

Modelos de regresión y análisis multivariante con R-Commander

Daniel Liviano Solís

Maria Pujol Jover

PID_00208275

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea éste eléctrico, químico, mecánico, óptico, grabación, fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del copyright.

Índice

Introducción	5
Objetivos	6
1. Modelos de regresión	7
1.1. Introducción	7
1.2. Modelo de regresión lineal simple (MRLS)	7
1.3. Modelo de regresión lineal múltiple (MRLM)	15
2. Análisis de la varianza (ANOVA) y tablas de contingencia	20
2.1. Análisis de la varianza (ANOVA)	20
2.2. Tablas de contingencia	24
3. Análisis de componentes principales y análisis clúster	29
3.1. Introducción	29
3.2. Análisis de componentes principales (ACP)	29
3.3. Análisis clúster	35
Bibliografía	39

Introducción

Hasta ahora, hemos visto cómo se utilizaban R y R-Commander para llevar a cabo un análisis univariante, es decir, de una sola variable. En este módulo, emplearemos R y R-Commander para el análisis de dos o más variables (bivariante y multivariante, respectivamente).

A grandes rasgos, las técnicas de análisis multivariante en sentido estricto, es decir, de más de una variable, se pueden dividir atendiendo a distintos criterios. La combinación de estos criterios es precisamente lo que hará que elijamos la técnica adecuada para resolver un problema.

- **Relación entre las variables:** puede haber una relación de dependencia o bien una interrelación entre las distintas variables con las que estemos trabajando.
- **Objetivo del estudio:** según el problema planteado, estaremos interesados en reducir variables agrupando aquellas que expliquen conceptos similares o que ayuden a explicar un mismo concepto; o bien en agrupar a individuos u observaciones con características similares.
- **Número de variables que intervienen en el análisis:** podemos trabajar únicamente con dos variables (análisis bivariante) o más (análisis multivariante).
- **Naturaleza de las variables:** las variables utilizadas pueden ser métricas (numéricas discretas o continuas, dicotómicas, procedentes de escalas de Likert, cualitativas ordinales) o categóricas.

El análisis bivariante no es más que un caso particular del análisis multivariante.



Así pues, se puede deducir que encontramos múltiples técnicas que contemplan el análisis de más de una variable, aunque en este módulo solo trataremos algunas de las mismas. En concreto, las que veremos las hemos dividido en tres apartados: en el primero se trabajan los modelos de regresión; en el segundo, el análisis de la varianza y las tablas de contingencia; y en el tercero, el análisis de componentes principales y el análisis clúster.

Objetivos

1. Estimar e interpretar los resultados que arroja R-Commander cuando llevamos a cabo un modelo de regresión lineal simple (MRLS) o un modelo de regresión lineal múltiple (MRLM).
2. Utilizar R-Commander para llevar a cabo un análisis de la varianza (ANOVA) que no es más que un contraste de igualdad de tres o más medias poblacionales ($\mu_1, \mu_2, \dots, \mu_k$).
3. Obtener tablas de contingencia (TC) o efectuar el contraste de la χ^2 con R-Commander.
4. Calcular e interpretar todos los resultados obtenidos con R-Commander de un análisis factorial de componentes principales (AFCP), incluyendo el gráfico de bivariente.
5. Calcular e interpretar todos los resultados obtenidos con R-Commander de un análisis clúster o de conglomerados, incluyendo el dendrograma.

1. Modelos de regresión

1.1. Introducción

En este apartado desarrollamos el modelo de regresión, en el que estudiamos la relación que se establece entre dos o más variables. En este punto, conviene hacer una distinción fundamental entre dos conceptos relacionados pero distintos entre sí.

- **Correlación:** este concepto hace referencia al grado de relación que hay entre dos variables, pero no establece ningún tipo de relación de causa ni efecto de una sobre otra. El indicador de correlación más simple es el coeficiente de correlación lineal de Pearson, el cual indica en qué medida la relación lineal entre dos variables es directa, inversa o nula.
- **Modelo de regresión:** al hacer este modelo, estamos suponiendo que no solo hay correlación entre las variables, sino que además se produce una relación de causalidad, es decir, una o más variables influyen en otra.

Encontramos dos tipos de modelos de regresión.

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

Para más información sobre el MRLM, podéis ver el material asociado a la asignatura *Econometría*.



1.2. 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 la 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:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1, \dots, N.$$

En esta ecuación, α es la constante, β es la pendiente y ε es una variable aleatoria denominada término de error o de perturbación. Al ser una ecuación teórica que engloba

La variable explicada

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

La variable explicativa

Esta variable se puede denominar de diferentes maneras alternativas: exógena, independiente o regresor.

a toda la población, los parámetros α y β son desconocidos. Así pues, el objetivo de la estimación del modelo será llevar a cabo inferencia sobre el mismo. Por tanto, el primer paso consistirá en 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 expresión siguiente.

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

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 manera siguiente:

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

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 pues, cuanto menores sean los e_i mejor será la estimación del modelo y más fiables serán las predicciones sobre el comportamiento de Y obtenidas con esta estimación.

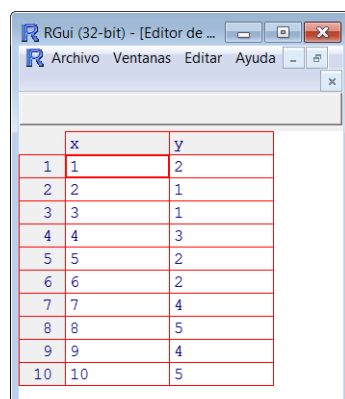
Como ejemplo, supongamos que disponemos de $N = 10$ observaciones de las variables X e Y :

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

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

Como vimos en el módulo dedicado al análisis descriptivo, es recomendable iniciar el análisis con estadísticos básicos de las variables. Una primera explotación estadística se obtiene siguiendo la ruta:

Estadísticos / Resúmenes / Conjunto de datos activo

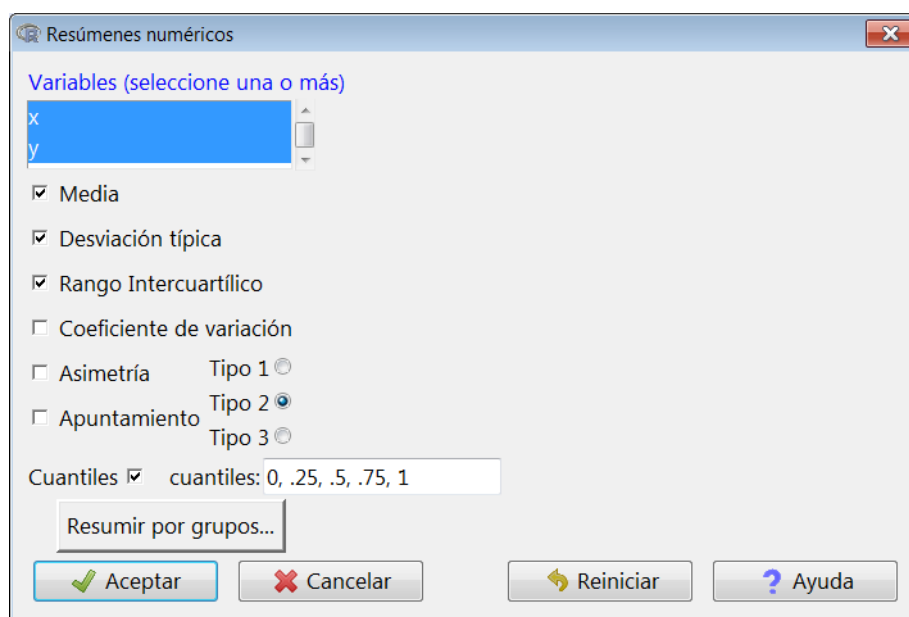
Con esto, el resultado será el siguiente:

```
> 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 esto, encontramos una opción en la que se puede elegir entre un conjunto de estadísticos. Para acceder a esta opción, la ruta que hay que seguir será esta:

Estadísticos / Resúmenes / Resúmenes numéricos

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



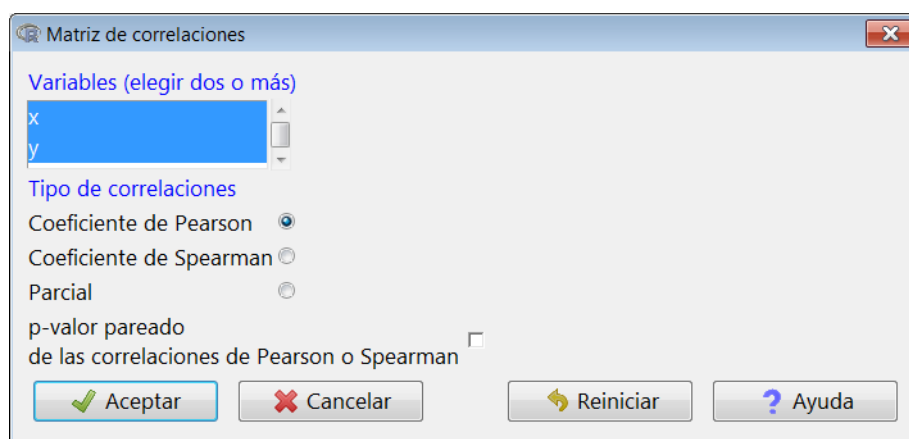
Este es el resultado que aparece en la ventana de resultados:

```
> 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 esta ruta:

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:



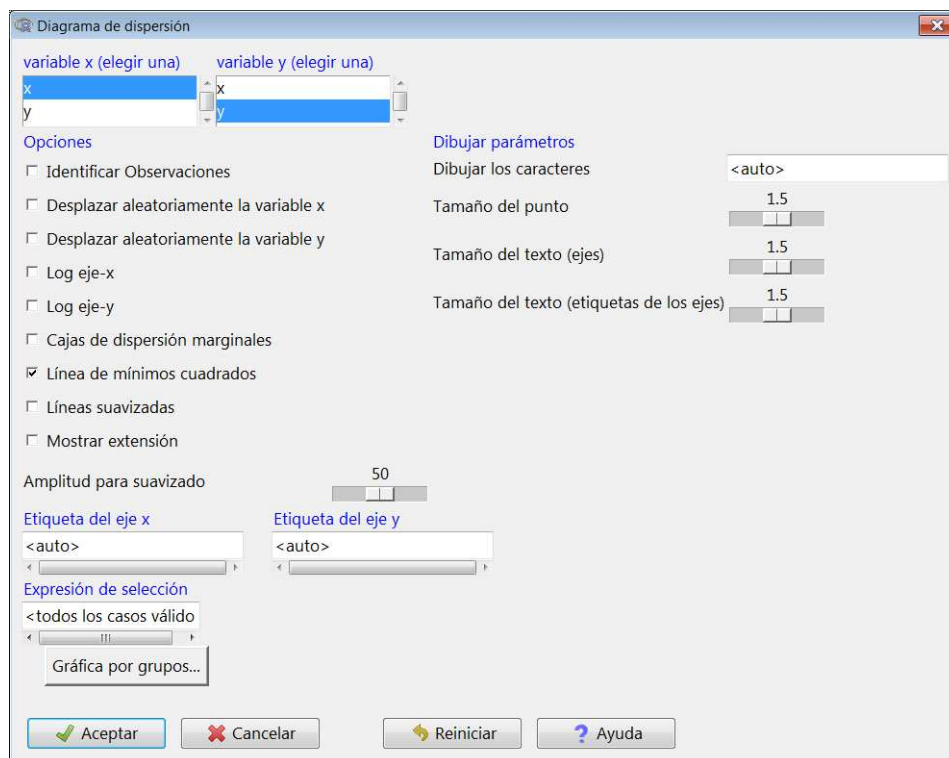
Como vemos, la correlación entre las dos variables es positiva y bastante elevada:

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

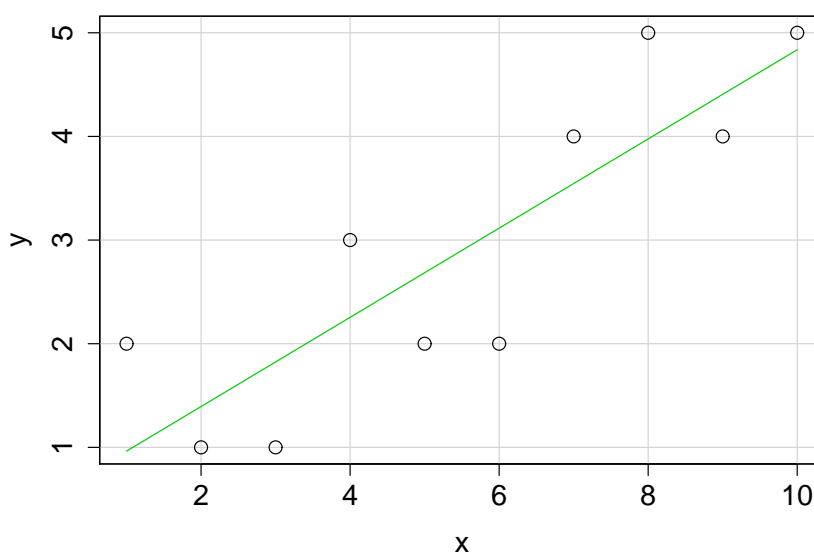
Visualmente, la correlación entre dos variables se puede comprobar mediante un diagrama de dispersión de las variables X e Y . Obtener este gráfico en R-Commander es algo inmediato si accedemos a esta ruta:

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

Observad 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 pleno 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 hay que 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 algo inmediato. Basta con tener en cuenta los operadores descritos en la tabla 1, ya vistos en el primer módulo.

Tabla 1. Operadores estadísticos básicos con R

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

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

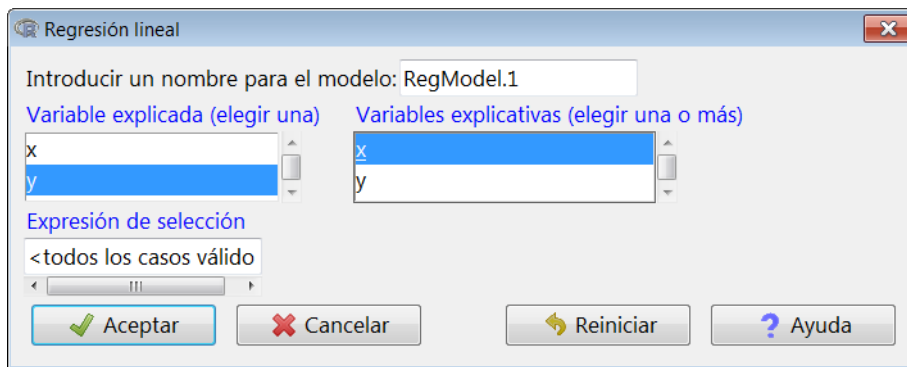
```
> attach(Datos)
> beta <- cov(x,y)/var(x)
> alpha <- mean(y)-beta*mean(x)
> coef.det <- cor(x,y)^2

> 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 las variables X e Y han sido introducidas, hay que estimar un modelo. La manera más sencilla consiste en seguir esta 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 (*Reg-Model.1*):



El resultado aparecerá en la ventana de resultados:

```
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 resumen de la regresión. Veamos sus principales componentes.

- **Residuals.** Mínimo, máximo y cuartiles de los residuos de la regresión, los cuales 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 hacer el contraste de significación individual de los parámetros estimados.
- **Pr(> |t|).** p -valor del contraste de significación individual de cada parámetro estimado, el cual 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.
- **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 1 % (p -valor < 0,01).

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 %), hay una probabilidad del 95 % de que, por ejemplo, el parámetro β esté incluido en el intervalo siguiente:

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

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

Modelos / Intervalos de confianza

Aparecerá un cuadro de diálogo en el que hay que especificar el nivel de confianza deseado y, tras pulsar *Aceptar*, obtendremos este 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 distinto de cero. Vemos que esto no ocurre en el caso de la pendiente.

