Models de regressió lineal simple i múltiple amb R

Daniel Liviano Solís
Maria Pujol Jover

PID_00211043

Universitat Oberta de Catalunya

Índex

Introducció

Aquest mòdul té com a principal objectiu introduir l'estudiant en l'econometria utilitzant l'entorn estadístic R i la seva interfície R-Commander. L'ús de l'econometria és imprescindible en àmbits com l'economia, l'empresa i el màrqueting. De fet, l'econometria es podria definir com l'estadística aplicada a l'economia, ja que es basa en l'ús de la modelització. Les aplicacions més comunes són les que es mostren a continuació:

1) L'anàlisi estructural.

2) La predicció.

3) L'avaluació de polítiques.

En aquest primer mòdul es tracten els aspectes més bàsics d'aquesta disciplina. El primer capítol és de caràcter teòric, i té com a objectiu introduir formalment el model de regressió i la seva estimació per mínims quadrats ordinaris (MCO). Tot seguit, els dos capítols següents il·lustren, mitjançant exemples, el model de regressió lineal simple (MRLS) i el model de regressió lineal múltiple (MRLM) per mitjà d'exemples amb R i R-Commander. El quart capítol està dedicat a la inserció de variables qualitatives politòmiques i dicotòmiques, tant en un MRLS com en un MRLM, i també a la creació de variables fictícies segons els diferents usos que els podem donar. Finalment, el cinquè i últim capítol aborda el tema de la introducció de restriccions a un model de regressió i la seva estimació posterior per mínims quadrats restringits (MCR).

Objectius

1. Saber especificar correctament models de regressió lineals simples i múltiples (MRLS i MRLM).

2. Estimar MRLS i MRLM per mínims quadrats ordinaris (MCO) amb R i R-Commander.

3. Estimar MRLM per màxima versemblança (MV) utilitzant R i R-Commander.

4. Verificar mitjançant els resultats de l'estimació d'un model que es compleixen totes les hipòtesis bàsiques de tot MRLM.

5. Validar un model i quantificar la bondat del seu ajust.

6. Fer prediccions puntuals i per interval amb un model de regressió amb l'ajuda de R i R-Commander.

7. Incorporar variables dicotòmiques i politòmiques en un MRLM amb R i R-Commander.

8. Crear variables fictícies amb l'objectiu de satisfer diferents objectius d'anàlisi.

9. Introduir restriccions lineals en l'estimació d'MRLM amb R i R-Commander.

10. Estimar per mínims quadrats restringits (MCR) un MRLM amb R i R-Commander.

11. Distingir entre les propietats dels estimadors per MCO i MCR.

1. Introducció als models de regressió

1.1. Marc general

Un investigador, a l'hora d'analitzar estadísticament una sèrie de variables, ha de tenir en compte una diferència fonamental entre dos conceptes relacionats però diferents entre si:

- La **correlació** fa referència al grau de relació que hi ha entre dues variables, però no estableix cap tipus de relació de causa ni efecte d'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.

- El **model de regressió** suposa que no solament hi ha correlació entre les variables, sinó que a més a més hi ha una relació de *causalitat*, és a dir, una o més variables influeixen en l'altra.

Específicament, definim una anàlisi economètrica com un estudi de les relacions que s'estableixen entre un conjunt de variables. L'anàlisi parteix d'una sèrie d'observacions empíriques de les variables que es volen estudiar:

$$\{(y_1, x_1), (y_2, x_2), ..., (y_i, x_i), ..., (y_n, x_n)\} = \{(y_i, x_i) : i = 1, ..., n\}$$

Cada parell $\{y_i, x_i\} \in R \times R^k$ correspon a una **observació** d'una unitat (individu, empresa, família, etc.). El conjunt d'observacions considerades configuren la **mostra**, que està composta de n elements. Cada parell (y_i, x_i) està compost per dos elements: la **variable dependent** y_i i el **vector de regressors** o **variables independents** x_i. Mentre que per a cada unitat mostral la variable dependent és un escalar (nombre), els regressors formen un vector. Aquest vector té com a primer element una constant (en concret, el nombre 1), ja que fa referència al terme independent del model. La notació del vector és:

La variable explicada

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

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}_{k\times 1} = \begin{pmatrix} 1 \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}_{k\times 1}$$

Les variables explicatives

Aquestes variables es poden denominar de diferents maneres alternatives: exògenes, independents o regressores.

Aquest vector correspon a l'i-èsim element de la mostra, i inclou les observacions dels seus $k - 1$ regressors més la constant.

L'objectiu d'una regressió és el de trobar la tendència central de la distribució condicional de y_i, donat x_i. Una mesura de tendència central per excel·lència d'una variable és la mitjana. Per al cas condicional, la mesura anàloga és la **mitjana condicional o esperança condicionada** $m(x) = E(y_i|x_i = x)$ que pot prendre qualsevol forma: lineal, quadràtica, logarítmica, etc. No obstant això, com veurem a continuació, la forma més usual és la lineal.

Un element important de la regressió és l'**error** o **terme de pertorbació** u_i, i és definit com la diferència entre y_i i la mitjana condicional corresponent, és a dir:

$$u_i = y_i - m(x)$$

Matemàticament, podem reorganitzar l'equació anterior i obtenir-ne la fórmula:

$$y_i = m(x) + u_i$$

Els únics supòsits en què ens basem a l'hora de construir aquesta nova equació són que les variables (y_i, x_i) tenen una distribució conjunta i que $E|y_i| < \infty$.

És fonamental establir les propietats de u_i, és a dir, del terme d'error o pertorbació de la regressió:

1. $E(u_i|x_i) = 0$
2. $E(u_i) = 0$
3. $E(h(x_i)u_i) = 0$ per a qualsevol funció $h(\cdot)$
4. $E(x_i u_i) = 0$

Sovint l'equació $y_i = m(x) + u_i$ i la propietat $E(u_i|x_i) = 0$ es consideren el marc d'una regressió però no un model. En aquest marc no se suposa cap restricció (és a dir, no se suposa cap característica de la distribució conjunta de les variables, com podria ser la linealitat, per exemple). Així, totes dues equacions són certes per definició.

1.2. Model de regressió lineal

En el moment en què assumim certes restriccions en la distribució conjunta de les variables, obtenim un model. La restricció més comuna en econometria, per simplicitat i per facilitat d'estimació i inferència, és la de **linealitat**, és a dir, s'assumeix *a priori* que $m(x)$ és una funció lineal de (x). Aplicant aquesta restricció al marc definit anteriorment s'obté el **model de regressió lineal**:

$$y_i = x_i'\beta + u_i$$

$$E(u_i|x_i) = 0.$$

La restricció de linealitat no necessàriament s'ha de complir en qualsevol aplicació específica. La validesa d'aquesta restricció (i de qualsevol altra) dependrà, en tot cas, de les característiques de la distribució conjunta de les variables considerades.

En l'equació $y_i = x_i'\beta + u_i$ apareix, a més de la variable dependent y_i, el vector de regressors transposat x_i' i el terme de pertorbació de la regressió u_i, un nou vector, el **vector de paràmetres** β:

La propietat de transposició...

... intercanvia files per columnes d'una matriu o vector amb la finalitat de poder fer operacions matemàtiques, per exemple, la multiplicació entre objectes.

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}_{k\times 1}.$$

L'equació $y_i = x_i'\beta + u_i$ mostra la notació compacta del model de regressió lineal. Una forma més detallada seria:

$$y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i, \qquad i = 1, \ldots, n.$$

Aquest model de regressió lineal és incomplet sense una descripció del terme d'error u_i. Suposarem, *a priori*, que té una esperança nul·la $E(u_i) = 0$, que té variància finita $E(u_i^2) < \infty$, i que no està correlacionat amb els regressors $E(x_i u_i) = 0$. Aquesta última condició es denomina **condició d'ortogonalitat**, ja que implica que el vector de regressors i el d'errors són ortogonals. Com veurem més endavant, aquesta condició és dèbil i en molts casos no s'ha de complir necessàriament.

Ortogonalitat

Algebraicament, diem que els vectors a i b són ortogonals si es compleix que $a'b = 0$.

1.3. Notació matricial

Abans de tractar l'estimació del model, és important conèixer la notació matricial del model de regrsesió, ja que a partir d'ara s'utilitzarà, segons convingui, aquesta notació. La matriu de variables dependents adquireix la forma:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{1\times n}$$

com ja hem vist, el vector de paràmetres es defineix de la manera següent:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}_{k\times 1}.$$

considerant que el vector transposat de regressors, per a cada element mostral, és:

$$x_i' = (1 \;\; x_{2i} \;\; \cdots \;\; x_{ki})_{1\times k}$$

definim la matriu de regressors o variables independents de la manera següent:

$$X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}_{n\times k} = \begin{pmatrix} 1 & x_{21} & x_{31} & \cdots & x_{k1} \\ 1 & x_{22} & x_{32} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & x_{3n} & \cdots & x_{kn} \end{pmatrix}_{n\times k}$$

i, finalment, la matriu d'errors adquireix la forma:

$$u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}_{1\times n}$$

Com es pot observar, Y i β, igual que u, són vectors, mentre que X és una matriu.

El model de regressió lineal $y_i = x_i'\beta + u_i$ només mostra, per simplicitat, una equació per a l'individu i. Alternativament, podem expressar aquest model com un sistema de n equacions, una per a cada observació:

$$y_1 = x_1'\beta + u_1$$

$$y_2 = x_2'\beta + u_2$$

$$\cdots$$

$$y_n = x_n'\beta + u_n$$

o, equivalentment, mitjançant notació matricial:

$$Y = X\beta + u$$

Els productes següents també s'expressen de manera matricial:

$$\sum_{i=1}^{n} x_i x_i' = X'X$$

$$\sum_{i=1}^{n} x_i y_i = X'Y$$

1.4. Especificació i estimació

A l'hora d'analitzar mitjançant l'econometria les relacions que s'estableixen entre diferents variables, s'ha de tenir en compte que l'estudi d'un model de regressió es compon de dues etapes fonamentals:

1) Especificació: aquesta fase fa referència a la construcció del model, és a dir, quines variables s'inclouen i en quina forma funcional (lineal, additiva, multiplicativa, etc.).

2) Estimació: una vegada construït el model, l'estimació fa referència a la tècnica que s'utilitza per a obtenir estimacions dels paràmetres del model. El coneixement previ del fenomen analitzat i la disponibilitat de contrastos ajudaran a triar correctament l'estimador.

Pel que fa al procés d'estimació, definirem el vector de coeficients estimats com a $\hat{\beta}$, diferent del vector de paràmetres β, que és desconegut.

Hi ha diversos procediments per a obtenir una estimació del model de regressió lineal. El més senzill i immediat és el mètode dels **mínims quadrats ordinaris (MCO)**. Aquest estimador consisteix a minimitzar la suma al quadrat dels errors (SCE). Si definim la funció SCE de la manera següent:

$$S_n(\beta) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

podem definir l'estimador MCO com aquell que minimitza la funció SCE:

$$\hat{\beta}_{MCO} = \arg\min_{\beta} S_n(\beta)$$

així doncs, aquest estimador té l'expressió següent:

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$$

Hi ha diverses maneres d'arribar a aquesta expressió. Una és considerar el model de regressió lineal i premultiplicar-lo per x_i, per després prendre esperances:

$$x_i y_i = x_i x_i'\beta + x_i u_i$$

$$E(x_i y_i) = E(x_i x_i')\beta + E(x_i u_i) = E(x_i x_i')\beta$$

D'aquesta manera obtenim que el paràmetre β és una funció dels segons moments poblacionals de (y_i, x_i):

Els moments ens proporcionen informació sobre la distribució de la variable considerada.

$$\beta = E(x_i x_i')^{-1}E(x_i y_i)$$

Noteu que aquesta expressió dels paràmetres es basa en la hipòtesi $E(x_i u_i) = 0$, que no s'ha de complir necessàriament en la realitat. Per a obtenir estimacions dels paràmetres a partir d'aquesta expressió, n'hi ha prou de substituir els moments poblacionals per moments mostrals, que són:

$$\hat{E}(x_i y_i) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$\hat{E}(x_i x_i') = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$$

L'estimador adquireix llavors la forma següent:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i y_i = (X'X)^{-1}X'Y$$

Una altra manera alternativa d'obtenir l'estimador és partint de la condició d'ortogonalitat. Considerant la notació matricial, i tenint en compte que aquesta notació és $E(X'e) = 0$, obtenim:

$$E(X'u) = E(X'(Y - X\beta)) = E(X'Y) - E(X'X)\beta = 0$$

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

Com sempre, com que desconeixem els vertaders valors poblacionals, els paràmetres s'estimen substituint els moments poblacionals pels mostrals.

