)
Model      7   62.744 4179.5  0.001 ***
Residual 36725   78.760
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taula XI-12: Anova del RDA

Bé, un cop visualitzada aquesta sortida es pot comprovar que el nombre de permutacions que duu a terme són 999 i que el valor de l'estadístic F coincideix. A més, el *p-valor* associat a aquest contrast és significatiu. Per tant, es pot procedir amb el càlcul del *PCA* de la matriu de variables resposta estimada.

11.2.3 Càlcul del AIC

Tal i com s'ha explicat a la teoria aquest estadístic es pot calcular de diverses maneres.

```{r}
k <- 7
(aic <- nrow(d)*log((1-r2)/nrow(d))+nrow(d)*log(tss)+2*k)
```

```
[1] 160404.2
```

Taula XI-13: Càlcul del AIC

Fent ús de la següent igualtat.

```r
(aic <- nrow(d)*log(rss/nrow(d))+2*k)
```

```
[1] 160404.2
```

Taula XI-14: Càlcul alternatiu del AIC

Aquest valor surt de tenir en compte que k és igual al rang de la matriu de disseny, és a dir, igual al nombre de variables explicatives més un. Aquest terme que es suma fa referència a la variància de l'error ja que també es desconegut.

I finalment el càlcul fet en R.

```r
extractAIC(RDA)
```

```
[1]      8.0 160406.2
```

Taula XI-15: Càlcul AIC amb R

Es comprova que independentment de la manera de calcular aquests estadístics s'obtenen els mateixos valors i per conseqüència s'arriben a les mateixes conclusions.

11.3 Anàlisi de les Components Principals per $\widehat{Y}$

L'anàlisi de les components principals *(PCA)* és una tècnica d'anàlisi multivariant que consisteix en reduir la dimensionalitat d'una matriu de dades, en aquest cas, de la matriu de respostes estimades $(\hat{Y})$ mitjançant la regressió multivariant.

Aquest mètode (Rodrigo, 2017) consisteix en construir una combinació lineal de variables que millor representin les estimacions $\hat{Y}_1, \dots, \hat{Y}_s$. Siguin $Z_1, \dots, Z_m$ les combinacions lineals de les s respostes estimades, és a dir,

$$Z_m = \sum_{i=1}^{s} \phi_{im} \hat{Y}_j$$

On $\phi_{1m}, \dots, \phi_{sm}$ són les constants de les components principals. Aquests valors informen sobre el pes que té una variable en cada component.

Normalment és imperatiu estandarditzar les variables, però en aquest cas no serà necessari ja que les dues variables resposta estan mesurades en les mateixes unitats. Tal i com s'indica a la taula V-1.

Prèviament una de les preguntes que un es pot fer alhora de realitzar un *PCA* és quanta informació es capaç de capturar cada una de les components principals obtingudes? I per donar resposta a aquesta qüestió s'usaran les següents definicions:

- La variància total es calcula a partir de la suma de la diagonal de la matriu de covariàncies de Y o bé, fent la suma de la variabilitat associada a les noves variables. És a dir,

$$V_T = \sum_{i=1}^{s} Var(Z_i) = \sum_{i=1}^{s} \lambda_i$$

- La variància explicada per la component s-èssima:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{s} \phi_{jm}\hat{y}_{ij}\right)$$

- La variància acumulada per la mateixa component:

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{s} \phi_{jm}\hat{y}_{ij}\right)^2}{\sum_{j=1}^{s}\sum_{i=1}^{n} \hat{y}_{ij}^2}$$

Tot seguit es procedeix a calcular el *PCA* de la matriu de valors ajustats per tal d'obtenir un vector de valors propis canònics fent ús de la funció `prcomp()`.

```
Importance of components:
                          PC1     PC2
Standard deviation     7.7623 1.57790
Proportion of Variance 0.9603 0.03968
Cumulative Proportion  0.9603 1.00000
```

Taula XI-16: Nombre components principals per $\hat{Y}$

La primera component principal (Z_1) és aquella que reflecteix la major variabilitat de les dades. A més, en aquesta taula es mostra com s'ha reduït la dimensionalitat, ja que només amb la primera s'explica un 96.03% de la variabilitat total de les prediccions.

Tot seguit el nombre màxim d'eixos canònics es pot obtenir a partir del resum fet anteriorment:

```
Nombre eixos canònics
[1] 60.253959  2.489755
```

Taula XI-17: Nombre màxim d'eixos

S'observen dos, els quals sí que compleixen les restriccions. És a dir, és menor al rang de les variables predictores, a la màxima dimensió no restringida de n objectes i és igual al nombre de variables resposta.

També és important destacar que aquests càlculs no reflecteixen els eixos no canònics com a representació dels residus. Per tant, per a seguir l'esquema descrit a l'apartat 3.2, es calculen les components principals dels errors usant la mateixa funció que abans.

```
Importance of components:
                            PC1     PC2
Standard deviation     8.7611 1.41555
Proportion of Variance 0.9746 0.02544
Cumulative Proportion  0.9746 1.00000
```

Taula XI-18: Resum dels eixos no canònics

Amb la *PC1* ja s'explica un 97.46% de la variabilitat total dels residus. I el nombre d'eixos no canònics:

```
Nombre eixos no canònics
[1] 76.756214  2.003784
```

Taula XI-19: Nombre d'eixos no canònics

S'observen dos i compleixen les restriccions ja que és més petit o igual que la dimensió de y i que el nombre d'observacions menys u.

Per observar les anteriors taules de forma més compacte s'usaran les següents instruccions:

- Primerament es calcula la matriu de covariàncies per tal de trobar els valors propis *(eigenvalues)* i els vectors propis *(eigenvectors)*.

```
             CO        NOX
CO   5.119683   9.000263
NOX  9.000263 136.384028
```

Taula XI-20: Matriu de covariàncies

Sigui A la matriu de covariàncies, el valors propis es troben solucionant la següent expressió:

$$det(A - \lambda I) = 0$$

Essent I la matriu identitat.

$$|A - \lambda I| = \left| \begin{pmatrix} 5.119683 & 9.000263 \\ 9.000263 & 136.384028 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right|$$

$$= \begin{vmatrix} 5.119683 - \lambda & 9.000263 \\ 9.000263 & 136.384028 - \lambda \end{vmatrix}$$

$$= (5.119683 - \lambda)(136.384028 - \lambda) - (9.000263^2)$$

$$= \lambda^2 - 141.503711\lambda + 617.23825 = 0$$

I resolent el polinomi característic,

$$\lambda_1 = 136.9982652 \; i \; \lambda_2 = 4.505445777$$

I els valors obtinguts amb R

```
[1] 136.998266   4.505446
```

Taula XI-21: Valors propis

Es comprova que coincideixen. I per calcular els vectors propis, s'ha de resoldre $(A - \lambda I)\eta = 0$ per cada valor propi. Obtenint així la següent matriu:

```
           [,1]       [,2]
[1,] 0.0680882 -0.9976793
[2,] 0.9976793  0.0680882
```

Taula XI-22: Vectors propis

- Seguidament pot ser interessant comprovar el valor de la proporció de variància i l'acumulada usant les fórmules esmentades abans:

```
```{r}
proportion_variance <- c(var(a$x[,1])/(var(a$x[,1]) + var(a$x[,2])),
                         var(a$x[,2])/(var(a$x[,1]) + var(a$x[,2])))
cumulative_proportion <- c(var(a$x[,1])/(var(a$x[,1]) + var(a$x[,2])),
                           var(a$x[,1])/(var(a$x[,1]) + var(a$x[,2]))+
                             var(a$x[,2])/(var(a$x[,1]) + var(a$x[,2])))
round(rbind(proportion_variance,cumulative_proportion), 5)
```
```

```
                          [,1]    [,2]
proportion_variance    0.96032 0.03968
cumulative_proportion 0.96032 1.00000
```

Taula XI-23: Proporció de variància i acumulada

S'obtenen els mateixos valors que a la taula X-16. Per trobar la variabilitat dels eixos no canònics el procediment és el mateix.

