Models de regressió i anàlisi multivariant amb R-Commander

Daniel Liviano Solís
Maria Pujol Jover

PID_00208269

Universitat Oberta de Catalunya

Índex

Introducció

Fins ara hem vist com s'utilitzen R i R-Commander per a fer una anàlisi univariant, és a dir, d'una sola variable. En aquest mòdul utilitzarem R i R-Commander per a l'anàlisi de dues o més variables (bivariant i multivariant, respectivament).

A grans trets, les tècniques d'anàlisi multivariant en sentit estricte, és a dir, de més d'una variable, es poden dividir atenent diversos criteris, i la combinació d'aquests criteris és precisament el que farà que triem la tècnica adequada per a resoldre un problema:

- **Relació entre les variables:** hi pot haver una relació de dependència o bé una interrelació entre les diferents variables amb què estiguem treballant.

- **Objectiu de l'estudi:** segons el problema plantejat estarem interessats a reduir variables agrupant les que expliquin conceptes similars o que ajudin a explicar un mateix concepte o bé a agrupar individus o observacions amb característiques similars.

- **Nombre de variables que intervenen en l'anàlisi:** podem treballar únicament amb dues variables (anàlisi bivariant) o amb més (anàlisi multivariant).

L'anàlisi bivariant no és més que un cas particular de l'anàlisi multivariant.

- **Naturalesa de les variables:** les variables utilitzades poden ser mètriques (numèriques discretes o contínues, dicotòmiques, procedents d'escales de Likert, qualitatives ordinals) o categòriques.

Així doncs, es pot deduir que hi ha múltiples tècniques que preveuen l'anàlisi de més d'una variable, encara que en aquest mòdul només en veurem algunes. En concret, les que veurem les hem dividit en tres capítols: en el primer es treballen els models de regressió; en el segon, l'anàlisi de la variància i les taules de contingència; i en el tercer, l'anàlisi de components principals i l'anàlisi clúster.

Objectius

1. Estimar i interpretar els resultats que llança R-Commander quan elaborem un model de regressió lineal simple (MRLS) o un model de regressió lineal múltiple (MRLM).

2. Utilitzar R-Commander per a fer una anàlisi de la variància (ANOVA), que no és més que un contrast d'igualtat de tres o més mitjanes poblacionals ($\mu_1, \mu_2, \ldots, \mu_k$).

3. Obtenir taules de contingència (TC) o fer el contrast de la χ^2 amb R-Commander.

4. Calcular i interpretar tots els resultats obtinguts amb R-Commander d'una anàlisi factorial de components principals (AFCP), incloent-hi el gràfic de bivariant.

5. Calcular i interpretar tots els resultats obtinguts amb R-Commander d'una anàlisi clúster o de conglomerats, incloent-hi el dendrograma.

1. Models de regressió

1.1. Introducció

En aquest capítol desenvoluparem el model de regressió, en el qual estudiem la relació que s'estableix entre dues o més variables. En aquest punt, convé fer una distinció fonamental entre dos conceptes relacionats però diferents:

- **Correlació:** aquest concepte fa referència al grau de relació que hi ha entre dues variables, però no estableix cap mena de relació de causa ni d'efecte de l'una sobre l'altra. L'indicador de correlació més simple és el coeficient de correlació lineal de Pearson, que indica en quina mesura la relació lineal entre dues variables és directa, inversa o nul·la.

- **Model de regressió:** en fer aquest model, suposem que no solament hi ha correlació entre les variables, sinó que a més hi ha una relació de *causalitat*, és a dir, una o més variables influeixen en una altra.

Hi ha dos tipus de models de regressió:

- **Model de regressió lineal simple (MRLS):** estudia el comportament d'una variable en funció d'una altra. Formalment es defineix com a $y = f(x)$.

- **Model de regressió lineal múltiple (MRLM):** estudia el comportament d'una variable en funció de més d'una variable. Formalment es defineix com a $y = f(x_1, x_2, \ldots)$.

Per a més informació sobre l'MRLM vegeu el material associat a l'assignatura d'*Econometria*.

1.2. Model de regressió lineal simple (MRLS)

En el model de regressió simple es vol estudiar el comportament de la variable explicada Y en funció de la variable explicativa X. Per a estimar el signe i la magnitud d'aquesta relació es pren una mostra de dimensió N, és a dir, s'obtenen N observacions de les variables X i Y.

La variable explicada

Aquesta variable també es pot denominar endògena, dependent o variable per explicar.

La variable explicativa

Aquesta variable es pot denominar de diferents maneres alternatives: exògena, independent o regressor.

El model que s'ha d'estimar és el següent:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = 1, \ldots, N.$$

En aquesta equació, α és la constant, β és la pendent i ε és una variable aleatòria denominada *terme d'error o de pertorbació*. En ser una equació teòrica que engloba tota la

població, els paràmetres α i β són desconeguts. Així doncs, l'objectiu de l'estimació del model serà poder fer inferència sobre aquest. Per tant, el primer pas serà obtenir els coeficients estimats dels paràmetres ($\hat{\alpha}$ i $\hat{\beta}$) a partir dels valors mostrals de X i Y. Un cop obtinguts, el model estimat tindrà l'expressió següent:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$$

La diferència entre els valors mostrals de la variable dependent (Y_i) i els seus valors estimats per la recta ($\hat{Y}_i$) són els *residus* o *errors* de l'estimació:

$$e_i = Y_i - \hat{Y}_i$$

Així doncs, el model estimat també es pot expressar de la manera següent:

$$Y_i = \hat{\alpha} + \hat{\beta}X_i + e_i$$

És important no confondre el terme de pertorbació ε amb els residus (e). El primer concepte és teòric i no observable, mentre que el segon depèn de la mostra i del mètode d'estimació escollit, de manera que és mesurable i analitzable.

Una bona estimació d'un model, és a dir, amb un bon ajust, serà la resultant de valors de e_i reduïts i distribuïts normalment. Així doncs, com més petits siguin els e_i millor serà l'estimació del model i més fiables seran les prediccions sobre el comportament de Y obtingudes amb aquesta estimació.

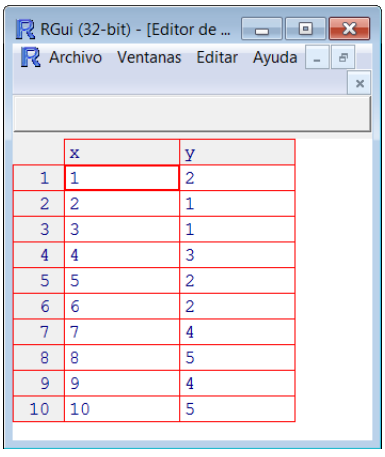
Com a exemple, suposem que disposem de $N = 10$ observacions de les variables X i Y:

X	1	2	3	4	5	6	7	8	9	10
Y	2	1	1	3	2	2	4	5	4	5

En R-Commander utilitzarem la ruta següent per a introduir aquestes dades:

Dades / Nova taula de dades

Un cop especificat un nombre per a aquest conjunt de dades, les introduïm en un full en què cada columna és una variable, tal com es mostra a continuació.



	x	y
1	1	2
2	2	1
3	3	1
4	4	3
5	5	2
6	6	2
7	7	4
8	8	5
9	9	4
10	10	5

Com hem vist en el mòdul dedicat l'anàlisi descriptiva, és recomanable iniciar l'anàlisi amb estadístics bàsics de les variables. Una primera explotació estadística s'obté seguint la ruta:

Estadístics / Resums / Taula de dades activa

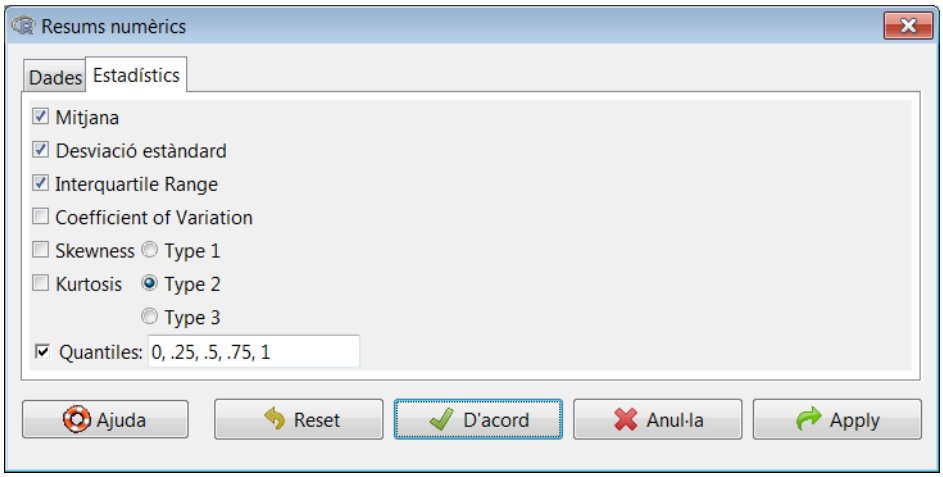
Amb això, el resultat serà el següent:

```
> summary(Dades)
       x                y
 Min.   : 1.00   Min.   :1.0
 1st Qu.: 3.25   1st Qu.:2.0
 Median : 5.50   Median :2.5
 Mean   : 5.50   Mean   :2.9
 3rd Qu.: 7.75   3rd Qu.:4.0
 Max.   :10.00   Max.   :5.0
```

Sovint no en tindrem prou amb els estadístics bàsics i voldrem obtenir mesures addicionals com l'asimetria, la curtosi, el coeficient de variació, la desviació típica o alguns quantils. Per a això hi ha una opció en la qual es pot triar entre un conjunt d'estadístics. Per a accedir a aquesta opció, la ruta que hem de seguir és la següent:

Estadístics / Resums / Resums numèrics

Obtindrem el menú següent, en el qual seleccionarem de les variables desitjades els estadístics que vulguem obtenir.