1.5. Interpretació del model

A l'hora d'interpretar un model de regressió hem de considerar diferents aspectes:

- Significació de cada un dels coeficients estimats. Es fa mitjançant el contrast de significació individual de cada paràmetre.

- Signe dels paràmetres. Si es tracta de coeficients significatius s'ha de veure si el signe obtingut en l'estimació té una lògica basada en teories.

- Magnitud dels paràmetres. En un model de regressió, cada un dels paràmetres que acompanyen les variables explicatives indica la variació que experimenta el valor esperat de la variable endògena davant un increment en una unitat de la variable explicativa a la qual fa referència el paràmetre estimat. En canvi, l'estimació de la constant indica quin és el valor esperat de la variable endògena quan totes les variables explicatives prenen valor nul.

- Significació general de tots els coeficients estimats. Es fa mitjançant el contrast de significació global dels paràmetres.

- Bondat de l'ajust. Amb els coeficients de determinació i de determinació ajustat es mesura quina part de la variable endògena queda explicada pel model estimat.

Encara que hi ha moltes classificacions diferents dels models de regressió, una classificació bàsica s'estableix en funció del nombre de regressors que conté el model:

- **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 $y = f(x_1, x_2, \ldots)$.

A continuació, s'ofereixen exemples d'estimació de models economètrics amb R i R-Commander atenent aquesta classificació. Posteriorment, s'analitza què hem de fer quan tenim alguna variable qualitativa que volem introduir com a regressor. Finalment, s'estudien els models de regressió quan s'imposen restriccions lineals *a priori*.

2. Model de regressió lineal simple (MRLS)

En el model de regressió simple s'estudia el comportament de la variable explicada (Y) en funció d'una sola 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.

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

$$y_i = \alpha + \beta x_i + u_i, \qquad i = 1, \ldots, N.$$

Encara que una explicació teòrica detallada del model de regressió s'inclou en el capítol precedent, recordem que en aquesta equació, α és la constant, β és el pendent i u és una variable aleatòria denominada *terme d'error* o *de pertorbació*. Com que és 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. Una vegada obtinguts aquests, 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 (u) 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 triat, amb la qual cosa 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, com més petits siguin els e_i més bona serà l'estimació del model i més fiables seran les prediccions sobre el comportament de Y obtingudes amb aquesta estimació.

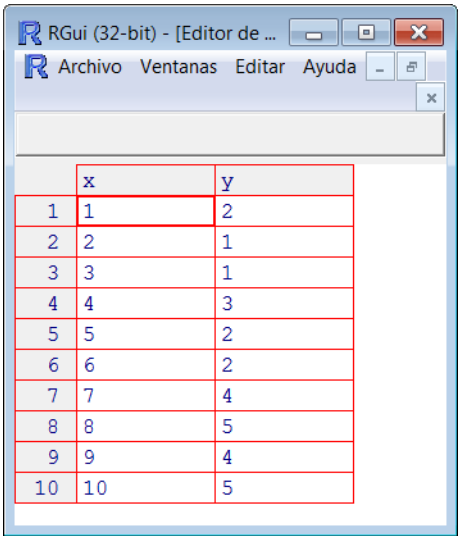
Com a exemple, suposem que disposem de $N = 10$ observacions de les variables X i Y:

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:

Datos / Nou conjunt de dades

Una vegada especificat un nom per a aquest conjunt de dades, les introduïm en un full on 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

Como vam veure en el mòdul dedicat a l'anàlisi descriptiva del manual *Matemáticas y Estadística con* R, és recomanable iniciar l'anàlisi amb estadístics bàsics de les variables. Una primera explotació estadística s'obté seguint les instruccions:

Estadístics / Resums / Conjunt de dades actiu

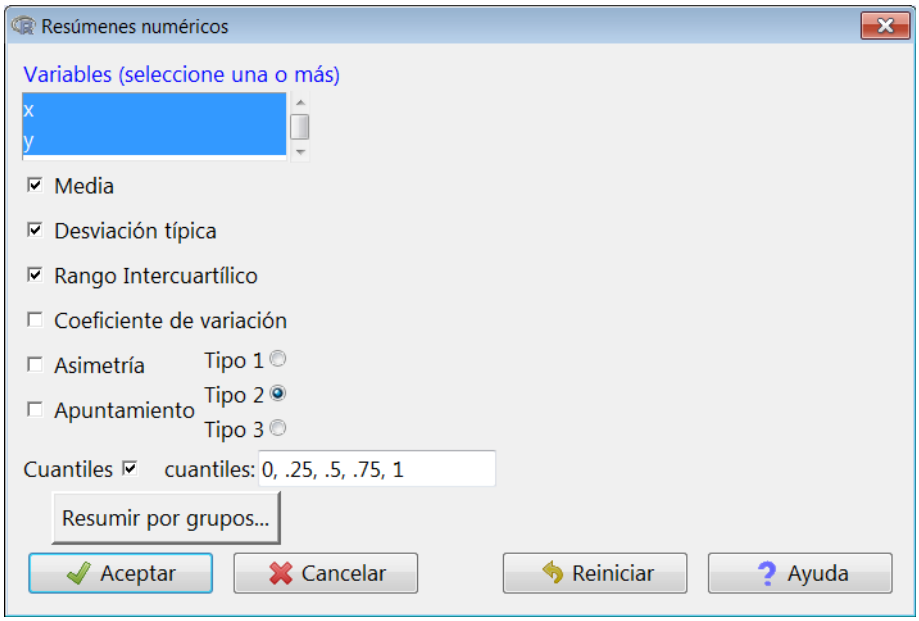
Amb això, el resultat serà el següent:

```
> summary(Datos)
       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
```

Moltes vegades 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 què es pot triar entre un conjunt d'estadístics; la ruta que haurem de seguir per a això serà:

Estadístics / Resums / Resums numèrics

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



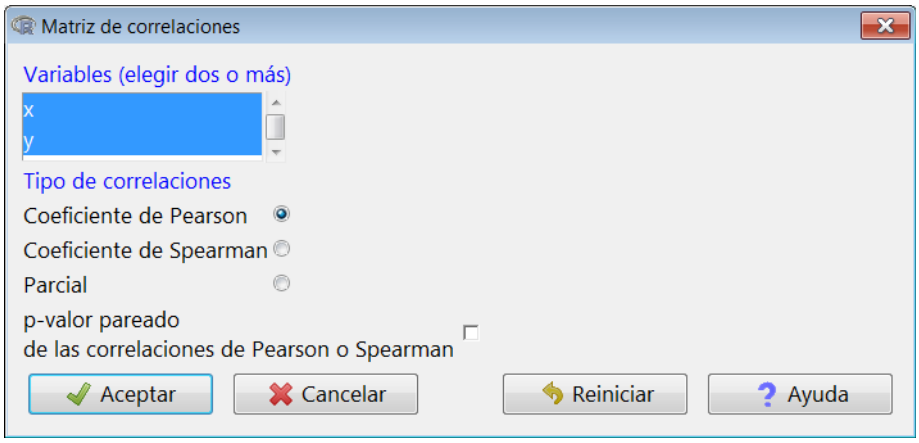
Aquest és el resultat que apareix en la *finestra de resultats*:

```
> numSummary(Datos[,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, s'ha de seguir la instrucció:

Estadístics / Resums / Matriu de correlacions

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



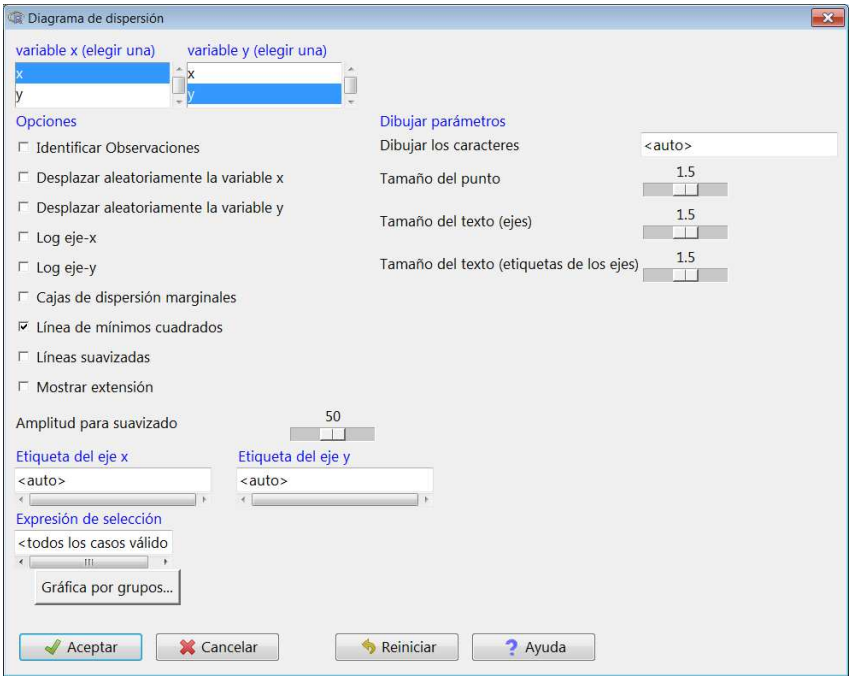
Com veiem, la correlació entre totes dues variables és positiva i bastant elevada:

```
> cor(Datos[,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 mitjançant un diagrama de dispersió de les variables X i Y. Obtenir aquest gràfic en R-Commander és immediat accedint a:

Gràfiques / Diagrama de dispersió

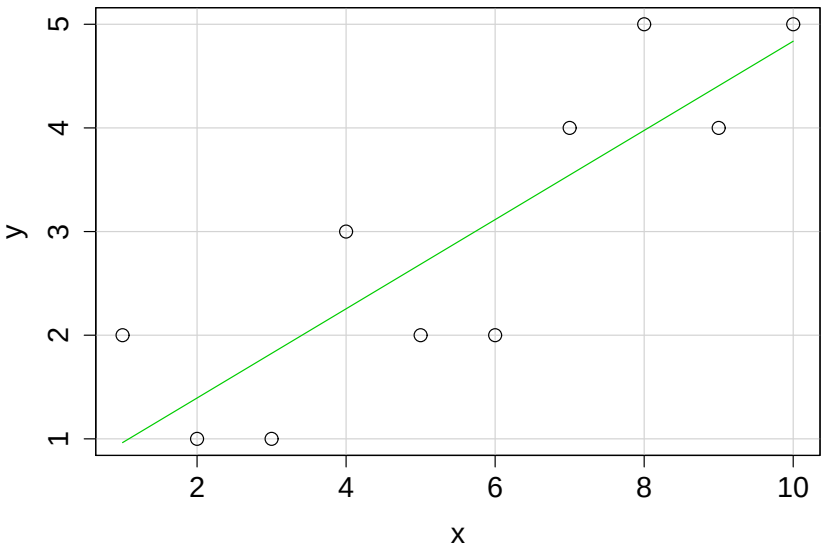
Apareixerà el menú següent, on especificarem la variable x (corresponent a l'eix horitzontal) i y (corresponent a l'eix vertical). A més a més, activarem l'opció *Línia de mínims quadrats*, que dibuixa la recta de regressió sobre els punts.



