Modelos de Regresión Lineal Simple y Múltiple con R

Daniel Liviano Solís
Maria Pujol Jover

PID_00211046

UOC
Universitat Oberta de Catalunya

Índice

Introducción

Este módulo tiene como principal objetivo introducir al estudiante en la econometría utilizando el entorno estadístico R y su interfaz R-Commander. El uso de la econometría resulta imprescindible en ámbitos como la economía, la empresa y el marketing. De hecho, la econometría podría definirse como la estadística aplicada a la economía, ya que se basa en el uso de la modelización. Las aplicaciones más comunes son las que se muestran a continuación:

1) El análisis estructural.

2) La predicción.

3) La evaluación de políticas.

En este primer módulo se tratan los aspectos más básicos de esta disciplina. El primer capítulo es de carácter teórico, y tiene como objetivo introducir formalmente el modelo de regresión y su estimación por Mínimos Cuadrados Ordinarios (MCO). Seguidamente, los dos siguientes capítulos ilustran, mediante ejemplos, el Modelo de Regresión Lineal Simple (MRLS) y el Modelo de Regresión Lineal Múltiple (MRLM) mediante ejemplos con R y R-Commander. El cuarto capítulo está dedicado a la inserción de variables cualitativas politómicas y dicotómicas, tanto en un MRLS como en un MRLM, así como a la creación de variables ficticias según los distintos usos que les podemos dar. Finalmente, el quinto y último capítulo abordan todo el tema de la introducción de restricciones a un modelo de regresión y su posterior estimación por Mínimos Cuadrados Restringidos (MCR).

Objetivos

1. Saber especificar correctamente modelos de regresión lineales simples y múltiples (MRLS y MRLM).

2. Estimar MRLS y MRLM por Mínimos Cuadrados Ordinarios (MCO) con R y R-Commander

3. Estimar MRLM por Máxima Verosimilitud (MV) utilizando R y R-Commander.

4. Verificar mediante los resultados de la estimación de un modelo que se cumplen todas las hipótesis básicas de todo MRLM.

5. Validar un modelo y cuantificar la bondad de su ajuste.

6. Realizar predicciones puntuales y por intervalo con un modelo de regresión con la ayuda de R y R-Commander.

7. Incorporar variables dicotómicas y politómicas en un MRLM con R y R-Commander.

8. Crear variables ficticias con objeto de satisfacer distintos objetivos de análisis.

9. Introducir restricciones lineales en la estimación de MRLM con R y R-Commander.

10. Estimar por Mínimos Cuadrados Restringidos (MCR) un MRLM con R y R-Commander.

11. Distinguir entre las propiedades de los estimadores por MCO y MCR.

1. Introducción a los modelos de regresión

1.1. Marco general

Un investigador, a la hora de realizar un análisis estadístico de una serie de variables, ha de tener en cuenta una diferencia fundamental entre dos conceptos relacionados pero diferentes entre sí:

- La **correlación** hace referencia al grado de relación que existe entre dos variables, pero no establece ningún tipo de relación de causa ni efecto de una sobre la otra. El indicador de correlación más simple es el coeficiente de correlación lineal de Pearson, que indica en qué medida la relación lineal entre dos variables es directa, inversa o nula.

- El **modelo de regresión** supone que no solo existe correlación entre las variables, sino que además existe una relación de *causalidad*, es decir, una o más variables influyen en otra.

Específicamente, definimos un análisis econométrico como un estudio de las relaciones que se establecen entre un conjunto de variables. El análisis parte de una serie de observaciones empíricas de las variables que se desea 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 par $\{y_i, x_i\} \in R \times R^k$ corresponde a una **observación** de una unidad (individuo, empresa, familia, etc.). El conjunto de observaciones consideradas configuran la **muestra**, que está compuesta de n elementos. Cada par (y_i, x_i) está compuesto por dos elementos: la **variable dependiente** y_i y el **vector de regresores** o **variables independientes** x_i. Mientras que para cada unidad muestral la variable dependiente es un escalar (número), los regresores forman un vector. Dicho vector tiene como primer elemento una constante (en concreto, el número 1), ya que hace referencia al término independiente del modelo. La notación del vector es:

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

La variable explicada

Esta variable también se puede denominar endógena, dependiente o variable por explicar.

Las variables explicativas

Estas variables se pueden denominar de diferentes maneras alternativas: exógenas, independientes o regresoras.

Dicho vector corresponde al i-ésimo elemento de la muestra, e incluye las observaciones de sus $k-1$ regresores más la constante.

El objetivo de una regresión es el de encontrar la tendencia central de la distribución condicional de y_i, dado x_i. Una medida de tendencia central por excelencia de una variable es su media. Para el caso condicional, la medida análoga es la **media condicional o esperanza condicionada** $m(x) = E(y_i|x_i = x)$ que puede tomar cualquier forma: lineal, cuadrática, logarítmica, etc. Sin embargo, como veremos a continuación, la forma más usual es la lineal.

Un elemento importante de la regresión es el **error** o **término de perturbación** u_i, y viene definido como la diferencia entre y_i y su media condicional, esto es:

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

Matemáticamente, podemos reorganizar la ecuación anterior, obteniendo la fórmula:

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

Los únicos supuestos en los que nos basamos a la hora de construir esta nueva ecuación son que las variables (y_i, x_i) tienen una distribución conjunta y que $E|y_i| < \infty$.

Es fundamental establecer las propiedades de u_i, es decir, del término de error o perturbación de la regresión:

1. $E(u_i|x_i) = 0$
2. $E(u_i) = 0$
3. $E(h(x_i)u_i) = 0$ para cualquier función $h(\cdot)$
4. $E(x_i u_i) = 0$

La ecuación $y_i = m(x)+u_i$ y la propiedad $E(u_i|x_i) = 0$ son consideradas con frecuencia el marco de una regresión pero no un modelo. En este marco no se está suponiendo ninguna restricción (es decir, no se supone ninguna característica de la distribución conjunta de las variables, como podría ser la linealidad, por ejemplo). Así, ambas ecuaciones son ciertas por definición.

1.2. Modelo de regresión lineal

En el momento en el que asumimos ciertas restricciones en la distribución conjunta de las variables, obtenemos un modelo. La restricción más común en econometría, por simplicidad y por facilidad de estimación e inferencia, es la de **linealidad**, es decir,

se asume a priori que $m(x)$ es una función lineal de (x). Aplicando esta restricción al marco definido anteriormente se obtiene el **modelo de regresión lineal**:

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

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

La restricción de linealidad no tiene por qué cumplirse en cualquier aplicación específica. La validez de esta restricción (y de cualquier otra) dependerá, en todo caso, de las características de la distribución conjunta de las variables consideradas.

En la ecuación $y_i = x_i'\beta + u_i$ aparece, además de la variable dependiente y_i, el vector de regresores traspuesto x_i' y el término de perturbación de la regresión u_i; un nuevo vector: el **vector de parámetros** β:

La propiedad de trasposición...

... intercambia filas por columnas de una matriz o vector con el fin de poder realizar operaciones matemáticas, por ejemplo, la multiplicación entre objetos.

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

La ecuación $y_i = x_i'\beta + u_i$ muestra la notación compacta del modelo de regresión lineal. Una forma más detallada sería:

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

Este modelo de regresión lineal es incompleto sin una descripción del término de error u_i. Supondremos, a priori, que tiene una esperanza nula $E(u_i) = 0$, que tiene varianza finita $E(u_i^2) < \infty$, y que no está correlacionado con los regresores $E(x_i u_i) = 0$. Esta última condición se denomina **condición de ortogonalidad**, ya que supone que el vector de regresores y el de errores son ortogonales. Como veremos más adelante, esta condición es débil y en muchos casos no tiene por qué cumplirse.

Ortogonalidad

Algebraicamente, decimos que los vectores a y b son ortogonales si se cumple que $a'b = 0$.

1.3. Notación matricial

Antes de tratar la estimación del modelo, es importante conocer la notación matricial del modelo de regresión, ya que a partir de ahora se utilizará, según conveniencia, dicha notación. La matriz de variables dependientes adquiere la forma:

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

como ya hemos visto, el vector de parámetros se define como:

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

considerando que el vector traspuesto de regresores, para cada elemento muestral, es:

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

definimos la matriz de regresores o variables independientes como:

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

y, finalmente, la matriz de errores adquiere la forma:

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

Como se puede observar, Y y β, al igual que u, son vectores, mientras que X es una matriz.

El modelo de regresión lineal $y_i = x_i'\beta + u_i$ muestra, por simplicidad, una sola ecuación para el individuo i. Alternativamente, podemos expresar dicho modelo como un sistema de n ecuaciones, una para cada observación:

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

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

$$\dots$$

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

o, equivalentemente, mediante notación matricial:

$$Y = X\beta + u$$

Los siguientes productos también se expresan 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. Especificación y estimación

A la hora de analizar mediante la econometría las relaciones que se establecen entre distintas variables, hay que tener en cuenta que el estudio de un modelo de regresión se compone de dos etapas fundamentales:

1) Especificación: Esta fase hace referencia a la construcción del modelo, es decir, qué variables se incluyen y en qué forma funcional (lineal, aditiva, multiplicativa, etc.).

2) Estimación: Una vez construido el modelo, la estimación hace referencia a la técnica utilizada para obtener estimaciones de los parámetros del modelo. El conocimiento previo del fenómeno analizado y la disponibilidad de contrastes ayudarán a realizar una elección correcta del estimador.

En lo que se refiere al proceso de estimación, definiremos el vector de coeficientes estimados como $\hat{\beta}$, diferente al vector de parámetros β, que es desconocido.

Existen diversos procedimientos para obtener una estimación del modelo de regresión lineal. El más sencillo e inmediato es el método de los **Mínimos Cuadrados Ordinarios (MCO)**. Este estimador consiste en minimizar la suma al cuadrado de los errores (SCE). Si definimos la función SCE como:

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

podemos definir el estimador MCO como aquel que minimiza la función SCE:

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

así pues, este estimador tiene la siguiente expresión:

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

Hay varias maneras de llegar a esta expresión. Una de ellas es considerar el modelo de regresión lineal y premultiplicarlo por x_i, para después tomar esperanzas:

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

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

De este modo obtenemos que el parámetro β es una función de los segundos momentos poblacionales de (y_i, x_i):

Los momentos nos proporcionan información sobre la distribución de la variable considerada.