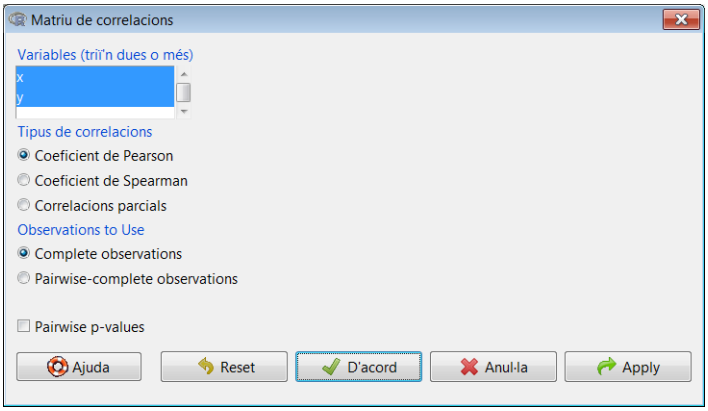
Aquest és el resultat que apareix en la finestra de resultats:

```
> numSummary(Dades[,c("x", "y")], statistics=c("mean", "sd",
+ "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
  mean       sd IQR 0%  25% 50%  75% 100%  n
x  5.5 3.027650 4.5  1 3.25 5.5 7.75   10 10
y  2.9 1.523884 2.0  1 2.00 2.5 4.00    5 10
```

Un estadístic rellevant quan es treballa amb més d'una variable és el coeficient de correlació lineal de Pearson. Per a calcular-lo, cal seguir la ruta següent:

Estadístics / Resums / Matriu de correlacions

Apareixerà el quadre de diàleg següent, en el qual seleccionarem les variables per a les quals volem calcular el coeficient de correlació:



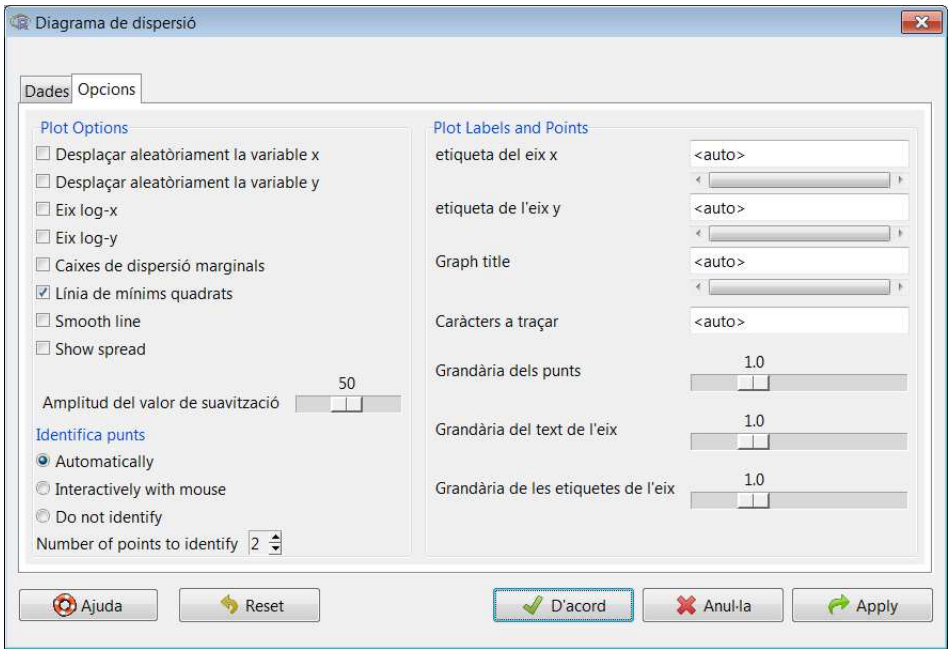
Com veiem, la correlació entre les dues variables és positiva i força elevada:

```
> cor(Dades[,c("x","y")], use="complete.obs")
          x         y
x 1.0000000 0.8549254
y 0.8549254 1.0000000
```

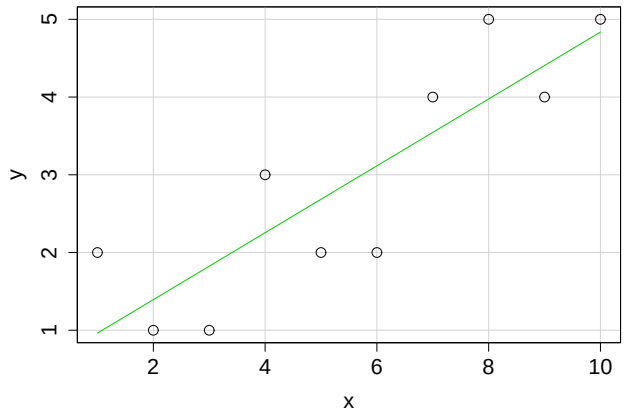
Visualment, la correlació entre dues variables es pot comprovar amb un diagrama de dispersió de les variables X i Y. Obtenir aquest gràfic en R-Commander és immediat accedint a la ruta següent:

Gràfics / Diagrama de dispersió

Apareixerà el menú següent, en què especificarem la variable x (corresponent a l'eix horitzontal) i y (corresponent a l'eix vertical). A més, activarem l'opció *Línia de mínims quadrats*, que dibuixa la recta de regressió sobre els punts.



En el gràfic resultant, les diferències verticals entre cada observació i la recta estimada són els residus (e_i). Com més reduïts siguin, millor serà l'ajust de l'estimació del model.



Recta de regressió

Fixeu-vos que, en aquesta recta de regressió estimada, el punt de tall amb l'eix vertical és $\hat{\alpha}$, mentre que el seu pendent és $\hat{\beta}$.

Entrant de ple en l'estimació del model de regressió, primer veurem com es calculen els coeficients del model i el seu coeficient de determinació (R^2) mitjançant codi de manera manual.

Les fórmules que hem d'aplicar són les següents:

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}$$

$$\hat{\alpha} = \bar{Y} - \bar{X}\hat{\beta}$$

$$R^2 = r^2 = \left(\frac{s_{xy}}{s_x s_y}\right)^2$$

Essent s la desviació estàndard, s^2 la variància, s_{xy} la covariància i r el coeficient de correlació lineal de Pearson. En R-Commander, fer aquests càlculs utilitzant la sintaxi del llenguatge propi d'R és immediat. N'hi ha prou de tenir en compte els operadors descrits en la taula 1, que ja hem vist en el primer mòdul.

Taula 1. Operadors estadístics bàsics amb R

Descripció	Instrucció	Resultat
Longitud	`length(x)`	`10`
Màxim	`max(x)`	`10`
Mínim	`min(x)`	`1`
Suma	`sum(x)`	`55`
Producte	`prod(x)`	`3628800`
Mitjana	`mean(x)`	`5.5`
Mediana	`median(x)`	`5.5`
Desviació estàndard	`sd(x)`	`3.02765`
Variància	`var(x)`	`9.166667`
Covariància	`cov(x,y)`	`3.944444`
Correlació	`cor(x,y)`	`0.8549254`
Producte escalar	`sum(x*y)`	`195`

Amb aquesta informació, calcularem $\hat{\alpha}$, $\hat{\beta}$ i R^2 introduint les fórmules respectives en la finestra d'instruccions, després seleccionarem el conjunt i premerem *Executar*:

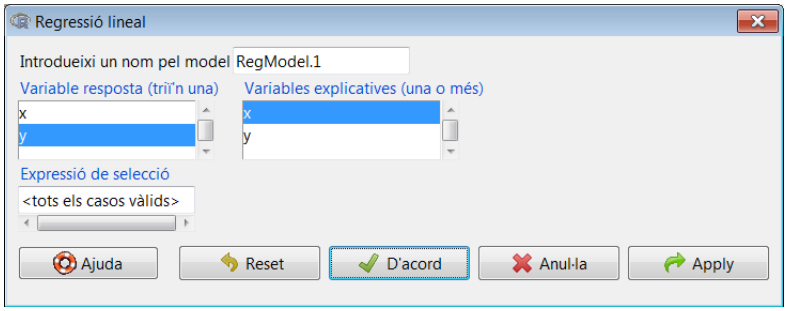
```
> attach(Dades)
> beta <- cov(x,y)/var(x)
> alpha <- mean(y)-beta*mean(x)
> coef.det <- cor(x,y)^2

> print(c(alpha,beta,coef.det))
[1] 0.5333333 0.4303030 0.7308975
```

Naturalment, R-Commander ofereix una manera més ràpida i immediata de calcular una recta de regressió, que a més inclou més informació estadística del model. Un cop les variables X i Y s'han introduït, cal estimar un model. La manera més senzilla és seguir aquesta ruta:

Estadístics / Ajustament de models / Regressió lineal

Apareixerà un quadre de diàleg en què especificarem quina és la variable dependent i la independent, a més d'introduir un nom per al model estimat (*RegModel.1*):



El resultat apareixerà en la finestra de resultats:

```
Call:
lm(formula = y ~ x, data = Dades)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1152 -0.6151 -0.1152  0.6727  1.0364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.53333    0.57278   0.931  0.37903
x            0.43030    0.09231   4.661  0.00162 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8385 on 8 degrees of freedom
Multiple R-squared: 0.7309, Adjusted R-squared: 0.6973
F-statistic: 21.73 on 1 and 8 DF,  p-value: 0.001621
```

Aquest resultat és un ampli sumari de la regressió. Vegem-ne els components principals:

- `Residuals`. Mínim, màxim i quartils dels residus de la regressió, els quals proporcionen informació sobre la seva distribució.

- `Coefficients`. Quadre en què apareix informació de l'estimació dels paràmetres (o coeficients) estimats.

- `Estimate`. Estimació de cada paràmetre (*intercept* significa constant).

- `Std.Error`. Desviació (o error) estàndard de cada paràmetre estimat.

- `t value`. Estadístic t de cada paràmetre estimat, obtingut dividint l'estimació del paràmetre entre la seva desviació estàndard. Aquest estadístic és el que utilitzem per a fer el contrast de significació individual dels paràmetres estimats.

- `Pr(>|t|)`. *p-valor* del contrast de significació individual de cada paràmetre estimat, el qual indica la seva significació estadística.

- `Signif. codes`. Mostra, amb asteriscos i punts, per a quins nivells de significació els coeficients estimats són significatius o no. En aquest cas, veiem que $\hat{\alpha} = 0,533$ no és significatiu i que $\hat{\beta} = 0,430$ és significatiu amb un nivell de significació de l'1% (`'**' 0.01`).

- `Residual standard error`. Desviació (o error) estàndard dels residus.

- `Multiple R − squared`. Coeficient de determinació.

- `Adjusted R − squared`. Coeficient de determinació ajustat.

- `F − statistic`. Estadístic F per al contrast de la significació global o conjunta dels paràmetres estimats del model.

- `p − value`. *p-valor* associat al contrast anterior. En aquest cas, veiem que el conjunt de paràmetres estimats és significatiu amb un nivell de significació de l'1% (*p-valor*$< 0,01$).

Una manera alternativa d'estudiar la significació individual dels paràmetres estimats és calculant intervals de confiança. Prenent un nivell de confiança del 95% (és a dir, una significació del 5%), hi ha una probabilitat del 95% que, per exemple, el paràmetre β estigui inclòs en l'interval següent:

$$\beta \in \left[\hat{\beta} \pm t_{0,025;\,8}\, s_{\hat{\beta}}\right].$$

En què $t_{0,025;\,8}$ és el valor en taules de l'estadístic t i $s_{\hat{\beta}}$ la desviació estàndard del coeficient estimat. R-Commander permet calcular conjuntament els intervals del confiança de tots els paràmetres estimats del model (en aquest cas dos). Un cop seleccionat el model, la ruta és la següent:

Models / Intervals de confiança

Apareixerà un quadre de diàleg en què cal especificar el nivell de confiança que volem, i en prémer *D'acord* obtindrem el resultat següent:

```
> Confint(RegModel.1, level=0.95)
                Estimate      2.5 %    97.5 %
(Intercept) 0.5333333 -0.7875075 1.8541742
x           0.4303030  0.2174303 0.6431758
```

Com que en el cas de la constant el valor zero està inclòs en l'interval de confiança (els extrems són de signe oposat), al 95% de confiança podem afirmar que el paràmetre estimat de la constant no és significatiu, la qual cosa equival a afirmar que no és estadísticament diferent de zero. Veiem que això no passa en el cas del pendent.

1.3. Model de regressió lineal múltiple (MRLM)

L'MRLM és una generalització del model simple a k variables explicatives i $k+1$ paràmetres (incloent-hi la constant). Per tant, la variable endògena s'explica per més d'una variable exògena:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon_i, \qquad i = 1, \ldots, N.$$

Considerem un exemple pràctic. Es vol explicar un model de regressió lineal per a estudiar els determinants del nivell d'atur a $N = 295$ municipis catalans:

$$ATUR_i = \beta_0 + \beta_1 MOTOR_i + \beta_2 RBFD_i + \beta_3 TEMP_i + \beta_4 UNIV_i + \varepsilon_i.$$

En què tenim les variables següents:

- *ATUR:* taxa d'atur.
- *MOTOR:* índex de motorització (turismes per habitant).
- *RBFD:* renda bruta familiar disponible, en milers d'euros.
- *TEMP:* taxa de temporalitat laboral.
- *UNIV:* percentatge d'estudiants universitaris a la població.