Finalment per generalitzar, incloure i posar tots els eixos i fórmules explicades anteriorment i alhora sintetitzar tota aquesta informació, pot ser útil construir la següent taula.

```
                          [,1]    [,2]     [,3]    [,4]
valors_propis          60.25396 2.48975 76.75621 2.00378
proprcio_var_explicada  0.42581 0.01759  0.54243 0.01416
propro_var_acumulada    0.42581 0.44341  0.98584 1.00000
```

Taula XI-24: Valors propis, proporció de variància i l'acumulada

La qual també es pot obtenir usant la funció `rda()` del paquet *vegan* i aquí s'adjunta una part del resum que genera.

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

Partitioning of variance:
              Inertia Proportion
Total          141.50     1.0000
Constrained     62.74     0.4434
Unconstrained   78.76     0.5566

Eigenvalues, and their contribution to the variance

Importance of components:
                          RDA1    RDA2     PC1     PC2
Eigenvalue            60.2540 2.48975 76.7562 2.00378
Proportion Explained   0.4258 0.01759  0.5424 0.01416
Cumulative Proportion  0.4258 0.44341  0.9858 1.00000
```

Taula XI-25: Resum de la contribució de la variància usant rda()

S'obtenen les mateixes taules i els mateixos valors. Tot i que fent servir aquesta funció el resum retorna la quantitat de variància restringida (62.74) i la no restringida (78.76). Aquest primer valor s'obté fent la suma de les variàncies de les components principals, i el darrer calculant la diferència de la variància total respecte la suma anterior. A més també cal esmentar que la proporció de variància restringida coincideix amb el coeficient de determinació calculat a l'apartat (10.2.1).

Els valors propis canònics, en aquest cas *RDA1* i *RDA2*, mesuren les quantitats de variància explicada pel model *RDA*, mentre que els valors propis residuals, en aquest cas *PC1* i *PC2*, mesuren les quantitats de variància representades pels eixos residuals però no explicats per cap model.

Per concloure s'ha pogut copsar que els valors propis estan en ordre decreixent i coincideixen amb els calculats de forma manual, tot i que hi ha un valor que no segueix aquest patró. El primer valor propi de la part no restringida és superior a l'últim valor propi de la part restringida. Això implica que l'estructura residual té més variància que una part de l'estructura que es pot explicar per les regressores. Per tant, és un clar indicador de que aquest model presenta marge de millora. Per exemple fent una selecció de variables, incorporant interaccions, etc.

11.4 Selecció de Variables

Tal i com s'ha vist a l'apartat 10.1.1 hi ha diverses variables que són candidates a contribuir poc a la millora del model. I, tenint en compte que els valors propis haurien d'estar en ordre decreixent són dues propostes que indiquen que una possible selecció de variables és oportuna.

En primer lloc s'estudia si el conjunt de variables escollit a l'apartat 10.1 presenta o no multicol·linealitat.

```
```{r}
vif.cca(RDA)
```
```

xAT	xAP	xAH	xAFDP	xTAT	xGTEP	xTIT
2.146895	1.321139	1.550018	2.470074	4.624270	14.008793	9.065257

Taula XI-26: Multicol·linealitat del model RDA

Es demostra que hi ha un problema greu de multicol·linealitat ja que hi ha valors que presenten uns *FIV's* superior a 5 i 10.

Tenint en compte els diferents mètodes de selecció explicats a l'apartat 9.3, aquí s'ha decidit utilitzar la funció `forward.sel()`, ja que no es tenen variables factors.

Els passos que desenvolupa aquesta funció són:

A. Calcula la importància de la prova global subjecte a totes les variables explicatives considerades.

B. Utilitza cada variable com a explicació en l'ordenació restringida i anota la variació explicada.

C. Ordena les variables segons les variacions explicades de forma decreixent.

D. Comprova a cada iteració si la variació explicada per la millor variable és significativa mitjançant la prova de permutació de *Monte-Carlo*. Si és així, la inclou, altrament s'atura.

E. Utilitza cadascuna de les variables explicatives restants i comprova quina variació explica si es posa com a explicativa (amb la variable ja seleccionada que actua com a covariable).

F. Les ordena de nou segons la variació decreixent explicada per elles i tria la que expliqui més. Si la variació és significativa s'afegeix al model.

G. Continua al pas E fins que la variació explicada per la millor variable deixi de ser significativa. És a dir, la regla d'aturada que s'usa és calcular la variació ajustada explicada pel model *RDA* i si durant la selecció el R^2_{adj} s'apropa al del model general, s'atura.

```r
R2a.all <- RsquareAdj(RDA)$adj.r.squared
forward.sel(y, x, adjR2thresh = R2a.all)
```

	variables	order	R2	R2Cum	AdjR2Cum	F	pval
	<S3: AsIs>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	AT	1	0.301384908	0.3013849	0.3013659	15845.8773	0.001
2	GTEP	6	0.039395074	0.3407800	0.3407441	2194.9896	0.001
3	AH	3	0.025911079	0.3666911	0.3666393	1502.7232	0.001
4	TIT	7	0.020558579	0.3872496	0.3871829	1232.2726	0.001
5	TAT	5	0.043163297	0.4304129	0.4303354	2783.1714	0.001
6	AP	2	0.010738847	0.4411518	0.4410605	705.7281	0.001
7	AFDP	4	0.002255055	0.4434068	0.4433007	148.7925	0.001

7 rows

Taula XI-27: Selecció variables RDA

S'observa que no es prescindeix de cap variable de les introduïdes, però tenint en compte el coeficient ajustat acumulat (columna 5) es veu que a partir de la variable 5 *(TAT)* la millora que presenta és pràcticament nul·la.

Cal esmentar que per obtenir un model alternatiu s'ha decidit que la diferència entre la variació explicada pel model global i la variació del model que es va generant sigui de 0.011 ja que la capacitat predictiva del model pràcticament no varia. El global presenta un $R^2_{adj} = 0.4433$ i el resultant un $R^2_{adj} = 0.4303$, essent aquesta diferència menyspreable, implica que el nombre de variables introduïdes inicialment es pot veure reduït.

```r
R2a.all <- RsquareAdj(RDA)$adj.r.squared
forward.sel(y, x, adjR2thresh = R2a.all, R2more = 0.011)
```

	variables	order	R2	R2Cum	AdjR2Cum	F	pval
	<S3: AsIs>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	AT	1	0.30138491	0.3013849	0.3013659	15845.877	0.001
2	GTEP	6	0.03939507	0.3407800	0.3407441	2194.990	0.001
3	AH	3	0.02591108	0.3666911	0.3666393	1502.723	0.001
4	TIT	7	0.02055858	0.3872496	0.3871829	1232.273	0.001
5	TAT	5	0.04316330	0.4304129	0.4303354	2783.171	0.001

5 rows

Taula XI-28: Model definitiu

A més, tal i com ja indicava prèviament l'anàlisi dels contrastos multivariants fets a l'apartat 10.1.1 s'ha prescindit de la variable *ambient pressure (AP)*. La variable *Air filter difference pressure (AFDP)* també ha estat eliminada donat que la seva contribució individual en termes del coeficient ajustat era menor.

Tot seguit, es calcula novament la multicol·linealitat d'aquest model.