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

Nótese que esta expresión de los parámetros se basa en la hipótesis $E(x_i u_i) = 0$, que no tiene por qué cumplirse en la realidad. Para obtener estimaciones de los parámetros a partir de esta expresión, basta sustituir los momentos poblacionales por momentos muestrales, que son:

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

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

El estimador adquiere entonces la siguiente forma:

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

Otro modo alternativo de obtener el estimador es partiendo de la condición de ortogonalidad. Considerando la notación matricial, y teniendo en cuenta que dicha notación es $E(X'e) = 0$, obtenemos:

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

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

Como siempre, dado que desconocemos los verdaderos valores poblacionales, la estimación de los parámetros se realiza sustituyendo los momentos poblacionales por los muestrales.

1.5. Interpretación del modelo

A la hora de interpretar un modelo de regresión debemos considerar distintos aspectos:

- Significación de cada uno de los coeficientes estimados.Se realiza mediante el contraste de significación individual de cada parámetro.

- Signo de los parámetros. En caso de tratarse de coeficientes significativos hay que ver si existe lógica respaldada por teorías acerca del signo obtenido en la estimación.

- Magnitud de los parámetros. En un modelo de regresión, cada uno de los parámetros que acompañan a las variables explicativas indica la variación que sufre el valor esperado de la variable endógena ante un incremento en una unidad de la variable explicativa a la que hace referencia el parámetro estimado. En cambio, la estimación de la constante indica cuál es el valor esperado de la variable endógena cuando todas las variables explicativas toman valor nulo.

- Significación general de todos los coeficientes estimados. Se realiza mediante el contraste de significación global de los parámetros.

- Bondad del ajuste. A través de los coeficientes de determinación y de determinación ajustado se mide qué parte de la variable endógena queda explicada por el modelo estimado.

Aunque existen muchas clasificaciones diferentes de los modelos de regresión, una clasificación básica se establece en función del número de regresores que contiene el modelo:

- **Modelo de Regresión Lineal Simple (MRLS)**: estudia el comportamiento de una variable en función de otra. Formalmente se define como $y = f(x)$.

- **Modelo de Regresión Lineal Múltiple (MRLM)**: estudia el comportamiento de una variable en función de más de una variable. Formalmente se define como $y = f(x_1, x_2, \ldots)$.

A continuación, se ofrecen ejemplos de estimación de modelos econométricos con R y R-Commander atendiendo a dicha clasificación. Posteriormente, se analiza qué debemos hacer cuando tenemos alguna variable cualitativa que queremos introducir como regresor. Finalmente, se estudian los modelos de regresión cuando se imponen restricciones lineales *a priori*.

2. Modelo de Regresión Lineal Simple (MRLS)

En el modelo de regresión simple se desea estudiar el comportamiento de la variable explicada (Y) en función de una sola variable explicativa (X). Para estimar el signo y la magnitud de esta relación se toma una muestra de dimensión N, es decir, se obtienen N observaciones de las variables X e Y.

El modelo que hay que estimar es el siguiente:

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

Aunque una explicación teórica detallada del modelo de regresión se incluye en el capítulo precedente, recordemos que en esta ecuación, α es la constante, β es la pendiente y u es una variable aleatoria denominada *término de error o de perturbación*. Al ser una ecuación teórica que engloba a toda la población, los parámetros α y β son desconocidos. Así pues, el objetivo de la estimación del modelo será poder realizar inferencia sobre este. Por tanto, el primer paso será obtener los coeficientes estimados de los parámetros ($\hat{\alpha}$ y $\hat{\beta}$) a partir de los valores muestrales de X e Y. Una vez obtenidos estos, el modelo estimado tendrá la siguiente expresión:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

La diferencia entre los valores muestrales de la variable dependiente (y_i) y sus valores estimados por la recta ($\hat{y}_i$) son los *residuos* o *errores* de la estimación:

$$e_i = y_i - \hat{y}_i$$

Así pues, el modelo estimado también se puede expresar de la siguiente manera:

$$y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$$

Es importante no confundir el término de perturbación (u) con los residuos (e). El primer concepto es teórico y no observable, mientras que el segundo depende de la muestra y del método de estimación elegido, con lo que es medible y analizable.

Una buena estimación de un modelo, es decir, con un buen ajuste, será la resultante de valores de e_i reducidos y distribuidos normalmente. Así, cuanto más pequeños sean los e_i mejor será la estimación del modelo y más fiables serán las predicciones sobre el comportamiento de Y obtenidas con dicha estimación.

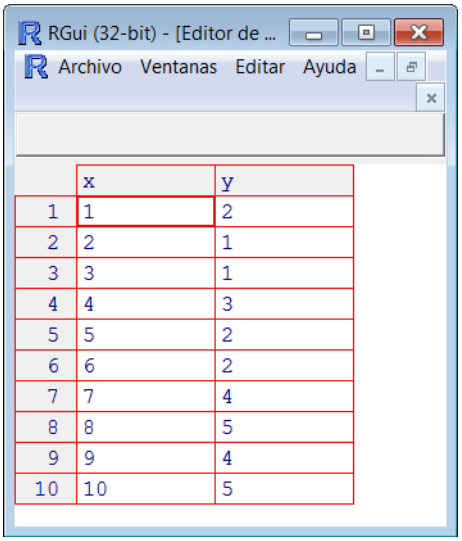
Como ejemplo, supongamos que disponemos de $N = 10$ observaciones de las variables X e Y:

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 utilizaremos la siguiente ruta para introducir estos datos:

Datos / Nuevo conjunto de datos

Una vez especificado un nombre para este conjunto de datos, los introducimos en una hoja donde cada columna es una variable, tal y como se muestra a continuación.



	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 vimos en el módulo dedicado al análisis descriptivo del manual *Matemáticas y Estadística con* R, es recomendable iniciar el análisis con estadísticos básicos de las variables. Una primera explotación estadística se obtiene siguiendo las instrucciones:

Estadísticos / Resúmenes / Conjunto de datos activo

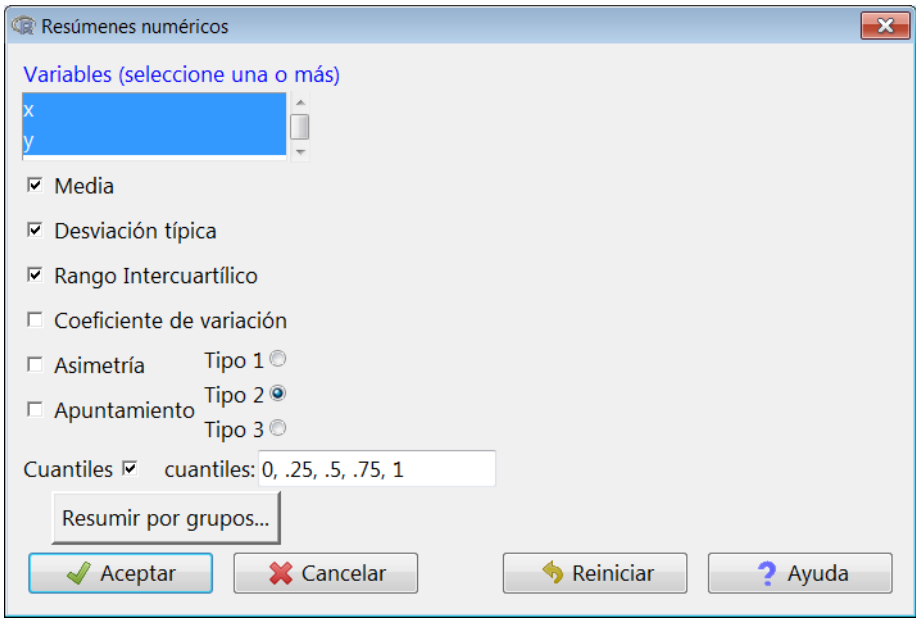
Con esto, el resultado será el siguiente:

```
> 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
```

Muchas veces no tendremos suficiente con los estadísticos básicos y desearemos obtener medidas adicionales, como la asimetría, la curtosis, el coeficiente de variación, la desviación típica o algunos cuantiles. Para ello existe una opción en la que se puede elegir entre un conjunto de estadísticos; la ruta que deberemos seguir para ello será:

Estadísticos / Resúmenes / Resúmenes numéricos

Obtendremos el siguiente menú, en el que seleccionaremos, de las variables deseadas, los estadísticos que queramos obtener.