Les dades, un cop importades del document d'Excel, són les següents:

Dades

	MUNICIPI	ATUR	TEMP	UNIV	RBFD	MOTOR
1	Abrera	13.98	80.85	7.92	13197	5.663600
2	Aguilar de Segarra	6.25	87.50	10.34	13814	7.579377
3	Alella	8.93	82.86	26.09	21201	5.386800
4	Alpens	5.92	50.00	13.88	15920	4.501600
5	Ametlla del Vallès, L'	10.43	85.71	22.59	19649	5.407000
6	Arenys de Mar	15.10	74.14	13.49	13464	4.456800
7	Arenys de Munt	14.92	85.56	11.50	13999	4.840000
8	Argençola	4.96	75.00	10.34	11449	5.041700
9	Argentona	13.35	86.57	13.31	15267	5.137100
10	Artés	15.30	93.64	8.03	13814	4.982500
11	Avià	10.15	80.00	9.47	15512	5.408000

Abans de fer una estimació, és molt útil fer una descripció estadística de les variables. El resum del conjunt de dades és el següent:

```
> summary(Dades)
                  MUNICIPI        ATUR            TEMP
 Abrera               :  1   Min.   : 0.00   Min.   :  0.00
 Aguilar de Segarra   :  1   1st Qu.:10.85   1st Qu.: 80.00
 Aiguafreda           :  1   Median :13.63   Median : 85.75
 Alella               :  1   Mean   :13.42   Mean   : 82.74
 Alpens               :  1   3rd Qu.:16.12   3rd Qu.: 91.67
 Ametlla del Vallès, L':  1   Max.   :24.87   Max.   :100.00
 (Other)              :289
     UNIV            RBFD           MOTOR
 Min.   : 2.89   Min.   : 8.49   Min.   : 1.528
 1st Qu.: 7.57   1st Qu.:12.35   1st Qu.: 4.673
 Median :10.09   Median :13.81   Median : 5.025
 Mean   :10.98   Mean   :14.04   Mean   : 5.138
 3rd Qu.:13.04   3rd Qu.:15.51   3rd Qu.: 5.452
 Max.   :33.61   Max.   :24.96   Max.   :12.573
```

Per a veure més estadístics de les variables, es pot seguir la ruta següent i seleccionar-ne entre una llista d'estadístics:

Estadístics / Resums / Resums numèrics

El resultat és el següent:

```
> numSummary(Dades[,c("ATUR", "MOTOR", "RBFD", "TEMP", "UNIV")],
+ statistics=c("mean", "sd"))
           mean         sd   n
ATUR  13.418305  4.1859261 295
MOTOR  5.138357  0.9198514 295
RBFD  14.044332  2.4227029 295
TEMP  82.737356 16.5445532 295
UNIV  10.981864  4.9144900 295
```

Si dues o més variables tenen entre elles una correlació alta, pot ser problemàtic incloure-les simultàniament com a variables explicatives. Per això mateix, resulta molt útil calcular la matriu de correlacions lineals de les variables explicatives:

En concret, com es veurà en l'assignatura d'*Econometria*, una correlació alta entre dos regressors pot donar lloc a problemes de multicolinealitat.

Estadístics / Resums / Matriu de correlacions

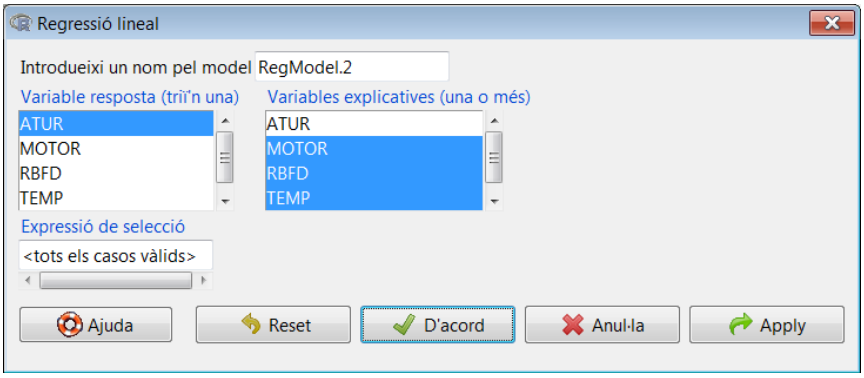
Seleccionant les variables que volem incloure, obtenim el resultat següent:

```
> cor(Dades[,c("ATUR","MOTOR","RBFD","TEMP","UNIV")],
+ use="complete")
         ATUR    MOTOR     RBFD     TEMP     UNIV
ATUR   1.0000 -0.38242 -0.41906  0.16408 -0.46244
MOTOR -0.3824  1.00000  0.05575 -0.25111 -0.02443
RBFD  -0.4190  0.05575  1.00000 -0.10259  0.58442
TEMP   0.1640 -0.25111 -0.10259  1.00000 -0.03479
UNIV  -0.4624 -0.02443  0.58442 -0.03479  1.00000
```

Anàlogament al cas de l'MRLS, la ruta següent ens permetrà estimar un model de regressió, seleccionant les variables explicada i explicatives:

Estadístics / Ajustament de models / Regressió lineal

En el quadre de diàleg resultant introduïm les variables explicada i les explicatives, a més del nom d'aquest model (*RegModel.2*):



En la finestra de resultats obtenim el següent:

```
> summary(RegModel.2)

Call:
lm(formula = ATUR ~ MOTOR + RBFD + TEMP + UNIV, data = Dades)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5297  -1.4544   0.4991   1.9571   7.7362

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.08206    2.05236  14.170  < 2e-16 ***
MOTOR       -1.68948    0.21566  -7.834 8.99e-14 ***
RBFD        -0.31369    0.09812  -3.197  0.00154 **
TEMP         0.01001    0.01201   0.833  0.40533
UNIV        -0.31007    0.04815  -6.440 4.96e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.284 on 290 degrees of freedom
Multiple R-squared:  0.3929,  Adjusted R-squared:  0.3845
F-statistic: 46.92 on 4 and 290 DF,  p-value: < 2.2e-16
```

Com veiem, tots els coeficients estimats menys el de la variable *TEMP* són significatius al 5% de significació, encara que l'ajust del model ($R^2 = 0,392$) és més aviat pobre.

Si calculem els intervals de confiança (IC) dels coeficients estimats obtenim:

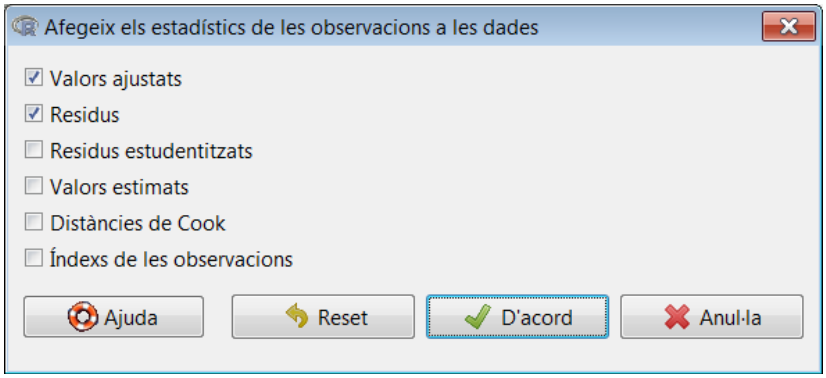
```
> Confint(RegModel.2, level=0.95)
                Estimate       2.5 %      97.5 %
(Intercept) 29.08205989 25.04264359 33.12147619
MOTOR       -1.68948313 -2.11394003 -1.26502623
RBFD        -0.31369304 -0.50681698 -0.12056910
TEMP         0.01000953 -0.01363066  0.03364972
UNIV        -0.31006949 -0.40483881 -0.21530016
```

Recordeu que per a obtenir els IC dels paràmetres hem de seleccionar el model, seguir la ruta *Models / Intervals de confiança* i seleccionar el nivell de confiança que volem.

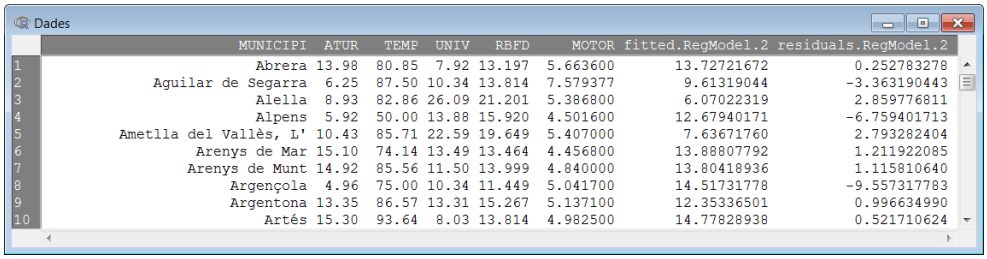
R-Commander ens dóna l'opció d'obtenir informació estadística addicional del model estimat. Entre altres indicadors, podem extreure els residus (e_i) i els valors ajustats a la recta ($\hat{Y}_i$). Per fer això, accedirem a:

Models / Afegeix els estadístics de les observacions a les dades

En el nostre exemple, només afegirem al nostre conjunt de dades els residus i els valors ajustats a la recta. Per a això, els activarem en el quadre de diàleg següent:



A continuació, si visualitzem el nostre conjunt de dades, observarem com s'han afegit aquestes dues variables:


Dades

	MUNICIPI	ATUR	TEMP	UNIV	RBFD	MOTOR	fitted.RegModel.2	residuals.RegModel.2
1	Abrera	13.98	80.85	7.92	13.197	5.663600	13.72721672	0.252783278
2	Aguilar de Segarra	6.25	87.50	10.34	13.814	7.579377	9.61319044	-3.363190443
3	Alella	8.93	82.86	26.09	21.201	5.386800	6.07022319	2.859776811
4	Alpens	5.92	50.00	13.88	15.920	4.501600	12.67940171	-6.759401713
5	Ametlla del Vallès, L'	10.43	85.71	22.59	19.649	5.407000	7.63671760	2.793282404
6	Arenys de Mar	15.10	74.14	13.49	13.464	4.456800	13.88807792	1.211922085
7	Arenys de Munt	14.92	85.56	11.50	13.999	4.840000	13.80418936	1.115810640
8	Argençola	4.96	75.00	10.34	11.449	5.041700	14.51731778	-9.557317783
9	Argentona	13.35	86.57	13.31	15.267	5.137100	12.35336501	0.996634990
10	Artés	15.30	93.64	8.03	13.814	4.982500	14.77828938	0.521710624