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

Recta de regressió

Noteu 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, en primer lloc 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 usant la sintaxi del llenguatge propi de R és immediat. N'hi ha prou de tenir en compte els operadors descrits en la taula 1, ja vistos en el primer mòdul del manual *Matemáticas y Estadística con* R.

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`
Media	`mean(x)`	`5.5`
Media	`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*:

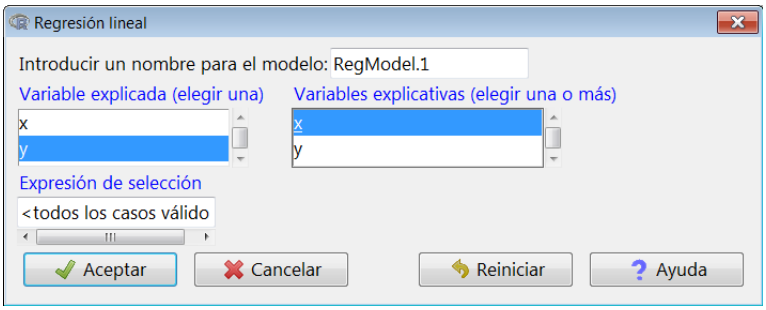
```
> attach(Datos)
> beta <- cov(x,y)/var(x)
> alpha <- mean(y)-beta*mean(x)
> coef.det <- cor(x,y)^2

> 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 a més inclou més informació estadística del model. Una vegada que les variables X i Y s'han introduït en una base de dades, s'ha d'estimar un model. La manera més senzilla és seguir la ruta següent:

Estadístics / Ajust de models / Regressió lineal

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



En la *Finestra de resultats* apareixerà:

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

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 principals components:

- `Residuals`: mínim, màxim i quartils dels residus de la regressió, que 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.

Contrast de significació individual...

... és un contrast les hipòtesis del qual són H_0 : *parámetro* $= 0$ i H_1 : *parámetro* $\neq 0$.

- `Pr(>|t|)`: *p-valor* del contrast de significació individual de cada paràmetre estimat, que n'indica la significació estadística.

- `Signif. codes`: mostra, amb asteriscos i punts, per a quins nivells de significació els coeficients estimats són o no significatius. 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`).

Contrast de significació global...

... és un contrast les hipòtesis del qual són H_0 : *todos los parámetros* $= 0$ i H_1 : *algún parámetro* $\neq 0$.

- `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.

- `DF`: graus de llibertat de l'estadístic F.

- `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 del 0.1% ($p\text{-valor} < 0,001$).

Una manera alternativa d'estudiar la significació individual dels paràmetres estimats és el càlcul d'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:

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

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

El valor en taules de t depèn del nivell de significació (α) i del nombre de paràmetres que s'han d'estimar (2, en particular per a un MRLS i k en general per a un MRLM), ja que aquest valor és $t_{\alpha/2;\,N-k}$.

Models / Intervals de confiança

Apareixerà un quadre de diàleg, on s'ha d'especificar el nivell de confiança que es vol aconseguir i després de prémer *Acceptar* 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 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, cosa que equival a afirmar que no és estadísticament diferent de zero. Veiem que això no passa en el cas del pendent.

3. Model de regressió lineal múltiple (MRLM)

L'MRLM és una generalització del model simple en k paràmetres (incloent-hi la constant). Per tant, la variable endògena s'explica per més d'una variable exògena:

$$y_i = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u_i, \qquad i = 1, \ldots, N.$$

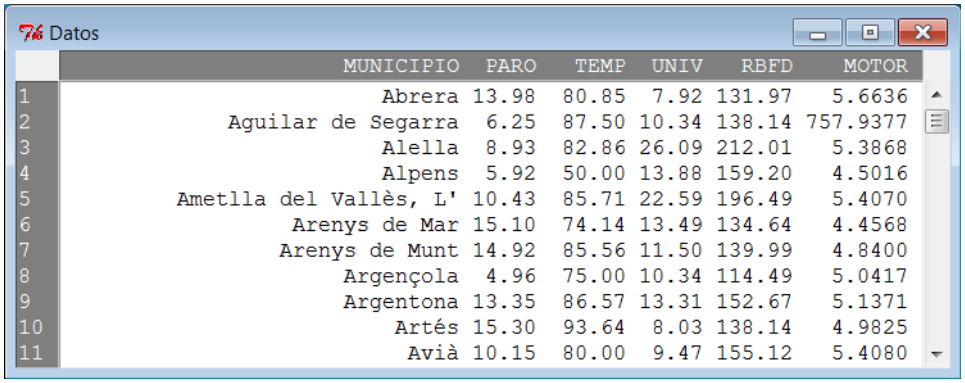
Considerem un exemple pràctic. Es vol estudiar un model de regressió lineal per a estudiar els determinants del nivell d'atur en diferents municipis catalans, en concret, $N = 295$. El model que hem d'especificar és:

$$PARO_i = \beta_1 + \beta_2 MOTOR_i + \beta_3 RBFD_i + \beta_4 TEMP_i + \beta_5 UNIV_i + u_i.$$

On tenim les variables següents:

- *PARO*: 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, una vegada importades del document d'Excel, són les següents:


Datos

	MUNICIPIO	PARO	TEMP	UNIV	RBFD	MOTOR
1	Abrera	13.98	80.85	7.92	131.97	5.6636
2	Aguilar de Segarra	6.25	87.50	10.34	138.14	757.9377
3	Alella	8.93	82.86	26.09	212.01	5.3868
4	Alpens	5.92	50.00	13.88	159.20	4.5016
5	Ametlla del Vallès, L'	10.43	85.71	22.59	196.49	5.4070
6	Arenys de Mar	15.10	74.14	13.49	134.64	4.4568
7	Arenys de Munt	14.92	85.56	11.50	139.99	4.8400
8	Argençola	4.96	75.00	10.34	114.49	5.0417
9	Argentona	13.35	86.57	13.31	152.67	5.1371
10	Artés	15.30	93.64	8.03	138.14	4.9825
11	Avià	10.15	80.00	9.47	155.12	5.4080

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

```
> summary(Datos)
                  MUNICIPIO        PARO            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.   : 84.9   Min.   :  2.527
 1st Qu.: 7.57   1st Qu.:123.5   1st Qu.:  4.679
 Median :10.09   Median :138.1   Median :  5.030
 Mean   :10.98   Mean   :140.4   Mean   : 10.934
 3rd Qu.:13.04   3rd Qu.:155.1   3rd Qu.:  5.460
 Max.   :33.61   Max.   :249.6   Max.   :784.879
```

Com hem fet abans, per a veure més estadístics de les variables es poden seguir les instruccions que es mostren a continuació i seleccionar el que més ens interessi calcular de la llista d'estadístics que ofereix R-Commander:

Estadístics / Resums / Resums numèrics

El resultat obtingut és el següent:

```
> numSummary(Datos[,c("MOTOR", "PARO", "RBFD", "TEMP", "UNIV")
   ],
statistics=c("mean", "sd"), quantiles=c())
           mean        sd %   n
MOTOR  10.93426 63.558945 0 295
PARO   13.41831  4.185926 0 295
RBFD  140.44332 24.227029 0 295
TEMP   82.73736 16.544553 0 295
UNIV   10.98186  4.914490 0 295
```

Si dues o més variables tenen entre elles una alta correlació, 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 el mòdul següent, una alta correlació entre dos regressors pot donar lloc a problemes de *multicol·linealitat*.

Estadístics / Resums / Matriu de correlacions

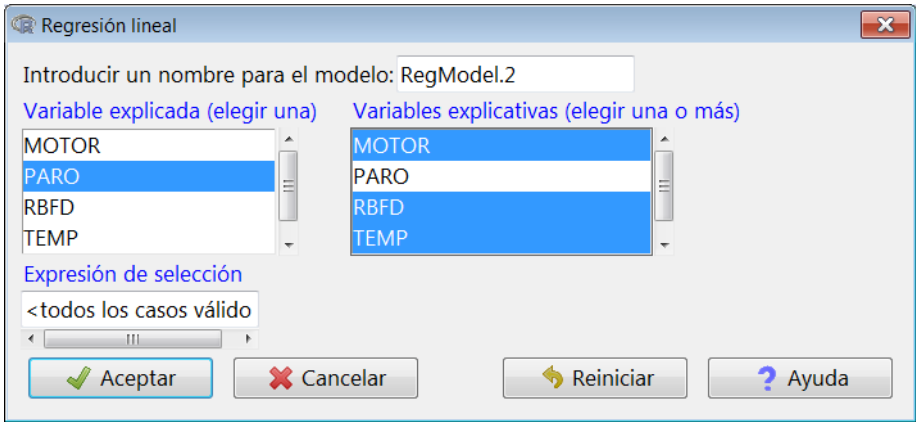
Seleccionant les variables que volem incloure, obtenim el resultat següent:

```
> cor(Datos[,c("MOTOR","PARO","RBFD","TEMP","UNIV")],
+ use="complete.obs")
            MOTOR        PARO         RBFD         TEMP         UNIV
MOTOR  1.00000000  -0.1292680  -0.01265723  -0.06906589   0.01746496
PARO  -0.12926800   1.0000000  -0.41906630   0.16408409  -0.46244755
RBFD  -0.01265723  -0.4190663   1.00000000  -0.10259722   0.58442097
TEMP  -0.06906589   0.1640841  -0.10259722   1.00000000  -0.03479639
UNIV   0.01746496  -0.4624475   0.58442097  -0.03479639   1.00000000
```

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

Estadístics / Ajust de models / Regressió lineal

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



En la finestra de resultats obtenim el següent:

```
> RegModel.2 <- lm(PARO~MOTOR+RBFD+TEMP+UNIV, data=Datos)
> summary(RegModel.1)
Call:
lm(formula = PARO ~ MOTOR + RBFD + TEMP + UNIV, data = Datos)

Residuals:
     Min       1Q   Median       3Q      Max
-12.4220  -1.6582   0.4862   2.2200   8.5622

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 19.244961   1.738272  11.071  < 2e-16 ***
```