```
     x2AT     x2GTEP       x2AH      x2TIT      x2TAT
1.555215 13.607197   1.409390   7.808360   4.012242
```

Taula XI-29: Multicol·linealitat

Amb aquestes variables com a regressores s'obté un model menys complexa, significatiu, i tot i que presenta una mica de multicol·linealitat en conjunt resolent els problemes de la regressió.

11.5 Ordenació per Y i X

Es possible representar les dades en els seus eixos canònics, en concret, en dimensió 2. Per tant, aquestes observacions es poden representar a l'espai de les variables resposta Y o bé en l'espai de la matriu X. Seguint l'esquema de l'apartat III aquesta ordenació es pot realitzar mitjançant la matriu F *o* Z, respectivament.

A partir del *PCA* realitzat a la matriu de resposta estimada i de l'obtenció dels valors propis, ara és necessari obtenir una matriu U de vectors propis canònics.

	PC1	PC2
CO	-0.07884	0.99689
NOX	-0.99689	-0.07884

Taula XI-30: Matriu de vectors propis canònics

Tot seguit pot ser quelcom il·lustratiu mostrar un gràfic de dues dimensions que mostri tant les variables resposta com les observacions o llocs en un espai definit pels eixos que es decideixi interpretar, en aquest cas, en dimensió 2.