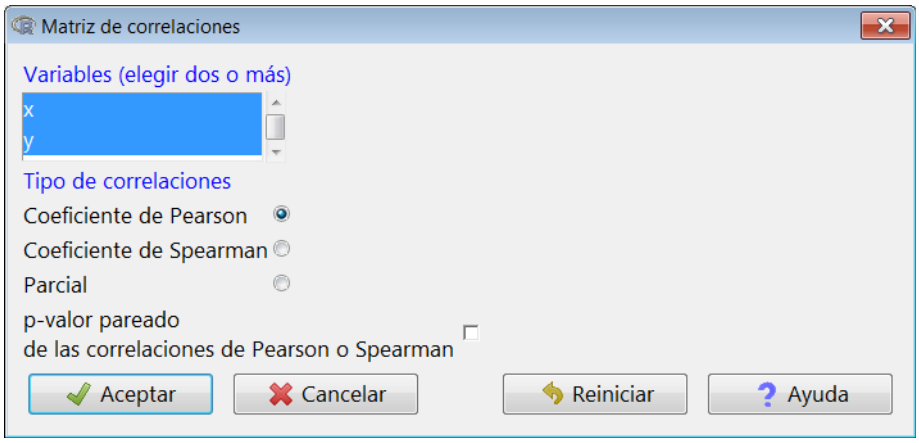
Este es el resultado que aparece en la *ventana de resultados*:

```
> 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 estadístico relevante, cuando se trabaja con más de una variable, es el coeficiente de correlación lineal de Pearson. Para calcularlo, hay que seguir la instrucción:

Estadísticos / Resúmenes / Matriz de correlaciones

Aparecerá el siguiente cuadro de diálogo, en el que seleccionaremos las variables para las que deseamos calcular el coeficiente de correlación:



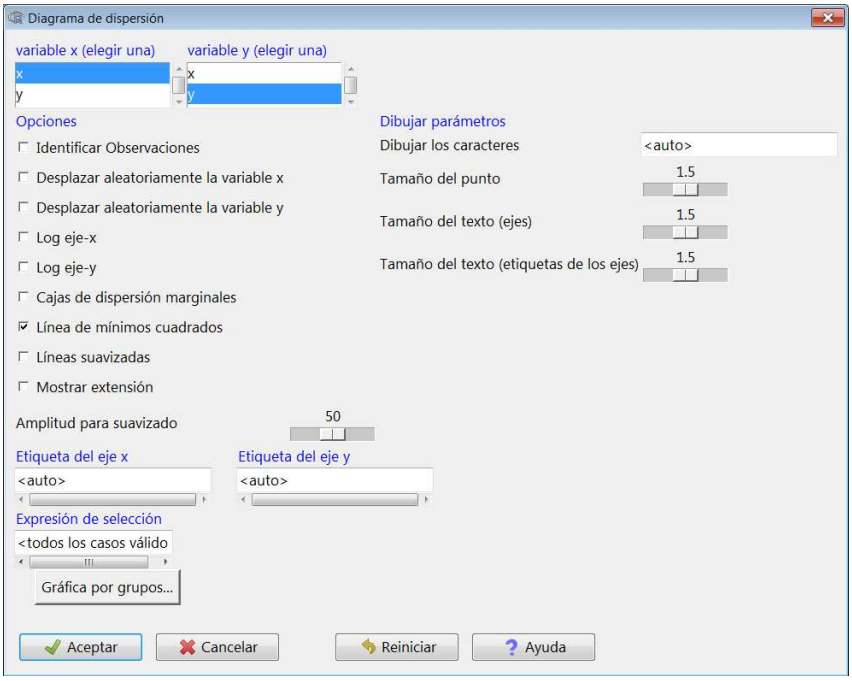
Como vemos, la correlación entre ambas variables es positiva y bastante elevada:

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

Visualmente, la correlación entre dos variables se puede comprobar mediante un diagrama de dispersión de las variables X e Y. Obtener este gráfico en R-Commander es inmediato accediendo a:

Gráficas / Diagrama de dispersión

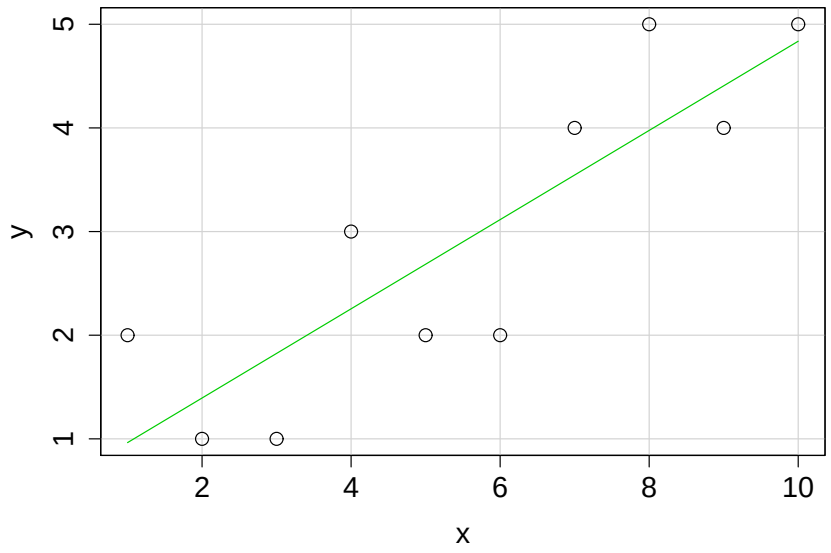
Aparecerá el siguiente menú, donde especificaremos la variable x (correspondiente al eje horizontal) e y (correspondiente al eje vertical). Además, activaremos la opción *Línea de mínimos cuadrados*, que dibuja la recta de regresión sobre los puntos.



En el gráfico resultante, las diferencias verticales entre cada observación y la recta estimada son los residuos (e_i). Cuanto más reducidos sean estos, mejor será el ajuste de la estimación del modelo.

Recta de regresión

Nótese que, en esta recta de regresión estimada, el punto de corte con el eje vertical es $\hat{\alpha}$, mientras que su pendiente es $\hat{\beta}$.



Entrando de lleno en la estimación del modelo de regresión, primero veremos cómo calcular los coeficientes del modelo y su coeficiente de determinación (R^2) mediante código de manera manual.

Las fórmulas que debemos aplicar son las siguientes:

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

$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta}$$

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

Siendo s la desviación estándar, s^2 la varianza, s_{xy} la covarianza y r el coeficiente de correlación lineal de Pearson. En R-Commander, hacer estos cálculos usando la sintaxis del lenguaje propio de R es inmediato. Basta con tener en cuenta los operadores descritos en la tabla 1, ya vistos en el primer módulo del manual *Matemáticas y Estadística con* R.

Tabla 1. Operadores estadísticos básicos con R

Descripción	Instrucción	Resultado
Longitud	`length(x)`	`10`
Máximo	`max(x)`	`10`
Mínimo	`min(x)`	`1`
Suma	`sum(x)`	`55`
Producto	`prod(x)`	`3628800`
Media	`mean(x)`	`5.5`
Media	`median(x)`	`5.5`
Desviación estándar	`sd(x)`	`3.02765`
Varianza	`var(x)`	`9.166667`
Covarianza	`cov(x,y)`	`3.944444`
Correlación	`cor(x,y)`	`0.8549254`
Producto escalar	`sum(x*y)`	`195`

Con esta información, calcularemos $\hat{\alpha}$, $\hat{\beta}$ y R^2 introduciendo sus respectivas fórmulas en la *Ventana de instrucciones*, luego seleccionaremos el conjunto y pulsaremos en *Ejecutar*:

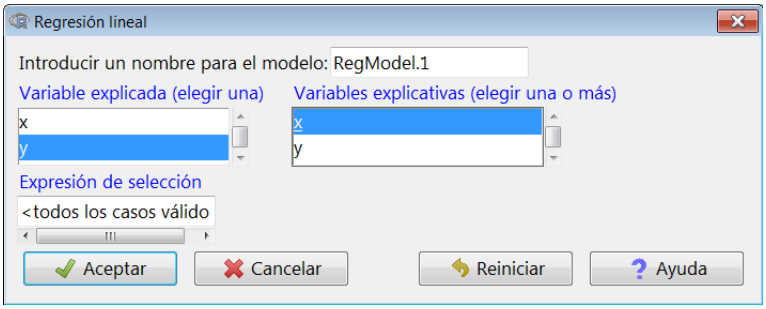
```
> attach(Datos)
> beta <- cov(x,y)/var(x)
> alpha <- mean(y)-beta*mean(x)
> coef.det <- cor(x,y)^2

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

Naturalmente, R-Commander ofrece una manera más rápida e inmediata de calcular una recta de regresión, que además incluye más información estadística del modelo. Una vez que las variables X e Y han sido introducidas en una base de datos, hay que estimar un modelo. La manera más sencilla es seguir la siguiente ruta:

Estadísticos / Ajuste de modelos / Regresión lineal

Aparecerá un cuadro de diálogo donde especificaremos cuál es la variable dependiente y la independiente, además de introducir un nombre para el modelo estimado (*RegModel.1*):



En la *ventana de resultados* aparecerá:

```
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
```

Este resultado es un amplio sumario de la regresión. Veamos sus principales componentes:

- `Residuals`: mínimo, máximo y cuartiles de los residuos de la regresión, que proporcionan información sobre su distribución.

- `Coefficients`: cuadro en el que aparece información de la estimación de los parámetros (o coeficientes) estimados.

- `Estimate`: estimación de cada parámetro (*intercept* significa constante).

- `Std.Error`: desviación (o error) estándar de cada parámetro estimado.

- `t value`: estadístico t de cada parámetro estimado, obtenido dividiendo la estimación del parámetro entre su desviación estándar. Este estadístico es el que utilizamos para realizar el contraste de significación individual de los parámetros estimados.

Contraste de significación individual...

... es un contraste cuyas hipótesis son H_0 : *parámetro* = 0 y H_1 : *parámetro* $\neq 0$.

- `Pr(>|t|)`: *p-valor* del contraste de significación individual de cada parámetro estimado, que indica su significación estadística.

- `Signif. codes`: muestra, con asteriscos y puntos, para qué niveles de significación los coeficientes estimados son o no significativos. En este caso, vemos que $\hat{\alpha} = 0,533$ no es significativo y que $\hat{\beta} = 0,430$ es significativo con un nivel de significación del 1 % ('**`**`'`0,01`).

- `Residual standard error`: desviación (o error) estándar de los residuos.

- `Multiple R-squared`: coeficiente de determinación.
- `Adjusted R-squared`: coeficiente de determinación ajustado.
- `F-statistic`: estadístico F para el contraste de la significación global o conjunta de los parámetros estimados del modelo.
- `DF`: grados de libertad del estadístico F.
- `p-value`: *p-valor* asociado al contraste anterior. En este caso, vemos que el conjunto de parámetros estimados es significativo con un nivel de significación del 0,1 % (*p-valor*$< 0,001$).

Contraste de significación global...

... es un contraste cuyas hipótesis son H_0 : *todos los parámetros* $= 0$ y H_1 : *algún parámetro* $\neq 0$.

Una manera alternativa de estudiar la significación individual de los parámetros estimados es el cálculo de intervalos de confianza. Tomando un nivel de confianza del 95 % (es decir, una significación del 5 %), existe una probabilidad del 95 % de que, por ejemplo, el parámetro β esté incluido en el siguiente intervalo:

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

Donde $t_{0{,}025;\,8}$ es el valor en tablas del estadístico t, a dos colas y con 8 grados de libertad, y $s_{\hat{\beta}}$ la desviación estándar del coeficiente estimado. R-Commander permite calcular conjuntamente los intervalos de confianza de todos los parámetros estimados del modelo (en este caso dos). Una vez seleccionado el modelo, la ruta es la siguiente:

El valor en tablas de t depende del nivel de significación (α) y del número de parámetros que estimar (2, en particular para un MRLS y k en general para un MRLM), ya que dicho valor es $t_{\alpha/2;\,N-k}$.

Modelos / Intervalos de confianza

Aparecerá un cuadro de diálogo, donde hay que especificar el nivel de confianza deseado y tras pulsar *Aceptar* obtendremos el siguiente resultado:

```
> 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
```

Como en el caso de la constante el valor cero está incluido en el intervalo de confianza (los extremos son de signo opuesto), al 95 % de confianza podemos afirmar que el parámetro estimado de la constante no es significativo, lo que equivale a afirmar que no es estadísticamente diferente de cero. Vemos que esto no ocurre en el caso de la pendiente.