```
MOTOR       -0.007755   0.003296  -2.353 0.019290 *
RBFD        -0.037110   0.010685  -3.473 0.000593 ***
TEMP         0.030971   0.012728   2.433 0.015569 *
UNIV        -0.281595   0.052421  -5.372  1.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.581 on 290 degrees of freedom
Multiple R-squared: 0.2782, Adjusted R-squared: 0.2682
F-statistic: 27.94 on 4 and 290 DF,  p-value: < 2.2e-16
```

Com veiem, tots els coeficients estimats són significatius, encara que l'ajust del model ($R^2 = 0,278$) és més aviat pobre.

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

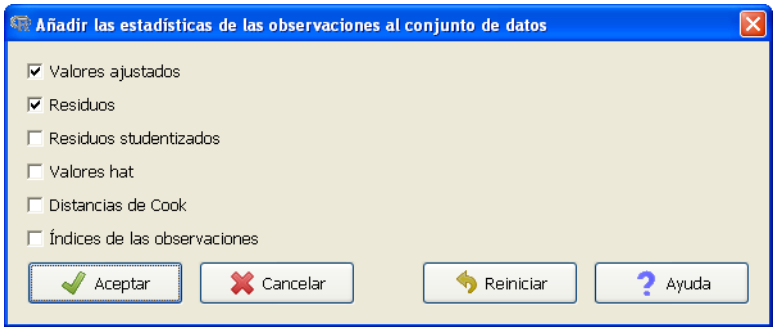
```
> Confint(RegModel.2, level=0.95)
                Estimate        2.5 %       97.5 %
(Intercept) 19.244961330 15.823732809 22.666189851
MOTOR       -0.007755427 -0.014242476 -0.001268379
RBFD        -0.037110206 -0.058139257 -0.016081156
TEMP         0.030971082  0.005919276  0.056022888
UNIV        -0.281595276 -0.384769255 -0.178421297
```

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

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 / Afegir les estadístiques de les observacions a les dades

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



Una vegada hem fet això, si visualitzem el nostre conjunt de dades, observarem com s'han afegit aquestes dues variables:

Datos

	MUNICIPIO	PARO	TEMP	UNIV	RBFD	MOTOR	fitted.RegModel.1	residuals.RegModel.1
1	Abrera	13.98	80.85	7.92	131.97	5.6636	14.577381	-0.59738113
2	Aguilar de Segarra	6.25	87.50	10.34	138.14	757.9377	8.038701	-1.78870105
3	Alella	8.93	82.86	26.09	212.01	5.3868	6.554893	2.37510738
4	Alpens	5.92	50.00	13.88	159.20	4.5016	10.942116	-5.02211630
5	Ametlla del Vallès, L'	10.43	85.71	22.59	196.49	5.4070	8.204537	2.22546259
6	Arenys de Mar	15.10	74.14	13.49	134.64	4.4568	12.711354	2.38864552
7	Arenys de Munt	14.92	85.56	11.50	139.99	4.8400	13.423907	1.49609265
8	Argençola	4.96	75.00	10.34	114.49	5.0417	14.368249	-9.40824924
9	Argentona	13.35	86.57	13.31	152.67	5.1371	12.472639	0.87736086
10	Artés	15.30	93.64	8.03	138.14	4.9825	14.718838	0.58116197

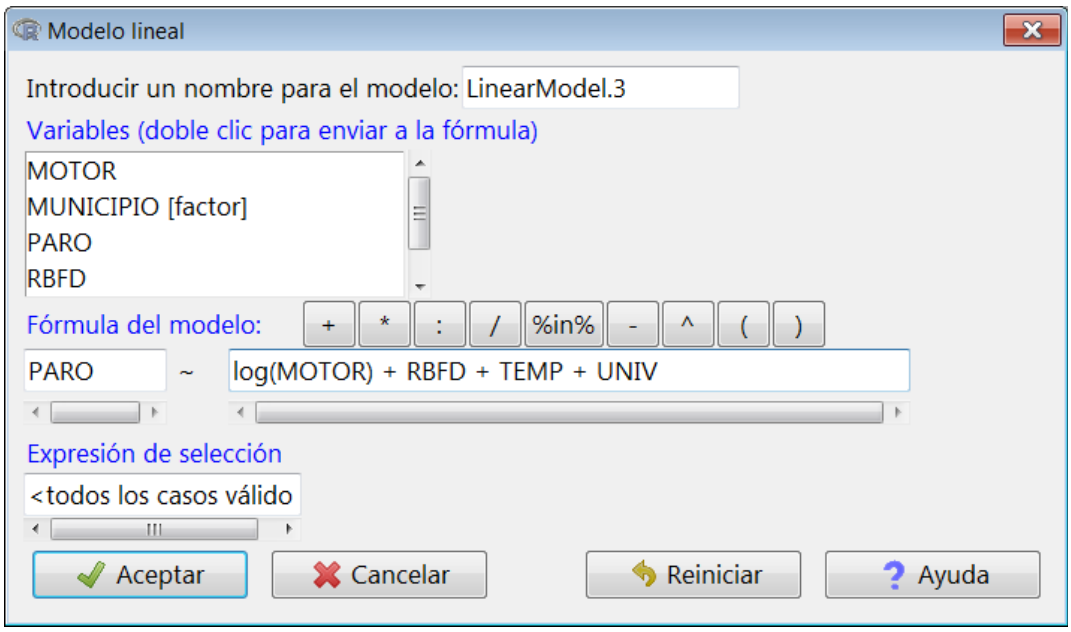
El més lògic és que no estiguem satisfets amb el model estimat i vulguem millorar-ne l'estimació. Podem optar pel següent model alternatiu, on la variable *MOTOR* apareix en logaritmes:

$$PARO_i = \beta_1 + \beta_2 log(MOTOR)_i + \beta_3 RBFD_i + \beta_4 TEMP_i + \beta_5 UNIV_i + u_i.$$

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. No obstant això, 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 / Ajust de models / Model lineal

Ens apareixerà un quadre de diàleg en què 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 a més, assignarem a aquest model el nom *LinearModel.3*:



El resultat que s'obté és el següent:

```
> LinearModel.3 <- lm(PARO ~ log(MOTOR) +TEMP +UNIV +RBFD, data
    =Datos)
> summary(LinearModel.2)
Call:
lm(formula = PARO ~ log(MOTOR) + TEMP + UNIV + RBFD, data =
    Datos)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9532  -1.5396   0.5311   2.0798   8.4970

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) 22.56227    1.87918  12.006  < 2e-16 ***
log(MOTOR)  -1.85324    0.41466  -4.469 1.13e-05 ***
TEMP         0.02791    0.01245   2.241 0.025764 *
UNIV        -0.27547    0.05121  -5.379 1.54e-07 ***
RBFD        -0.03788    0.01043  -3.631 0.000333 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.496 on 290 degrees of freedom
Multiple R-squared: 0.3118, Adjusted R-squared: 0.3023
F-statistic: 32.85 on 4 and 290 DF,  p-value: < 2.2e-16
```

3.1. Comparació dels dos models

Comprovem que l'ajust del model ha millorat respecte al model anterior. És important destacar que el resultat de l'estimació sempre mostra dos valors del coeficient de determinació; un 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 pujarà, 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 aquell que tingui un valor més gran de la R^2 ajustada.

4. Variables exògenes qualitatives

En qualsevol estudi, és habitual trobar-nos amb models de regressió en què es preveuen situacions amb variables explicatives; són atributs o variables de caràcter qualitatiu. La codificació d'aquestes variables suposa identificar cada categoria amb un valor. Per aquest motiu, quan les categories no tenen una ordenació clara i recorrem a criteris arbitraris per fer la codificació, hem de prestar especial atenció. Per exemple, sense adonar-nos-en, en ordenar les destinacions turístiques amb més afluència de passatgers per ordre alfabètic, automàticament estem establint una prioritat que pot distar molt de la realitat.

Per norma general, per no tenir problemes d'interpretació, entre d'altres, evitarem introduir directament com a variable explicativa d'un model de regressió una variable qualitativa politòmica, la codificació numèrica de la qual indueixi un ordre (i, com a conseqüència, una distància) entre les diferents categories que no s'ajustin a la realitat.

Així doncs, el primer pas és classificar totes les possibles variables explicatives d'un model de regressió. Ja hem vist que quan es tracta de variables quantitatives o qualitatives amb ordre implícit, no tenim cap problema i les podem introduir directament. El problema el tenim quan ens trobem amb atributs o variables qualitatives. És crucial distingir aquí si són variables dicotòmiques o politòmiques, ja que en aquest últim cas les haurem de desglossar en diferents variables dicotòmiques.

Variables dicotòmiques, fictícies o *dummies*

Són variables que solament poden prendre dos valors, generalment 0 i 1. La categoria codificada amb un 0 se sol denominar **categoria de referència**.

4.1. Variables dicotòmiques i qualitatives politòmiques en l'MRLM

Ja hem esmentat que els valors més utilitzats en la codificació de les variables dicotòmiques són el 0 i l'1. La raó és perquè simplifica enormement la interpretació dels resultats obtinguts després d'estimar el model de regressió. Sabem que en un model de regressió lineal totes les variables explicatives tenen associat un paràmetre que hem d'estimar. Si aquesta variable pren el valor 0, l'impacte sobre el valor esperat de la variable dependent s'anul·la. En aquest cas, el paràmetre associat a una variable fictícia mostra quant varia el valor esperat de la variable endògena quan un individu posseeixi una determinada característica (identificada amb el valor 1 de la *dummy*), respecte a un individu que no la tingui (identificada amb el valor 0 de la *dummy*).

Podem introduir variables fictícies en un model de regressió de diferents maneres:

- Additiva: com una variable explicativa més.

- Multiplicativa: multiplicant una altra variable explicativa que ja existeix en el model.

- Mixta: tant additivament com multiplicativament.

Lògicament, segons com decidim incorporar les *dummies* al model la interpretació que en farem serà diferent. No obstant això, considerem un model molt senzill per il·lustrar el que acabem de comentar. Suposem que tenim una classe de 90 alumnes d'econometria i que volem explicar la qualificació obtinguda al final de curs (Y_i) en funció de les hores d'estudi (X_{1i}) i del gènere (X_{2i}).

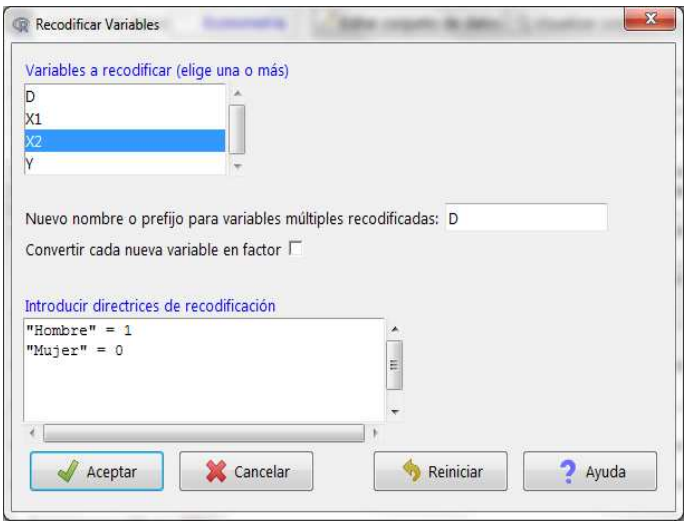
En primer lloc llegim la base de dades i la visualitzem, veiem que la variable X_2 és textual i per tant ens interessa recodificar-la com una variable fictícia, que denominarem D_i. Per a fer això amb R-Commander utilitzarem la ruta següent:

Dades / Modificar variables dins del conjunt de dades actiu / Recodificar variables

Ens sortirà un quadre de diàleg, que omplirem segons els nostres criteris de definició de la *dummy*. Per exemple, volem definir:

$$D_i = \begin{cases} 1 & \text{si} \quad X_{2i} = Hombre \\ 0 & \text{si} \quad X_{2i} = Mujer \end{cases}$$

Per tant:



Ara, en visualitzar el conjunt de dades, veiem que ens apareix la nova variable fictícia que acabem de crear. Així doncs, si suposem que la nostra base de dades està ordenada pel gènere dels estudiants (primer els homes i després les dones), podem definir un

model general per a tots els nostres alumnes independentment del gènere:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_i + u_i \qquad i = 1, \ldots, N.$$

un model per als homes, suposant que les primeres N_1 observacions corresponen a homes:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 + u_i = \beta_0 + \beta_2 + \beta_1 X_{1i} + u_i \qquad i = 1, \ldots, N_1.$$

i un altre per a les dones, suposant que des de l'observació $N_1 + 1$ a la N són dones:

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i \qquad i = N_1 + 1, \ldots, N.$$

Observem que en introduir la variable fictícia de manera additiva la diferència entre l'especificació dels homes i de les dones recau sobre la constant, que per als homes és $\beta_0 + \beta_2$, mentre que per a les dones es limita a β_0. Cal destacar també que l'efecte produït en la qualificació de les hores d'estudi és el mateix independentment del gènere de l'estudiant.

El resultat de l'estimació dels tres models utilitzant la ruta:

Estadístics / Ajust de models / Regressió lineal

i omplint els corresponents quadres de diàleg és:

Fixeu-vos que si valorem els models per MCO per separat no ens donen exactament els mateixos resultats; la raó és que en estimar el model complet assumim que la variància del terme de pertorbació és la mateixa per als homes que per a les dones. En canvi, quan valorem dos models per separat, la variància no ha de ser necessàriament la mateixa.