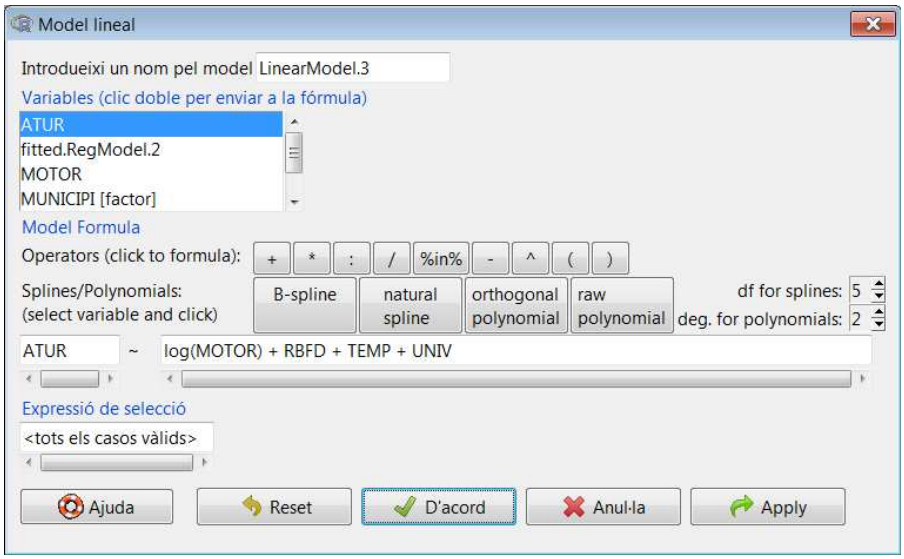
El més lògic és que no estiguem satisfets amb el model estimat i que en vulguem millorar l'estimació. Podem optar pel model alternatiu següent, en què la variable $MOTOR$ apareix en logaritmes:

$$ATUR_i = \beta_0 + \beta_1 log(MOTOR)_i + \beta_2 RBFD_i + \beta_3 TEMP_i + \beta_4 UNIV_i + \varepsilon_i.$$

Per a estimar aquest nou model, una possible solució seria crear una nova variable, $log(MOTOR)$, afegir-la al conjunt de dades i estimar el nou model com hem fet abans. Tanmateix, tenim a la nostra disposició una alternativa més ràpida i eficient: un quadre de diàleg complet que ens permet introduir variables transformades (aplicar logaritmes a una variable, elevar-la al quadrat, etc.) o multiplicades entre elles; fins i tot podem seleccionar una mostra de la nostra base de dades, etc. És a dir, la solució consisteix a estimar directament un model mitjançant la ruta alternativa següent:

Estadístics / Ajustament de models / Model lineal

Ens apareixerà el quadre de diàleg següent, en el qual introduirem la fórmula del model que té dues parts: la variable dependent i el conjunt de regressors o variables explicatives. En el nostre exemple, introduïm la variable *Motor* en logaritmes. A més, assignarem a aquest model el nom *LinearModel.3*:



El resultat que s'obté és el següent:

```
> LinearModel.3 <- lm(ATUR ~ log(MOTOR) + RBFD + TEMP + UNIV, data=Dades)

> summary(LinearModel.3)

Call:
lm(formula = ATUR ~ log(MOTOR) + RBFD + TEMP + UNIV, data = Dades)

Residuals:
     Min       1Q   Median       3Q      Max
-12.7937  -1.5613   0.5654   2.0256   7.7790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.42739    2.55005  12.716  < 2e-16 ***
log(MOTOR)  -7.95744    1.15789  -6.872 3.86e-11 ***
RBFD        -0.28664    0.10056  -2.850  0.00468 **
TEMP         0.01762    0.01210   1.456  0.14638
UNIV        -0.32186    0.04933  -6.525 3.03e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.352 on 290 degrees of freedom
Multiple R-squared:  0.3674,  Adjusted R-squared:  0.3587
F-statistic: 42.11 on 4 and 290 DF,  p-value: < 2.2e-16
```

Comprovem que l'ajust del model ha empitjorat respecte al model anterior. És important destacar que el resultat de l'estimació sempre mostra dos valors del coeficient de determinació, un dels quals es denomina *coeficient de determinació ajustat*. El motiu és que sempre que s'afegeixin noves variables explicatives a un model, el valor de R^2 augmentarà, encara que aquestes noves variables no aportin res de nou al model. Per això mateix, el valor ajustat de R^2 inclou una penalització pel nombre de regressors que el model conté.

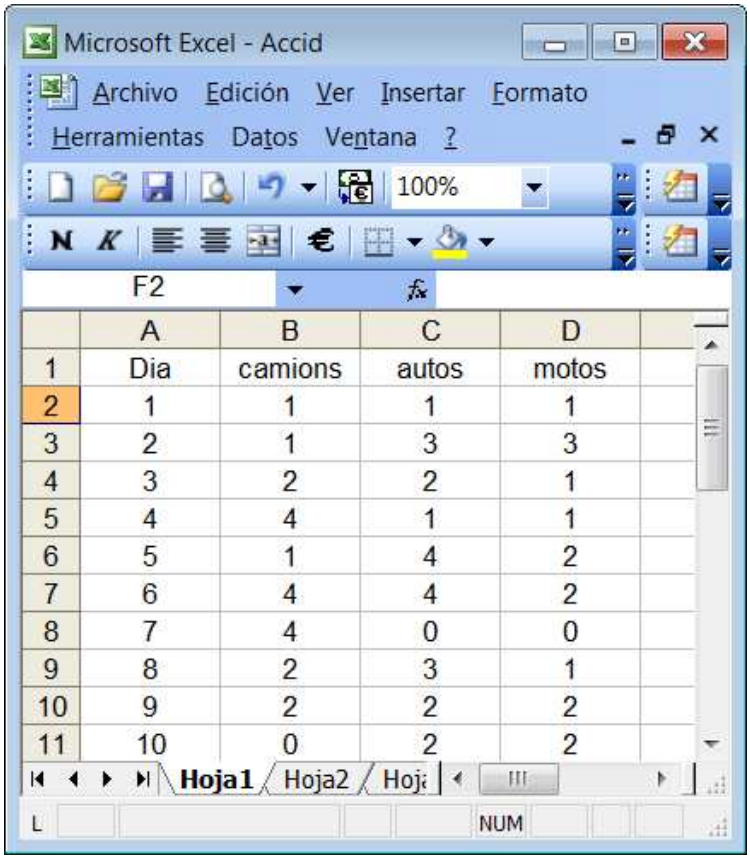
Quan vulguem comparar dos models que tinguin la mateixa variable endògena però amb diferent nombre de variables explicatives, triarem el que tingui un valor de la R^2 ajustada.

2. Anàlisi de la variància (ANOVA) i taules de contingència

Aquest capítol està dedicat a l'anàlisi de la variància (ANOVA) i a l'estudi de taules de contingència (TC). L'anàlisi ANOVA divideix la variació observada o dispersió d'una determinada variable en components atribuïbles a diferents fonts de variació. En la forma més simple, una ANOVA proporciona una prova estadística per a contrastar si les mitjanes de diferents grups són iguals o no. Per la seva banda, les TC acostumen a analitzar la distribució d'una base de dades en funció de dues variables.

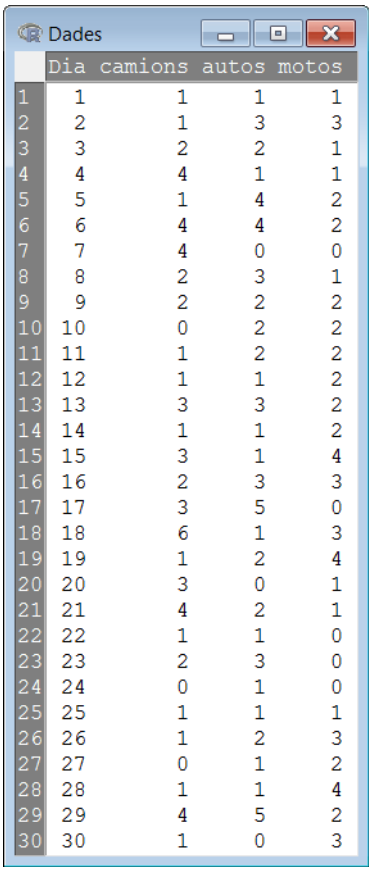
2.1. Anàlisi de la variància (ANOVA)

Per a fer una anàlisi empírica, considerarem una base de dades fictícia sobre el nombre d'accidents diaris durant un mes per diferents carreteres i tipus de vehicle (C = Camions, A = Automòbils, M = Motocicletes). Inicialment, les dades estan disponibles en un arxiu en format de Microsoft Excel (extensió *xls*):



	A	B	C	D
1	Dia	camions	autos	motos
2	1	1	1	1
3	2	1	3	3
4	3	2	2	1
5	4	4	1	1
6	5	1	4	2
7	6	4	4	2
8	7	4	0	0
9	8	2	3	1
10	9	2	2	2
11	10	0	2	2

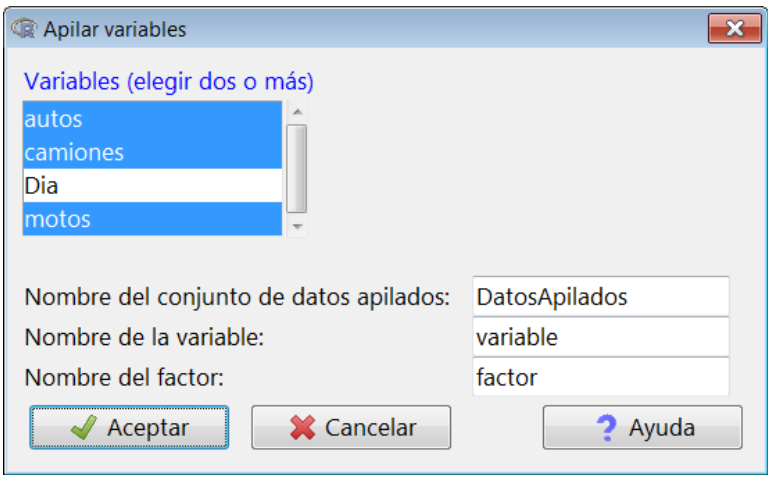
Comprovem que s'han carregat correctament les dades visualitzant el conjunt de dades actiu:



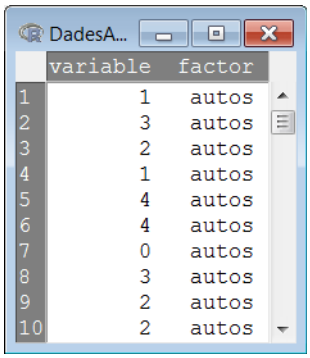
El primer pas per a fer l'anàlisi que ens proposem és crear una variable apilada. Això és: de tres variables que tenim (cada una amb 30 observacions), en creem una de sola amb $30 \times 3 = 90$ observacions, amb una variable qualitativa associada que indiqui el tipus de vehicle. Per a fer això, cal seguir la ruta següent:

Dades / Taula de dades activa / Apila variables a la taula de dades activa

Quan accedim a aquesta ruta, ens apareixerà un quadre de diàleg en el qual hem d'introduir quines variables volem apilar. A més, introduirem el nom de la variable apilada i el nom de la variable factor que defineix, per a cada observació, el tipus de vehicle:



Fet això, també haurem creat un nou conjunt de dades, que per defecte té el nom de *DadesApilades*, encara que si volem el podem canviar. Si ara anem a l'opció de *Visualitzar la taula de dades*, veurem que té dues variables de 90 observacions: una de numèrica (nombre de vehicles) i un factor (tipus de vehicle).



Entre altres coses, l'ANOVA permet comparar les mitjanes de diferents grups. En el nostre cas, definirem els tipus de vehicles de la manera següent: $V_1 = C$, $V_2 = A$ i $V_3 = M$. L'objectiu és fer el contrast d'hipòtesi següent:

$$\begin{array}{ll} H_0: & \mu_1 = \mu_2 = \mu_3 \\ H_1: & H_0 \text{ no és certa} \end{array}$$

És a dir, estadísticament, la mitjana poblacional d'accidents (μ) de cotxes, motos i camions és la mateixa? Per a fer aquest contrast, primer hem de definir el nombre de grups, $k = 3$ grups, i la mida de cada grup, que serà $n_1 = n_2 = n_3 = 30$, essent $N = k \cdot 30 = 90$.