3. Modelo de Regresión Lineal Múltiple (MRLM)

El MRLM es una generalización del modelo simple en k parámetros (incluyendo la constante). Por tanto, la variable endógena se explica por más de una variable exógena:

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

Consideremos un ejemplo práctico. Se desea estudiar un modelo de regresión lineal para estudiar los determinantes del nivel de paro en diferentes municipios catalanes, en concreto, $N = 295$. El modelo que debemos especificar es:

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

Donde tenemos las siguientes variables:

- *PARO*: tasa de paro.
- *MOTOR*: índice de motorización (turismos por habitante).
- *RBFD*: renta bruta familiar disponible, en miles de euros.
- *TEMP*: tasa de temporalidad laboral.
- *UNIV*: porcentaje de estudiantes universitarios en la población.

Los datos, una vez importados del documento de Excel, son los siguientes:

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

Antes de realizar una estimación, es muy útil hacer una descripción estadística de las variables. El resumen del conjunto de datos es el siguiente:

```
> 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
```

Como hemos hecho antes, para ver más estadísticos de las variables se pueden seguir las instrucciones que se muestran a continuación y seleccionar lo que más nos interese calcular de la lista de estadísticos que ofrece R-Commander:

Estadísticos / Resúmenes / Resúmenes numéricos

El resultado obtenido es el siguiente:

```
> 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 dos o más variables tienen entre ellas una alta correlación, puede ser problemático incluirlas simultáneamente como variables explicativas. Por eso mismo, resulta muy útil calcular la matriz de correlaciones lineales de las variables explicativas:

En concreto, como se verá en el siguiente módulo, una alta correlación entre dos regresores puede dar lugar a problemas de *multicolinealidad*.

Estadísticos / Resúmenes / Matriz de correlaciones

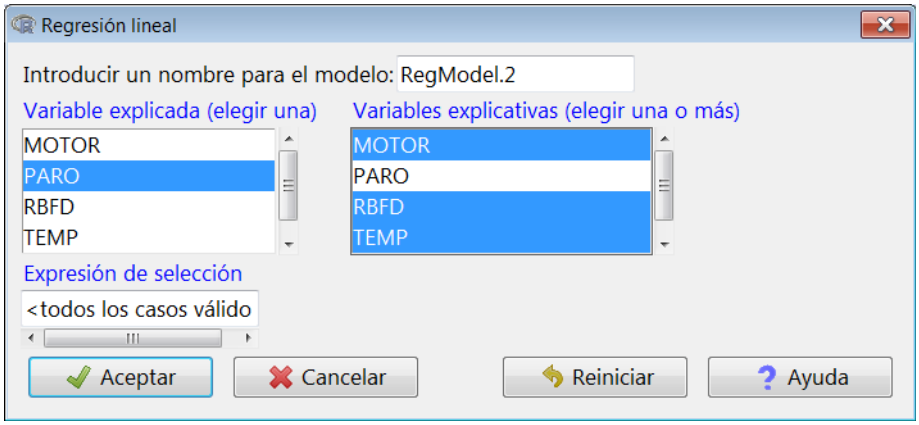
Seleccionando las variables que queremos incluir, obtenemos el siguiente resultado:

```
> 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
```

Análogamente al caso del MRLS, la siguiente ruta nos permitirá estimar un modelo de regresión, seleccionando las variables explicada y explicativas:

Estadísticos / Ajuste de modelos / Regresión lineal

En el cuadro de diálogo resultante introducimos las variables explicada y las explicativas, además del nombre de este modelo (*RegModel.2*):



En la ventana de resultados obtenemos lo siguiente:

```
> 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
```

Como vemos, todos los coeficientes estimados son significativos, aunque el ajuste del modelo ($R^2 = 0,278$) es más bien pobre.

Si calculamos los intervalos de confianza (IC) de los coeficientes estimados obtenemos:

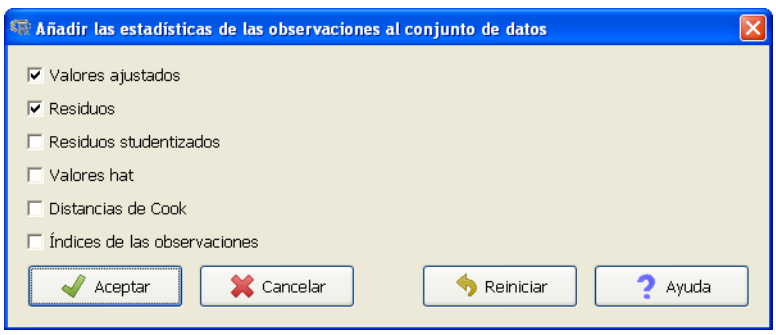
```
> Confint(RegModel.2, level=0.95)
                Estimate        2.5 %       97.5 %
(Intercept) 19.244961330 15.823732809 22.666189851
MOTOR       -0.007755427 -0.014242476 -0.001268379
RBFD        -0.037110206 -0.058139257 -0.016081156
TEMP         0.030971082  0.005919276  0.056022888
UNIV   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

Recordad que para obtener los IC de los parámetros hemos de seleccionar el modelo, seguir la ruta *Modelos / Intervalos de confianza* y seleccionar el nivel de confianza deseado.

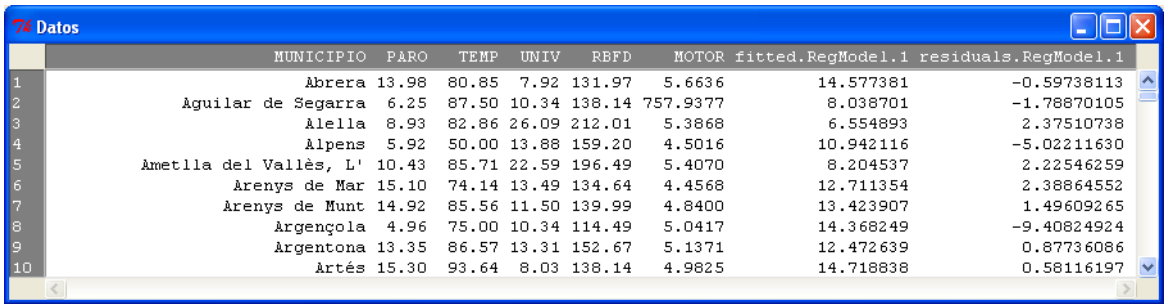
R-Commander nos da la opción de obtener información estadística adicional del modelo estimado. Entre otros indicadores, podemos extraer los residuos (e_i) y los valores ajustados a la recta ($\hat{y}_i$). Para hacer esto, accederemos a:

Modelos / Añadir las estadísticas de las observaciones a los datos

En nuestro ejemplo, solo añadiremos al conjunto de datos los residuos y los valores ajustados a la recta. Para ello, los activaremos en el cuadro de diálogo:



Habiendo hecho esto, si visualizamos nuestro conjunto de datos, observaremos cómo se han añadido estas dos 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

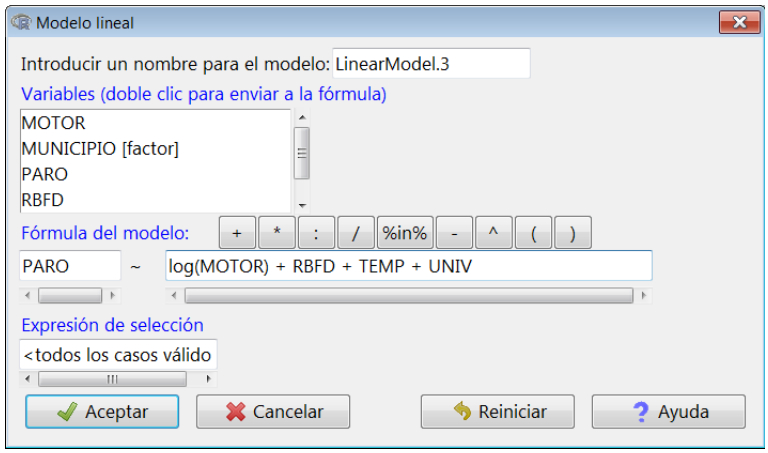
Lo más lógico es que no estemos satisfechos con el modelo estimado y queramos mejorar su estimación. Podemos optar por el siguiente modelo alternativo, donde la variable $MOTOR$ aparece en logaritmos:

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

Para estimar este nuevo modelo, una posible solución sería crear una nueva variable, $log(MOTOR)$, añadirla al conjunto de datos y estimar el nuevo modelo como hemos hecho antes. Sin embargo, tenemos a nuestra disposición una alternativa más rápida y eficiente: un completo cuadro de diálogo que nos permite introducir variables trasformadas (aplicar logaritmos a una variable, elevarla al cuadrado, etc.) o multiplicadas entre ellas; incluso podemos seleccionar una muestra de nuestra base de datos, etc. Es decir, la solución consiste en estimar directamente un modelo mediante la siguiente ruta alternativa:

Estadísticos / Ajuste de modelos / Modelo lineal

Nos aparecerá un cuadro de diálogo en el que introduciremos la fórmula del modelo, que tiene dos partes: la variable dependiente y el conjunto de regresores o variables explicativas. En nuestro ejemplo, introducimos la variable *Motor* en logaritmos. Además, le asignaremos a este modelo el nombre *LinearModel.3*:



El resultado que se obtiene es el siguiente:

```
> 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. Comparación de ambos modelos

Comprobamos que el ajuste del modelo ha mejorado respecto al modelo anterior. Es importante destacar que el resultado de la estimación siempre muestra dos valores del coeficiente de determinación; uno de ellos se denomina coeficiente de determinación ajustado. El motivo es que siempre que se añadan nuevas variables explicativas a un modelo, el valor de R^2 subirá, aun cuando estas nuevas variables no aporten nada nuevo al modelo. Por eso mismo, el valor ajustado de R^2 incluye una penalización por el número de regresores que el modelo contiene.

Cuando queramos comparar dos modelos que tengan la misma variable endógena pero con distinto número de variables explicativas elegiremos aquel que tenga un mayor valor de la R^2 ajustada.

4. Variables exógenas cualitativas

En cualquier estudio, es habitual encontrarnos con modelos de regresión en los que se contemplan situaciones con variables explicativas; son atributos o variables de carácter cualitativo. La codificación de estas variables supone identificar cada categoría con un valor. Por este motivo, cuando las categorías no tienen una ordenación clara y recurrimos a criterios arbitrarios para realizar la codificación, hemos de prestar especial atención. Por ejemplo, sin darnos cuenta, al ordenar los destinos turísticos con más afluencia de pasajeros por orden alfabético, estamos automáticamente estableciendo una prioridad que puede distar mucho de la realidad.

Por norma general, para no tener problemas de interpretación, entre otros, evitaremos introducir directamente como variable explicativa de un modelo de regresión una variable cualitativa politómica, cuya codificación numérica induzca un orden (y, como consecuencia, una distancia) entre las diferentes categorías que no se ajusten a la realidad.

Así pues, el primer paso es clasificar todas las posibles variables explicativas de un modelo de regresión. Ya hemos visto que cuando se trata de variables cuantitativas o cualitativas con orden implícito, no tenemos ningún problema y las podemos introducir directamente. El problema lo tenemos cuando nos encontramos con atributos o variables cualitativas. Es crucial distinguir aquí si son variables dicotómicas o politómicas, ya que en este último casos deberemo desglosarlas en distintas variables dicotómicas.

Variables dicotómicas, ficticias o *dummies*

Son variables que solo pueden tomar dos valores, generalmente 0 y 1. La categoría codificada con un 0 suele denominarse **categoría de referencia**.

4.1. Variables dicotómicas y cualitativas politómicas en el MRLM

Ya hemos mencionado que los valores más utilizados en la codificación de las variables dicotómicas son el 0 y el 1. La razón es porque simplifica enormemente la interpretación de los resultados obtenidos tras la estimación del modelo de regresión. Sabemos que en un modelo de regresión lineal todas las variables explicativas tienen asociado un parámetro que debemos estimar. Si esta variable toma el valor 0, su impacto sobre el valor esperado de la variable dependiente se anula. En este caso, el parámetro asociado a una variable ficticia muestra cuánto varía el valor esperado de la variable endógena cuando un individuo posee una determinada característica (identificada con el valor 1 de la *dummy*), respecto a un individuo que no la posea (identificada con el valor 0 de la *dummy*.

Podemos introducir variables ficticias en un modelo de regresión de distintas formas, de manera:

- Aditiva: como una variable explicativa más.
- Multiplicativa: multiplicando a otra variable explicativa ya existente en el modelo.
- Mixta: tanto aditiva como multiplicativamente.

Lógicamente, según cómo decidamos incorporar las *dummies* al modelo la interpretación que hagamos será distinta. No obstante, consideremos un modelo muy sencillo para ilustrar lo que acabamos de comentar. Supongamos que tenemos una clase de 90 alumnos de econometría y que queremos explicar la calificación obtenida al final de curso (Y_i) en función de sus horas de estudio (X_{1i}) y de su género (X_{2i}).

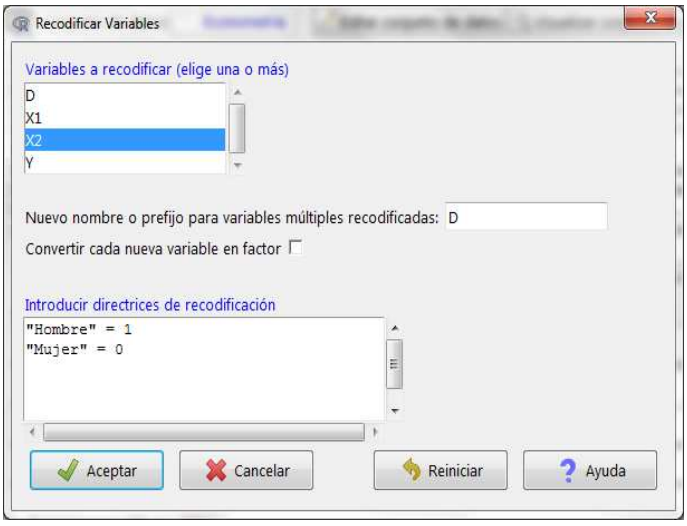
Primero leemos la base de datos y la visualizamos, vemos que la variable X_2 es textual y por tanto nos interesa recodificarla como una variable ficticia, que denominaremos D_i. Para hacer esto con R-Commander utilizaremos la siguiente ruta:

Datos / Modificar variables dentro del conjunto de datos activo / Recodificar variables

Nos saldrá un cuadro de diálogo, que rellenaremos según los nuestros criterios de definición de la *dummy*. Por ejemplo, queremos definir:

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

Por tanto:



Ahora, al visualizar el conjunto de datos vemos que nos aparece la nueva variable ficticia que acabamos de crear. Así pues, si suponemos que nuestra base de datos está ordenada por el género de los estudiantes (primero los hombres y luego las mujeres), podemos definir un modelo general para todos nuestros alumnos independientemente de su género:

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

un modelo para los hombres, suponiendo que las primeras N_1 observaciones corresponden a hombres:

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

y otro para las mujeres, suponiendo que desde la observación N_1+1 a la N son mujeres:

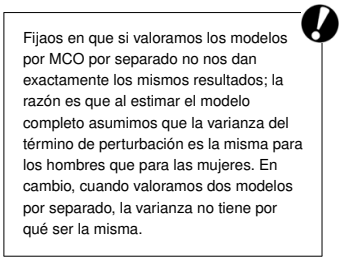
$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i \qquad i = N_1 + 1, \ldots, N.$$

Observemos que al introducir la variable ficticia de manera aditiva la diferencia entre la especificación de los hombres y de las mujeres recae sobre la constante, que para los hombres es $\beta_0 + \beta_2$, mientras que para las mujeres se limita a β_0. Cabe destacar también que el efecto producido en la calificación de las horas de estudio es el mismo independientemente del género del estudiante.

El resultado de la estimación de los tres modelos utilizando la ruta:

Estadísticos / Ajuste de modelos / Regresión lineal

y rellenando los correspondientes cuadros de diálogo es:



Fijaos en que si valoramos los modelos por MCO por separado no nos dan exactamente los mismos resultados; la razón es que al estimar el modelo completo asumimos que la varianza del término de perturbación es la misma para los hombres que para las mujeres. En cambio, cuando valoramos dos modelos por separado, la varianza no tiene por qué ser la misma.

```
> 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
```

Veamos ahora un caso en el que una de las variables explicativas tuviera más de dos categorías, por ejemplo el itinerario de estudios que sigue el estudiante, este puede ser:

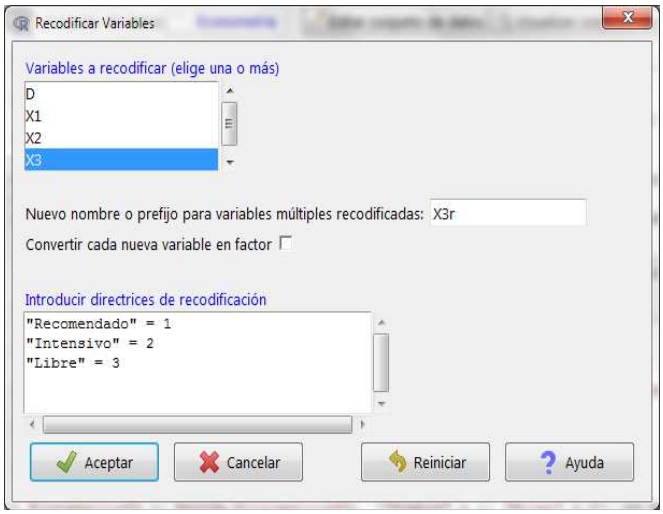
- Recomendado: el que se aconseja a los estudiantes para obtener el grado en 10 semestres.
- Intensivo: el que se pauta con el fin de obtener el grado en 8 semestres.
- Libre: cualquier otro que haya escogido el estudiante bajo su propia elección pero que no está marcado por la universidad.

Como siempre, primero leeremos la base de datos y, en caso de que esta sea un factor, la recodificaremos como hemos hecho antes pero ahora asignando un 1 a la primera categoría, un 2 a la segunda y así sucesivamente hasta j, que será el máximo de categorías posibles. Nuestro ejemplo sigue siendo la misma clase de 90 alumnos de econometría y queremos explicar la calificación obtenida a final de curso (Y_i) en función de sus horas de estudio (X_{1i}) y, en este caso, según el itinerario que hayan elegido (X_{3i}). Para realizar esta acción con R-Commander seguiremos la ruta:

Datos / Modificar variables dentro del conjunto de datos activo / Recodificar variables

Nos saldrá un cuadro de diálogo que rellenaremos según nuestros criterios de definición de la variable que queremos recodificar:

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



No obstante, no podemos utilizar directamente esta nueva variable porque implícitamente estaríamos imponiendo un orden y una proporcionalidad que no es cierta. Por tanto, debemos definir tantas variables dicotómicas como el número total de categorías excepto una. En nuestro caso basta con definir dos variables ficticias:

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

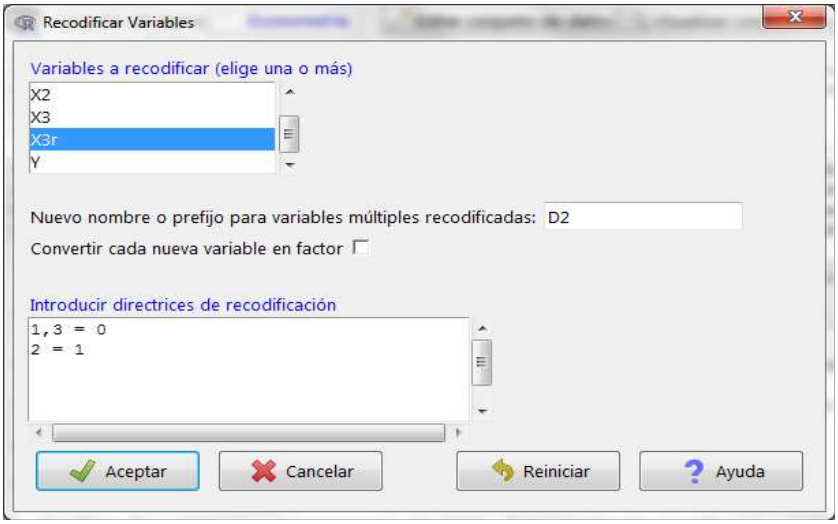
Fijaos en que D_{2i} toma valor 1 cuando los estudiantes siguen el itinerario intensivo y 0 en otro caso, es decir, si optan por el itinerario recomendado o el libre. En cambio, D_{3i} valdrá 1 cuando el itinerario elegido sea libre y 0 si este es el recomendado o el intensivo.

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

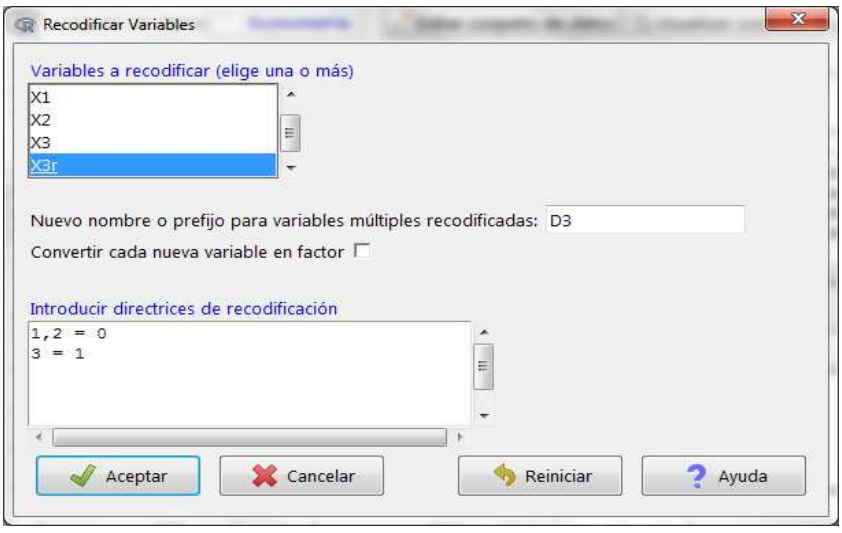
Esto lo haremos siguiendo la siguiente ruta:

Datos / Modificar variables dentro del conjunto de datos activo / Recodificar variables

y rellenando los dos cuadros de diálogo correspondientes a las dos nuevas variables:



Un estudiante que siga el itinerario recomendado tendrá asignado un cero tanto en D_{2i} como D_{3i}.



Observemos que la información respecto al itinerario de los estudiantes de econometría la podemos facilitar con una única variable codificada con tres valores (X_{3r}) o con dos variables dicotómicas (D_2 y D_3). No obstante, para introducir dicha información en un modelo de regresión se utiliza el criterio de definir tantas variables ficticias como el número total de categorías menos una para evitar caer en la *trampa de las variables ficticias*.

Categoría base o de referencia...

es aquella categoría que resulta cuando todas las *dummies* introducidas en el modelo toman el valor cero. En nuestro caso serían o bien aquellos estudiantes hombres que siguen el itinerario recomendado si introducimos tanto el género como el itinerario en el modelo o bien los estudiantes que siguen el itinerario recomendada en caso de no tener en cuenta el género.

El modelo que especificaríamos sería pues:

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

y su estimación nos daría:

Trampa de las ficticias...

... se produce cuando introducimos una *dummy* para cada categoría en un modelo de regresión con término independiente. En este caso, como veremos en el módulo siguiente, incurriríamos en un problema de multicolinealidad perfecta y no podríamos estimar el modelo.