1.3. Modelo de regresión lineal múltiple (MRLM)

El MRLM es una generalización del modelo simple a k variables explicativas y $k + 1$ parámetros (incluyendo la constante). Por tanto, la variable endógena se explica por más de una variable exógena:

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

Consideremos un ejemplo práctico. Se desea estudiar un modelo de regresión lineal para estudiar los determinantes del nivel de paro en $N = 295$ municipios catalanes:

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

Donde tenemos estas 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:



	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 efectuar 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

Para ver más estadísticos de las variables, se puede seguir esta ruta y seleccionar entre una lista de estadísticos:

Estadísticos / Resúmenes / Resúmenes numéricos

El resultado obtenido es el siguiente:

```
> numSummary(Datos[,c("MOTOR", "PARO", "RBF", "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
RBF	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 sí una alta correlación, puede ser problemático incluirlas de manera simultánea como variables explicativas. Por esto mismo, resulta muy útil calcular la matriz de correlaciones lineales de las variables explicativas:

En concreto, como se verá en la asignatura *Econometría*, una alta correlación entre dos regresores puede dar lugar a problemas de multicolinealidad.

Estadísticos / Resúmenes / Matriz de correlaciones

Si seleccionamos las variables que queremos incluir, obtenemos el resultado siguiente:

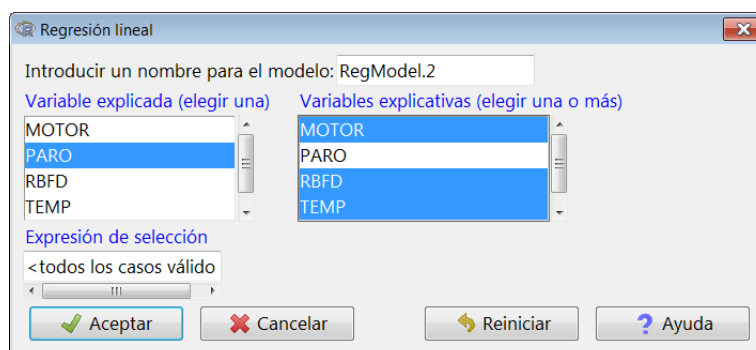
```
> cor(Datos[,c("MOTOR", "PARO", "RBF", "TEMP", "UNIV")],
+ use="complete.obs")
```

	MOTOR	PARO	RBF	TEMP	UNIV
MOTOR	1.00000000	-0.1292680	-0.01265723	-0.06906589	0.01746496
PARO	-0.12926800	1.00000000	-0.41906630	0.16408409	-0.46244755
RBF	-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

De manera análoga al caso del MRLS, esta ruta nos permitirá estimar un modelo de regresión, seleccionando las variables explicadas y explicativas:

Estadísticos / Ajuste de modelos / Regresión lineal

En el cuadro de diálogo resultante, introducimos las variables explicadas 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:

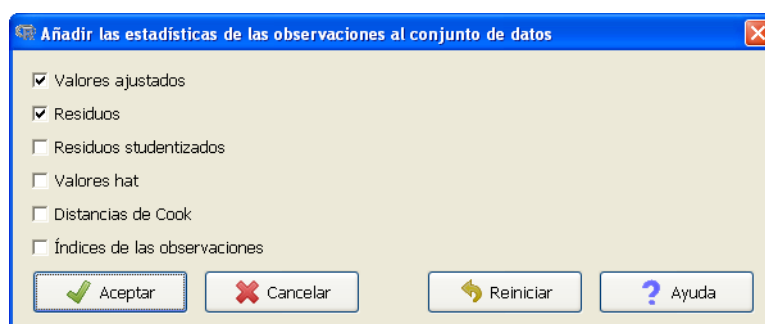
```
> Confint(RegModel.2, level=0.95)
              Estimate      2.5 %      97.5 %
(Intercept) 19.244961330 15.823732809 22.666189851
MOTOR       -0.007755427 -0.014242476 -0.001268379
RBFD        -0.037110206 -0.058139257 -0.016081156
TEMP         0.030971082  0.005919276  0.056022888
UNIV        -0.281595276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

Recordad que para obtener los IC de los parámetros tenemos que seleccionar el modelo, seguir la ruta *Modelos / Intervalos de confianza* y elegir 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 a nuestro conjunto de datos los residuos y los valores ajustados a la recta. Para esto, los activaremos en el cuadro de diálogo siguiente:



Una vez hecho esto, si visualizamos nuestro conjunto de datos, observaremos cómo se han añadido estas dos variables:

	MUNICIPIO	PARO	TEMP	UNIV	RBFD	MOTOR	fitted.RegModel.1	residuals.RegModel.1
1	Abdera	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	Àmetlla 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:

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

Para calcular 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 transformadas (aplicar logaritmos a una variable, elevarla al cuadrado, etc.) o multiplicadas entre sí; o incluso podemos seleccionar una muestra de nuestra base de datos. Es decir, la solución consiste en estimar directamente un modelo mediante la ruta alternativa siguiente:

Estadísticos / Ajuste de modelos / Modelo lineal

Nos aparecerá el siguiente 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, asignaremos a este modelo el nombre *LinearModel.3*:

Modelo lineal

Introducir un nombre para el modelo: LinearModel.3

Variables (doble clic para enviar a la fórmula)

MOTOR
MUNICIPIO [factor]
PARO
RBFD

Fórmula del modelo: + * : / %in% - ^ ()