```r
```{r}
set.seed(123)
f <- y %*% matriu_u
df1 <- data.frame(f) # PC1 and PC2
PC1 <- sample(df1$PC1,20,replace = FALSE)
PC2 <- sample(df1$PC2,20,replace = FALSE)
df1 <- data.frame(cbind(PC1,PC2))
rda.plot <- ggplot(df1, aes(x=PC1, y=PC2)) +
  geom_text(aes(label=rownames(df1)), size=4) +
  geom_hline(yintercept=0, linetype="dotted") +
  geom_vline(xintercept=0, linetype="dotted") +
  coord_fixed()
const <- attributes(summary(RDA))$const
df2 <- data.frame(matriu_u)*const # loadings for PC1 and PC2
rda.biplot <- rda.plot +
  geom_segment(data=df2, aes(x=0, xend=PC1-3, y=0, yend=PC2),
               color="red", arrow=arrow(length=unit(0.01,"npc"))) +
  geom_text(data=df2,
            aes(x=PC1,y=PC2,label=rownames(df2),
                hjust=0.5*(1-sign(PC1)),vjust=0.5*(1-sign(PC2))),
            color="red", size=4)
rda.biplot
```
```

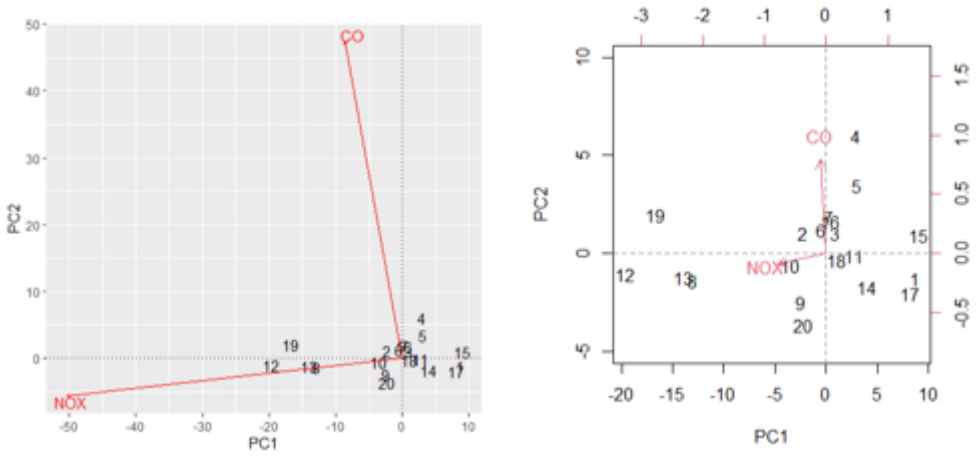
Il·lustració XI-3: Codi Biplot

Com que l'ordenació que s'ha programat ha estat per Y, inicialment s'ha calculat la matriu *F* ($F = YU$). Tot seguit s'ha desat aquesta com un *data.frame* anomenat *df1*. Aquest objecte conté les noves posicions de totes les observacions en termes de *PC1 i PC2*, però com hi ha moltes, s'ha decidit representar gràficament una submostra de només 20 per apreciar i explicar d'una manera més amena la seva interpretació.

Seguidament a partir de la funció `ggplot()` s'ha afegit aquest nou *data.frame* reduït i, als eixos se'ls ha anomenat *PC1 i PC2*.

Finalment, es crea un nou *data.frame (df2)* en el qual l'únic que s'afegeix és la matriu U multiplicada per una constant arbitrària que retorna el paquet vegan per representar millor les variables resposta (*vector vermell*).

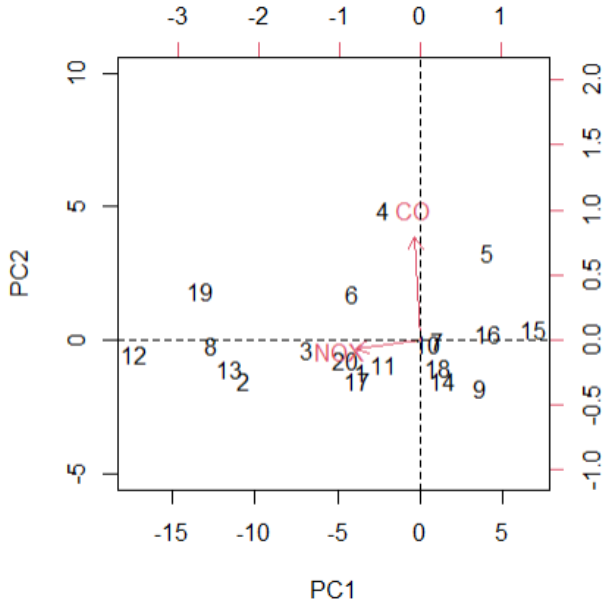
A continuació el gràfic reproduït a partir del codi anterior i l'obtingut amb R:



Il·lustració XI-4: Biplots per l'ordenació de Y

En termes generals, són els mateixos gràfics i s'interpreten igual. Però cal esmentar que partint del gràfic dret, l'eix vertical de l'esquerra i l'horitzontal inferior es fa servir per ordenar les observacions. En canvi, l'eix vertical de la dreta i l'horitzontal superior fa referència a l'ordenació per les variables resposta. Es pot apreciar que la longitud dels vectors resposta són menors que al gràfic de l'esquerra. Això es degut a que els vectors vermells no es multipliquen per cap constant.

En termes de X el procediment i el càlcul és el mateix, l'única diferència existent és que per aquesta ordenació en lloc de multiplicar la matriu resposta per la de vectors propis, es multiplica l'estimació de la resposta pels vectors propis, és a dir, $Z = \hat{Y}U$. I el gràfic que s'obté:



Il·lustració XI-5: Biplot ordenació per X

En principi no s'aprecien grans canvis independentment de l'ordenació feta, però si s'observa l'observació 4, es pot veure com aquesta està distorsionada. Això mostra que realitzant l'ordenació per Y, els punts obtinguts s'ajusten al màxim als valors reals de les dades observades, en canvi, per l'ordenació de X es visualitza la projecció de les dades ajustades per les diferents regressions múltiples, és a dir, que s'aproximen únicament als valors ajustats. Cal esmentar que qualsevol dels dos criteris aplicats serveixen de la mateixa manera alhora d'extreure les conclusions.

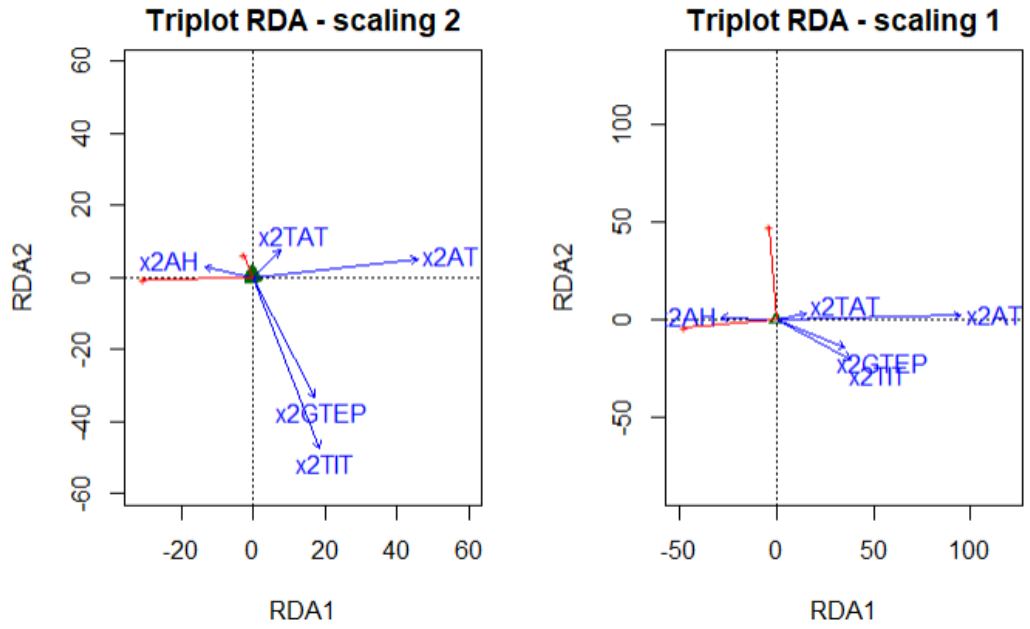
Per tant, si s'usa l'escala 1, s'estudia el gràfic de les distàncies. En conseqüència, la distància entre els objectes són aproximacions de les seves distàncies euclidianes i l'angle entre els dos vectors projectats manquen de significat.

Finalment utilitzant l'escala 2, la que usa l'R per defecte (Il·lustracions X-4 i X-5), el més rellevant és l'angle que separa el *monòxid de carboni (CO) i el NOx* donat que és pràcticament de 90º, i sabent que *cos(90º) = 0*, la correlació entre ambdues respostes és nul·la. Si l'objectiu és interpretar possibles relacions entre objectes s'escull l'escala 1, però si el interès principal es centra en les relacions entre els vectors, s'utilitzarà tal i com s'ha vist l'escala 2.

11.6 Ordenació conjunta

Per a representar tota la informació de manera més sintetitzada s'utilitzen els anomenats *triplots* donat que aquests poden representar en un mateix gràfic els conjunts d'observacions, variables resposta i variables explicatives.

En aquest cas en lloc d'usar el *PCA* realitzat anteriorment, s'usa l'objecte *RDA* que conté les variables resposta i les 5 variables regressores òptimes obtingudes mitjançant la selecció de variables.



Il·lustració XI-6: Triplots usant escala 1 i 2

En el codi emprat, si es vol utilitzar els *lc* per als llocs caldrà detallar en el paràmetre *display* el tipus per a les tres entitats fent *display = c ("sp", "lc", "cn")*. On *"Sp"* són les variables respostes, *"lc"* les observacions i *"cn"* les variables explicatives. Pel que fa a la interpretació, per a les observacions i les respostes és la mateixa que en el *PCA* amb els dos tipus d'escala. Per representar les variables explicatives en aquest gràfic s'usa la correlació entre la matriu d'explicatives i el producte entre la matriu de vectors propis canònics i els valors ajustats. A més també es diferencien dues escales:

- Escala 1: Es modifiquen les correlacions multiplicant cada variable *j* per $\sqrt{\frac{\lambda_j}{Variància\ de\ Y}}$.

- Escala 2: Els angles entre les diferents variables independentment que siguin explicatives o resposta representen la seva correlació i es mostren a partir del càlcul clàssic.

En primer lloc independentment de l'escala emprada, els gràfics presenten una similitud. Allò que els diferencia és la manera en que s'ha realitzat el càlcul de les correlacions, tal i com s'esmentava abans. La variable resposta *NOx* es troba correlacionada amb un alt nivell d'humitat ambiental *(AH)* ja que l'angle entre les dues variables és 0. I sabent que cos(0º) = 1, aquestes dues variables estan molt correlacionades. A més, temperatura ambient *(AH)* també mostra un angle proper a 180º, i sabent que *cos(180) = -sin(90) = -1*, implica que existeix una correlació prou forta negativa.

Mirant els dos *triplots* es pot extreure que les variables regressores humitat ambiental *(AH)* i temperatura ambient *(AT)* juguen un paper molt rellevant respecte la dispersió de les diferents observacions al llarg del primer eix canònic *(RDA1)*. En canvi, temperatura després de la turbina *(TAT)*, la temperatura d'entrada a la turbina *(TIT)* i la pressió exhaustiva de la turbina *(GTEP)* són les variables detonants a l'hora d'explicar la dispersió en el segon eix canònic *(RDA2)*. Per tant, en aquest gràfic es pot observar com el *RDA1* conté les variables ambientals i el segon eix les que fan referència al procediment de la turbina.

A més, si les observacions no estiguessin tan concentrades es podria veure amb quina/quines variable/es estarien caracteritzades. Però tenint en compte aquest cas, es pot concloure que les observacions estan presents o relacionades amb condicions ambientals intermèdies.

XII Conclusions

En aquest treball s'ha volgut mostrar com la regressió ordinària pot ser millorada mitjançant tècniques d'anàlisi multivariant.

A la pràctica les hipòtesis que sustenten els models lineals normalment no es verifiquen. Per tant, s'ha de tenir cura d'aquests fets.

Amb les dades estudiades les regressions no complien amb la normalitat esperada i tampoc amb l'homoscedasticitat. És per això que s'havia de pensar una solució alternativa, com ara, els models lineals generalitzats, però finalment es va descartar per tal de poder fer una comparació entre els models ordinaris i el *RDA*.

Per utilitzar aquests models hem provat diferents aproximacions, com per exemple la transformació òptima per cada variable resposta. En concret, es va decidir per *CO* i *NOx* usar el logaritme i la inversa, respectivament. Aquestes van produir uns residus més estables en termes d'esperança i variància, però amb l'inconvenient que les variables explicatives presentaven multicol·linealitat. Per tant, quan es manifesta aquest problema les estimacions en sí no es veuen afectades però sí els seus intervals fent-los més amplis.