Aquest test es basa a descompondre la suma total de quadrats (SQT) en dues parts: la dispersió explicada amb els grups (SQE) i la dispersió explicada per mitjà d'altres factors diferents dels grups en què hem dividit la població (SQD). Per tant, sempre es complirà que:

$$SQT = SQE + SQD$$

A més, com més s'aproximi SQE a SQT, més voldrà dir que hi ha factors dins de cada grup que fan que siguin diferents entre ells (mitjanes diferenciades). En canvi, un SQD proporcionalment elevat significa que la variació de les dades es deu a factors externs no relacionats amb els grups.

L'estadístic que s'ha de calcular i la seva distribució són:

$$F^* = \frac{\frac{SQE}{k-1}}{\frac{SQD}{N-k}} = \frac{SQE}{2} \cdot \frac{87}{SQD} \sim F_{\alpha,k-1,N-k}$$

La decisió final es prendrà adoptant el criteri següent:

$F^* > F_{\alpha,k-1,N-k}$	$\Rightarrow$	Rebuig de H_0
$F^* \leq F_{\alpha,k-1,N-k}$	$\Rightarrow$	No rebuig de H_0

Per a aplicar aquest contrast amb R-Commander, anirem al menú i accedirem a la ruta següent:

Estadísticas / Medias / ANOVA de un factor

Aleshores ens apareixerà el quadre de diàleg següent:

En aquest exemple, com que només hi ha una variable numèrica i un factor, només cal prémer *D'acord*, i apareixerà el resultat que es mostra a continuació:

```
> AnovaModel.1 <- aov(variable ~ factor, data=DadesApilades)

> summary(AnovaModel.1)
             Df Sum Sq Mean Sq F value Pr(>F)
factor        2   0.62  0.3111   0.168  0.845
Residuals    87 160.67  1.8467

> numSummary(DadesApilades$variable , groups=DadesApilades$factor,
+ statistics=c("mean", "sd"))
             mean       sd data:n
autos    1.933333 1.362891     30
camions  2.000000 1.485563     30
motos    1.800000 1.214851     30
```

Primer apareix el resultat del contrast ANOVA, i tot seguit la mitjana i la desviació estàndard dels tres tipus de vehicles. Quant al primer resultat, veiem que la dispersió explicada mitjançant els grups és molt reduïda ($SQE = 0,62$), mentre que la dispersió residual, és a dir, explicada per altres factors diferents, és majoritària ($SQD = 160,67$). La suma d'aquestes magnituds és $SQT = 161,29$. El valor de l'estadístic F és de $0,168$, i el *p-valor* és $0,845$.

És molt útil tenir present el criteri següent:

- Si el *p-valor* és inferior a $0,05$ rebutgem la hipòtesi nul·la, i podem assegurar que la mitjana del nombre d'accidents depèn del tipus de vehicle.

- Si el *p-valor* és superior a $0,05$ no rebutgem la hipòtesi nul·la i, per tant, no podem assegurar que la mitjana del nombre d'accidents depengui del tipus de vehicle.

En el nostre exemple, clarament no es rebutja la H_0, i concloem que no podem assegurar que la mitjana del nombre d'accidents depengui del tipus de vehicle.

2.2. Taules de contingència

En aquesta secció veurem un exemple d'una taula de contingència. Seguint amb l'exemple que hem vist en el capítol anterior, per a cada tipus de vehicle, crearem una variable dicotòmica que prengui el valor 0 en cas que el nombre d'accidents diaris sigui inferior o igual a 2, i valor 1 en cas contrari. L'objectiu és contrastar si aquesta variable dicotòmica depèn del tipus de vehicle, o si és independent.

El primer pas consisteix a crear tres variables dicotòmiques (que només poden prendre els valors 0 i 1), de manera que obtenim $\{\tilde{C}_j\, \tilde{A}_j\, , \tilde{M}_j; \quad j = 1, \ldots, n\}$. Aquestes variables es calculen de la manera següent (només es mostra el cas de la variable C, la resta son anàlegs):

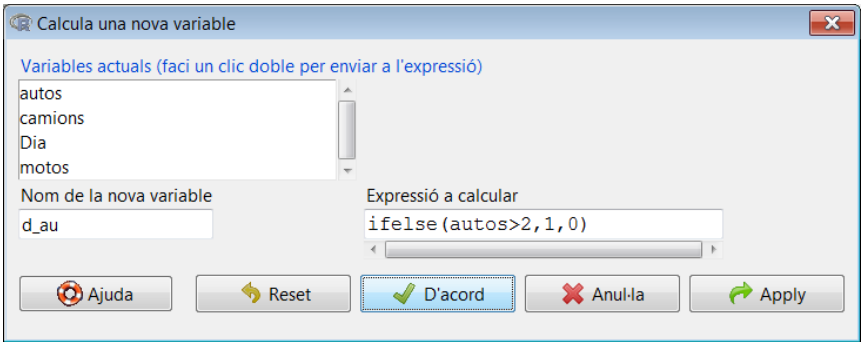
$$\tilde{C}_j = \begin{cases} 0 & \text{si } C_j \leq 2 \\ 1 & \text{si } C_j > 2 \end{cases}$$

En R-Commander, el primer pas serà crear aquestes tres variables dicotòmiques, que anomenarem *d_au*, *d_ca* i *d_mo*. Abans, haurem de canviar el conjunt de dades actiu i seleccionar el conjunt *Dades*, que és on hi ha les tres variables originals (és a dir, ja no ens interessen les variables apilades que hem fet servir amb l'anàlisi ANOVA).

Per crear variables, seguirem aquesta ruta:

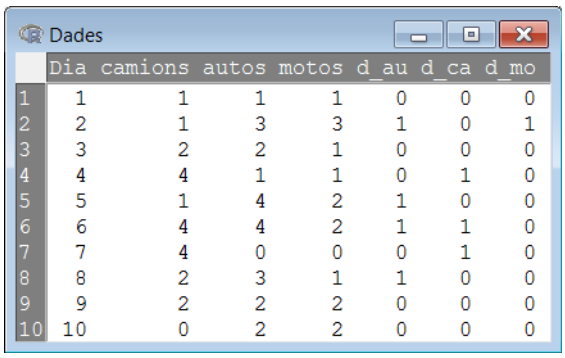
Dades / Modifica variables de la taula de dades activa / Calcula una nova variable

Apareixerà un quadre de diàleg, en el qual introduirem el nom de la nova variable i la seva expressió (és a dir, com es calcula). De manera similar a com es fa amb el full de càlcul *Excel*, introduirem:



La funció *ifelse* té, en primer lloc, una expressió lògica ($autos > 2$), després el valor que prendrà la nova variable si aquesta condició es compleix (1), i després el valor que prendrà si és falsa (0). Això ho farem tres vegades, una per cada variable. Un cop fet això, si visualitzem el conjunt de dades actiu, comprovarem que s'han creat tres

variables consistents en uns i zeros. Són aquestes tres variables les que utilitzarem per a fer l'anàlisi de contingència.



	Dia	camions	autos	motos	d_au	d_ca	d_mo
1	1	1	1	1	0	0	0
2	2	1	3	3	1	0	1
3	3	2	2	1	0	0	0
4	4	4	1	1	0	1	0
5	5	1	4	2	1	0	0
6	6	4	4	2	1	1	0
7	7	4	0	0	0	1	0
8	8	2	3	1	1	0	0
9	9	2	2	2	0	0	0
10	10	0	2	2	0	0	0

Un cop tenim aquestes variables calculades, es tractaria d'analitzar si són independents o si depenen del tipus de vehicle. Aleshores, utilitzarem el test χ^2 d'independència, que permet comparar, a partir d'una taula de contingència, les freqüències observades amb les quals, en teoria, s'haurien de donar si hi hagués independència. En aquesta taula, en les files hi haurà dues categories (G_1 i G_2), la primera per als valors u i la segona per als valors zero. I en les columnes hi haurà el tipus de vehicle, és a dir, $V_1 = C$, $V_2 = A$ i $V_3 = M$. A més, en aquest cas s'ha de diferenciar el nombre de dies pels quals es tenen dades ($n = 30$) i el nombre total d'observacions, que és $N = 3 \cdot n = 90$. Aleshores, les freqüències observades són:

Categoria	V_1	V_2	V_3	**Total**
G_1	N_{11}	N_{12}	N_{13}	N_{G_1}
G_2	N_{21}	N_{22}	N_{23}	N_{G_2}
Total	N_{V_1}	N_{V_2}	N_{V_3}	N

En R-Commander, el pas següent serà calcular G_1 i G_2, és a dir, la suma d'uns i zeros per a cada tipus de vehicle. Una manera de fer això és, en la finestra d'instruccions, introduir les fórmules següents:

```
> sum(Dades$d_au)
> sum(Dades$d_ca)
> sum(Dades$d_mo)
```

Seleccionant aquestes tres instruccions, si premem *Executar* obtindrem el resultat següent en la finestra de resultats:

```
> sum(Dades$d_au)
[1] 9

> sum(Dades$d_ca)
[1] 10

> sum(Dades$d_mo)
[1] 8
```

Amb aquesta informació, sabent que hi ha 30 dades per a cada tipus de vehicle, ja sabem que la taula de contingència que hem d'analitzar serà la següent:

Categoria	V_1	V_2	V_3	Total
G_1	9	10	8	27
G_2	21	20	22	63
Total	30	30	30	90

$$N'_{ij} = \frac{N_{G_i} N_{V_j}}{N}$$

La qüestió consisteix a determinar si les diferències $n_{ij} - n'_{ij}$ són aleatòries, per la qual cosa hi ha independència, o si al contrari, aquestes diferències són massa grans i cal admetre que hi ha algun tipus d'associació entre les dues estratificacions. Formalment, es tracta de contrastar les hipòtesis:

H_0:	hi ha independència.
H_1:	no hi ha independència, i hi ha algun tipus d'associació.

