Regresión múltiple

El modelo de regresión múltiple es la extensión a *k* variables explicativas del modelo de regresión simple estudiado en el apartado anterior. En general, una variable de interés *y* depende de varias variables $x_{1,\ldots,}x_k$ y no sólo de una únicavariable de predicción *x*. Por ejemplo, para estudiar la variación del precio de una vivienda, parece razonable considerarmás de una variable explicativa, como pueden ser el precio del suelo, la superficie del piso, el número de cuartos de baño, la edad de la vivienda, etc. Además de las variables observables, la variablede interés puede depender de otras desconocidas para el investigador. Un modelo de regresión representa el efecto de estas variables en lo que se conoce como error aleatorio o perturbación.

Si suponemos un modelo de regresión teórico en el que las variables se pueden relacionar mediante una función de tipo lineal, éste puede escribirse

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \ ,$$

donde $\beta_0, \beta_1, \ldots, \beta_k$ son los parámetros desconocidos que vamos a estimar y ε es el error aleatorio o perturbación. *y* es la variable de interés que queremos predecir, también llamada variable respuesta o variable dependiente. Las variables $x_{1,\ldots,}x_k$ se llaman variables independientes, explicativas o de predicción. El error ε representa el efecto de todas las variables que pueden afectar a la variable dependiente y no están incluidas en el modelo de regresión.

Algunos ejemplos de modelos de regresión múltiple pueden ser:

- El consumo de combustible de un vehículo, cuya variación puede ser explicada por la velocidad media del mismo y por eltipo de carretera. Podemos incluir en el término de error, variables como el efecto del conductor, las condiciones meteorológicas, etc.

- El presupuesto de una universidad, cuya variación puede serexplicada por el número de alumnos. También podríamos considerar en el modelo variables como el número de profesores, el número de laboratorios, la superficie disponible de instalaciones, personal de administración, etc.

Si se desea explicar los valores de una variable aleatoria *y,* mediante *k* variables, que a su vez toman *n* valores, tenemos entonces

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \ , \quad i = 1, \ldots, n.$$

Las perturbaciones deben verificar las siguientes hipótesis:

- Su esperanza es cero
- Su varianza es constante
- Son independientes entre sí
- Su distribución es normal

Los parámetros desconocidos son estimados por mínimos cuadrados, resultando la ecuación estimada de regresión dada por

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki} + \varepsilon_i \ ,$$

donde cada coeficiente $\hat{\beta}_i$ representa el efecto sobre la respuesta cuando la variable aumenta en una unidad y las demás variables permanecen constantes. Puede interpretarse como el efecto diferencial de esta variable sobre la variable respuesta cuando controlamos los efectos de las otras variables. $\hat{\beta}_0$ es el valor de la respuesta ajustada cuando todas las variables explicativas toman el valor cero.

Descomposición de la variabilidad y contrastes de hipótesis

La variabilidad de la respuesta puede descomponerse de igual forma que en regresión simple

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Esta descomposición la notamos por:

$$SCT = SCE + SCR_{eg} \ ,$$

donde *SCT* es la suma de cuadrados total y representa la variabilidad total, SCR_{eg} es la suma de cuadrados de la regresión y representa la variabilidad explicada por el modelo de regresión. *SCE* es la suma de cuadrados residual y representa la variabilidad que queda sin explicar. Esta descomposición se resume en la siguiente tabla

Tabla ANOVA

Fuente de variación	Suma de cuadrados	g.l.	Cuadrados medios	F
Regresión	$B^t X^t Y^t - \frac{1}{n}\left(\sum y_i\right)^2$	$k = m-1$	$\frac{S.C.R.}{m-1}$	$F_{\exp} = \frac{\frac{S.C.R.}{m-1}}{\frac{S.C.E.}{n-m}}$
Error	$Y^t Y - B^t X^t Y$	$n-m$	$\frac{S.C.E}{n-m}$	
Total	$Y^t Y - \frac{1}{n}\left(\sum y_i\right)^2$	$n-1$		

El valor del estadístico $F_{\exp}$ permite resolver el contraste de regresión, dado por

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ H_1 : \beta_j \neq 0 \ \text{ para algún } \ j = 1, \ldots, k \end{cases}$$

Fijado un nivel de significación α se rechaza H_0 si $F_{\exp} > F_{\alpha,k,n-k-1}$. En la prácticaSPSS proporciona el *p-valor* o nivel mínimo designificación para el rechazo de H_0, que permite resolver elcontraste de hipótesis fijado un nivel de significación.

Si $p-valor < \alpha$, entonces se rechaza H_0
Si $p-valor \geq \alpha$, entonces no se rechaza H_0

Si estamos interesados en estudiar el efecto individual de una variable explicativa sobre la variable respuesta se considera el siguiente contraste

$$\begin{cases} \mathrm{H}_0 : \beta_i = 0 \\ \mathrm{H}_1 : \beta_i \neq 0 \end{cases}$$

En este caso el estadístico de contraste sigue una *F* de Snedecor con 1 y $n-k-1$ grados de libertad. Este contraste es equivalenteal contraste de regresión con una única variable explicativa, estudiado en el apartado anterior. El rechazo de la hipótesis nula implica admitir la validez de la variable explicativa x_i para predecir la variable de interés y.

Coeficiente de determinación

Para construir una medida descriptiva del ajuste global de un modelo de regresión se emplea el coeficiente de determinación, dado por

$$R^2 = \frac{S.C.R.}{S.C.T.} = 1 - \frac{S.C.E.}{S.C.T.}.$$

R^2 representa la proporción de variación de y explicada por el modelo de regresión. Por construcción, es evidente que $0 \leq R^2 \leq 1$.