Una de les solucions proposades va ser estudiar un model penalitzat, en concret, la regressió *Ridge*. Però aquest mètode no reduïa el problema de la multicol·linealitat i no es podia comparar directament amb el *RDA,* sinó amb un *RDA* penalitzat no estudiat. Tot seguit, per observar si era necessària la inclusió de totes les regressores, es va procedir fent servir la funció `bestsubset()`, i un algorisme ad-hoc que es va programar per tal d'obtenir uns subconjunts sense multicol·linealitat. El subconjunt obtingut en el cas del *NOx,* la capacitat predictiva $\left(R^2_{adj}\right)$ presenta una lleugera disminució. En canvi, per l'altre variable *(CO),* la capacitat predictiva del model resultant va augmentar respecte l'original.

Per tal d'aplicar el *RDA* es va construir el model multivariant considerant el conjunt de totes les explicatives obtingudes mitjançant la unió de les dues regressions ordinàries múltiples. I un cop creat l'objecte *rda* i a partir de la funció `anova()` s'ha conclòs que el model és significatiu, i en conseqüència, s'ha pogut dur a terme l'anàlisi de components principals per la matriu d'estimacions.

Cal esmentar que un dels punts més rellevants d'aquesta part ha estat que a partir dels valors propis obtinguts de la part no restringida i de la restringida, sumat als resultats dels contrastos multivariants, s'ha pogut copsar que el model tenia marge de millora. A més, com també presentava multicol·linealitat s'ha dut a terme una selecció de variables.

Per últim, i per concloure, hem volgut destacar l'aplicabilitat d'aquesta tècnica de la mateixa forma que s'ha vist els darrers anys demostrant l'eficiència que se'n deriva, ja que no només permet estudiar diferents variables resposta en un mateix model, sinó que proporciona, tot i tenir una mica de multicol·linealitat, un model menys complex donat que és capaç de prescindir de dues variables predictores: l'aire després de passar pel filtre de la turbina *(AFDP)* i de la pressió ambiental *(AP)*. A més, també és rellevant la interpretació trobada ja que només amb 5 variables explicatives s'ha pogut veure que el primer eix canònic s'explica mitjançant variables de caire ambiental i, el segon eix per les variables referents al procediment pre i post turbina.

XIII Bibliografia I Webgrafia

Borcard, D. (2011). *et al. Numerical Ecology with R.* Montréal, Québec: Springer.

Carmona, F. (2005). *Modelos Lineales. Barcelona: Publicacions i Edicions de la Universitat de Barcelona, 2005.*

Christensen, R. (2019). *Advanced Linear Modelling. Third Edition, University of New Mexico, Alburquerque USA .*

Faraway, J. J. (2015). *Linear Models with R, Second Edition.* Chapman & Hall/CRC Press.

GUSTA ME. *GUide to STatistical Analysis in Microbial Ecology*. Recollit de https://mb3is.megx.net/gustame/constrained-analyses/rda

JOHNSON, R. A. (2007). *Applied multivariate statistical analysis.* New York: Springer.

Johnson, R. A. (2007). *Applied multivariate statistical analysis, Sixth Edition.* Madison: Prentice Hall.

Kaya, H. (2019). *archive*. Recollit de http://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set

Lane, D. M. *et al. onlinestatsbook*. Recollit de https://onlinestatbook.com/2/index.html

Legendre, P. (2012). *et al. Numerical Ecology, Third English Edition.* Montréal, Québec: ELSEVIER.

Lodz University. (2017). *researchgate.net.* Recollit de https://www.researchgate.net/publication/323123713_On_the_Use_of_Redundancy_Analysis_to_Study_the_Property_Crime_in_Poland

Martínez, C. G. (Maig / 2018). *pubs*. Recollit de https://rstudio-pubs-static.s3.amazonaws.com/392216_5adc98e84ff246e3b6418f3426c8fe38.html#m%C3%A9todos_de_reducci%C3%B3n_dimensional

Misztal, M. (11 / 6 / 2017). *researchgate.net.* Recollit de https://www.researchgate.net/publication/323123713_On_the_Use_of_Redundancy_Analysis_to_Study_the_Property_Crime_in_Poland/fulltext/5d7be7b6299bf1d5a97d5d96/On-the-Use-of-Redundancy-Analysis-to-Study-the-Property-Crime-in-Poland.pdf

Peres-Neto. (2006). *et al. Variation partitioning of species data matrices: estimation and comparison of fractions .* Ecology.

Rawlings, J. O. (1998). *Applied Regression Analysis: a research tool.* New York: Springer.

Rodrigo, J. A. (Junio / 2017). *rpubs.* Recollit de https://rpubs.com/Joaquin_AR/287787

S., J. C.-R. (21 / 10 / 2020). *bookdown.* Recollit de https://bookdown.org/stephi_gascon/bookdown-demo-master_-_multivariant/_book/ordination.html

XIV Apèndix I: Taules apartat 9.3

	Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
	<int>	<int>	<chr>	<dbl>	<dbl>	<dbl>
1	1	1	AT	0.311557731	0.311538988	15703.29773
6	2	1	TIT	0.045738440	0.045712460	35948.30138
5	3	1	GTEP	0.040654620	0.040628502	36335.48901
2	4	1	AP	0.036840085	0.036813863	36626.00700
4	5	1	AFDP	0.035437062	0.035410801	36732.86233
9	6	1	CDP	0.029328595	0.029302168	37198.08792
3	7	1	AH	0.027098899	0.027072412	37367.90324
8	8	1	TEY	0.013485592	0.013458734	38404.70329
7	9	1	TAT	0.008610257	0.008583266	38776.01262
13	10	2	AT GTEP	0.342612056	0.342576260	13340.17624

1-10 of 511 rows Previous 1 2 3 4 5 6 ... 52 Next

Taula XIV-1: Contribució de cada variable per a construir un model

Best Subsets Regression

Model Index	Predictors
1	AT
2	AT GTEP
3	AT AH GTEP
4	AT TIT TAT TEY
5	AT AH TIT TAT TEY
6	AT AP AH TIT TAT TEY
7	AT AP AH AFDP TIT TAT TEY
8	AT AP AH AFDP TIT TAT TEY CDP
9	AT AP AH AFDP GTEP TIT TAT TEY CDP

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.3118	0.3115	0.3115	15703.2977	271095.5647	166850.4495	271121.0990	3449048.2022	93.9002	0.0026	0.6885
2	0.3426	0.3426	0.3425	13340.1762	269402.0744	165156.5963	269436.1201	3293557.8497	89.6694	0.0024	0.6575
3	0.3691	0.3691	0.369	11323.5228	267892.3559	163646.6227	267934.9130	3160851.2841	86.0587	0.0023	0.6310
4	0.4641	0.4640	0.4639	4093.1934	261902.0100	157657.2513	261953.0786	2685148.7266	73.1090	0.0020	0.5361
5	0.5031	0.5031	0.5029	1119.7296	259123.5836	154879.4924	259183.1637	2489471.5844	67.7831	0.0018	0.4970
6	0.5168	0.5167	0.5165	84.5844	258104.7448	153860.9797	258172.8363	2421305.6490	65.9289	0.0018	0.4834
7	0.5176	0.5175	0.5173	22.4135	258042.6411	153798.9000	258119.2440	2417149.6661	65.8176	0.0018	0.4826
8	0.5178	0.5177	0.5174	12.1254	258032.3531	153788.6177	258117.4674	2416407.0113	65.7991	0.0018	0.4825
9	0.5178	0.5177	0.5175	10.0000	258030.2268	153786.4940	258123.8525	2416201.3824	65.7953	0.0018	0.4824

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error
HSP: Hocking's Sp

Taula XIV-2: Models òptims segons les variables incloses i justificació analítica

Description: df[,6] [84 x 6]

	Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
	<int>	<int>	<chr>	<dbl>	<dbl>	<dbl>
398	382	6	AT AP AH TIT TAT TEY	0.5167646	0.5166856	84.58438
431	383	6	AT AH TIT TAT TEY CDP	0.5055352	0.5054544	939.82247
423	384	6	AT AH AFDP TIT TAT TEY	0.5039388	0.5038578	1061.40551
427	385	6	AT AH GTEP TIT TAT TEY	0.5031527	0.5030715	1121.27818
399	386	6	AT AP AH TIT TAT CDP	0.4873448	0.4872610	2325.22174
424	387	6	AT AH AFDP TIT TAT CDP	0.4867531	0.4866692	2370.28478
428	388	6	AT AH GTEP TIT TAT CDP	0.4858353	0.4857513	2440.18332
412	389	6	AT AP GTEP TIT TAT TEY	0.4684051	0.4683182	3767.68358
416	390	6	AT AP TIT TAT TEY CDP	0.4680354	0.4679485	3795.83758
408	391	6	AT AP AFDP TIT TAT TEY	0.4678702	0.4677832	3808.42163

1-10 of 84 rows Previous 1 2 3 4 5 6 ... 9 Next

Taula XIV-3: Conjunt de models amb 6 explicatives

	Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
	<int>	<int>	<chr>	<dbl>	<dbl>	<dbl>
383	406	6	AT AP AH AFDP GTEP TAT	0.3995562	0.3994581	9011.269
389	422	6	AT AP AH AFDP TAT TEY	0.3859853	0.3858850	10044.833
390	423	6	AT AP AH AFDP TAT CDP	0.3858938	0.3857935	10051.803
386	438	6	AT AP AH AFDP TIT TAT	0.3532859	0.3531803	12535.247

4 rows

Taula XIV-4: Models amb 6 variables sense multicol·linealitat

XV Apèndix II: Taules apartat 9.5

	rstudent <dbl>	unadjusted p-value <dbl>	Bonferroni p <dbl>
3615	9.780186	1.4593e-22	5.3606e-18
11359	9.018552	1.9958e-19	7.3311e-15
35098	8.947771	3.7930e-19	1.3933e-14
14417	8.688186	3.8321e-18	1.4076e-13
5752	6.164056	7.1641e-10	2.6316e-05
14266	5.894747	3.7853e-09	1.3904e-04
14227	5.813498	6.1686e-09	2.2659e-04
14326	5.141603	2.7380e-07	1.0057e-02

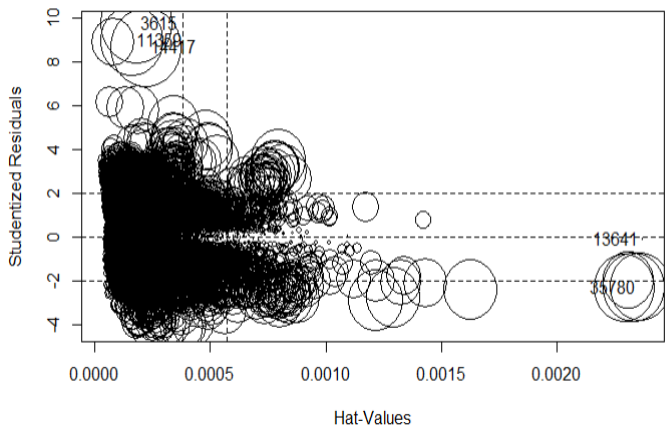
8 rows

Taula XV-1: Individus amb residu alt, el seu p-valor i l'ajustat associat

	StudRes <dbl>	Hat <dbl>	CookD <dbl>
3615	9.78018553	0.0001793690	2.445140e-03
11359	9.01855203	0.0001617121	1.875165e-03
13641	-0.06663243	0.0023692601	1.506357e-06
14417	8.68818583	0.0002198399	2.366368e-03
35780	-2.26362390	0.0023576591	1.729688e-03

5 rows

Taula XV-2: Individus amb influència real



Il·lustració XV-1: Individus amb influència real

XVI Apèndix III: Codi R