La manera de procedir és molt similar a la de l'exercici anterior. En aquest cas, l'estadístic és:

$$E = \sum_{\forall G,V} \frac{(N_{ij} - N'_{ij})^2}{N'_{ij}} \sim \chi^2_{\alpha,(L-1)\cdot(K-1)}$$

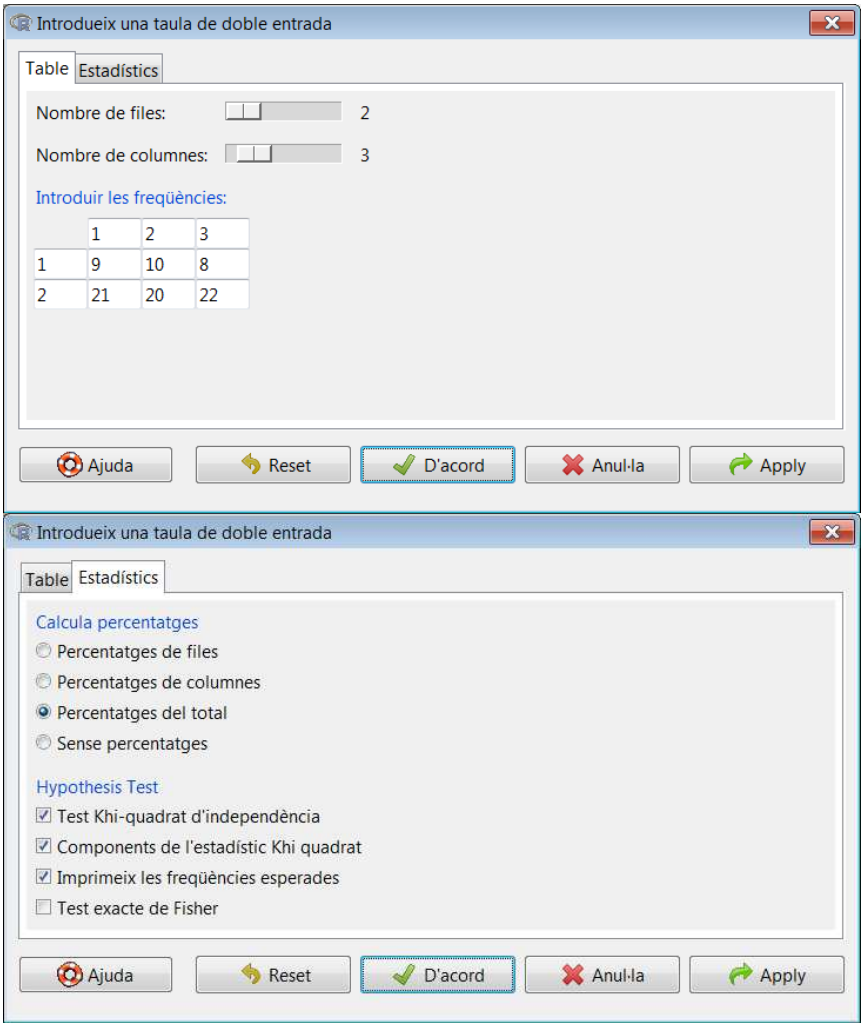
En el nostre cas, $L = 2$ és el nombre de dimensions horitzontals i $K = 3$ és el nombre de dimensions verticals. El valor crític del contrast amb graus de llibertat $\alpha = 0.05$ i $(L-1)\cdot(K-1) = 2$ és $\chi^2_{0.05,2} \simeq 5,99$. La resolució del contrast serà, doncs:

$E > \chi^2_{0.05,2}$	$\Rightarrow$	Rebuig de H_0
$E \leq \chi^2_{0.05,2}$	$\Rightarrow$	No rebuig de H_0

Per fer aquesta anàlisi amb R-Commander, anirem a la ruta següent del menú:

Estadístics / Taules de contingència / Introdueix i analitza una taula de doble entrada

Apareixerà el quadre de diàleg següent, en el qual introduirem les nostres dades:



En prémer *D'acord*, obtindrem el resultat següent:

```
> .Table <- matrix(c(9,10,8,21,20,22), 2, 3, byrow=TRUE)

> rownames(.Table) <- c('1', '2')

> colnames(.Table) <- c('1', '2', '3')

> .Table  # Counts
   1  2  3
1  9 10  8
2 21 20 22

> totPercents(.Table) # Percentage of Total
         1    2    3 Total
1     10.0 11.1  8.9    30
2     23.3 22.2 24.4    70
Total 33.3 33.3 33.3   100

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

  Pearson's Chi-squared test

data:  .Table
X-squared = 0.3175, df = 2, p-value = 0.8532
```

```
> .Test$expected # Expected Counts
   1  2  3
1  9  9  9
2 21 21 21

> round(.Test$residuals^2, 2) # Chi-square Components
  1    2    3
1 0 0.11 0.11
2 0 0.05 0.05
```

En aquest cas el valor de l'estadístic és de $0,317$, i amb una confiança del 95% el *p-valor*$> 0,05$ indica que la H_0 no es pot rebutjar, és a dir, hi ha independència i no hi ha relació entre que hi hagi més de 2 accidents i el tipus de vehicle.

3. Anàlisi de components principals i anàlisi clúster

3.1. Introducció

En el model de regressió que hem analitzat en el primer capítol d'aquest mòdul assumia una relació de dependència entre una variable dependent i un grup de regressors. Tanmateix, en aquest capítol, analitzarem alguns mètodes d'interdependència, és a dir, aquells en què, a priori, no s'assumeix cap relació entre les diferents variables que participen en un estudi.

L'objectiu de l'anàlisi es basa a simplificar la complexa estructura d'una base de dades. Això ho podem fer eliminant les variables menys representatives i agrupant les més rellevants, però també es podria optar per agrupar observacions amb característiques similars.

D'una banda, entre les anàlisis de la primera opció, la reducció de variables, el que més s'utilitza és l'anàlisi factorial que, segons la naturalesa de les dades, es deriva en diverses anàlisis. Les més conegudes són:

- l'anàlisi de components principals (ACP)
- l'anàlisi de correspondències simple (ACS)
- l'anàlisi de correspondències múltiple (ACM)

D'altra banda, entre les anàlisis basades en l'agrupació d'individus destaquen les anàlisis clúster o de conglomerats.

Finalment, hem de deixar clar que la reducció de variables (columnes) i l'agrupació d'observacions (files) no són opcions excloents. Hi ha infinitat de situacions en què estarem interessats a fer totes dues coses, reduir variables i agrupar individus. Els exemples més clars els trobem en l'àmbit de la investigació de mercats quan fem una segmentació de clients.

3.2. Anàlisi de components principals (ACP)

És important destacar que els nous components o factors que obtinguem després de qualsevol anàlisi factorial **sempre** estaran incorelacionats.

L'ACP és un procediment estadístic que té com a objectiu agrupar un conjunt de variables en un o més components mitjançant una transformació ortogonal, de manera que la múltiple informació que contenen aquestes variables quedi sintetitzada en menys dimensions. El nombre de components principals és més petit o igual que el nombre de variables originals. Aquesta transformació es defineix de tal manera que el primer component principal té la variància més gran possible (és a dir, recull la major proporció de la variabilitat possible de les dades), i cada component subsegüent té, al seu

torn, la màxima variància possible en virtut de la restricció que sigui ortogonal (és a dir, no correlacionada amb els components anteriors).

En l'ACP, els valors propis d'una matriu tenen un paper fonamental. Per a veure-ho suposem que disposem de k variables, cada una amb n observacions: $\{X_1, X_2, \ldots, X_n\}$. La solució ideal per a l'anàlisi és que cada variable reculli un aspecte únic i específic de cada individu, amb la qual cosa totes les variables serien rellevants per a l'anàlisi i no hi hauria cap variable redundant. Si això fos així, es compliria que $Cov(X_i, X_j) = 0$ i, com a conseqüència, la $Cor(X_i, X_j) = 0$, $\forall i \neq j$. És a dir, la matriu de variàncies i covariàncies (MVC) tindria la forma següent[1]:

$$MVC = \begin{pmatrix} Var(X_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Var(X_k) \end{pmatrix}_{k \times k}$$

Matemàticament, els k valors propis $\lambda_1, \ldots, \lambda_k$ d'una matriu quadrada compleixen les propietats següents:

1) La suma dels valors propis **sempre** és igual a la suma dels elements de la diagonal:

$$\sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} Var(X_i)$$

2) Si tots els elements de la diagonal són igual a zero, els valors propis seran igual als elements de la diagonal en ordre invers:

$$\lambda_1 = Var(X_k), \ldots, \lambda_k = Var(X_1)$$

Ara bé, en la realitat això passa rarament, ja que gairebé sempre hi ha un grau de covariància i correlació entre les variables.

Per aquest motiu, i especialment quan el grau de correlació entre algunes variables és elevat, l'ús de la tècnica de l'ACP està més que justificada, ja que el que fa és reduir les k variables a un nombre més petit de **components principals (CP)**, entre els quals no hi hagi cap correlació (és a dir, que siguin ortogonals).

Vegem un exemple empíric amb R i R-Commander. Disposem d'una base de dades amb $N = 100$ observacions corresponents als alumnes d'una escola que conté informació que ha d'ajudar a categoritzar i estudiar millor el comportament global d'aquests estudiants[2]. La informació que es té per a cada alumne es resumeix en les 6 variables definides en la taula 2.

[1] Anàlogament, la matriu de correlacions serà la matriu identitat amb dimensió $k \times k$.

[2] Aquesta base de dades és totalment fictícia i s'ha generat artificialment per a il·lustrar el concepte d'ACP.

Taula 2. Dades de rendiment acadèmic

Variable	Definició
n_mates	Nota en matemàtiques
n_llengua	Nota en llengua
n_angles	Nota en anglès
absent	Nombre d'absències a classe
indis	Nombre d'actes d'indisciplina
penal	Nombre de penalitzacions rebudes

El primer pas serà efectuar la importació de la base de dades i, posteriorment, visualitzar els resultats per a comprovar que aquesta importació s'ha efectuat correctament:

Visualitzant les dades

Encara que la mostra contingui $N = 100$ observacions, aquí només mostrem les 10 primeres per una qüestió d'espai.

Dades

	n_mates	n_llengua	n_angles	absent	indis	penal
1	0.7	1.6	0.5	1	2	3
2	1.2	2.2	0.9	1	3	4
3	0.7	2.9	1.3	1	2	3
4	2.0	1.6	0.7	2	3	4
5	1.5	2.2	1.4	0	1	2
6	1.0	2.4	1.4	2	3	4
7	1.7	2.8	1.5	0	0	1
8	1.9	2.4	2.1	1	3	3
9	1.9	3.6	1.0	0	0	1
10	1.6	2.6	2.4	0	0	0

A continuació, és important que fem un resum del conjunt de dades que conté informació estadística bàsica:

```
> summary(Dades)
    n_mates         n_llengua        n_angles
 Min.   :0.700   Min.   :1.600   Min.   :0.500
 1st Qu.:3.000   1st Qu.:3.900   1st Qu.:2.775
 Median :5.000   Median :5.300   Median :4.700
 Mean   :5.054   Mean   :5.488   Mean   :5.000
 3rd Qu.:7.100   3rd Qu.:7.050   3rd Qu.:7.150
 Max.   :9.400   Max.   :9.800   Max.   :9.400
     absent          indis           penal
 Min.   :0.00   Min.   :0.00   Min.   :0.00
 1st Qu.:0.00   1st Qu.:1.00   1st Qu.:1.00
 Median :1.00   Median :2.00   Median :2.00
 Mean   :0.81   Mean   :1.86   Mean   :2.64
 3rd Qu.:1.00   3rd Qu.:3.00   3rd Qu.:3.25
 Max.   :3.00   Max.   :6.00   Max.   :8.00
```

Després de fer-nos una idea sobre la informació que contenen les variables haurem d'estudiar la correlació entre totes elles. Fins ara hem vist que la manera de fer aquest càlcul amb R-Commander consistia a seguir la ruta següent i seleccionar les variables d'interès:

Estadístics / Matriu de correlacions

De manera que el resultat és el següent:

```
> cor(Data[,c("absent","indis","n_angles","n_llengua","n_mates",
+ "penal")],use="complete.obs")
          absent  indis  n_angles  n_llengua  n_mates  penal
absent      1.00   0.93      0.17       0.08     0.18   0.91
indis       0.93   1.00      0.20       0.12     0.20   0.95
n_angles    0.17   0.20      1.00       0.93     0.96   0.18
n_llengua   0.08   0.12      0.93       1.00     0.93   0.12
n_mates     0.18   0.20      0.96       0.93     1.00   0.20
penal       0.91   0.95      0.18       0.12     0.20   1.00
```

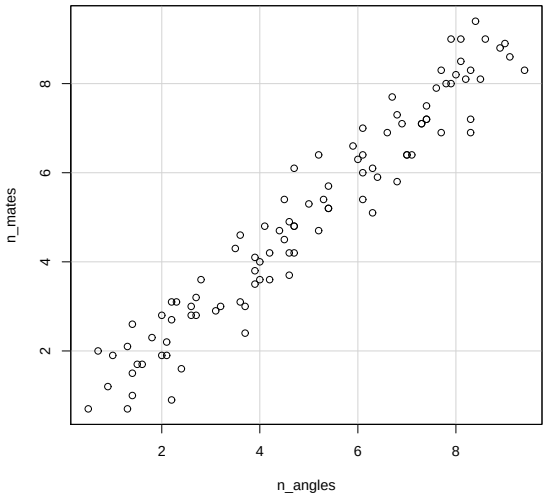
Resultat simplificat

Per una qüestió d'espai, hem suprimit part dels decimals de les correlacions.