PARO ~ log(MOTOR) + RBFD + TEMP + UNIV

Expresión de selección

< todos los casos válido

Aceptar Cancelar Reiniciar Ayuda

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
```

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 los cuales 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 este mismo motivo, el valor ajustado de R^2 incluye una penalización por el número de regresores que el modelo contiene.

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

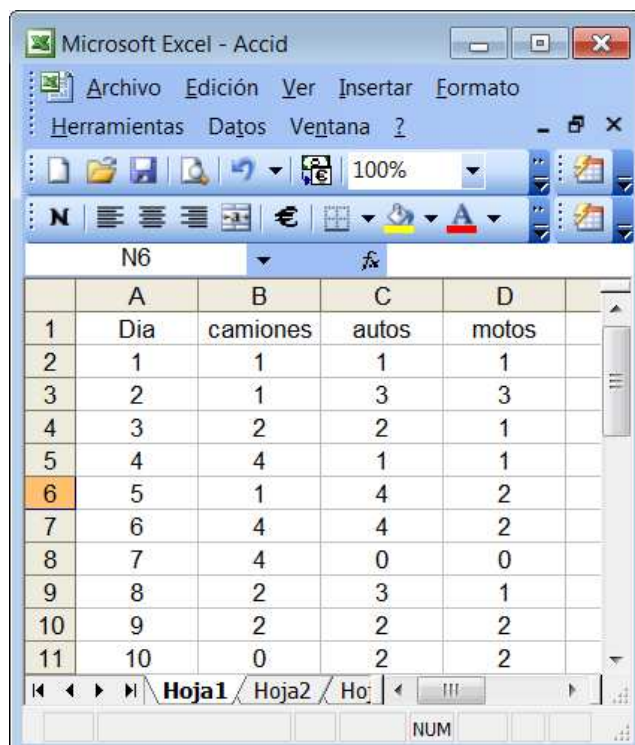


2. Análisis de la varianza (ANOVA) y tablas de contingencia

Este apartado está dedicado al análisis de la varianza (ANOVA) y al estudio de tablas de contingencia (TC). El análisis ANOVA divide la variación observada o dispersión de una determinada variable en componentes atribuibles a distintas fuentes de variación. En su forma más simple, un ANOVA proporciona una prueba estadística para contrastar si las medias de varios grupos son iguales o no. Por su parte, las TC acostumbra a analizar la distribución de una base de datos en función de dos variables.

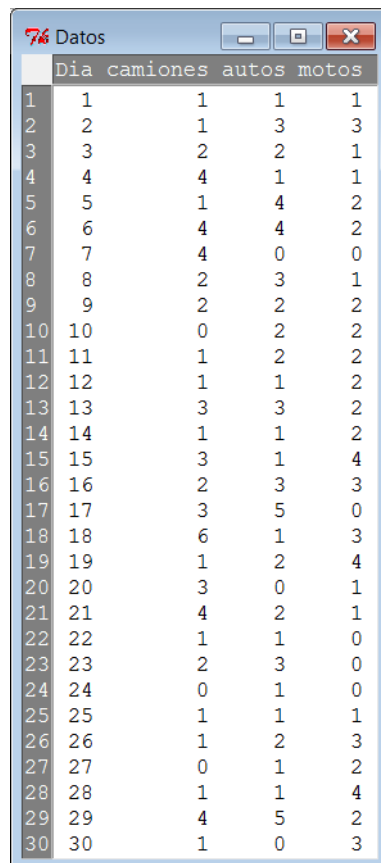
2.1. Análisis de la varianza (ANOVA)

Para llevar a cabo un análisis empírico, consideraremos una base de datos ficticia sobre el número de accidentes diarios durante un mes por diferentes carreteras y tipos de vehículo (C = camiones, A = automóviles, M = motocicletas). Inicialmente, los datos están disponibles en un archivo en formato de Microsoft Excel (extensión *xls*):



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

Una vez cargados los datos, comprobamos que se han cargado correctamente mediante la visualización del conjunto de datos activo:

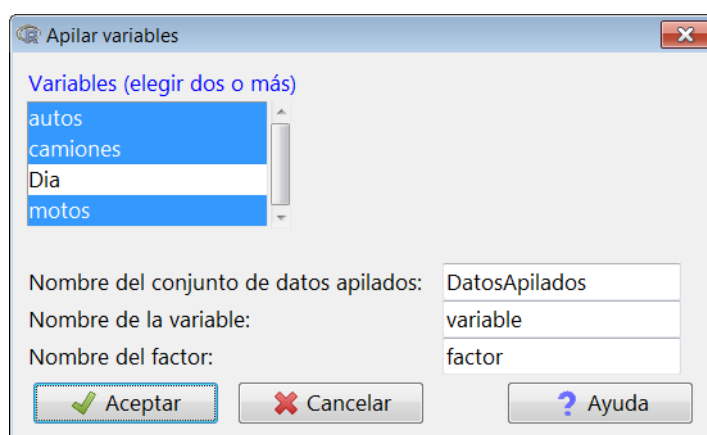


	Dia	camiones	autos	motos
1	1	1	1	1
2	2	1	3	3
3	3	2	2	1
4	4	4	1	1
5	5	1	4	2
6	6	4	4	2
7	7	4	0	0
8	8	2	3	1
9	9	2	2	2
10	10	0	2	2
11	11	1	2	2
12	12	1	1	2
13	13	3	3	2
14	14	1	1	2
15	15	3	1	4
16	16	2	3	3
17	17	3	5	0
18	18	6	1	3
19	19	1	2	4
20	20	3	0	1
21	21	4	2	1
22	22	1	1	0
23	23	2	3	0
24	24	0	1	0
25	25	1	1	1
26	26	1	2	3
27	27	0	1	2
28	28	1	1	4
29	29	4	5	2
30	30	1	0	3

El primer paso para hacer el análisis que nos proponemos es crear una variable apilada. Es decir, de tres variables que tenemos (cada una con 30 observaciones), creamos una sola con $30 \times 3 = 90$ observaciones, con una variable cualitativa asociada que indique el tipo de vehículo. Para hacer esto, hay que seguir la ruta siguiente:

Datos / Conjunto de datos activo / Apilar variables en el conjunto de datos activo

Al acceder a esta ruta, nos aparecerá un cuadro de diálogo en el que tenemos que introducir qué variables deseamos apilar. Además, introduciremos el nombre de la variable apilada y el nombre de la variable factor que define, para cada observación, el tipo de vehículo:



Apilar variables

Variables (elegir dos o más)

- autos
- camiones
- Dia
- motos

Nombre del conjunto de datos apilados: DatosApilados

Nombre de la variable: variable

Nombre del factor: factor

Una vez hecho esto, habremos creado también un nuevo conjunto de datos, que por defecto tiene el nombre de *DatosApilados*, aunque podemos cambiarlo si lo deseamos. Si ahora vamos a la opción *Visualizar conjunto de datos*, veremos que tiene dos variables de 90 observaciones: una numérica (número de vehículos) y un factor (tipo de vehículo).

	variable	factor
1	1	autos
2	3	autos
3	2	autos
4	1	autos
5	4	autos
6	4	autos
7	0	autos
8	3	autos
9	2	autos
10	2	autos

El ANOVA permite, entre otras cosas, comparar las medias de varios grupos. En nuestro caso, definiremos los tipos de vehículos de esta manera: $V_1 = C$, $V_2 = A$ y $V_3 = M$. El objetivo es llevar a cabo el siguiente contraste de hipótesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : H_0 \text{ no es cierta}$$

Es decir, ¿estadísticamente, la media poblacional de accidentes (μ) de autos, motos y camiones es la misma? Para hacer este contraste, de manera previa debemos definir el número de grupos, $k = 3$ grupos, y el tamaño de cada grupo, que será $n_1 = n_2 = n_3 = 30$, y $N = k \cdot 30 = 90$.

Este test se basa en descomponer la suma total de cuadrados (SCT) en dos partes: la dispersión explicada mediante los grupos (SCE) y la dispersión explicada por medio de otros factores distintos de los grupos en los que hemos dividido la población (SCD). Por lo tanto, siempre se cumplirá que:

$$SCT = SCE + SCD$$

Además, cuanto más se aproxime SCE a SCT , querrá decir que hay factores dentro de cada grupo que hacen que sean diferentes entre sí (medias diferenciadas). En cambio, un SCD proporcionalmente elevado significa que la variación de los datos se debe a factores externos no relacionados con los grupos.

El estadístico que hay que calcular y la distribución del mismo son:

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

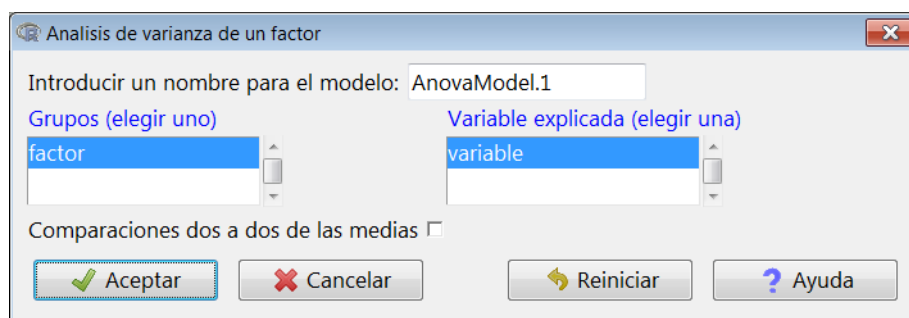
La decisión final se tomará siguiendo este criterio:

$$\begin{array}{ll} F^* > F_{\alpha, k-1, N-k} & \Rightarrow \text{Rechazo de } H_0 \\ F^* \leq F_{\alpha, k-1, N-k} & \Rightarrow \text{No rechazo de } H_0 \end{array}$$

Para aplicar este contraste con R-Commander, iremos al menú y accederemos a la ruta siguiente:

Estadísticas / Medias / ANOVA de un factor

Entonces nos aparecerá este cuadro de diálogo:



En este ejemplo, ya que solo hay una variable numérica y un factor, basta con pulsar en *Aceptar*, con lo que aparecerá el resultado que se muestra a continuación:

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

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

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

Primero aparece el resultado del contraste ANOVA, y a continuación la media y la desviación estándar de los tres tipos de vehículos. En cuanto al primer resultado, vemos que la dispersión explicada mediante los grupos es muy reducida ($SCE = 0,62$), mientras que la dispersión residual, es decir, explicada por otros factores distintos, es mayoritaria ($SCD = 160,67$). La suma de estas magnitudes es $SCT = 161,29$. El valor del estadístico F es de 0,168, y el *p-valor* es 0,845.

Resulta muy útil tener presente el criterio siguiente.

- Si el *p-valor* es inferior a 0,05 rechazamos la hipótesis nula, y podemos asegurar que la media del número de accidentes depende del tipo de vehículo.
- Si el *p-valor* es superior a 0,05 no rechazamos la hipótesis nula y, por tanto, no podemos asegurar que la media del número de accidentes dependa del tipo de vehículo.

En nuestro ejemplo, claramente no se rechaza la H_0 , y concluimos que no podemos asegurar que la media del número de accidentes dependa del tipo de vehículo.

2.2. Tablas de contingencia

En esta sección, veremos un ejemplo de una tabla de contingencia. Siguiendo con el ejemplo visto en el apartado anterior, para cada tipo de vehículo crearemos una variable dicotómica que tome valor 0 en caso de que el número de accidentes diarios sea inferior o igual a 2, y valor 1 en caso contrario. El objetivo es contrastar si esta variable dicotómica depende del tipo de vehículo, o si es independiente.

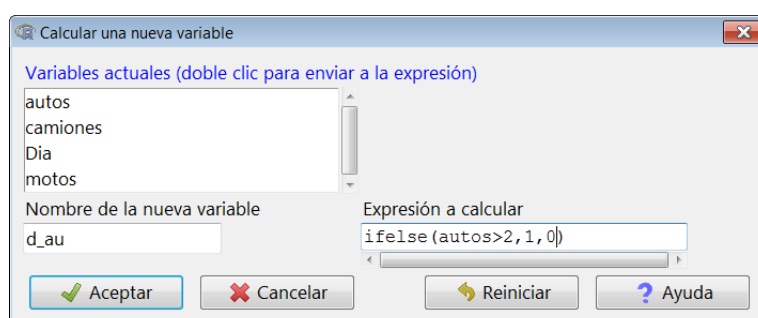
El primer paso consiste en crear tres variables dicotómicas (que solo pueden tomar los valores 0 y 1), tal que obtenemos $\{\tilde{C}_j, \tilde{A}_j, \tilde{M}_j; j = 1, \dots, n\}$. Estas variables se calculan de esta manera (se muestra solo el caso de la variable C ; el resto de los casos son análogos):

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

En R-Commander, el primer paso consistirá en crear estas tres variables dicotómicas, a las que denominaremos d_au , d_ca y d_mo . Previamente, tendremos que cambiar el conjunto de datos activo, y seleccionar el conjunto *Datos*, donde están las tres variables originales (es decir, ya no nos interesan las variables apiladas que hemos usado con el análisis ANOVA). Para crear variables, seguiremos esta ruta:

Datos / Medias / Manejar variables en el conjunto de datos activo

Aparecerá un cuadro de diálogo, en el que introduciremos el nombre de la nueva variable y su expresión (es decir, cómo se calcula). De manera similar a como se hace con la hoja de cálculo Excel, introduciremos:



La función *ifelse* tiene, en primer lugar, una expresión lógica (*autos* > 2), después el valor que tomará la nueva variable si esta condición se cumple (1) y, por último, el valor que tomará si es falsa (0). Esto lo haremos tres veces, una por cada variable. Una vez hecho esto, si visualizamos el conjunto de datos activo, comprobaremos que se han creado tres variables consistentes en unos y ceros. Son estas tres variables las que utilizaremos para hacer el análisis de contingencia.

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

Una vez tenemos estas variables calculadas, se trataría de analizar si son independientes o si dependen del tipo de vehículo. Entonces utilizaremos el test χ^2 de independencia, que permite comparar, a partir de una tabla de contingencia, las frecuencias observadas con las que, en teoría, deberían darse si hubiera independencia. En esta tabla, en las filas habrá dos categorías (G_1 y G_2), la primera por los valores uno y la segunda por los valores cero. Y en las columnas tendremos el tipo de vehículo, es decir, $V_1 = C$, $V_2 = A$ y $V_3 = M$. Además, en este caso se debe diferenciar el número de días por los que se tienen datos ($n = 30$) y el número total de observaciones, que es $N = 3 \cdot n = 90$. Entonces, las frecuencias observadas son:

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

En R-Commander, el paso siguiente consistirá en calcular G_1 y G_2 , es decir, la suma de unos y ceros para cada tipo de vehículo. Una manera de hacer esto consiste en introducir en la ventana de instrucciones las fórmulas siguientes:

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

Seleccionando estas tres instrucciones, si pulsamos en *Ejecutar* obtendremos el siguiente resultado en la ventana de resultados:

```
> sum(Datos$d_au)
[1] 9
> sum(Datos$d_ca)
[1] 10
> sum(Datos$d_mo)
[1] 8
```

Con esta información, teniendo en cuenta que hay 30 datos para cada tipo de vehículo, ya sabemos que la tabla de contingencia para analizar será esta: En caso de indepen-

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

dencia, las frecuencias se calculan a partir de la tabla anterior, aplicando esta fórmula:

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

La cuestión consiste en determinar si las diferencias $n_{ij} - n'_{ij}$ son aleatorias, por lo que hay independencia, o si por el contrario estas diferencias son demasiado grandes y se debe admitir que hay algún tipo de asociación entre las dos estratificaciones. Formalmente, se trata de contrastar las hipótesis:

H_0 : hay independencia.

H_1 : no hay independencia, y encontramos algún tipo de asociación.

La manera de proceder es muy similar a la del ejercicio anterior. El estadístico es, en este caso:

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

En nuestro caso, $L = 2$ es el número de dimensiones horizontales y $K = 3$ es el número de dimensiones verticales. El valor crítico del contraste con grados de libertad $\alpha = 0,05$ y $(L - 1) \cdot (K - 1) = 2$ es $\chi^2_{0,05,2} \simeq 5,99$. La resolución del contraste será, pues:

$$\begin{array}{ll} E > \chi^2_{0,05,2} & \Rightarrow \text{Rechazo de } H_0 \\ E \leq \chi^2_{0,05,2} & \Rightarrow \text{No rechazo de } H_0 \end{array}$$

Para hacer este análisis con R-Commander, iremos a la siguiente ruta del menú:

Estadísticas / Tablas de contingencias / Entrar e introducir una tabla de doble entrada

Aparecerá este cuadro de diálogo, en el que introduciremos nuestros datos:

Introducir una tabla de doble entrada

Número de filas: 2
Número de columnas: 3

Introducir las frecuencias:

	1	2	3
1	9	10	8
2	21	20	22

Calcular porcentajes

Porcentajes por filas ☐
 Porcentajes por columnas ☐
 Porcentajes totales ☒
 Sin porcentajes ☐

Test de hipótesis

☒ Test de independencia Chi-cuadrado
☒ Componentes del estadístico Chi-cuadrado
☒ Imprimir las frecuencias esperadas
☐ Test exacto de Fisher

Aceptar Cancelar Reiniciar Ayuda

Al pulsar en *Aceptar*, obtendremos el resultado siguiente:

```
> .Table <- matrix(c(9,10,8,21,20,22), 2, 3, byrow=TRUE)
> rownames(.Table) <- c('1', '2')
> colnames(.Table) <- c('1', '2', '3')
> .Table # Counts
  1  2  3
1  9 10  8
2 21 20 22
> totPercents(.Table) # Percentage of Total
      1      2      3 Total
1  10.0 11.1  8.9    30
2  23.3 22.2 24.4    70
Total 33.3 33.3 33.3   100
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test