```
# llibraries
library(dplyr)
library(ggplot2)
library(GGally)
library(Hmisc)
library(corrplot)
library(PerformanceAnalytics)
library(olsrr)
library(nortest)
library(car)
library(knitr)
library(moments)
library(base)
library(MASS)
library(ModelMetrics)
library(glmnet)
library(stats)
library(rcompanion)
library(gghalfnorm)
library(faraway)
library(vegan)
library(olsrr)
library(stringr)
library(vegan)
library(ggplot2)
library(packfor)
library(rrcov)

# lectura base de dades
d <- read.csv2("gt_2015.csv",sep = ",",header = TRUE)
```

```
d$AT   <- as.numeric(d$AT)

d$AP   <- as.numeric(d$AP)

d$AH   <- as.numeric(d$AH)

d$AFDP <- as.numeric(d$AFDP)

d$GTEP <- as.numeric(d$GTEP)

d$TIT  <- as.numeric(d$TIT)

d$TAT  <- as.numeric(d$TAT)

d$TEY  <- as.numeric(d$TEY)

d$CDP  <- as.numeric(d$CDP)

d$CO   <- as.numeric(d$CO)

d$NOX  <- as.numeric(d$NOX)

head(d)

# resum

summary(d)

# taula resum

Nom_var <- names(d)

minim <- c(min(d$AT),min(d$AP),min(d$AH),min(d$AFDP),min(d$GTEP),min(d$TIT),min(d$TAT),min(d$TEY),min(d$CDP),min(d$CO),min(d$NOX))

maxim <- c(max(d$AT),max(d$AP),max(d$AH),max(d$AFDP),max(d$GTEP),max(d$TIT),max(d$TAT),max(d$TEY),max(d$CDP),max(d$CO),max(d$NOX))

mitjana <- c(mean(d$AT),mean(d$AP),mean(d$AH),mean(d$AFDP),mean(d$GTEP),mean(d$TIT),mean(d$TAT),mean(d$TEY),mean(d$CDP),mean(d$CO),mean(d$NOX))

sd <- c(sd(d$AT),sd(d$AP),sd(d$AH),sd(d$AFDP),sd(d$GTEP),sd(d$TIT),sd(d$TAT),sd(d$TEY),sd(d$CDP),sd(d$CO),sd(d$NOX))

kable(data.frame(Variables = c("Ambient temperature", "Ambient pressure","Ambient humidity","Air filter difference pressure","Gas turbine exhaust pressure", "Turbine inlet temperature","Turbine after temperature","Turbine energy yield","Compressor discharge pressure","Carbon monoxide","Nitrogen oxides"), Abre. = Nom_var, Unitats = c("ºC","mbar","(%)","mbar","mbar","ºC","ºC","MWH","mbar","mg/m^3","mg/m^3"),Min= round(minim,2),Max = round(maxim,2), Mean = round(mitjana,2), Des.Tip = round(sd,2)))