Un cop disposem del codi en R, una opció interessant és assignar un nom a aquesta matriu de correlacions, canviar l'ordre de les variables i arrodonir les xifres a dos dígits:

```
> MC <- cor(Data[,c("n_angles","n_llengua","n_mates","penal",
+ "absent","indis")],use="complete.obs")
> round(MC,digits=2)
          n_angles n_llengua n_mates penal absent indis
n_angles      1.00      0.94    0.97  0.19   0.17  0.20
n_llengua     0.94      1.00    0.93  0.12   0.09  0.12
n_mates       0.97      0.93    1.00  0.20   0.19  0.21
penal         0.19      0.12    0.20  1.00   0.92  0.96
absent        0.17      0.09    0.19  0.92   1.00  0.93
indis         0.20      0.12    0.21  0.96   0.93  1.00
```

Intuïtivament, aquesta matriu de correlacions permet observar que, d'una banda, les tres variables de notes estan molt correlacionades (per sobre de $0,9$) i, de l'altra, les tres variables relacionades amb el comportament també ho estan. D'alguna manera, es pot afirmar que hi ha variables redundants, és a dir, que amb menys variables es podria explicar el mateix. Vegem un gràfic de dispersió de les variables *nota d'anglès* i *nota de matemàtiques* per a obtenir una evidència visual:



Una manera sofisticada de calcular el nivell de correlació entre variables és calculant els vectors propis de la matriu de correlacions (MC):

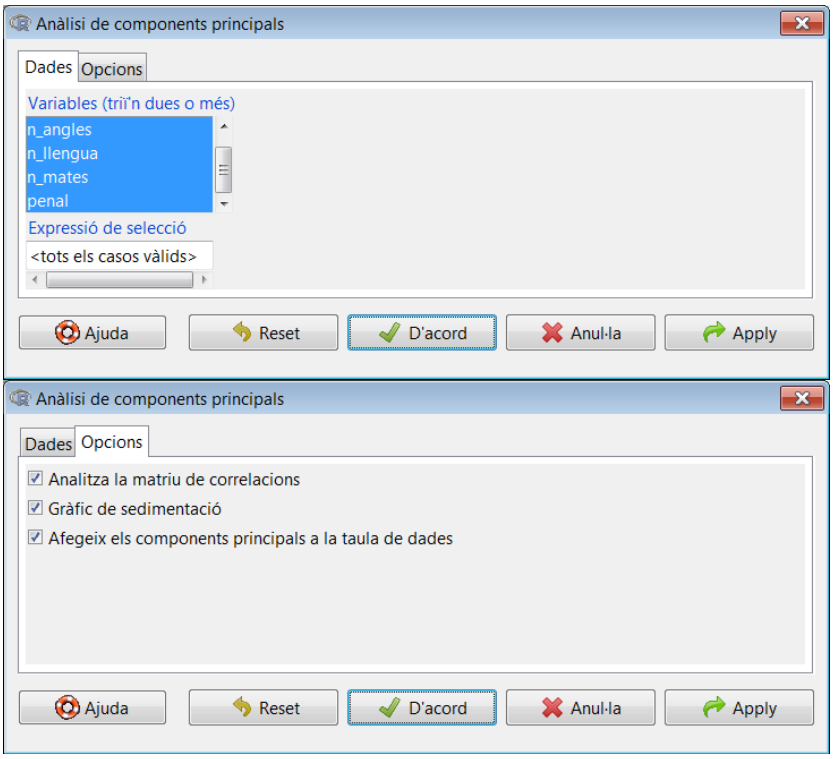
El càlcul dels valors propis de la matriu de variàncies i covariàncies donaria un resultat similar.

```
> eigen(MC)$values
[1] 3.383908 2.386493 0.092356 0.063587 0.043842 0.029810
```

Com veiem, la magnitud dels dos primers valors propis comparats amb la resta és enorme. Això es tradueix en el fet que les sis variables de l'estudi es poden reduir a dues dimensions, i aquest fet el constatarà l'ACP. Per a fer aquesta anàlisi en R-Commander cal seguir la ruta següent:

Estadístics / Anàlisi dimensional / Anàlisi de components principals

Aleshores, ens apareixerà el quadre de diàleg següent, en el qual activarem les opcions disponibles.



Això fa que obtinguem, d'una banda, el resultat següent en la finestra de resultats:

1) *Component loadings.* És la matriu factorial, que mostra la correlació entre els components calculats i les variables objecte d'estudi.

```
> .PC <-
+   princomp(~absent+indis+n_angles+n_llengua+n_mates+penal,
+   cor=TRUE, data=Dades)

> unclass(loadings(.PC))  # component loadings
              Comp.1     Comp.2     Comp.3      Comp.4
absent     -0.3952811  0.4166928  0.7069336 -0.40148060
indis      -0.4115994  0.4072536 -0.2407877  0.23451814
n_angles   -0.4213028 -0.3948581  0.2019751  0.32225209
n_llengua  -0.3881367 -0.4316741 -0.3552555 -0.71965433
n_mates    -0.4245466 -0.3886644  0.1744042  0.39043337
penal      -0.4073731  0.4088962 -0.4948220  0.09812031
               Comp.5      Comp.6
absent     -0.09288941  0.02344405
indis       0.63503349 -0.38533804
n_angles    0.32825831  0.64358148
n_llengua   0.09755816 -0.09685559
n_mates    -0.43463287 -0.54492149
penal      -0.53096815  0.36117505
```

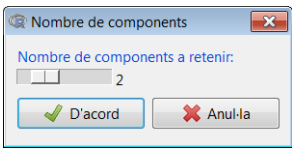
2) *Component variances.* Les variàncies dels components, és a dir, els valors propis de la matriu de correlacions calculats anteriorment. Com veiem, la magnitud dels dos primers representa gairebé el total de la suma de valors propis.

```
> .PC$sd^2  # component variances
  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6
 3.38421 2.38447 0.09285 0.06453 0.04426 0.02966
```

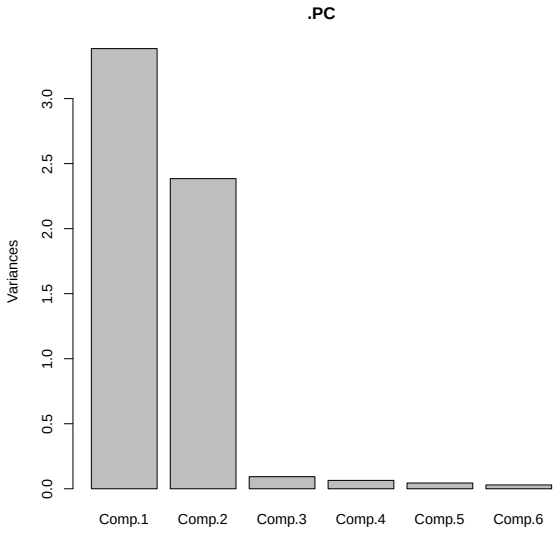
3) *Proportions of variance.* Proporcions de la variància (o variabilitat) total de les dades recollides pels components. Com veiem, els dos primers components recullen respectivament el 56, 4% i el 39, 7% de la proporció de la variància total. L'última fila ens mostra que entre els dos components es recull el 96, 1% de la variabilitat total de les sis variables.

```
> summary(.PC) # proportions of variance
Importance of components:
                          Comp.1    Comp.2     Comp.3
Standard deviation     1.8396223 1.5441756 0.30471529
Proportion of Variance 0.5640351 0.3974131 0.01547523
Cumulative Proportion  0.5640351 0.9614481 0.97692336
                           Comp.4      Comp.5      Comp.6
Standard deviation     0.25402956 0.210396015 0.172227515
Proportion of Variance 0.01075517 0.007377747 0.004943719
Cumulative Proportion  0.98767853 0.995056281 1.000000000
```

Abans de prémer *D'acord* en el quadre de diàleg anterior ens haurà aparegut l'opció d'escollir quants components principals volem emmagatzemar. Considerant l'anàlisi de valors propis que hem fet anteriorment, seleccionem dos components:



D'altra banda, en la consola d'R apareixerà el gràfic amb els valors propis dels components principals. El *criteri de Kaiser* estableix que considerarem els components principals que tinguin un valor propi superior a u. En aquest cas, és obvi que els dos primers components són suficients per a explicar la variabilitat de les dades.

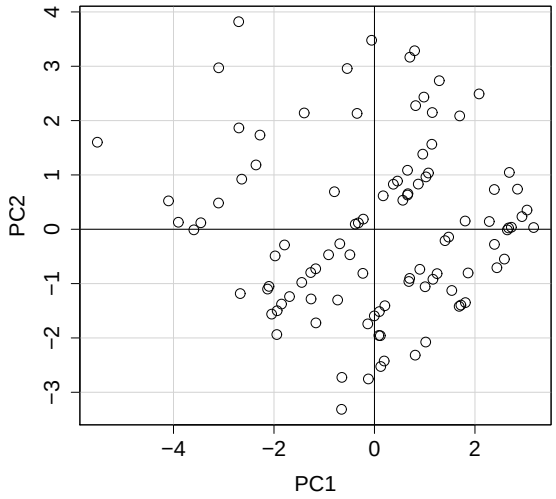


En haver emmagatzemat els dos components principals, si tornem a visualitzar el nostre conjunt de dades, apareixeran aquestes dues noves variables.

Dades

	n_mates	n_llengua	n_angles	absent	indis	penal	PC1	PC2
1	0.7	1.6	0.5	1	2	3	2.081771682	2.49082765
2	1.2	2.2	0.9	1	3	4	1.289605833	2.73413289
3	0.7	2.9	1.3	1	2	3	1.695992542	2.08611713
4	2.0	1.6	0.7	2	3	4	0.798502136	3.28556074
5	1.5	2.2	1.4	0	1	2	2.683860417	1.04593558
6	1.0	2.4	1.4	2	3	4	0.704036434	3.16657841
7	1.7	2.8	1.5	0	0	1	3.036055410	0.35317480
8	1.9	2.4	2.1	1	3	3	1.150089859	2.14821611
9	1.9	3.6	1.0	0	0	1	2.935431639	0.23309443
10	1.6	2.6	2.4	0	0	0	3.167854521	0.03170310

Per a tenir una evidència visual de com els dos components principals seleccionats expliquen la variabilitat de les dades, és recomanable fer una gràfica de dispersió dels dos components i veure com es distribueixen les observacions. Amb aquest gràfic es pot comprovar si hi ha alguna agrupació entre estudiants a través d'aquestes dues dimensions, és a dir, si hi ha diferents grups amb combinacions de bones/males notes i bon/mal comportament.