```
> EcoModel.1 <- lm(Y~X1+D, data=Econometria)
> summary(EcoModel.1)
Call:
lm(formula = Y ~ X1 + D, data = Econometria)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5879 -0.2385  0.0011  0.1892  3.3932

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -0.419483   0.121693  -3.447 0.000875 ***
X1           0.176411   0.003034  58.154  < 2e-16 ***
D            0.028347   0.096139   0.295 0.768808
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4554 on 87 degrees of freedom
Multiple R-squared:  0.975, Adjusted R-squared:  0.9744
F-statistic:  1694 on 2 and 87 DF,  p-value: < 2.2e-16
```

```
> EcoModel.2 <- lm(Y~X1, data=Econometria, subset=D==1)
> summary(EcoModel.2)
Call:
lm(formula = Y ~ X1, data = Econometria, subset = D == 1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5936 -0.3258 -0.0238  0.1379  3.3927

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -0.398787   0.222335  -1.794   0.0801 .
X1           0.176650   0.006348  27.829   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6018 on 42 degrees of freedom
Multiple R-squared:  0.9486,  Adjusted R-squared:  0.9473
F-statistic: 774.4 on 1 and 42 DF,  p-value: < 2.2e-16
```

```
> EcoModel.3 <- lm(Y~X1, data=Econometria, subset=D==0)
> summary(EcoModel.3)
Call:
lm(formula = Y ~ X1, data = Econometria, subset = D == 0)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58211 -0.18030  0.02911  0.19660  0.43663

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) -0.41417    0.08198  -5.052 8.11e-06 ***
X1           0.17625    0.00218  80.844  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2538 on 44 degrees of freedom
Multiple R-squared:  0.9933,  Adjusted R-squared:  0.9932
F-statistic:  6536 on 1 and 44 DF,  p-value: < 2.2e-16
```

Vegem ara un cas en què una de les variables explicatives tingui més de dues categories, per exemple l'itinerari d'estudis que segueix l'estudiant. Aquest pot ser:

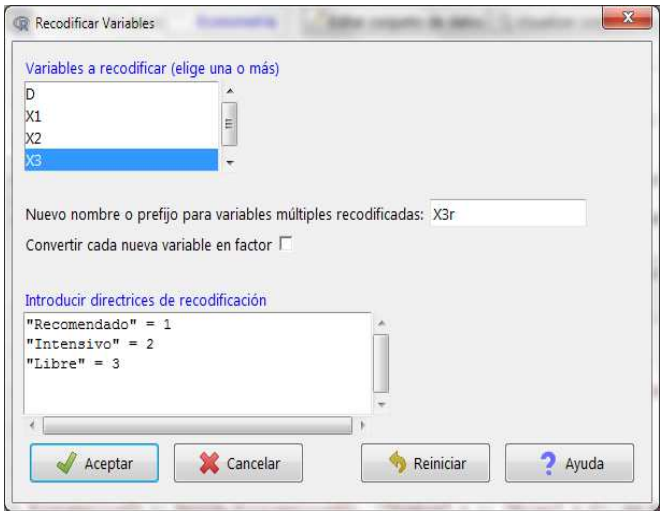
- Recomanat: el que s'aconsella als estudiants per a obtenir el grau en 10 semestres.
- Intensiu: el que es pauta amb la finalitat d'obtenir el grau en 8 semestres.
- Lliure: qualsevol altre que hagi escollit l'estudiant però que no estigui marcat per la universitat.

Com sempre, en primer lloc llegirem la base de dades i, en cas que aquesta sigui un factor, la recodificarem com hem fet abans però ara assignant un 1 a la primera categoria, un 2 a la segona i així successivament fins a j, que serà el màxim de categories possibles. El nostre exemple continua essent la mateixa classe de 90 alumnes d'econometria i volem explicar la qualificació obtinguda a final de curs (Y_i) en funció de les hores d'estudi (X_{1i}) i, en aquest cas, segons l'itinerari que hagin triat (X_{3i}). Per fer aquesta acció amb R-Commander seguirem la ruta:

Dades / Modificar variables dins del conjunt de dades actiu / Recodificar variables

Ens sortirà un quadre de diàleg que omplirem segons els nostres criteris de definició de la variable que volem recodificar:

$$X_{3ri} = \begin{cases} 1 & \text{si} \quad X_{3i} = Recomendado \\ 2 & \text{si} \quad X_{3i} = Intensivo \\ 3 & \text{si} \quad X_{3i} = Libre \end{cases}$$



No obstant això, no podem utilitzar directament aquesta nova variable perquè implícitament estaríem imposant un ordre i una proporcionalitat que no és certa. Per tant, hem de definir tantes variables dicotòmiques com el nombre total de categories excepte una. En el nostre cas n'hi ha prou de definir dues variables fictícies:

$$D_{2i} = \begin{cases} 1 & \text{si} \quad X_{3ri} = 2 \\ 0 & \text{si} \quad X_{3ri} \neq 2 \end{cases}$$

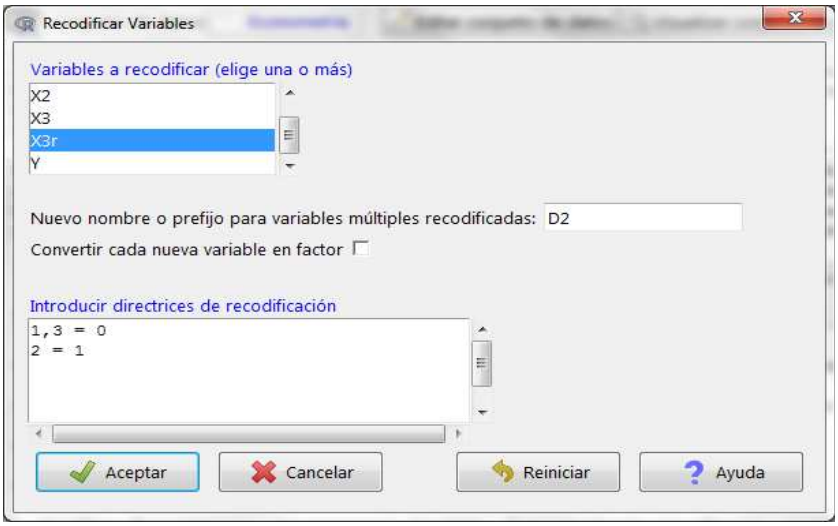
Fixeu-vos que D_{2i} pren valor 1 quan els estudiants segueixen l'itinerari intensiu i 0 altrament, és a dir, si opten per l'itinerari recomanat o el lliure. En canvi, D_{3i} valdrà 1 quan l'itinerari escollit sigui lliure i 0 si aquest és el recomanat o l'intensiu.

$$D_{3i} = \begin{cases} 1 & \text{si} \quad X_{3ri} = 3 \\ 0 & \text{si} \quad X_{3ri} \neq 3 \end{cases}$$

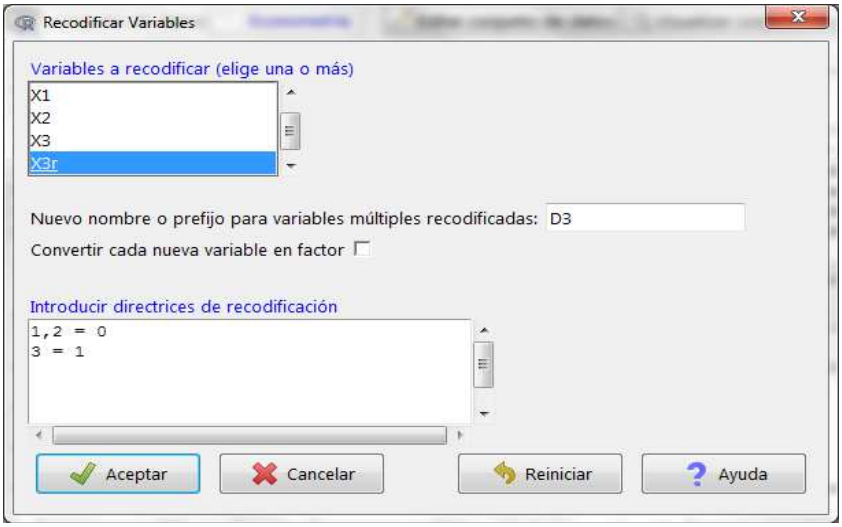
Farem això seguint la ruta següent:

Dades / Modificar variables dins del conjunt de dades actiu / Recodificar variables

i omplint els dos quadres de diàleg corresponents a les dues noves variables:



Un estudiant que segueixi l'itinerari recomanat tindrà assignat un zero tant a D_{2i} com a D_{3i}.



Categoria base o de referència...

és aquella categoria que resulta quan totes les *dummies* introduïdes en el model prenen el valor zero. En el nostre cas serien o bé aquells estudiants homes que segueixen l'itinerari recomanat si introduïm tant el gènere com l'itinerari en el model o bé els estudiants que segueixen l'itinerari recomanat en cas de no tenir en compte el gènere.

Observem que la informació respecte a l'itinerari dels estudiants d'econometria la podem facilitar amb una única variable codificada amb tres valors (X_{3r}) o amb dues variables dicotòmiques (D_2 i D_3). No obstant això, per a introduir aquesta informació en un model de regressió s'utilitza el criteri de definir tantes variables fictícies com el nombre total de categories menys una per a evitar caure en la *trampa de les variables fictícies*.

Trampa de les fictícies...

... es produeix quan introduïm una *dummy* per a cada categoria en un model de regressió amb terme independent. En aquest cas, com veurem en el mòdul següent, incorriríem en un problema de multicol·linealitat perfecta i no podríem estimar el model.

El model que especificaríem seria, doncs:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \qquad i = 1, \ldots, N.$$

i l'estimació ens donaria:

```
> EcoModel.4 <- lm(Y~X1+D2+D3, data=Econometria)
> summary(EcoModel.4)
Call:
lm(formula = Y ~ X1 + D2 + D3, data = Econometria)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6183 -0.2379 -0.0107  0.1790  3.3724

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -0.342012   0.131838  -2.594   0.0111 *
X1           0.175576   0.003115  56.358   <2e-16 ***
D2          -0.018524   0.110708  -0.167   0.8675
D3          -0.161528   0.132805  -1.216   0.2272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4542 on 86 degrees of freedom
Multiple R-squared:  0.9754,  Adjusted R-squared:  0.9745
F-statistic:  1136 on 3 and 86 DF,  p-value: < 2.2e-16
```

Per tant, sempre que tinguem informació recollida en variables qualitatives haurem d'utilitzar *dummies* per a introduir-la en qualsevol model de regressió. Sempre podrem fer això independentment del tipus d'informació que vulguem recollir: podem crear variables fictícies espacials per identificar, per exemple, la zona de residència dels nostres estudiants; variables fictícies temporals, per a identificar si han nascut en un determinat període, etc.

4.2. Interpretació dels coeficients de les variables fictícies

La interpretació dels paràmetres de les *dummies* variarà segons la manera com hàgim introduït aquestes variables en el model:

Si utilitzéssim una altra codificació diferent del 0 i l'1, la interpretació del model es complicaria molt.

- Additiva
- Multiplicativa
- Mixta

4.2.1. Introducció de *dummies* en un model de manera additiva

El paràmetre estimat d'una variable fictícia incorporada a un model de regressió de manera additiva s'interpreta com la variació que es produirà en el valor esperat de la variable endògena quan l'individu pertanyi a la categoria identificada amb valor 1 respecte al que tindria si l'individu pertanyés a la categoria complementària (identificada

amb un 0). Aquesta variació serà siempre la mateixa, al marge del valor que adquireixi la resta de les variables explicatives que conté el model.

En el capítol següent s'aprofundirà sobre com es fan restriccions lineals en un model de regressió.

Fixeu-vos que si el que interpretem és un contrast estadístic com els que podríem fer en el model general de l'exemple dels estudiants d'econometria, primer hem de reflexionar sobre el significat dels coeficients que acompanyen cada una de les variables explicatives del model.

Per exemple, recuperem el primer model especificat en la secció anterior, on volíem explicar la qualificació obtinguda al final de curs (Y_i) de 90 estudiants d'econometria en funció de les hores d'estudi (X_{1i}) i del gènere recollit en la variable fictícia (D_i):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_i + u_i \qquad i = 1, \ldots, N.$$

l'estimació de la qual era:

```
> EcoModel.1 <- lm(Y~X1+D, data=Econometria)
> summary(EcoModel.1)
Call:
lm(formula = Y ~ X1 + D, data = Econometria)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5879 -0.2385  0.0011  0.1892  3.3932

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -0.419483   0.121693  -3.447 0.000875 ***
X1           0.176411   0.003034  58.154  < 2e-16 ***
D            0.028347   0.096139   0.295 0.768808
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4554 on 87 degrees of freedom
Multiple R-squared:  0.975, Adjusted R-squared:  0.9744
F-statistic:  1694 on 2 and 87 DF,  p-value: < 2.2e-16
```

Vegem les interpretacions de cada un dels paràmetres del model mitjançant els contrastos de significació individual (CSI) d'aquests:

La interpretació del paràmetre β_1 que acompanya la variable X_1 és la mateixa que en un model de regressió, és a dir, quan un estudiant incrementa en una unitat les hores d'estudi, la qualificació esperada incrementa en uns 0.18 punts.

- El paràmetre β_2 compara si havent estudiat les mateixes hores, la qualificació esperada d'un home és igual o diferent de la d'una dona. En el nostre exemple, no rebutgem la hipòtesi nul·la, per tant, la qualificació esperada hauria de ser la mateixa per als homes i les dones.

- El paràmetre β_0 mostra si la qualificació mitjana de les dones és significativament diferent de zero quan no s'ha estudiat cap hora (X_1). El nostre exemple mostra que la qualificació mitjana de les dones que no han estudiat cap hora és de gairebé mig punt menys que la que han obtingut aquelles que han invertit hores d'estudi. Si es volgués fer el mateix amb els homes, hauríem de fer un contrast de restriccions lineals (que veurem en el capítol següent) amb $\beta_0 + \beta_2 = 0$.

Si ara ens centrem en el model on explicàvem la qualificació d'econometria (Y_i) en funció de les hores d'estudi (X_{1i}) i l'itinerari que han seguit els estudiants (D_2 i D_3):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \qquad i = 1, \ldots, N.$$

els resultats d'estimació del qual eren:

```
> EcoModel.4 <- lm(Y~X1+D2+D3, data=Econometria)
> summary(EcoModel.4)
Call:
lm(formula = Y ~ X1 + D2 + D3, data = Econometria)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6183 -0.2379 -0.0107  0.1790  3.3724

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -0.342012   0.131838  -2.594   0.0111 *
X1           0.175576   0.003115  56.358   <2e-16 ***
D2          -0.018524   0.110708  -0.167   0.8675
D3          -0.161528   0.132805  -1.216   0.2272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4542 on 86 degrees of freedom
Multiple R-squared:  0.9754,  Adjusted R-squared:  0.9745
F-statistic:  1136 on 3 and 86 DF,  p-value: < 2.2e-16
```

En aquest cas, la interpretació dels CSI dels paràmetres serien:

- El paràmetre β_3 indica si la qualificació esperada d'un estudiant que segueix un itinerari lliure és igual o diferent de la d'un altre que seguia qualsevol altre itinerari quan les hores d'estudi es mantinguin constants. En aquest exemple, el fet de seguir un itinerari lliure no afecta la qualificació esperada de l'estudiant.

- El paràmetre β_2 compara si, havent estudiat les mateixes hores, la qualificació esperada d'un estudiant que segueix un itinerari intensiu és igual o diferent de la d'un altre que seguia qualsevol altre itinerari. Una altra vegada, la qualificació esperada de l'estudiant no queda alterada perquè aquest hagi triat seguir un itinerari intensiu.

- El paràmetre β_0 mostra si la qualificació mitjana dels estudiants que han seguit l'itinerari recomanat és significativament diferent de zero quan no s'ha estudiat cap hora (X_1). Veiem que si un estudiant que segueix l'itinerari recomanat no estudia cap hora, la seva qualificació es redueix en uns 0.34 punts.

- La diferència de qualificació esperada entre un estudiant que segueix un itinerari intensiu i un altre que en segueix un de lliure, amb les mateixes hores d'estudi, estaria determinada per $\beta_2 - \beta_3$. Si no atenem la significació dels coeficients, en el nostre exemple aquesta diferència seria de $-0.018524 - (-0.161528) = 0.143004$.

Imaginem que en lloc del model anterior hem optat per especificar-ne un altre sense terme independent però incorporant una altra nova variable fictícia que reculli si un estudiant opta per seguir l'itinerari recomanat. Si definim D_4 com a:

$$D_{4i} = \begin{cases} 1 & \text{si} \quad X_{3ri} = 1 \\ 0 & \text{si} \quad X_{3ri} \neq 1 \end{cases}$$

El model que especificaríem seria:

$$Y_i = \delta_1 X_{1i} + \delta_2 D_{2i} + \delta_3 D_{3i} + \delta_4 D_{4i} + u_i \qquad i = 1, \ldots, N.$$

i l'estimació resultaria:

```
> EcoModel.5 <- lm(Y~0+X1+D2+D3+D4, data=Econometria)
> summary(EcoModel.5)
Call:
lm(formula = Y ~ 0 + X1 + D2 + D3 + D4, data = Econometria)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6183 -0.2379 -0.0107  0.1790  3.3724

Coefficients:
    Estimate Std. Error t value Pr(>t)
X1  0.175576   0.003115  56.358  < 2e-16 ***
D2 -0.360537   0.128545  -2.805 0.006225 **
D3 -0.503540   0.137827  -3.653 0.000444 ***
D4 -0.342012   0.131838  -2.594 0.011144 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4542 on 86 degrees of freedom
Multiple R-squared:  0.9946,  Adjusted R-squared:  0.9944
F-statistic:  3996 on 4 and 86 DF,  p-value: < 2.2e-16
```

Ara, la interpretació dels CSI dels paràmetres seria:

- El paràmetre δ_4 recull la qualificació esperada dels estudiants que han seguit l'itinerari quan no s'ha estudiat cap hora (X_1). Veiem que aquest paràmetre és de -0.34; així, si un estudiant que segueix l'itinerari recomanat no estudia cap hora, la seva qualificació es redueix en uns 0.34 punts.

- El paràmetre δ_3 indica la qualificació esperada d'un estudiant que segueix un itinerari lliure en cas que no hagi estudiat gens. En aquest exemple, veiem que el paràmetre és de -0.5.

- El paràmetre δ_2 mostra la qualificació esperada d'un estudiant que segueix un itinerari intensiu quan aquest no ha estudiat cap hora, és a dir, -0.36 punts.

Vegem l'equivalència que hi ha entre les dues solucions proposades dels dos models:

Taula 2. Equivalència dels models EcoModel.4 i EcoModel.5

Valor esperat de la qualificació quan $X_{1i} = 0$		
Modelo	EcoModel.4	EcoModel.5
Recomanat	$\beta_0 = -0.342012$	$\delta_4 = -0.342012$
Intensiu	$\beta_0 + \beta_2 = -0.360536$	$\delta_2 = -0.360537$
Lliure	$\beta_0 + \beta_3 = -0.50354$	$\delta_3 = -0.503540$

Continuant amb l'exemple dels estudiants d'econometria suposem que compliquem una mica el model: esperem que l'impacte esperat que tenen les hores d'estudi sobre la qualificació no sigui el mateix per als homes que per a les dones.

Hem de decidir si seguim suposant que, quan un estudiant no ha estudiat gens, la qualificació esperada és la mateixa per als homes que per a les dones (introducció d'una *dummy* en el model multiplicativament), o és diferent (introducció d'una *dummy* en el model additivament i multiplicativament). En els dos casos hem d'incorporar una variable dicotòmica de manera multiplicativa.

4.2.2. Introducció de *dummies* en un model de manera multiplicativa

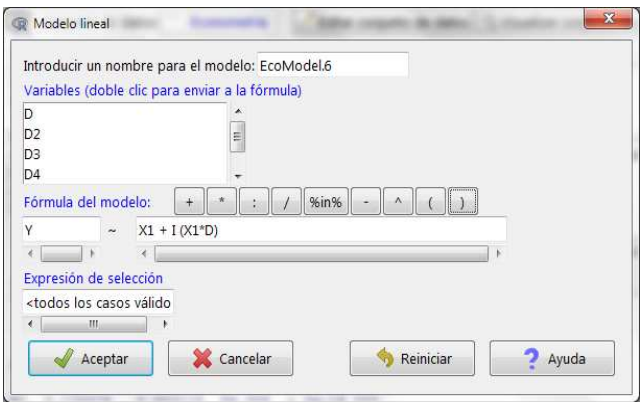
El model que especificaríem seria:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i} D_i + u_i \qquad i = 1, \ldots, N.$$

i l'estimació la faríem seguint la ruta:

Estadístics / Ajust de models / Model lineal

omplint el quadre de diàleg tal com es mostra a continuació:



La funció `I()` ens permet fer operacions de diverses variables element a element, que és el que pretenem en introduir la variable fictícia de manera multiplicativa.

i obtenim el resultat següent:

```
> EcoModel.6 <- lm(Y ~ X1 + I (X1*D), data=Econometria)
> summary(EcoModel.6)
Call:
lm(formula = Y ~ X1 + I(X1 * D), data = Econometria)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5976 -0.2459  0.0042  0.1882  3.3937

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -0.407508   0.110749  -3.680 0.000404 ***
X1           0.176094   0.003171  55.530  < 2e-16 ***
I(X1 * D)    0.000783   0.002665   0.294 0.769604
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4554 on 87 degrees of freedom
Multiple R-squared:  0.975, Adjusted R-squared:  0.9744
F-statistic:  1694 on 2 and 87 DF,  p-value: < 2.2e-16
```

En el model multiplicatiu EcoModel.6, el gènere de l'estudiant afecta l'impacte de les hores d'estudi sobre la qualificació esperada; és a dir, per als homes aquest impacte serà igual a $\beta_1 + \beta_2 = 0.176877$, mentre que per a les dones serà igual a $\beta_1 = 0.176094$. Cal destacar que, en aquest exemple, el paràmetre β_2 no resulta estadísticament significatiu i, per tant, el supòsit realitzat inicialment no seria cert.

4.2.3. Introducció de *dummies* en un model de manera mixta (additiva i multiplicativa)

En aquest cas el model que s'ha d'especificar és:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i} D_i + +\beta_3 D_i + u_i \qquad i = 1, \ldots, N.$$

farem l'estimació seguint la mateixa ruta que abans, i ompliríem el quadre de diàleg tal com hem fet. El resultat seria:

```
> EcoModel.7 <- lm(Y ~ X1 + I (X1*D) + D, data=Econometria)
> summary(EcoModel.7)
Call:
lm(formula = Y ~ X1 + I(X1 * D) + D, data = Econometria)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5936 -0.2415  0.0029  0.1894  3.3927