# histogrames vars. resposta

par(mfrow = c(1,2))

hist(d$CO, main = "Histograma del CO",xlab = "CO")
```

```
hist(d$NOX,main = "Histograma del NOx",xlab = "NOx")
# correlació
correlacio <- round(cor(d),2)
corrplot(correlacio, method="number", type="upper")
# creació models lineals
lmod <- lm(CO ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP, data = d)
#summary(lmod)
mod <- lm(NOX ~ AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP, data = d)
#summary(mod)
# normalitat
par(mfrow=c(2,2))
qqnorm(residuals(lmod),ylab="Residuals",main="CO")
qqline(residuals(lmod))
hist(residuals(lmod),xlab="Residuals",main="CO")
qqnorm(residuals(mod),ylab="Residuals",main="NOx")
qqline(residuals(mod))
hist(residuals(mod),xlab="Residuals",main="NOx")
# homocedasticitat
par(mfrow=c(2,2))
plot(lmod,c(1,3))
plot(mod,c(1,3))
# transformació
par(mfrow=c(1,2))
bm_co = boxcox(lmod,plotit = TRUE,lambda = seq(0,0.4,by = 0.01))
bm_nox = boxcox(mod,plotit = TRUE,lambda = seq(-0.5,-0.8,by = -0.01))
# lambdes òptimes
lamvalue <- c()
lamvalue <- c(lamvalue,bm_co$x[which(bm_co$y==max(bm_co$y))],
              bm_nox$x[which(bm_nox$y==max(bm_nox$y))])
names(lamvalue) <- c("Lmax CO", "Lmax NOx")
lamvalue
# reducció observacions per usar la funció transformTuckey
```

```
set.seed(123)
co_reduida <- sample(d$CO,5000,replace = FALSE)
transformTukey(co_reduida)
no_reduida <- sample(d$NOX,5000,replace = FALSE)
transformTukey(no_reduida)
transformTukey(co_reduida)
# tots els models possibles
models <- ols_step_all_possible(lmod)
head(models,10)
# els millors subconjunts
ols_step_best_subset(lmod)
# tria de les variables a usar
k <- ols_step_best_subset(lmod)
plot(k)
# possibles models amb 6 variables
models_6_variables <- models[models$n == 6,]
head(models_6_variables,10)
# algorisme per seleccionar models amb 6 vars i sense Multicol·linealitat
vif_efectiu <- c()
for(i in 1:nrow(models_6_variables)){
  var_exp <- unlist(strsplit(models_6_variables$predictors[i], " "))
  b <- vif(lm(d[,"CO"] ~ d[,var_exp[1]]+d[,var_exp[2]]+
                d[,var_exp[3]]+d[,var_exp[4]]+d[,var_exp[5]]+d[,var_exp[6]]))
  if((b[1]<=5) & (b[2]<=5) &  (b[3]<=5) & (b[4]<=5) & (b[5]<=5) & (b[6]<=5)){
    vif_efectiu <- c(vif_efectiu,i)
  }
}
models_6_variables[vif_efectiu,]
models2 <- ols_step_all_possible(mod)
head(models2,10)
ols_step_best_subset(mod)
k <- ols_step_best_subset(mod)
```

```
plot(k)
models2_6_variables <- models2[models2$n == 6,]
models2_6_variables
vif_efectiu <- c()
for(i in 1:nrow(models2_6_variables)){
  var_exp <- unlist(strsplit(models2_6_variables$predictors[i], " "))
  b <- vif(lm(d[,"CO"] ~ d[,var_exp[1]]+d[,var_exp[2]]+
                d[,var_exp[3]]+d[,var_exp[4]]+d[,var_exp[5]]+d[,var_exp[6]]))
  if((b[1]<=5) & (b[2]<=5) &  (b[3]<=5) & (b[4]<=5) & (b[5]<=5) & (b[6]<=5)){
    vif_efectiu <- c(vif_efectiu,i)
  }
}
models2_6_variables[vif_efectiu,]
# segmentació BBDD mitjançant crosvalidació
n <- nrow(d)
learn <- 1:round(0.67*n)
nlearn <- length(learn)
ntesting <- n - nlearn
training <- d[1:nlearn,] # training
testing <- d[24612:nrow(d),]# testing
# ridge regression
x_vars <- as.matrix(training[,-c(10,11)]) #es treuen les variables respostes
y_var <- training$CO
set.seed(123)
# busca la millor lambda sense posar un interval de lambdes
cv_output <- cv.glmnet(x_vars, y_var, alpha = 0)
# # alpha = 1 lasso i alpha entre zero i u = elastic regression
cv_output$lambda.min
best_lam <- cv_output$lambda.min
ridge_best <- glmnet(x_vars, y_var, alpha = 0, lambda = best_lam)
ypred <- predict(ridge_best, s = best_lam , newx = as.matrix(testing[,-c(10,11)]))
co <- rmse(ypred, testing$CO)
```

```
x_vars <- as.matrix(training[,-c(10,11)])
y_var <- training$NOX
set.seed(123)
cv_output <- cv.glmnet(x_vars, y_var, alpha = 0)
cv_output$lambda.min
best_lam <- cv_output$lambda.min
ridge_best <- glmnet(x_vars, y_var, alpha = 0, lambda = best_lam)
ypred <- predict(ridge_best, s = best_lam , newx = as.matrix(testing[,-c(10,11)]))
nox <- rmse(ypred, testing$NOX)
# rmse dels dos models ridge
f <- c(co,nox)
#row.names(f) <- "RMSE"
names(f) <- c("CO","NOx")
f
# definició dels models resultants
model_co <- lm(log(CO) ~ AT + AP + AH + AFDP + TIT + TAT,d)# potser s'ha d'ampliar
model_no <- lm(1/NOX ~ AT + AP + AH + AFDP + GTEP + TAT,d)
# observacions amb un alt leverage
par(mfrow=c(1,2))
hatv <- hatvalues(model_co)
plot(1:length(d$CO),hatv,type='h', main="CO")
abline(h=2*(6+1)/length(d$CO),col='red')
hatv <- hatvalues(model_no)
plot(1:length(d$NOX),hatv,type='h',main="NOx")
abline(h=2*(6+1)/length(d$NOX),col='red')
# observacions outliers
outlierTest(model_co)
outlierTest(model_no)
# obs amb influència real
influencePlot(model_co)
influencePlot(model_no)
## Model Multivariant
```

```
### Traça de Pillai
mmlm <- lm(cbind(CO,NOX) ~ AT + AP + AH + AFDP + GTEP + TIT + TAT, data = d)
anova(mmlm)
# wilks
anova(mmlm,test="Wilks")
# hotelling-lawley
anova(mmlm,test="Hotelling-Lawley")
# rang matriu
matriu_disseny <- d[,c(1,2,3,4,7,5,6)]
matriu_resposta <- d[,c(10,11)]
qr(matriu_disseny)$rank
# centrar variables
center_apply <- function(k) {
  apply(k, 2, function(y) y - mean(y))
}
x <- center_apply(matriu_disseny)
head(x)
center_apply <- function(k) {
  apply(k, 2, function(y) y - mean(y))
}
y <- center_apply(matriu_resposta)
head(y)
# prediccions i erros
x_t_x <- t(x)%*%x #x'x
xtxi <- solve(t(x) %*% x) #(x'x)^-1
betes <- xtxi %*% t(x) %*% y # hat{B}
pred <- x %*% betes # hat{Y}
e <- y - pred
head(e)
# suma quadrats residuals i variància model
rss <- sum(e^2)
```