```
> 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
```

Por tanto, siempre que tengamos información recogida en variables cualitativas deberemos utilizar *dummies* para introducirla en cualquier modelo de regresión. Esto lo podremos hacer siempre independientemente del tipo de información que queramos recoger: podemos crear variables ficticias espaciales para identificar, por ejemplo, la zona de residencia de nuestros estudiantes; variables ficticias temporales, para identificar si han nacido en un determinado período, etc.

4.2. Interpretación de los coeficientes de las variables ficticias

La interpretación de los parámetros de las *dummies* variará según el modo como hayamos introducido dichas variables en el modelo:

Si utilizáramos otra codificación distinta al 0 y 1, la interpretación del modelo se complicaría mucho.

- Aditiva
- Multiplicativa
- Mixta

4.2.1. Introducción de *dummies* en un modelo de forma aditiva

El parámetro estimado de una variable ficticia incorporada a un modelo de regresión de manera aditiva se interpreta como la variación que se producirá en el valor esperado de la variable endógena cuando el individuo pertenezca a la categoría identificada con valor 1 respecto al que tendría si el individuo perteneciese a la categoría complementaria (identificada con un 0). Esta variación será siempre la misma, al margen del valor que adquiera el resto de las variables explicativas contenidas en el modelo.

En el siguiente capítulo se profundizará sobre cómo realizar restricciones lineales en un modelo de regresión.

Fijaos en que si lo que interpretamos es un contraste estadístico como los que podríamos realizar en el modelo general del ejemplo de los estudiantes de econometría, primero debemos reflexionar sobre el significado de los coeficientes que acompañan a cada una de las variables explicativas del modelo.

Por ejemplo, recuperemos el primer modelo especificado en la sección anterior, donde queríamos explicar la calificación obtenida al final de curso (Y_i) de 90 estudiantes de econometría en función de sus horas de estudio (X_{1i}) y de su género recogido en la variable ficticia (D_i):

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

cuya estimación 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
```

Veamos las interpretaciones de cada uno de los parámetros del modelo mediante los contrastes de significación individual (CSI) de estos:

La interpretación del parámetro β_1 que acompaña a la variable X_1 es el mismo que en un modelo de regresión, es decir, cuando un estudiante incrementa en una unidad sus horas de estudio, su calificación esperada incrementa en unos $0{,}18$ puntos.

- El parámetro β_2 compara si habiendo estudiado las mismas horas, la calificación esperada de un hombre es igual o diferente a la de una mujer. En nuestro ejemplo, no rechazamos la hipótesis nula, por tanto, la calificación esperada debería ser la misma para los hombres y las mujeres.

- El parámetro β_0 muestra si la calificación media de las mujeres es significativamente diferente de cero cuando no se ha realizado ninguna hora de estudio (X_1). Nuestro ejemplo muestra que la calificación media de las mujeres que no han estudiado ninguna hora es de casi medio punto menos a la obtenida por aquellas que han invertido horas de estudio. Si se quisiese hacer lo mismo con los hombres, deberíamos realizar un contraste de restricciones lineales que veremos en el siguiente capítulo con $\beta_0 + \beta_2 = 0$.

Si ahora nos centramos en el modelo donde explicábamos la calificación de econometría (Y_i) en función de las horas de estudio (X_{1i}) y el itinerario seguido por los estudiantes (D_2 y D_3):

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

cuyos resultados de estimación eran:

```
> 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 este caso, la interpretación de los CSI de los parámetros serían:

- El parámetro β_3 indica si la calificación esperada de un estudiante que sigue un itinerario libre es igual o diferente a la de otro que seguía cualquier otro itinerario cuando las horas de estudio permanecen constantes. En este ejemplo, el hecho de seguir un itinerario libre no afecta a la calificación esperada del estudiante.

- El parámetro β_2 compara si, habiendo estudiado las mismas horas, la calificación esperada de un estudiante que sigue un itinerario intensivo es igual o diferente a la de otro que seguía cualquier otro itinerario. Otra vez, la calificación esperada del estudiante no queda alterada porque este haya elegido seguir un itinerario intensivo.

- El parámetro β_0 muestra si la calificación media de los estudiantes que han seguido el itinerario recomendado es significativamente diferente de cero cuando no se ha realizado ninguna hora de estudio (X_1). Vemos que si un estudiante que sigue el itinerario recomendado no realiza ninguna hora de estudio, su calificación se reduce en unos 0,34 puntos.

- La diferencia de calificación esperada entre un estudiante que sigue un itinerario intensivo y otro que sigue uno libre, con las mismas horas de estudio, estaría determinada por $\beta_2 - \beta_3$. Si no atendemos a la significación de los coeficientes, en nuestro ejemplo esta diferencia sería de $-0{,}018524 - (-0{,}161528) = 0{,}143004$.

Imaginemos que en lugar del modelo anterior hemos optado por especificar otro sin término independiente pero incorporando otra nueva variable ficticia que recoja si un estudiante opta por seguir el itinerario recomendado. Si definimos D_4 como:

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

El modelo que especificaríamos sería:

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

y su estimación resultaría:

```
> 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
```

Ahora, la interpretación de los CSI de los parámetros serían:

- El parámetro δ_4 recoge la calificación esperada de los estudiantes que han seguido el itinerario cuando no se ha realizado ninguna hora de estudio (X_1). Vemos que dicho parámetro es de $-0{,}34$; así, si un estudiante que sigue el itinerario recomendado no realiza ninguna hora de estudio, su calificación se reduce en unos 0,34 puntos.

- El parámetro δ_3 indica la calificación esperada de un estudiante que sigue un itinerario libre en el caso de que no haya estudiado nada. En este ejemplo, vemos que el parámetro es de $-0{,}5$.

- El parámetro δ_2 muestra la calificación esperada de un estudiante que sigue un itinerario intensivo cuando este no ha estudiado ninguna hora, es decir, $-0{,}36$ puntos.

Veamos la equivalencia que existe entre las dos soluciones propuestas de ambos modelos:

Tabla 2. Equivalencia de los modelos EcoModel.4 y EcoModel.5

Valor esperado de la calificación cuando $X_{1i} = 0$		
Modelo	EcoModel.4	EcoModel.5
Recomendado	$\beta_0 = -0{,}342012$	$\delta_4 = -0{,}342012$
Intensivo	$\beta_0 + \beta_2 = -0{,}360536$	$\delta_2 = -0{,}360537$
Libre	$\beta_0 + \beta_3 = -0{,}50354$	$\delta_3 = -0{,}503540$

Continuando con el ejemplo de los estudiantes de econometría supongamos que complicamos un poco el modelo: esperamos que el impacto esperado que tienen las horas de estudio sobre la calificación no sea el mismo para los hombres que para las mujeres.

Deberemos decidir si seguimos suponiendo que, cuando un estudiante no ha estudiado nada, la calificación esperada es la misma para los hombres que para las mujeres (introducción de una *dummy* en el modelo multiplicativamente), o es diferente (introducción de una *dummy* en el modelo aditiva y multiplicativamente). En ambos casos hemos de incorporar una variable dicotómica de forma multiplicativa.

4.2.2. Introducción de *dummies* en un modelo de forma multiplicativa

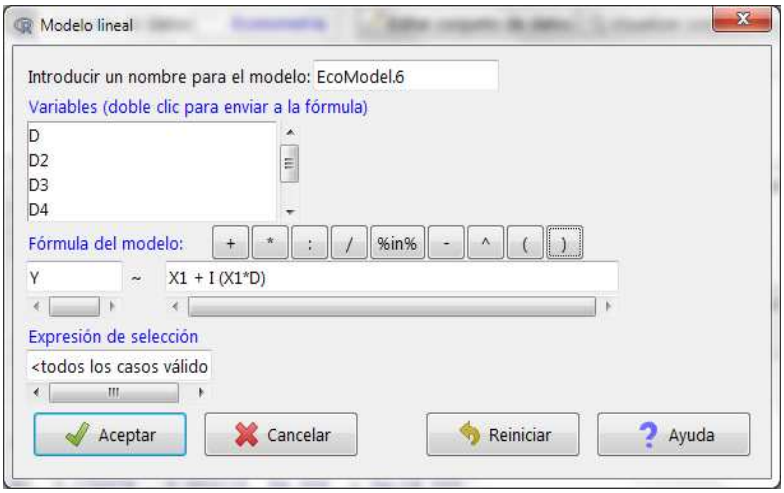
El modelo que especificaríamos sería:

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

y su estimación la haríamos siguiendo la ruta:

Estadísticos / Ajuste de modelos / Modelo lineal

rellenando el cuadro de diálogo tal y como se muestra a continuación:



La función `I()` nos permite realizar operaciones de varias variables elemento a elemento, que es lo que pretendemos al introducir la variable ficticia de manera multiplicativa.

obteniendo el siguiente resultado:

```
> 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 modelo multiplicativo EcoModel.6, el género del estudiante afecta al impacto de las horas de estudio sobre la calificación esperada; es decir, para los hombres este impacto será igual a $\beta_1 + \beta_2 = 0{,}176877$, mientras que para las mujeres será igual a $\beta_1 = 0{,}176094$. Cabe destacar que, en este ejemplo, el parámetro β_2 no resulta estadísticamente significativo y, por tanto, el supuesto realizado inicialmente no sería cierto.