Pearson's Chi-squared test

data: .Table
X-squared = 0.3175, df = 2, p-value = 0.8532
> .Test$expected # Expected Counts
  1  2  3
1  9  9  9
2 21 21 21
> round(.Test$residuals^2, 2) # Chi-square Components
  1  2  3
1 0 0.11 0.11
2 0 0.05 0.05
```

En este caso, el valor del estadístico es de 0,317, y con una confianza del 95 % el $p\text{-valor} > 0,05$ indica que la H_0 no se puede rechazar, es decir, que hay independencia y no encontramos relación entre que haya más de 2 accidentes y el tipo de vehículo.

3. Análisis de componentes principales y análisis clúster

3.1. Introducción

El modelo de regresión analizado en el primer apartado de este módulo asumía una relación de dependencia entre una variable dependiente y un grupo de regresores. En este apartado, sin embargo, analizaremos algunos métodos de interdependencia, es decir, aquellos en los que, *a priori*, no se asume relación alguna entre las diferentes variables que participan en un estudio.

El objetivo del análisis se basa en simplificar la compleja estructura de una base de datos. Esto lo podemos hacer eliminando aquellas variables menos representativas y agrupando las más relevantes, pero también se podría optar por agrupar observaciones con características similares entre sí.

Por una parte, entre los análisis de la primera opción (la reducción de variables), lo que más se utiliza es el análisis factorial que, según sea la naturaleza de los datos, se deriva en distintos análisis. Los más conocidos son:

- El análisis de componentes principales (ACP).
- El análisis de correspondencias simple (ACS).
- El análisis de correspondencias múltiple (ACM).

Por otra parte, entre los análisis basados en la agrupación de individuos destaca el análisis clúster o de conglomerados.

Finalmente, debemos dejar claro que la reducción de variables (columnas) y la agrupación de observaciones (filas) no son opciones excluyentes. En infinidad de ocasiones, estaremos interesados en hacer las dos cosas, reducir variables y agrupar individuos. Los ejemplos más claros los encontramos en el ámbito de la investigación de mercados, cuando llevamos a cabo una segmentación de clientes.

3.2. Análisis de componentes principales (ACP)

El ACP es un procedimiento estadístico que tiene como objetivo, mediante una transformación ortogonal, agrupar un conjunto de variables en uno o más componentes, de manera que la múltiple información contenida en estas variables quede sintetizada en menos dimensiones. El número de componentes principales es menor o igual al

Es importante denotar que los nuevos componentes o factores que obtengamos tras cualquier análisis factorial siempre estarán incorrelacionados.



número de variables originales. Esta transformación se define de tal modo que el primer componente principal tiene la mayor varianza posible (es decir, incluye la mayor proporción de la variabilidad posible de los datos), y cada componente subsiguiente, a su vez, tiene la mayor varianza posible en virtud de la restricción que sea ortogonal (es decir, no correlacionada con los componentes anteriores).

En el ACP, los valores propios de una matriz tienen un papel fundamental. Para verlo, supongamos que disponemos de k variables, cada una de estas con n observaciones: $\{X_1, X_2, \dots, X_n\}$. La solución ideal para el análisis es que cada variable incluyera un aspecto único y específico de cada individuo, con lo que todas las variables serían relevantes para el análisis y no habría ninguna variable redundante. Si esto fuera así, se cumpliría que $Cov(X_i, X_j) = 0$ y, como consecuencia, la $Cor(X_i, X_j) = 0, \forall i \neq j$. Es decir, la matriz de varianzas y covarianzas (MVC) tendría la siguiente forma¹:

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

Desde el punto de vista matemático, los k valores propios $\lambda_1, \dots, \lambda_k$ de una matriz cuadrada cumplen estas propiedades.

1) La suma de los valores propios **siempre** será igual a la suma de los elementos de la diagonal:

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

2) Si todos los elementos de la diagonal son iguales a cero, los valores propios serán iguales a los elementos de la diagonal en orden inverso:

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

Ahora bien, esto rara vez ocurre en la realidad, ya que casi siempre hay un grado de covarianza y correlación entre las variables.

Por este motivo, y especialmente cuando el grado de correlación entre algunas variables sea elevado, el uso de la técnica del ACP está más que justificado, ya que lo que hace es reducir las k variables a un número menor de **componentes principales (CP)**, entre los cuales no haya ninguna correlación (es decir, que sean ortogonales).

¹ De manera análoga, la matriz de correlaciones será la matriz identidad con dimensión $k \times k$.

Veamos un ejemplo empírico con R y R-Commander. Disponemos de una base de datos con $N = 100$ observaciones correspondientes a los alumnos de una escuela, y que contiene información que ha de ayudar a categorizar y estudiar mejor el comportamiento global de estos estudiantes². La información que se tiene para cada alumno se resume en las seis variables definidas en la tabla 2.

Tabla 2. Datos de rendimiento académico

Variable	Definición
n_mates	Nota en matemáticas
n_lengua	Nota en lengua
n_ingles	Nota en inglés
absent	Número de ausencias en clase
indis	Número de actos de indisciplina
penal	Número de penalizaciones recibidas

El primer paso consistirá en llevar a cabo la importación de la base de datos y, posteriormente, visualizar los resultados para comprobar que esta importación se ha hecho correctamente:

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

Visualizando los datos

Aunque la muestra contenga $N = 100$ observaciones, aquí solo mostramos las diez primeras por cuestión de espacio.

A continuación, es importante que hagamos un resumen del conjunto de datos que contiene información estadística básica:

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

² Esta base de datos es totalmente ficticia y se ha generado de manera artificial para ilustrar el concepto de ACP.

Tras habernos hecho una idea sobre la información que contienen las variables, deberemos estudiar la correlación que hay entre todas las mismas. Hasta ahora, hemos visto que la manera de hacer este cálculo con R-Commander consistía en seguir esta ruta y seleccionar las variables de interés:

Estadísticos / Matriz de correlaciones

Con lo que el resultado obtenido es el siguiente:

```
> cor(Data[,c("absent", "indis", "n_ingles", "n_lengua", "n_mates",
+ "penal")], use="complete.obs")
```

	absent	indis	n_ingles	n_lengua	n_mates	penal
absent	1.00	0.93	0.17	0.08	0.18	0.91
indis	0.93	1.00	0.20	0.12	0.20	0.95
n_ingles	0.17	0.20	1.00	0.93	0.96	0.18
n_lengua	0.08	0.12	0.93	1.00	0.93	0.12
n_mates	0.18	0.20	0.96	0.93	1.00	0.20
penal	0.91	0.95	0.18	0.12	0.20	1.00

Resultado simplificado

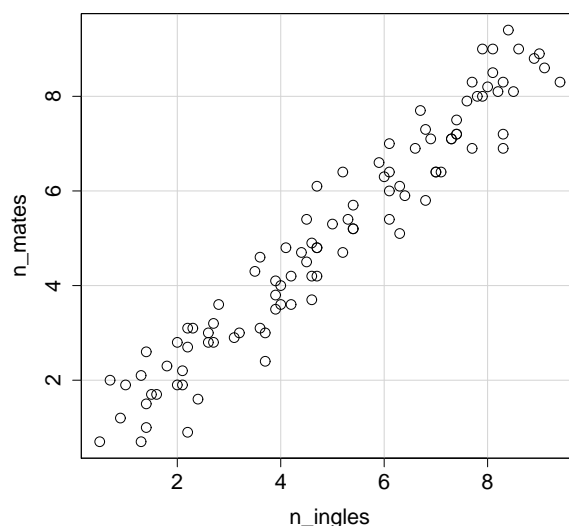
Por cuestión de espacio, hemos suprimido parte de los decimales de las correlaciones.

Una vez disponemos del código en R, una opción interesante es la de asignarle un nombre a esta matriz de correlaciones, cambiar el orden de las variables y redondear las cifras a dos dígitos:

```
> MC <- cor(Data[,c("n_ingles", "n_lengua", "n_mates", "penal",
+ "absent", "indis")], use="complete.obs")
> round(MC, digits=2)
```

	n_ingles	n_lengua	n_mates	penal	absent	indis
n_ingles	1.00	0.94	0.97	0.19	0.17	0.20
n_lengua	0.94	1.00	0.93	0.12	0.09	0.12
n_mates	0.97	0.93	1.00	0.20	0.19	0.21
penal	0.19	0.12	0.20	1.00	0.92	0.96
absent	0.17	0.09	0.19	0.92	1.00	0.93
indis	0.20	0.12	0.21	0.96	0.93	1.00

De manera intuitiva, esta matriz de correlaciones permite observar que, por una parte, las tres variables de notas están muy correlacionadas (por encima de 0,9) y, por otra parte, las tres variables relacionadas con el comportamiento también lo están. De algún modo, se puede afirmar que hay variables redundantes, es decir, que con menos variables se podría explicar lo mismo. Veamos un gráfico de dispersión de las variables nota de inglés y nota de matemáticas para obtener una evidencia visual:



Una manera sofisticada de calcular el nivel de correlación entre variables es el cálculo de los vectores propios de la matriz de correlaciones (MC):

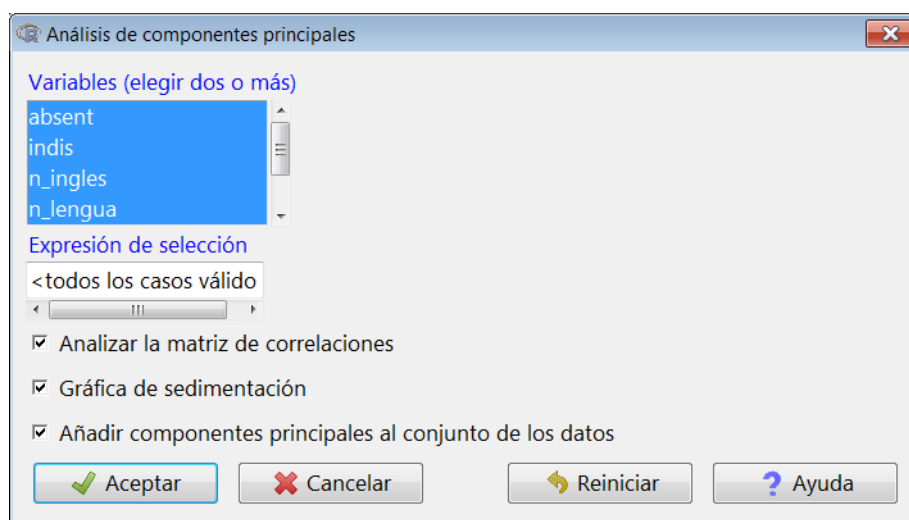
```
> eigen(MC)$values
[1] 3.383908 2.386493 0.092356 0.063587 0.043842 0.029810
```

El cálculo de los valores propios de la matriz de varianzas y covarianzas daría un resultado similar.

Como vemos, la magnitud de los dos primeros valores propios comparados con el resto es enorme. Esto se traduce en que las seis variables del estudio se pueden reducir a dos dimensiones, algo que constatará el ACP. Para llevar a cabo este análisis en R-Commander, hay que seguir esta ruta:

Estadísticos / Análisis dimensional / Análisis de componentes principales

Entonces nos aparecerá el siguiente cuadro de diálogo, donde activaremos las opciones disponibles.



Esto hace que obtengamos, por una parte, lo siguiente en la ventana de resultados.

1) *Component loadings*: es la matriz factorial, que muestra la correlación entre los componentes calculados y las variables objeto de estudio.

```
> .PC <-
+ princomp(~absent+indis+n_ingles+n_lengua+n_mates+penal,
+ cor=TRUE, data=Datos)

> unclass(loadings(.PC)) # component loadings
```

	Comp.1	Comp.2	Comp.3	Comp.4
absent	-0.3952811	0.4166928	0.7069336	-0.40148060
indis	-0.4115994	0.4072536	-0.2407877	0.23451814
n_ingles	-0.4213028	-0.3948581	0.2019751	0.32225209
n_lengua	-0.3881367	-0.4316741	-0.3552555	-0.71965433
n_mates	-0.4245466	-0.3886644	0.1744042	0.39043337
penal	-0.4073731	0.4088962	-0.4948220	0.09812031