```
# calculo MSE:
mse <- rss/(length(y)-qr(matriu_disseny)$rank)
# vari?ncia del model:
var.model <- rss/(length(y)-qr(matriu_disseny)$rank)
var.model
# creació model RDA usant funcions
RDA <- rda(y~x)
# estadístics del RDA
tss <- sum(y^2)
rss <- sum(e^2)
(r2 <- 1- rss/tss)
(r2_ajustat <- 1- (((nrow(y)-1)/(nrow(y)-ncol(matriu_disseny)-1))*(1-r2)))
RsquareAdj(RDA)
num <- r2/ncol(matriu_disseny)
den <- (1-r2)/(nrow(d)-ncol(matriu_disseny)-1)
(f_conjunt <- num/den)
anova(RDA)
k <- 7
(aic <- nrow(d)*log((1-r2)/nrow(d))+nrow(d)*log(tss)+2*k)
(aic <- nrow(d)*log(rss/nrow(d))+2*k)
extractAIC(RDA)
# anàlisi components principals Y_est
pca_y_est <- prcomp(pred)
a <- summary(pca_y_est)
a
cat("Nombre eixos canònics","\n")
a$sdev^2
matriu_u <- pca_y_est$rotation # matriu U (matriu canònica de vectors propis)
round(pca_y_est$rotation, 5)
#anàlisi components principals errors
pca_y_residual <- prcomp(e)
b <- summary(pca_y_residual)
```

```
cat("Nombre eixos no canònics","\n")
b$sdev^2
matriu_u_residual <- pca_y_residual$rotation
round(pca_y_residual$rotation, 5)
(vartotalY <- sum(diag(cov(y))))# var(y[,1]) + var(y[,2])OK, però crec que no és correcte
#cov(y) == cor(y) matriun y no estandarditzada => aquesta igualtat no es compleix.
# matriu covariància de la y_centrada
(covariancia <- cov(y))
# valors i vectors propis
eigen <- eigen(covariancia)
eigen$values
eigen$vectors
# proproció de variància
proportion_variance <- c(var(a$x[,1])/(var(a$x[,1]) + var(a$x[,2])),
                         var(a$x[,2])/(var(a$x[,1]) + var(a$x[,2])))
cumulative_proportion <- c(var(a$x[,1])/(var(a$x[,1]) + var(a$x[,2])),
                           var(a$x[,1])/(var(a$x[,1]) + var(a$x[,2]))+
                            var(a$x[,2])/(var(a$x[,1]) + var(a$x[,2])))
round(rbind(proportion_variance,cumulative_proportion), 5)
valors_propis <- c(pca_y_est$sdev^2, pca_y_residual$sdev^2)
proprcio_var_explicada <- c(pca_y_est$sdev^2,pca_y_residual$sdev^2)/vartotalY
propro_var_acumulada <- cumsum(proprcio_var_explicada)
round(rbind(valors_propis,proprcio_var_explicada,propro_var_acumulada), 5)
# càlcul multicol·linealitat i selecció variables
vif.cca(RDA)
R2a.all <- RsquareAdj(RDA)$adj.r.squared
forward.sel(y, x, adjR2thresh = R2a.all)
R2a.all <- RsquareAdj(RDA)$adj.r.squared
forward.sel(y, x, adjR2thresh = R2a.all, R2more = 0.011)
# conjunt resultants a usar i RDA d'aquestes
x2 <- x[,c(1,6,3,7,5)]
RDA2 <- rda(y ~ x2)
```

```
vif.cca(RDA2)
# Creació gràfic Biplot ordenat per Y
set.seed(123)
f <- y %*% matriu_u#### ffffff
df1 <- data.frame(f) # PC1 and PC2
PC1 <- sample(df1$PC1,20,replace = FALSE)
PC2 <- sample(df1$PC2,20,replace = FALSE)
df1 <- data.frame(cbind(PC1,PC2))
rda.plot <- ggplot(df1, aes(x=PC1, y=PC2)) +
  geom_text(aes(label=rownames(df1)), size=4) +
  geom_hline(yintercept=0, linetype="dotted") +
  geom_vline(xintercept=0, linetype="dotted") +
  coord_fixed()
const <- attributes(summary(RDA))$const
df2 <- data.frame(matriu_u)*const # loadings for PC1 and PC2
rda.biplot <- rda.plot +
  geom_segment(data=df2, aes(x=0, xend=PC1-3, y=0, yend=PC2),
               color="red", arrow=arrow(length=unit(0.01,"npc"))) +
  geom_text(data=df2,
            aes(x=PC1,y=PC2,label=rownames(df2),
                hjust=0.5*(1-sign(PC1)),vjust=0.5*(1-sign(PC2))),
            color="red", size=4)
rda.biplot
#biplots amb R ordenació per Y
biplot(df1,matriu_u,xlab = "PC1", ylab = "PC2", ylim= c(-5,10))
abline(h=0,v=0, lty=2, col=gray(0.6))
# ordenació per X, biplot
set.seed(123)
z <- pred %*% matriu_u### ZZZZZ
df1 <- data.frame(z) # PC1 and PC2
PC1 <- sample(df1$PC1,20,replace = FALSE)
PC2 <- sample(df1$PC2,20,replace = FALSE)
```

```
df1 <- data.frame(cbind(PC1,PC2))
rda.plot <- ggplot(df1, aes(x=PC1, y=PC2)) +
  geom_text(aes(label=rownames(df1)), size=4) +
  geom_hline(yintercept=0, linetype="dotted") +
  geom_vline(xintercept=0, linetype="dotted") +
  coord_fixed()
const <- attributes(summary(RDA))$const
df2 <- data.frame(matriu_u)*const # loadings for PC1 and PC2
rda.biplot <- rda.plot +
  geom_segment(data=df2, aes(x=0, xend=PC1-3, y=0, yend=PC2),
               color="red", arrow=arrow(length=unit(0.01,"npc"))) +
  geom_text(data=df2,
            aes(x=PC1,y=PC2,label=rownames(df2),
                hjust=0.5*(1-sign(PC1)),vjust=0.5*(1-sign(PC2))),
            color="red", size=4)
rda.biplot
# biplot per ordenació X
biplot(df1,matriu_u,xlab = "PC1", ylab = "PC2", ylim= c(-5,10))
abline(h=0,v=0, lty=2,)
# Triplots
par(mfrow = c(1,2))
plot(RDA2, display=c("sp","lc","cn"),
     main="Triplot RDA - scaling 2",ylim =c(-1,1),xlim=c(-32,60))
spe2.sc <- scores(RDA2, choices = 1:2, display = "sp")
arrows(0, 0, spe2.sc[,1], spe2.sc[,2], length=0, lty=1, col="red")

plot(RDA2, scaling = 1,display=c("sp","lc","cn"),
     main="Triplot RDA - scaling 1",xlim = c(-50,120))
spe1.sc <- scores(RDA2, choices = 1:2, scaling = 1, display = "sp")
arrows(0, 0, spe1.sc[,1], spe1.sc[,2], length=0, lty=1, col="red")
```