4.2.3. Introducción de *dummies* en un modelo de forma mixta (aditiva y multiplicativa)

En este caso el modelo que se ha de especificar es:

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

su estimación la realizaríamos siguiendo la misma ruta que antes, y rellenaríamos el cuadro de diálogo tal y como hemos hecho. El resultado sería:

```
> 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 modelo mixto (aditivo y multiplicativo) EcoModel.7, el género del estudiante afecta al impacto de las horas de estudio sobre la calificación esperada; es decir, para los hombres este impacto será igual a $\beta_1 + \beta_2 = 0{,}17665$, mientras que para las mujeres será igual a $\beta_1 = 0{,}1762521$. Además, la calificación esperada de un hombre que no haya estudiado nada vendrá explicada por $\beta_0 + \beta_3 = -0{,}3987868$ y la de una mujer que tampoco haya estudiado, será de $\beta_0 = -0{,}4141740$. Del mismo modo que ocurría antes, los parámetros β_2 y β_3 no son significativos. Así, estos impactos serían iguales independientemente del género del estudiante.

Otra observación es que el modelo EcoModel.7 es el modelo más general, y si realizamos en él distintos contrastes, llegamos a modelos más sencillos. Por ejemplo, si realizamos el contraste de $\beta_3 = 0$ estamos contrastando el modelo EcoModel.6 en la hipótesis nula. Esto nos llevaría a pensar erróneamente que es mejor especificar el modelo más general. No obstante, no debemos olvidar que si el número de categorías es elevado y además pretendemos considerar varios efectos multiplicativos, nuestro modelo tendrá muchos parámetros y podremos tener problemas de estimación.

4.2.4. Interpretación de las interacciones

En el supuesto de que queramos introducir distintas variables explicativas cualitativas en un modelo de regresión, lo más normal es que se hayan supuesto varios efectos multiplicativos entre ellas. En este caso, estos efectos multiplicativos se denominan **efectos cruzados o interacciones**.

Como nos pasaba cuando introducíamos una variable cualitativa, se nos presentan tres posibilidades: incorporación aditiva, multiplicativa o mixta.

Para ilustrar esto utilizaremos la base de datos de Econometría, que recogía información sobre dicha asignatura en una clase de 90 estudiantes. Las variables de las que disponemos son:

- Y_i: calificación del estudiante.
- X_{1i}: horas de estudio.
- D_i: variable ficticia indicadora del género del estudiante.
- D_{2i}: variable ficticia que identifica a aquellos estudiantes que han seguido un itinerario intensivo.
- D_{3i}: variable ficticia que recoge a los estudiantes que siguen un itinerario libre.

Uno de los modelos que podríamos definir sería:

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

Este modelo solo presenta efectos aditivos, por tanto:

- β_0 es la calificación esperada de una mujer que sigue el itinerario recomendado, es decir, el valor esperado de la variable endógena para la categoría base o de referencia.

- β_1 indica la diferencia entre la calificación esperada de los hombres y las mujeres que siguen el mismo itinerario de estudios.

- β_2 representa la diferencia entre la calificación esperada de un estudiante que sigue un itinerario intensivo y uno que sigue el itinerario recomendado.

- β_3 muestra la diferencia que existe entre la calificación esperada de un estudiante que sigue un itinerario libre y otro que opta por el itinerario recomendado.

Esto mismo lo podemos resumir en la tabla siguiente:

Tabla 3. Varias variables cualitativas con *dummies* de forma aditiva

Modelo sin interacciones		
	$E(Y_i)$	
Itinerario	Hombres	Mujeres
Recomendado	$\beta_0 + \beta_1$	β_0
Intensivo	$\beta_0 + \beta_1 + \beta_2$	$\beta_0 + \beta_2$
Libre	$\beta_0 + \beta_1 + \beta_3$	$\beta_0 + \beta_3$

Otro modelo podría ser:

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

cuyas interpretaciones se resumen en la tabla que se muestra a continuación:

Tabla 4. Varias variables cualitativas con *dummies* de forma mixta

Modelo con interacciones		
	$E(Y_i)$	
Itinerario	Hombres	Mujeres
Recomendado	$\beta_0 + \beta_1$	β_0
Intensivo	$\beta_0 + \beta_1 + \beta_2 + \beta_4$	$\beta_0 + \beta_2$
Libre	$\beta_0 + \beta_1 + \beta_3 + \beta_5$	$\beta_0 + \beta_3$

Recordad que para estimar un modelo con interacciones hemos de seguir la ruta:

Estadísticos / Ajuste de modelos / Modelo lineal

e introducir las interacciones mediante la función I() dentro del cuadro de diálogo.

4.3. Otros usos de las variables ficticias

Como ya hemos estudiado anteriormente, los usos más comunes de las variables ficticias son la introducción en un modelo de regresión de variables cualitativas no ordinales que nos permitan segmentar nuestra muestra en distintos grupos (sexo, estado civil, tipo de vivienda, etc.), o distinguir periodos temporales (nacidos antes o después de un cierto año, etc.) o localizaciones (comunidad de residencia, país de origen, etc.). No obstante, estas no son las únicas aplicaciones de dichas variables en un modelo de regresión, ya que también las podemos utilizar para:

- Datos atípicos
- Cambio estructural
- Estacionalidad
- Modelo de efectos fijos

4.3.1. Datos atípicos

En el siguiente módulo se analizará con más detalle lo que son los datos atípicos, aquí únicamente nos centraremos en el uso de las variables ficticias para, una vez detectados, mantenerlos dentro de la muestra sin que estos afecten a la estimación del modelo.

¿Qué son los datos atípicos?

Una observación o dato se considera atípica cuando hemos cometido un error de medida, cuando se trata de un individuo fuera de lo común o cuando se mide en un momento temporal ligado a una situación extraordinaria.

Medidas de detección de datos atípicos

Los instrumentos más usuales para detectar si en una muestra existen datos atípicos son el *leverage*, el residuo, el residuo estudentizado, el residuo estudentizado con omisión y la distancia de Cook.

Hay que tener en cuenta dos aspectos relevantes en este contexto:

- Introducción de una variable ficticia para cada observación atípica. En este caso, las estimaciones de los parámetros del modelo serán las mismas que si eliminásemos las observaciones atípicas pero con el agravio de que estamos incluyendo más regresores en el modelo. La única ventaja es que si valoramos por MCO el residuo de la estimación para cada observación atípica será cero y el coeficiente que acompañe a la variable ficticia recogerá el efecto de cada observación atípica sobre la endógena. Al crear una variable ficticia que es igual a 1 para la observación atípica y cero para el resto de las observaciones, la estimación MCO del parámetro que la acompaña no es más que el error de predicción si utilizáramos el modelo para realizar una predicción de la variable endógena con esta observación. Por tanto, también podríamos calcular los errores de predicción y sus intervalos de confianza.

- Introducción de una variable ficticia que recoja todas las observaciones atípicas. Las condiciones ahora son distintas porque tenemos una variable ficticia que vale 1 cuando se trata de una observación atípica y cero para el resto de las observaciones. En este caso, los residuos obtenidos de la estimación por MCO no tendrían por qué ser cero. Si eso fuera así, la información que nos aportaría sería mucho

más limitada que en el caso anterior. Además, no tendríamos evidencias sobre el impacto de cada una de las observaciones atípicas sino un efecto global de todas ellas.

Disponemos de una base de datos con N observaciones de la que sabemos que la observación $i_0 - sima$ es un dato atípico. Ante esta circunstancia definimos la variable ficticia D_i, que valdrá 1 para la observación $i_0 - sima$ y 0 en el resto de las observaciones:

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

4.3.2. Cambio estructural

¿A qué nos referimos cuando hablamos de cambio estructural?

Cuando trabajamos con series temporales un cambio de estructura dentro del período muestral indica que se ha producido uno o más cambios que han provocado alguna variación en el proceso generador de datos. Sin embargo, cuando trabajamos con datos de corte transversales un cambio estructural indica la existencia de dos o más grupos que se comportan de distinta manera entre ellos.

Aunque el cambio o la permanencia estructural se trate con más detalle en el módulo siguiente, su aplicación con variable ficticias es similar al realizado con observaciones atípicas pero cambiando radicalmente la definición de la variable ficticia a introducir en el modelo. Ahora nos interesa crear una variable ficticia que sea cero para todos los periodos anteriores al cambio y uno para todos aquellos en los que el cambio ya se haya producido. El coeficiente estimado que acompañe a dicha variable recogerá el efecto del cambio estructural en el valor esperado de la variable endógena. El contraste de significación individual de dicho coeficiente corroborará la existencia de dicho cambio. Si resulta que el parámetro es estadísticamente significativo habrá cambio estructural y no lo habrá en caso contrario.

El test de Chow

El contraste que realizamos para ver si en una muestra existe cambio o permanencia estructural se denomina test de Chow.

Todo cambio estructural puede afectar a la constante (si introducimos la variable ficticia de forma aditiva), a la pendiente (si introducimos la variable ficticia de forma multiplicativa) o a ambas (cuando la variable ficticia se introduce aditiva y multiplicativamente). Dado el siguiente modelo:

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

Se desea contrastar si los parámetros de la población se mantienen iguales a lo largo de los dos subperiodos: el primero desde 1 hasta T_1 y el segundo desde $T_1 + 1$ hasta T. Para ello definiremos una variable dicotómica D_t, que será igual a 0 para el primer subperíodo e igual a 1 para el segundo:

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

Una vez creada la variable ficticia la introduciríamos de forma multiplicativa en la especificación para recoger el supuesto cambio de pendiente que se produce en la segunda submuestra.

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

4.3.3. Estacionalidad

Cuando trabajamos con series temporales en las que los datos son estacionales (meses, trimestres, cuatrimestres o semestres) podemos estar interesados en aislar los efectos de una determinada estación para esclarecer mejor qué variables afectan más a la variable endógena. Así pues, podemos estar interesados en crear una variable ficticia (*dummy*) e introducirla en el modelo para evaluar el impacto producido de estación concreta en el modelo especificado. Si estamos interesados en ver las consecuencias de la estación s_i podemos definir una *dummy* de la siguiente manera:

El valor de s

s recoge la estacionalidad de una serie temporal e indica el número de estaciones dentro de un año natural. Así, si hablamos de meses $s = 12$ y por tanto si cada estación se denomina s_i, i tomará los valores desde 1 hasta 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. Modelo de efectos fijos

Cuando trabajamos con un panel de datos podemos estar interesados en especificar un modelo de regresión utilizando toda la información disponible ($N \bullet T$ observaciones) eliminando los efectos temporales y de discrepancia entre las unidades analizadas. Por tanto, tenemos que definir $N - 1$ variables ficticias que recojan las unidades objeto de estudio y otras $T - 1$ variables ficticias que recojan los periodos de los que disponemos información e introducirlas de forma aditiva en nuestro modelo de regresión. Este nuevo modelo se suele denominar modelo de efectos fijos.

¿Qué es un panel de datos?

Un panel de datos no es más que una base de datos donde se recoge información sobre las características de N unidades objeto de estudio (individuos, hogares, empresas, etc.) durante T periodos distintos.

Así, en un modelo de efectos fijos, los parámetros estimados de las de *dummies* que recogen unidades y períodos se entienden como los desplazamientos sufridos en la recta de regresión como consecuencia a los efectos fijos de las variables no observables. El mayor inconveniente de trabajar con un panel de datos y realizar un modelos de efectos fijos es que incrementamos muchísimo el número de coeficientes y podemos correr el riesgo de quedarnos sin grados de libertad para poder estimarlos.

Observad que incluimos $N - 1 \bullet T - 1$ *dummies*, ya que en ambos casos conservamos una categoría de referencia, una unidad y un período.

5. Restricciones lineales en el modelo de regresión

Una vez tenemos un modelo de regresión lineal, es muy común que deseemos contrastar determinados supuestos alternativos postulados por alguna teoría. Dichos supuestos suelen ser mucho más complejos que la significación estadística de las variables, y por eso hemos de recurrir a la formulación general de las restricciones lineales. Un ejemplo sería contrastar si el valor de un parámetro determinado se corresponde a un valor específico, o también contrastar si una determinada relación entre parámetros se puede corroborar estadísticamente.

Formalmente, consideremos un conjunto de q restricciones lineales:

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

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

$$\ldots$$

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

De forma compacta, estas restricciones lineales se pueden expresar en una única ecuación:

$$R\beta = r$$

La matriz R tiene una dimensión $q \times k$, donde q es el número de restricciones lineales que contrastar y k el número de parámetros del modelo. Por tanto, la matriz R tiene tantas columnas como parámetros del modelo y tantas filas como restricciones que contrastar. El vector r tiene una dimensión $q \times 1$, y está formado por los términos independientes de las restricciones lineales. Así pues, al igual que la matriz R, el vector r tendrá tantas filas como restricciones.

El vector β...

... recordemos que es una columna de dimensión $k \times 1$, es decir, tiene una sola columna y tantas filas como parámetros.

Una vez especificadas las restricciones, procedemos a la realización del contraste de hipótesis sobre los parámetros del modelo. El contraste que plantear toma la siguiente forma:

$$H_0 : R\beta = r$$

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

Esto es, bajo la hipótesis nula no se pueden rechazar las restricciones impuestas. El estadístico del contraste F_0 se distribuye según una distribución F de Fischer, donde el valor crítico vendrá definido por $F_{q,N-k;\,\alpha}$, siendo α el nivel de significación del contraste. La decisión del contraste seguirá el siguiente esquema:

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

Veamos un ejemplo derivado del caso planteado en el capítulo 3. Supongamos que una teoría establece que el parámetro β_5, asociado a la variable *UNIV* (porcentaje de estudiantes universitarios), tiene un valor de $-0,5$, y queremos contrastar esta teoría en nuestra estimación. Formalmente, la restricción que contrastar toma la siguiente forma:

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

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

Por tanto, el constaste que debemos realizar es:

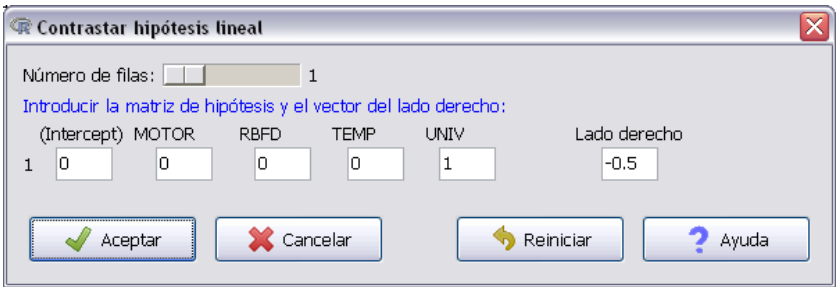
$$H_0 : \beta_5 = -0,5$$

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

Una vez hemos seleccionado como modelo activo *RegModel.2*, que fue el primer modelo que estimamos en el capítulo 3, para realizar el contraste con R-Commander acudimos a la siguiente ruta:

Modelos / Test de hipótesis / Hipótesis lineal

Aparecerá un cuadro de diálogo, donde introduciremos las restricciones R y r como se muestra a continuación:



El resultado es el siguiente:

```
> .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
```

Como vemos, el *p-valor* es prácticamente cero. De esta manera, con cualquier nivel de confianza rechazamos la hipótesis nula (H_0), es decir, podemos afirmar que la restricción $\beta_5 = -0,5$ no es cierta, al menos con estos datos y este modelo.

Introduzcamos ahora una restricción un poco más compleja. Supongamos que deseamos introducir dos restricciones: $\beta_5 = -0,3$ y $\beta_3 = -\beta_4$ (o lo que es lo mismo, $\beta_3 + \beta_4 = 0$). Formalmente tenemos las siguientes restricciones:

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

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

Específicamente:

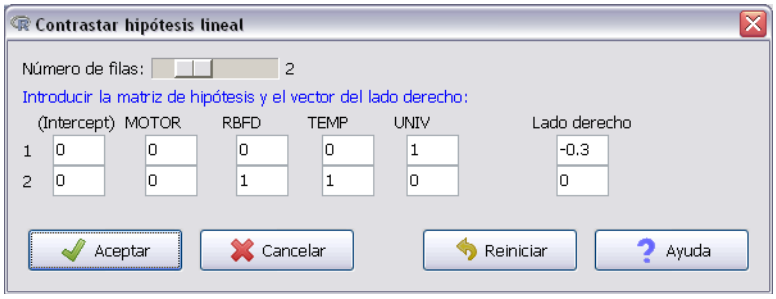
$$1 \cdot \beta_5 = -0,3$$

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

Al igual que en el caso anterior, accedemos a:

Modelos / Test de hipótesis / Hipótesis lineal

pero en este caso seleccionamos $q = 2$ filas, que equivale al número de restricciones. Además, introducimos las restricciones igual que en el caso anterior.



En este caso, observamos que el estadístico de prueba cae en la región de no rechazo de H_0, es decir, no podemos rechazar las dos restricciones introducidas. Como consecuencia, con estos datos y este modelo un incremento de un punto porcentual de universitarios (UNIV) hace reducir el valor esperado de la tasa de paro (PARO) en 0,3 puntos. Por otro lado, la renta familiar bruta disponible (RFBD) y la tasa de temporalidad (TEMP) tienen el mismo efecto en el valor esperado de la tasa de paro (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
```

Supongamos ahora que queremos estimar un modelo que incluya la restricción $\beta_3 = -\beta_4$. Haciendo una sencilla transformación algebraica, el modelo queda de la siguiente manera:

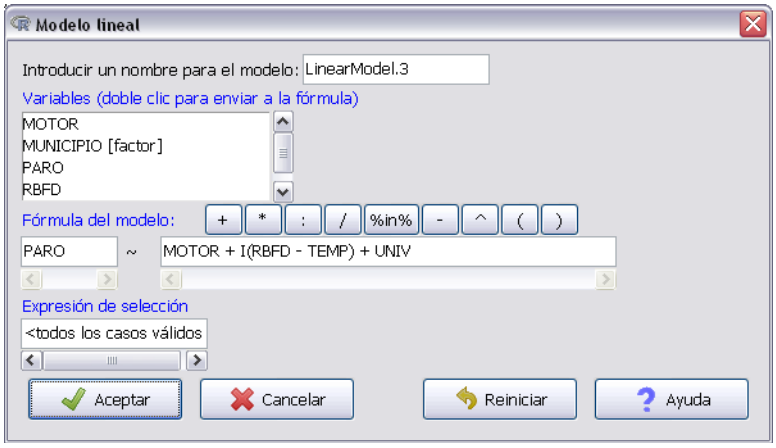
$$PARO_i = \beta_1 + \beta_2 MOTOR_i + \beta_3 RBFD_i + \beta_4 TEMP_i + \beta_5 UNIV_i + u_i =$$

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

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

Para efectuar esta estimación tenemos dos opciones: la más laboriosa es crear una nueva variable $RBFD_i - TEMP_i$, introducirla en el conjunto de datos y efectuar la estimación. Sin embargo, si solo estamos interesados en la estimación y no en la variable en sí misma, la podemos introducir en la especificación mediante el operador `I()`, como se muestra a continuación:

Estadísticos / Ajuste de modelos / Modelo lineal



Recordemos que el operador I() ya lo habíamos utilizado en el capítulo anterior cuando incorporábamos variables ficticias de forma multiplicativa en un modelo de regresión.

El resultado de la estimación es el siguiente:

```
> 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
```

Fijaos en que hemos vuelto a dar el mismo nombre al último modelo que habíamos estimado en el capítulo 3 del presente módulo. Así pues, como R es un programa orientado a la asignación de objetos, cuando damos el mismo nombre a un objeto existente, lo estamos sobreescribiendo. Es decir, la próxima vez que queramos utilizar el objeto LinearModel.3 recuperaremos este último modelo y habremos perdido el creado anteriormente.

Bibliografía

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

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