Visualitzant els components principals

En anglès aquest gràfic es denomina *scores plot*.

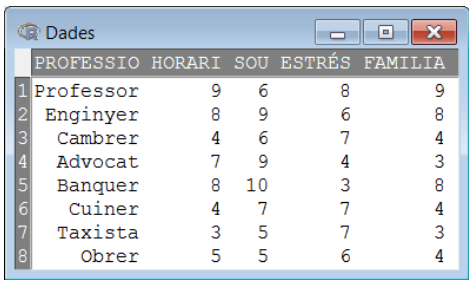
3.3. Anàlisi clúster

L'anàlisi de conglomerats o clúster engloba una àmplia gamma de mètodes numèrics que tenen com a objectiu detectar grups o conglomerats d'observacions homogenis (molt similars entre ells) i heterogenis entre ells (molt dispars entre grups). És a dir, les observacions han d'estar molt juntes per a formar part d'un mateix grup i molt separades per a formar part de diferents grups. Així doncs, els grups o clústers s'identifiquen per l'avaluació de les distàncies relatives entre els punts i això permet calcular l'homogeneïtat relativa de cada grup i el grau de separació entre els diferents grups.

És molt important no confondre l'ACP amb l'anàlisi clúster; el primer agrupa variables i el segon agrupa individus.

Posem un exemple amb dades fictícies. A partir d'una enquesta, es vol agrupar diferents professions segons la percepció que tenen els seus treballadors sobre diferents aspectes. Per a mesurar aquesta percepció s'ha utilitzat una escala de Likert; per tant, la resposta dels treballadors va des de 0 (mínima satisfacció) fins a 10 (màxima satisfacció).

La puntuació mitjana sobre les mostres de les diferents professions es recullen en la base de dades que es mostra a continuació:



	PROFESSIO	HORARI	SOU	ESTRÉS	FAMILIA
1	Professor	9	6	8	9
2	Enginyer	8	9	6	8
3	Cambrer	4	6	7	4
4	Advocat	7	9	4	3
5	Banquer	8	10	3	8
6	Cuiner	4	7	7	4
7	Taxista	3	5	7	3
8	Obrer	5	5	6	4

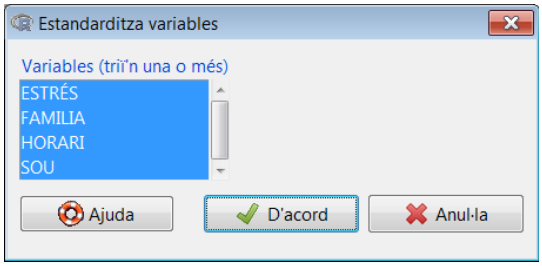
En aquest cas, l'objectiu de l'anàlisi clúster serà estudiar la distància entre les diferents professions a partir de les similituds entre les quatre variables de l'enquesta. La funció `dist` proporciona la matriu de distàncies entre diferents variables, essent la distància euclidiana l'opció per defecte. La informació entre parèntesis $[, 2:5]$ selecciona, del conjunt de dades, les columnes de la segona a la cinquena, que es corresponen amb les variables que volem analitzar.

```
> dist(Dades[,2:5])
         1        2        3        4        5        6        7
2 3.872983
3 7.141428 6.480741
4 8.062258 5.477226 5.291503
5 6.557439 3.162278 8.000000 5.291503
6 7.211103 6.082763 1.000000 4.795832 7.549834
7 8.602325 8.185353 1.732051 6.403124 9.539392 2.449490
8 6.782330 6.403124 1.732051 5.000000 7.681146 2.449490 2.449490
```

Com veiem, els oficis 3, 6, 7 i 8 són els més similars segons aquest criteri. Una eina visual adequada per a això és el *dendrograma*, la funció del qual és mostrar la formació de conglomerats, i també les distàncies entre ells. Abans de fer el dendrograma, és aconsellable tipificar les variables, perquè no hi hagi efectes d'escala. Per a això, com hem vist en el mòdul dedicat a l'anàlisi descriptiva, accedim a la ruta següent:

Dades / Modifica variables de la taula de dades activa / Estandarditza les variables

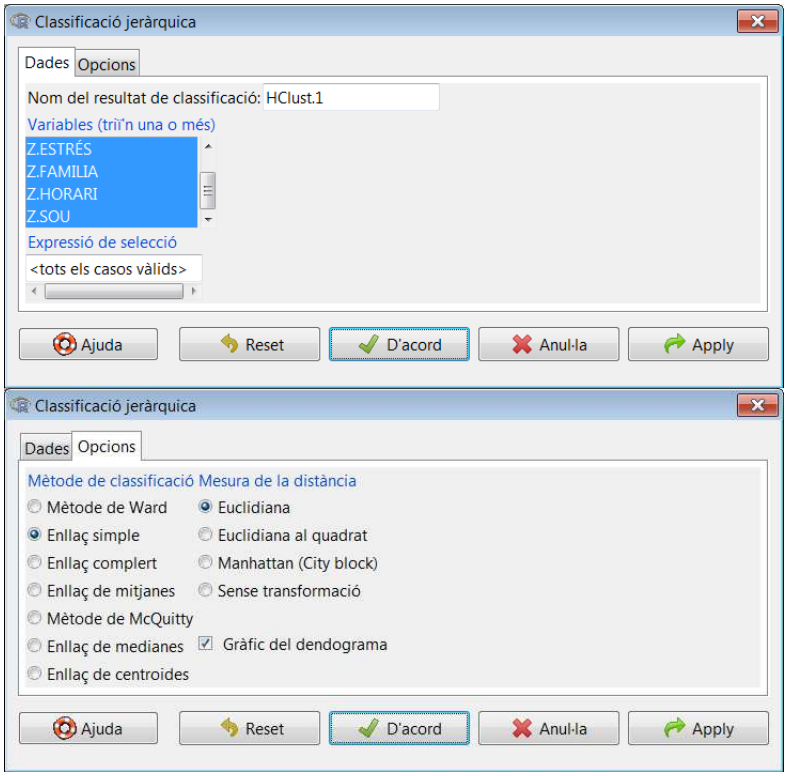
Ens apareixerà un menú en el qual seleccionarem les variables a tipificar:



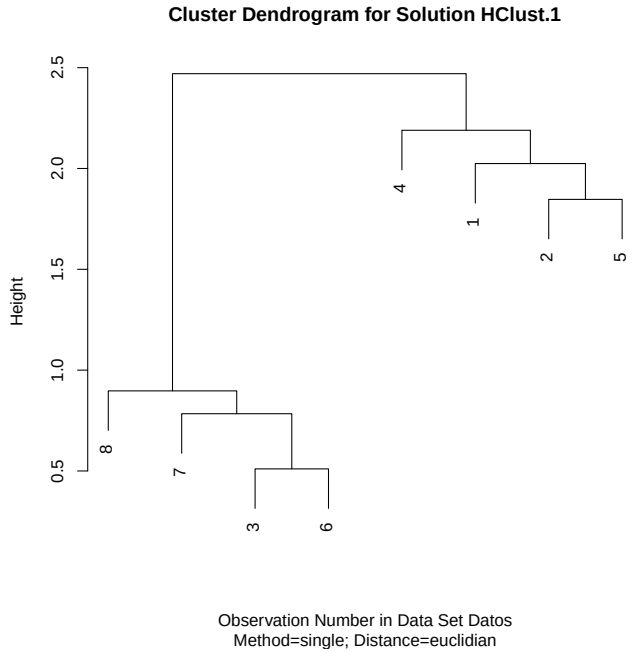
Un cop fet això, per a fer el dendrograma amb R-Commander, cal seguir aquesta ruta:

Estadístics / Anàlisi dimensional / Anàlisi de clústers / Classificació jeràrquica

Ens apareixerà el quadre de diàleg següent, en el qual seleccionarem les variables d'interès, el mètode d'agrupació, la mesura de distància i si volem visualitzar gràficament el dendrograma.



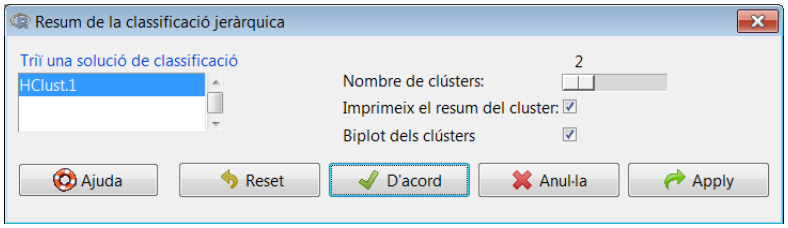
El resultat apareix en el gràfic que hi ha a continuació, en el qual es pot veure que hi ha dos grans grups d'oficis: a) professor, enginyer, advocat i banquer, i b) cambrer, cuiner, taxista i obrer.



Podem obtenir més informació de l'agrupació jeràrquica que hem fet accedint a la ruta següent:

Estadístics / Anàlisi dimensional / Anàlisi de clústers / Resumeix la classificació jeràrquica

Ens apareixerà el quadre de diàleg següent, en el qual seleccionarem el nombre de grups en el qual vulguem organitzar les dades (en el nostre cas, 2), a més del resum numèric i el gràfic associat.

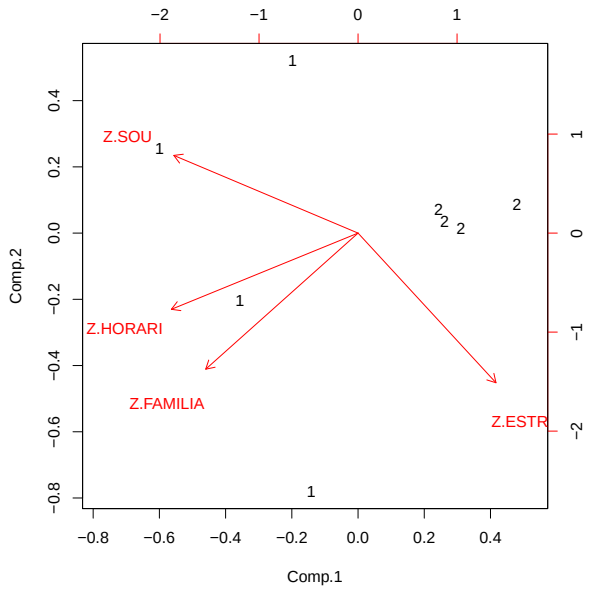


D'una banda, obtindrem el resum numèric següent:

```
> summary(as.factor(cutree(HClust.1, k = 2))) # Cluster Sizes
1 2
4 4

> by(model.matrix(~-1 + Z.ESTRÉS + Z.FAMILIA + Z.HORARI + Z.SOU, Dades),
+ as.factor(cutree(HClust.1, k = 2)), colMeans) # Cluster Centroids
INDICES: 1
  Z.ESTRÉS  Z.FAMILIA   Z.HORARI      Z.SOU
-0.4437060  0.6490734  0.8819171  0.7017420
------------------------------------------------------------------------
INDICES: 2
  Z.ESTRÉS  Z.FAMILIA   Z.HORARI      Z.SOU
 0.4437060 -0.6490734 -0.8819171 -0.7017420
```

I, de l'altra, obtindrem un gràfic, en el qual podem observar visualment les vuit observacions (quatre pertanyen a l'agrupació 1 i les altres quatre a l'agrupació 2), en un pla cartesià de components principals:



Bibliografia

Gibernans Bàguena, J.; Gil Estallo, À. J.; Rovira Escofet, C. (2009). *Estadística*. Barcelona: Material didàctic UOC.