```

      Comp.5      Comp.6
absent -0.09288941 0.02344405
indis   0.63503349 -0.38533804
n_ingles 0.32825831 0.64358148
n_lengua 0.09755816 -0.09685559
n_mates -0.43463287 -0.54492149
penal   -0.53096815 0.36117505
```

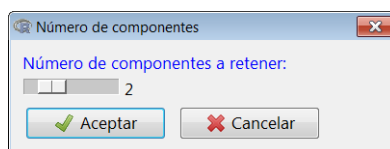
2) *Component variances*: las varianzas de los componentes, es decir, los valores propios de la matriz de correlaciones calculados anteriormente. Como vemos, la magnitud de los dos primeros representa casi el total de la suma de valores propios.

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

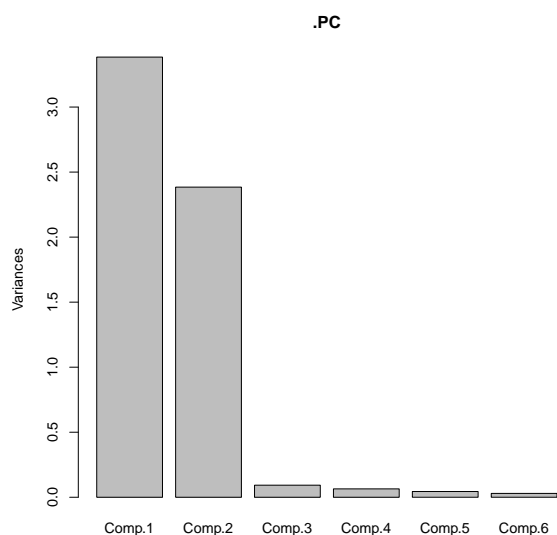
3) *Proportions of variance*: proporciones de la varianza (o variabilidad) total de los datos reunidas por los componentes. Como vemos, los dos primeros componentes recopilan el 56,4 % y el 39,7 % de la proporción de la varianza total, respectivamente. La última fila nos muestra que entre los dos componentes se reúne el 96,1 % de la variabilidad total de las seis variables.

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

Antes de pulsar *Aceptar*, en el cuadro de diálogo anterior nos habrá aparecido la opción de elegir cuántos componentes principales deseamos almacenar. Considerando el análisis de valores propios hecho anteriormente, seleccionamos dos componentes:



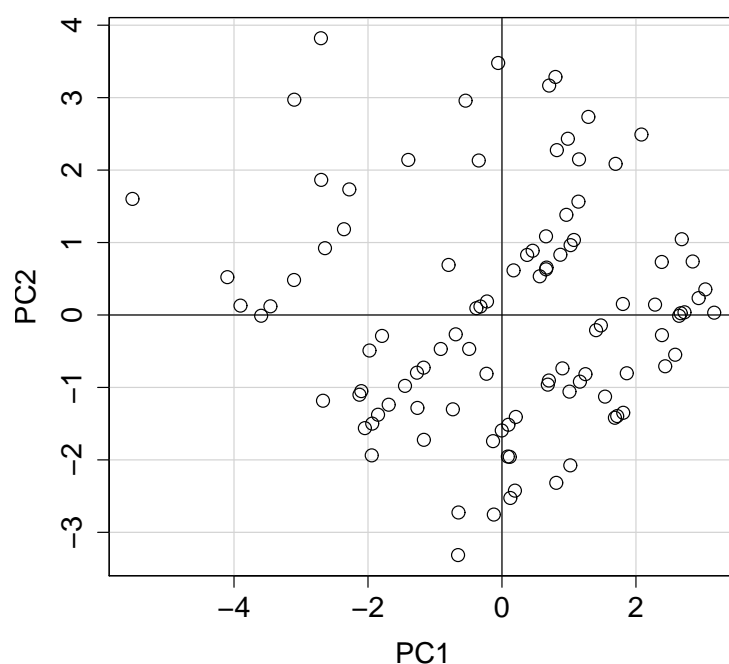
Por otra parte, en la consola de R aparecerá el gráfico con los valores propios de los componentes principales. El criterio de Kaiser establece que consideraremos aquellos componentes principales cuyo valor propio sea superior a uno. En este caso, es obvio que los dos primeros componentes bastan para explicar la variabilidad de los datos.



Al haber almacenado los dos componentes principales, si visualizamos otra vez nuestro conjunto de datos aparecerán estas dos nuevas variables.

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

Para tener una evidencia visual de cómo los dos componentes principales seleccionados explican la variabilidad de los datos, es recomendable hacer una gráfica de dispersión de los dos componentes y ver cómo se distribuyen las observaciones. Mediante este gráfico, se puede comprobar si hay alguna agrupación entre estudiantes a través de estas dos dimensiones, es decir, si hay diferentes grupos con combinaciones de buenas/malas notas y buen/mal comportamiento.



Visualizar los componentes principales

En inglés, este gráfico se denomina *scores plot*.

3.3. Análisis clúster

El análisis de conglomerados o clúster engloba una amplia gama de métodos numéricos que tienen como objetivo detectar grupos o conglomerados de observaciones homogéneos (muy similares entre sí) y heterogéneos (muy dispares entre grupos). Es decir, las observaciones han de estar muy juntas para formar parte de un mismo grupo

Es muy importante no confundir el ACP con el análisis clúster; el primero agrupa variables y el segundo agrupa a individuos.

y muy separadas para formar parte de distintos grupos. Así pues, los grupos o clústeres se identifican por la evaluación de las distancias relativas entre los puntos, y esto permite calcular la homogeneidad relativa de cada grupo y el grado de separación entre los distintos grupos.

Supongamos un ejemplo con datos ficticios. A partir de una encuesta, se pretende agrupar diferentes profesiones según la percepción que tienen sus trabajadores sobre distintos aspectos. Para medir esta percepción, se ha utilizado una escala de Likert; por tanto, la respuesta de los trabajadores va desde 0 (mínima satisfacción) hasta 10 (máxima satisfacción).

La puntuación media sobre las muestras de las distintas profesiones se incluye en la base de datos que se muestra a continuación:

	PROFESION	HORARIO	SUELDO	ESTRES	FAMILIA
1	Profesor	9	6	8	9
2	Ingeniero	8	9	6	8
3	Camarero	4	6	7	4
4	Abogado	7	9	4	3
5	Banquero	8	10	3	8
6	Cocinero	4	7	7	4
7	Taxista	3	5	7	3
8	Obrero	5	5	6	4

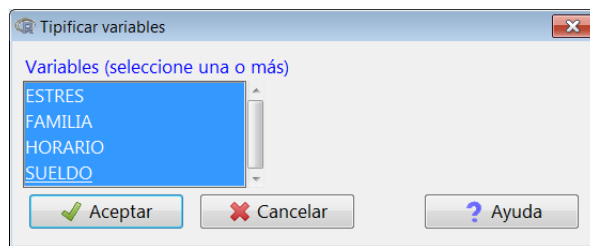
En este caso, el objetivo del análisis clúster será estudiar la distancia entre las diferentes profesiones a partir de las similitudes entre las cuatro variables de la encuesta. La función `dist` proporciona la matriz de distancias entre distintas variables, y la distancia euclidiana es la opción por defecto. La información entre paréntesis `[,2 : 5]` selecciona, del conjunto de datos, las columnas de la segunda a la quinta, que se corresponden con las variables que deseamos analizar.

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

Como vemos, los oficios 3, 6, 7 y 8 son los más similares según este criterio. Una herramienta visual adecuada para esto es el dendrograma, cuya función es mostrar la formación de conglomerados, así como las distancias entre los mismos. Antes de efectuar el dendrograma, es aconsejable tipificar las variables para que no haya efectos de escala. Para esto, como vimos en el módulo dedicado al análisis descriptivo, accedemos a la ruta siguiente:

Datos / Modificar variables del conjunto de datos activo / Tipificar variables

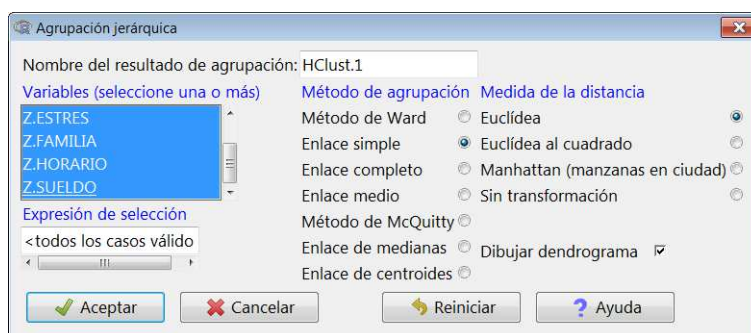
Nos aparecerá un menú en el que seleccionaremos las variables que hay que tipificar:



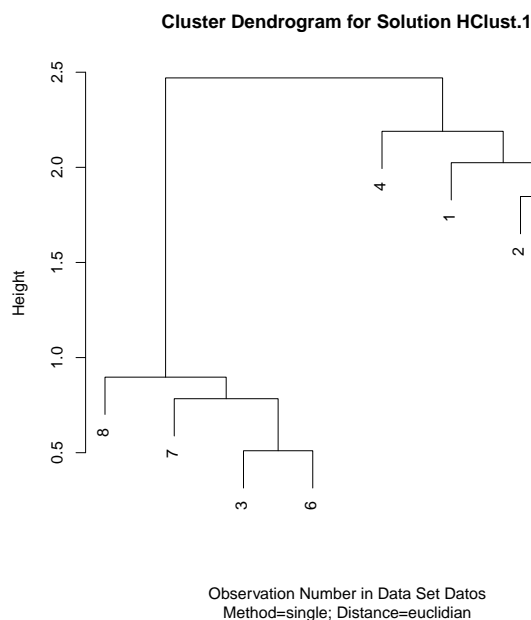
Una vez hecho esto, para llevar a cabo el dendrograma con R-Commander, hay que seguir la ruta siguiente.

Estadísticos / Análisis dimensional / Análisis de agrupación / Agrupación jerárquica

Nos aparecerá el siguiente cuadro de diálogo, en el que seleccionaremos las variables de interés, el método de agrupación, la medida de distancia y si queremos visualizar gráficamente el dendrograma.



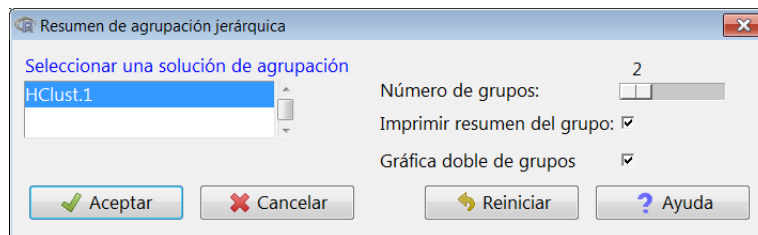
El resultado gráfico aparece a continuación, y en el mismo se puede observar que hay dos grandes grupos de oficios: a) profesor, ingeniero, abogado y banquero; y b) camarero, cocinero, taxista y obrero.



Podemos obtener más información de la agrupación jerárquica llevada a cabo si accedemos a la ruta siguiente:

Estadísticos / Análisis dimensional / Análisis de agrupación / Resumir la agrupación jerárquica

Nos aparecerá el siguiente cuadro de diálogo, en el que seleccionaremos el número de grupos en el que queramos organizar los datos (en nuestro caso, dos), además del resumen numérico y el gráfico asociado.

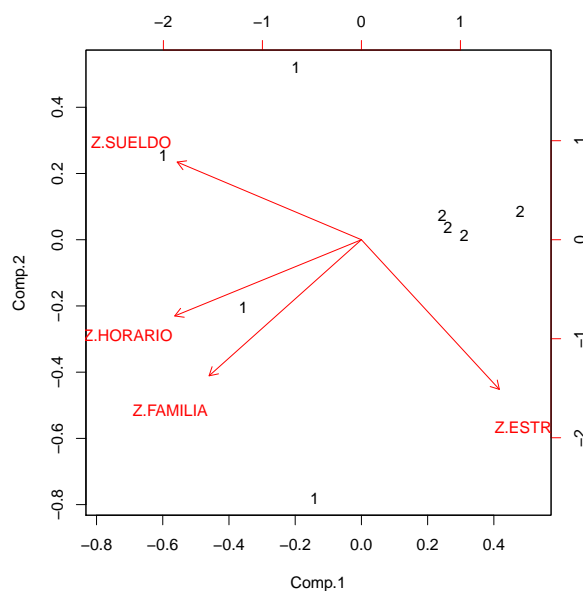


Por una parte, obtendremos este resumen numérico:

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

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

Y por otra parte obtendremos un gráfico, en el que podemos observar de manera visual las ocho observaciones (cuatro pertenecen a la agrupación 1 y otras cuatro, a la agrupación 2), en un plano cartesiano de componentes principales:



Bibliografía

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