Coefficients:
              Estimate Std. Error t value Pr(>t)
```

```
(Intercept) -0.4141740  0.1479553  -2.799  0.00632 **
X1           0.1762521  0.0039347  44.794  < 2e-16 ***
I(X1 * D)    0.0003979  0.0062312   0.064  0.94923
D            0.0153872  0.2247908   0.068  0.94559
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.458 on 86 degrees of freedom
Multiple R-squared:  0.975, Adjusted R-squared:  0.9741
F-statistic:  1116 on 3 and 86 DF,  p-value: < 2.2e-16
```

En el model mixt (additiu i multiplicatiu) EcoModel.7, el gènere de l'estudiant afecta l'impacte de les hores d'estudi sobre la qualificació esperada; és a dir, per als homes aquest impacte serà igual a $\beta_1 + \beta_2 = 0.17665$, mentre que per a les dones serà igual a $\beta_1 = 0.1762521$. A més a més, la qualificació esperada d'un home que no hagi estudiat gens s'explicarà per $\beta_0 + \beta_3 = -0.3987868$ i la d'una dona que tampoc hagi estudiat serà de $\beta_0 = -0.4141740$. De la mateixa manera que abans, els paràmetres β_2 i β_3 no són significatius. Així, aquests impactes serien iguals independentment del gènere de l'estudiant.

Una altra observació és que el model EcoModel.7 és el model més general, i si hi fem diferents contrastos, arribem a models més senzills. Per exemple, si fem el contrast de $\beta_3 = 0$ estem contrastant el model EcoModel.6 en la hipòtesi nul·la. Això ens portaria a pensar erròniament que és millor especificar el model més general. No obstant això, no hem d'oblidar que si el nombre de categories és elevat i a més pretenem considerar diversos efectes multiplicatius, el nostre model tindrà molts paràmetres i podrem tenir problemes d'estimació.

4.2.4. Interpretació de les interaccions

En el supòsit que vulguem introduir diferents variables explicatives qualitatives en un model de regressió, el més normal és que s'hagin suposat diversos efectes multiplicatius entre elles. En aquest cas, aquests efectes multiplicatius es denominen **efectes encreuats o interaccions**.

Com ens passava quan introduíem una variable qualitativa, se'ns presenten tres possibilitats: incorporació additiva, multiplicativa o mixta.

Per a il·lustrar això utilitzarem la base de dades d'Econometria, que recollia informació sobre aquesta assignatura en una classe de 90 estudiants. Les variables de què disposem són:

- Y_i: qualificació de l'estudiant.
- X_{1i}: hores d'estudi.

- D_i: variable fictícia indicadora del gènere de l'estudiant.

- D_{2i}: variable fictícia que identifica aquells estudiants que han seguit un itinerari intensiu.

- D_{3i}: variable fictícia que recull els estudiants que segueixen un itinerari lliure.

Un dels models que podríem definir seria:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i \qquad i = 1, \ldots, N.$$

Aquest model solament presenta efectes additius, per tant:

- β_0 és la qualificació esperada d'una dona que segueix l'itinerari recomanat, és a dir, el valor esperat de la variable endògena per a la categoria base o de referència.

- β_1 indica la diferència entre la qualificació esperada dels homes i les dones que segueixen el mateix itinerari d'estudis.

- β_2 representa la diferència entre la qualificació esperada d'un estudiant que segueix un itinerari intensiu i un que segueix l'itinerari recomanat.

- β_3 mostra la diferència que hi ha entre la qualificació esperada d'un estudiant que segueix un itinerari lliure i un altre que opta per l'itinerari recomanat.

Tot plegat ho podem resumir en la taula següent:

Taula 3. Diverses variables qualitatives amb *dummies* de manera additiva

Model sense interaccions		
	$E(Y_i)$	
Itinerari	Homes	Dones
Recomanat	$\beta_0 + \beta_1$	β_0
Intensiu	$\beta_0 + \beta_1 + \beta_2$	$\beta_0 + \beta_2$
Libre	$\beta_0 + \beta_1 + \beta_3$	$\beta_0 + \beta_3$

Un altre model podria ser:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_i D_{2i} + \beta_5 D_i D_{3i} + u_i \qquad i = 1, \ldots, N.$$

les interpretacions del qual es resumeixen en la taula que es mostra a continuació:

Recordeu que per estimar un model amb interaccions hem de seguir la ruta:

Estadístics / Ajust de models / Model lineal

Taula 4. Diverses variables qualitatives amb *dummies* de manera mixta

Model amb interaccions		
	$E(Y_i)$	
Itinerari	Homes	Dones
Recomanat	$\beta_0 + \beta_1$	β_0
Intensiu	$\beta_0 + \beta_1 + \beta_2 + \beta_4$	$\beta_0 + \beta_2$
Lliure	$\beta_0 + \beta_1 + \beta_3 + \beta_5$	$\beta_0 + \beta_3$

i introduir les interaccions mitjançant la funció I() dins del quadre de diàleg.

4.3. Altres usos de les variables fictícies

Com ja hem estudiat anteriorment, els usos més comuns de les variables fictícies són la introducció en un model de regressió de variables qualitatives no ordinals que ens permetin segmentar la nostra mostra en diferents grups (sexe, estat civil, tipus d'habitatge, etc.), o distingir períodes temporals (nascuts abans o després d'un determinat any, etc.) o localitzacions (comunitat de residència, país d'origen, etc.). No obstant això, aquestes no són les úniques aplicacions d'aquestes variables en un model de regressió, ja que també les podem utilitzar per al següent:

- Dades atípiques
- Canvi estructural
- Estacionalitat
- Model d'efectes fixos

4.3.1. Dades atípiques

En el mòdul següent s'analitzarà més detalladament el que són les dades atípiques, aquí únicament ens centrarem en l'ús de les variables fictícies per, una vegada detectades, mantenir-les dins de la mostra sense que aquests afectin l'estimació del model.

Què són les dades atípiques?

Una observació o dada es considera atípica quan hem comès un error de mesura, quan es tracta d'un individu fora de sèrie o quan es mesura en un moment temporal lligat a una situació extraordinària.

S'han que tenir en compte dos aspectes rellevants en aquest context:

- Introducció d'una variable fictícia per a cada observació atípica. En aquest cas, les estimacions dels paràmetres del model seran les mateixes que si eliminéssim les observacions atípiques però amb l'agreujant que estem incloent més regressors en el model. L'únic avantatge és que si valorem per MCO el residu de l'estimació per a cada observació atípica serà zero i el coeficient que acompanyi la variable fictícia recollirà l'efecte de cada observació atípica sobre l'endògena. En crear una variable fictícia que és igual a 1 per a l'observació atípica i zero per a la resta de les observacions, l'estimació MCO del paràmetre que l'acompanya no és més que l'error de predicció si utilitzéssim el model per a predir la variable endògena amb aquesta observació. Per tant, també podríem calcular els errors de predicció i els intervals de confiança.

Mesures de detecció de dades atípiques

Els instruments més usuals per a detectar si en una mostra hi ha dades atípiques són el *leverage*, el residu, el residu studentitzat, el residu studentitzat amb omissió i la distància de Cook.

- Introducció d'una variable fictícia que reculli totes les observacions atípiques. Les condicions ara són diferents perquè tenim una variable fictícia que val 1 quan es tracta d'una observació atípica i zero per a la resta de les observacions. En aquest cas, els residus obtinguts de l'estimació per MCO no haurien de ser necessàriament zero. Si això fos així, la informació que ens aportaria seria molt més limitada que en el cas anterior. A més a més, no tindríem proves sobre l'impacte de cada una de les observacions atípiques, sinó un efecte global de totes.

Disposem d'una base de dades amb N observacions de la qual sabem que l'observació $i_0 - sim$ és una dada atípica. Davant aquesta circumstància definim la variable fictícia D_i, que valdrà 1 per a l'observació $i_0 - sima$ i 0 en la resta de les observacions:

$$D_i = \begin{cases} 1 \quad \text{si} & i = i_0 \\ 0 \quad \text{si} & i \neq i_0 \end{cases}$$

4.3.2. Canvi estructural

Encara que el canvi o la permanència estructural es tracti amb més detall en el mòdul següent, la seva aplicació amb variables fictícies és similar al fet amb observacions atípiques però canviant radicalment la definició de la variable fictícia que s'ha d'introduir en el model. Ara ens interessa crear una variable fictícia que sigui zero per a tots els períodes anteriors al canvi i un per a tots aquells en què el canvi ja s'hagi produït. El coeficient estimat que acompanyi aquesta variable recollirà l'efecte del canvi estructural en el valor esperat de la variable endògena. El contrast de significació individual d'aquest coeficient corroborarà l'existència d'aquest canvi. Si resulta que el paràmetre és estadísticament significatiu hi haurà canvi estructural i no n'hi haurà en cas contrari.

A què ens referim quan parlem de canvi estructural?

Quan treballem amb sèries temporals un canvi d'estructura dins del període mostral indica que s'ha produït un o més canvis que han provocat alguna variació en el procés generador de dades. No obstant això, quan treballem amb dades de tall transversals un canvi estructural indica l'existència de dos o més grups que es comporten de diferent manera entre ells.

Tot canvi estructural pot afectar la constant (si introduïm la variable fictícia de manera additiva), el pendent (si introduïm la variable fictícia de forma multiplicativa) o tots dos (quan la variable fictícia s'introdueix additivament i multiplicativament). Donat el model següent:

El test de Chow

El contrast que fem per veure si en una mostra hi ha canvi o permanència estructural es denomina test de Chow.

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad t = 1, \ldots, T.$$

Es vol contrastar si els paràmetres de la població es mantenen iguals al llarg dels dos subperíodes: el primer des d'1 fins a T_1 i el segon des de $T_1 + 1$ fins a T. Per a això definirem una variable dicotòmica D_t, que serà igual a 0 per al primer subperíode i igual a 1 per al segon:

$$D_t = \begin{cases} 1 \quad \text{si} & t = 1, \ldots, T_1 \\ 0 \quad \text{si} & t = T_1 + 1, \ldots, T \end{cases}$$

Una vegada creada la variable fictícia la introduiríem de forma multiplicativa en l'especificació per a recollir el suposat canvi de pendent que es produeix en la segona submostra.

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_t D_t + u_t \qquad t = 1, \ldots, T.$$

4.3.3. Estacionalitat

Quan treballem amb sèries temporals en què les dades són estacionals (mesos, trimestres, quadrimestres o semestres) podem estar interessats a aïllar els efectes d'una determinada estació per a esclarir millor quines variables afecten més la variable endògena. Així doncs, podem estar interessats a crear una variable fictícia (*dummy*) i introduir-la en el model per a avaluar l'impacte produït d'estació concreta en el model especificat. Si estem interessats a veure les conseqüències de l'estació s_i podem definir una *dummy* de la manera següent:

El valor de s

s recull l'estacionalitat d'una sèrie temporal i indica el nombre d'estacions dins d'un any natural. Així, si parlem de mesos $s = 12$ i per tant si cada estació es denomina s_i, i prendrà els valors des d'1 fins a 12.

$$D_s = \begin{cases} 1 \quad \text{si} & s = s_i \\ 0 \quad \text{si} & s \neq s_i \end{cases}$$

4.3.4. Model d'efectes fixos

Quan treballem amb un panel de dades podem estar interessats a especificar un model de regressió utilitzant tota la informació disponible ($N \bullet T$ observacions) eliminant els efectes temporals i de discrepància entre les unitats analitzades. Per tant, hem de definir $N - 1$ variables fictícies que recullin les unitats objecte d'estudi i altres $T - 1$ variables fictícies que recullin els períodes dels quals disposem d'informació i introduir-les de manera additiva en el nostre model de regressió. Aquest nou model se sol denominar *model d'efectes fixos*.

Què és un panel de dades?

Un panel de dades no és més que una base de dades on es recull informació sobre les característiques de N unitats objecte d'estudi (individus, llars, empreses, etc.) durant T períodes diferents.

Així, en un model d'efectes fixos, els paràmetres estimats de les de *dummies* que recullen unitats i períodes s'entenen com els desplaçaments patits en la recta de regressió com a conseqüència dels efectes fixos de les variables no observables. L'inconvenient més important de treballar amb un panel de dades i realitzar un model d'efectes fixos és que incrementem moltíssim el nombre de coeficients i podem córrer el risc de quedar-nos sense graus de llibertat per poder-los estimar.

Observeu que incloem $N - 1 \bullet T - 1$ *dummies*, ja que en els dos casos conservem una categoria de referència, una unitat i un període.

5. Restriccions lineals en el model de regressió

Una vegada tinguem un model de regressió lineal, és molt habitual que vulguem contrastar determinats supòsits alternatius postulats per alguna teoria. Aquests supòsits solen ser molt més complexos que la significació estadística de les variables, i per això hem de recórrer a la formulació general de les restriccions lineals. Un exemple seria contrastar si el valor d'un paràmetre determinat es correspon amb un valor específic, o també contrastar si una determinada relació entre paràmetres es pot corroborar estadísticament.

Formalment, considerem un conjunt de q restriccions lineals:

$$R_{11}\beta_1 + \ldots + R_{1k}\beta_k = r_1$$

$$R_{21}\beta_1 + \ldots + R_{2k}\beta_k = r_2$$

$$\ldots$$

$$R_{q1}\beta_1 + \ldots + R_{qk}\beta_k = r_q$$

De manera compacta, aquestes restriccions lineals es poden expressar en una única equació:

$$R\beta = r$$

La matriu R té una dimensió $q \times k$, on q és el nombre de restriccions lineals per contrastar i k el nombre de paràmetres del model. Per tant, la matriu R té tantes columnes com paràmetres del model i tantes files com restriccions per contrastar. El vector r té una dimensió $q \times 1$, i està format pels termes independents de les restriccions lineals. Així doncs, igual que la matriu R, el vector r tindrà tantes files com restriccions.

El vector β...

... recordem que és una columna de dimensió $k \times 1$, és a dir, té una sola columna i tantes files com paràmetres.

Una vegada especificades les restriccions, procedim al contrast d'hipòtesi sobre els paràmetres del model. El contrast que hem de plantejar pren la forma següent:

$$H_0 : R\beta = r$$

$$H_1 : R\beta \neq r$$

Això és, des de la hipòtesi nul·la no es poden rebutjar les restriccions imposades. L'estadístic del contrast F_0 es distribueix segons una distribució F de Fischer, on el valor crític serà definit per $F_{q,N-k;\,\alpha}$, en què α és el nivell de significació del contrast. La decisió del contrast seguirà l'esquema següent:

$F_0 \geq F_{q,N-k;\,\alpha}$	Rebuig de H_0
$F_0 < F_{q,N-k;\,\alpha}$	No-rebuig de H_0

Vegem un exemple derivat del cas plantejat en el capítol 3. Suposem que una teoria estableix que el paràmetre β_5, associat a la variable *UNIV* (percentatge d'estudiants universitaris), té un valor de $-0,5$, i volem contrastar aquesta teoria en la nostra estimació. Formalment, la restricció que hem de contrastar pren la forma següent:

$$R_{15} \cdot \beta_5 = r_1$$

$$1 \cdot \beta_5 = -0,5.$$

Per tant, el contrast que hem de fer és:

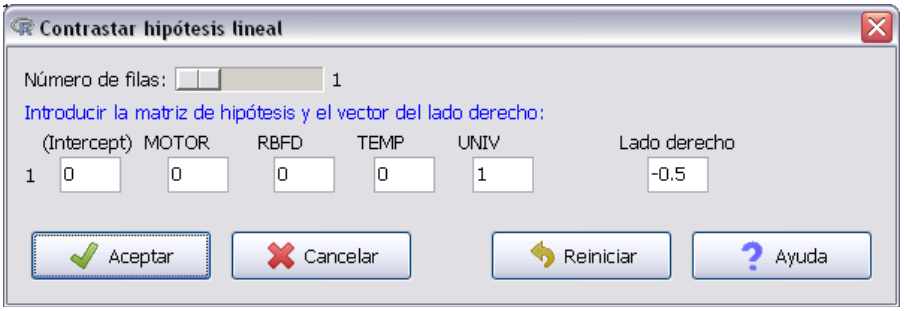
$$H_0 : \beta_5 = -0,5$$

$$H_1 : \beta_5 \neq -0,5.$$

Una vegada hem seleccionat com a model actiu *RegModel.2*, que va ser el primer model que vam estimar en el capítol 3, per fer el contrast amb R-Commander anem a la ruta següent:

Models / Test d'hipòtesi / Hipòtesi lineal

Apareixerà un quadre de diàleg, on introduirem les restriccions R i r com es mostra a continuació:



El resultat és el següent:

```
> .Hypothesis <- matrix(c(0,0,0,0,1), 1, 5, byrow=TRUE)

> .RHS <- c(-0.5)

> linearHypothesis(RegModel.2, .Hypothesis, rhs=.RHS)
Linear hypothesis test

Hypothesis:
UNIV = - 0.5

Model 1: restricted model
Model 2: PARO ~ MOTOR + RBFD + TEMP + UNIV

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    291 3940.9
2    290 3718.3  1    222.57 17.358 4.088e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com veiem, el *p-valor* és pràcticament zero. D'aquesta manera, amb qualsevol nivell de confiança rebutgem la hipòtesi nul·la (H_0), és a dir, podem afirmar que la restricció $\beta_5 = -0,5$ no és certa, almenys amb aquestes dades i aquest model.

Introduïm ara una restricció una mica més complexa. Suposem que volem introduir dues restriccions: $\beta_5 = -0,3$ i $\beta_3 = -\beta_4$ (o el que és el mateix, $\beta_3 + \beta_4 = 0$). Formalment tenim les restriccions següents:

$$R_{25} \cdot \beta_5 = r_1$$

$$R_{13} \cdot \beta_3 + R_{14} \cdot \beta_4 = r_2$$

Específicament:

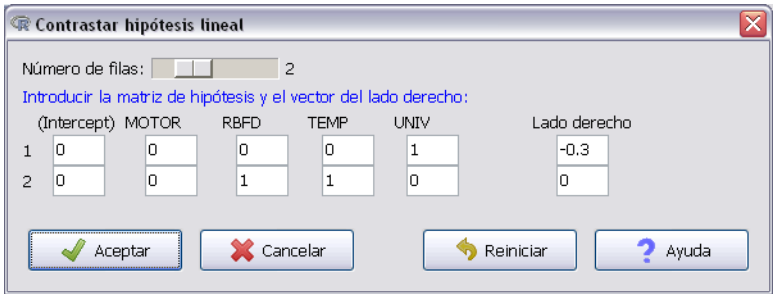
$$1 \cdot \beta_5 = -0,3$$

$$1 \cdot \beta_3 + 1 \cdot \beta_4 = +0,0$$

Igual que en el cas anterior, accedim a:

Models / Test d'hipòtesi / Hipòtesi lineal

però en aquest cas seleccionem $q = 2$ files, que equival al nombre de restriccions. A més a més, introduïm les restriccions igual que en el cas anterior.



En aquest cas, observem que l'estadístic de prova cau en la regió de no-rebuig de H_0, és a dir, no podem rebutjar les dues restriccions introduïdes. Com a conseqüència, amb aquestes dades i aquest model un increment d'un punt percentual d'universitaris (UNIV) fa reduir el valor esperat de la taxa d'atur (PARO) en 0.3 punts. D'altra banda, la renda familiar bruta disponible (RFBD) i la taxa de temporalitat (TEMP) tenen el mateix efecte en el valor esperat de la taxa d'atur (PARO).

```
> .Hypothesis <- matrix(c(0,0,0,0,1,0,0,1,1,0), 2, 5, byrow=
    TRUE)

> .RHS <- c(-0.3,0)

> linearHypothesis(RegModel.2, .Hypothesis, rhs=.RHS)
Linear hypothesis test

Hypothesis:
UNIV = - 0.3
RBFD  + TEMP = 0

Model 1: restricted model
Model 2: PARO ~ MOTOR + RBFD + TEMP + UNIV

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    292 3720.6
2    290 3718.3  2    2.2916 0.0894 0.9145
```

Suposem ara que volem estimar un model que inclogui la restricció $\beta_3 = -\beta_4$. Fent una senzilla transformació algebraica, el model queda de la manera següent:

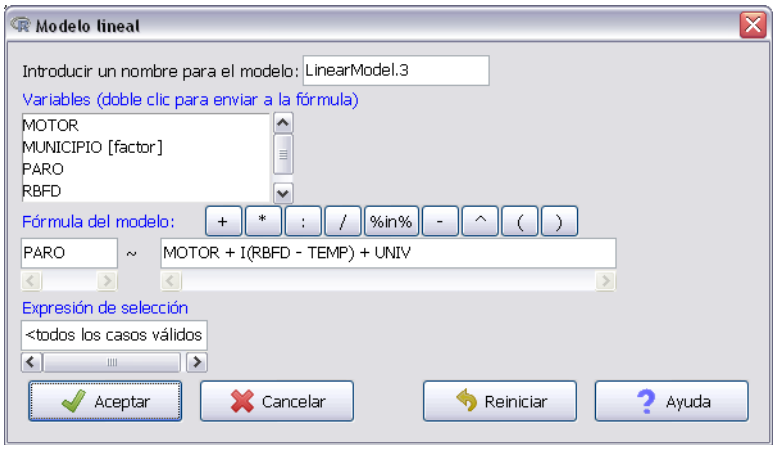
$$PARO_i = \beta_1 + \beta_2 MOTOR_i + \beta_3 RBFD_i + \beta_4 TEMP_i + \beta_5 UNIV_i + u_i =$$

$$= \beta_1 + \beta_2 MOTOR_i + \beta_3 RBFD_i + -\beta_3 TEMP_i + \beta_5 UNIV_i + u_i =$$

$$= \beta_1 + \beta_2 MOTOR_i + \beta_3 (RBFD_i - TEMP_i) + \beta_5 UNIV_i + u_i$$

Per fer aquesta estimació tenim dues opcions: la més laboriosa és crear una nova variable $RBFD_i - TEMP_i$, introduir-la en el conjunt de dades i fer-ne l'estimació. Tanmateix, si solament estem interessats en l'estimació i no en la variable en si mateixa, la podem introduir en l'especificació mitjançant l'operador `I()`, com es mostra a continuació:

Estadístics / Ajust de models / Model lineal



Recordem que ja havíem utilitzat l'operador I() en el capítol anterior quan incorporràvem variables fictícies de forma multiplicativa en un model de regressió.

El resultat de l'estimació és el següent:

```
> LinearModel.3 <- lm(PARO ~ MOTOR + I(RBFD - TEMP) + UNIV, data=
    Datos)

> summary(LinearModel.3)

Call:
lm(formula = PARO ~ MOTOR + I(RBFD - TEMP) + UNIV, data = Datos)

Residuals:
     Min       1Q   Median       3Q      Max
-12.4452  -1.6636   0.4475   2.2034   8.4552

Coefficients:
                Estimate Std. Error t value Pr(>t)
(Intercept)    18.664297   0.545701  34.202  < 2e-16 ***
MOTOR          -0.007670   0.003282  -2.337   0.0201 *
I(RBFD - TEMP) -0.034523   0.007742  -4.459 1.17e-05 ***
UNIV           -0.288652   0.048361  -5.969 6.94e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.575 on 291 degrees of freedom
Multiple R-squared:  0.2779,  Adjusted R-squared:  0.2705
F-statistic: 37.33 on 3 and 291 DF,  p-value: < 2.2e-16
```

Fixeu-vos que hem anomenat de la mateixa manera l'últim model que havíem estimat en el capítol 3 d'aquest mòdul. Així doncs, com que R és un programa orientat a l'assignació d'objectes, quan donem el mateix nom a un objecte existent, l'estem sobreescrivint. És a dir, la pròxima vegada que vulguem utilitzar l'objecte LinearModel.3 recuperarem aquest últim model i haurem perdut el creat anteriorment.

Bibliografia

Artís Ortuño, M.; del Barrio Castro, T.; Clar López, M.; Guillén Estany, M.; Suriñach Caralt, J. (2011). *Econometría*. Barcelona. Material didàctic UOC.

Liviano Solís, D.; Pujol Jover, M. (2013). *Matemáticas y Estadística con R*. Barcelona. Material didàctic UOC.