- Si $R^2 = 1$ entonces $SCR_{eg}=SCT$, por lo que toda la variación de y es explicada por el modelo de regresión.
- Si $R^2 = 0$ entonces *SCT=SCE*, por loque toda la variación de y queda sin explicar.

En general, cuanto más próximo esté a 1, mayor es la variación de y explicada por el modelo de regresión.

Sin embargo, en regresión múltiple, el coeficiente dedeterminación presenta el inconveniente de que su valor aumentaal añadir nuevas variables al modelo de regresión, independientemente de que éstas contribuyan de forma significativaa la explicación de la variable respuesta. Para evitar un aumentoinjustificado de este coeficiente, se introduce el coeficiente dedeterminación corregido, que notamos por $\bar{R}^2$ y que se obtiene a partir de R^2 en la forma

$$\bar{R}^2 = 1 - \frac{\frac{\sum e_i^2}{n-k-1}}{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

Este coeficiente no aumenta su valor cuando se añadennuevas variables, sino que en caso de añadir variables superfluasal modelo, el valor de $\bar{R}^2$ disminuye considerablemente respecto al valor del coeficiente R^2 .

Ejercicio:

Una empresa fabricante de cereales para el desayuno desea conocer la ecuación que permita predecir las ventas (en miles de euros) en función de los gastos en publicidad infantil en televisión (en miles de euros), la inversión en publicidad en radio (en miles de euros) y la inversión en publicidad en los periódicos (en miles de euros). Se realiza un estudio en el que se reúnen los datos mensuales correspondientes a los últimos 20 meses. Estos datos aparecen en la siguiente tabla

Ventas	Pub. en tv	Pub. en radio	Pub. en per.
10,0	1,30	56	,40
12,0	1,40	55	,40
11,0	1,50	60	,42
13,0	1,70	65	,50
12,0	1,75	69	,40
14,0	1,30	67	,44
16,0	1,45	68	,40
12,0	,90	67	,44
14,0	,80	97	,46
11,0	,90	66	,46
10,0	,80	65	,45
19,0	1,00	60	1,10
8,5	1,70	70	,30
8,0	1,80	110	,50
9,0	1,85	75	,45
13,0	1,90	80	,40
16,0	2,00	85	,80
18,0	2,00	90	,90
20,0	1,30	56	,90
22,0	1,40	55	1,10

Se pide:

a) Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación
b) Contrastar la significación del modelo propuesto.
c) ¿Puede eliminarse alguna variable del modelo? Realiza los contrastes de significación individuales
d) Coeficiente de determinación y de determinación corregido

Solución:

a) Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación

Notamos ventas**, publ_tv**, **publ_rad** y **publ_per** las variables que intervienen en el ejercicio. La variable **ventas** es la variable dependiente, mientras que **publ_tv**, **publ_rad** y **publ_per** son las variables explicativas.

Introducimos dichas variables en la Vista de Variables de SPSS, como se muestra

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
1	ventas	Numérico	8	1		Ninguna	Ninguna	8	Derecha	Escala
2	publ_tv	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala
3	publ_rad	Numérico	8	0		Ninguna	Ninguna	8	Derecha	Escala
4	publ_per	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala

Ajustamos un modelo de regresión que responde a una expresión deltipo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \ ,$$

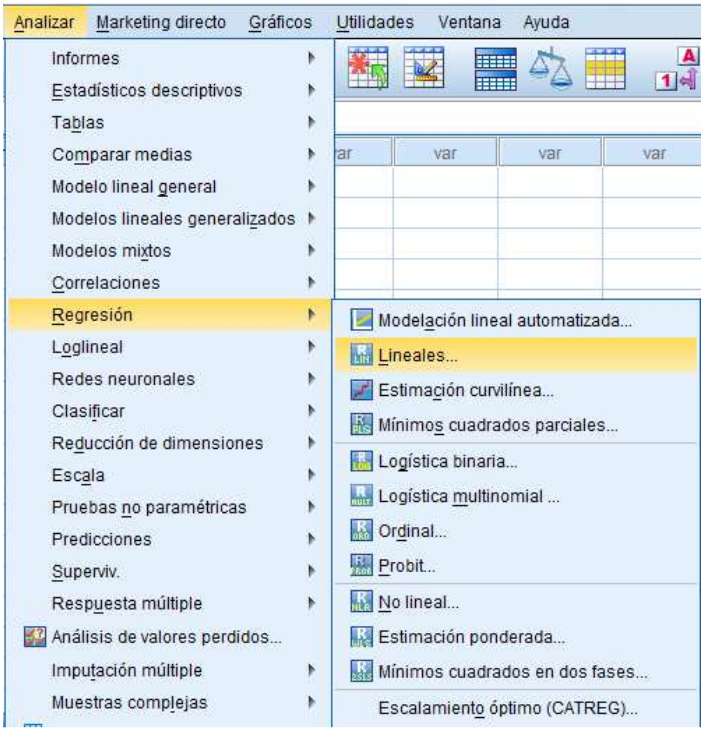
donde y representa las ventas de cereales (en miles de euros), x_1 es la publicidad en televisión (en miles de euros), x_2 es el coste de la publicidad en radio (en miles de euros) y x_3 es la publicidad en periódicos (en miles de euros).

De nuevo, los parámetros desconocidos β_0, β_1, β_2 y β_3 son estimados por mínimos cuadrados. La ecuación estimada de regresión está dada por:

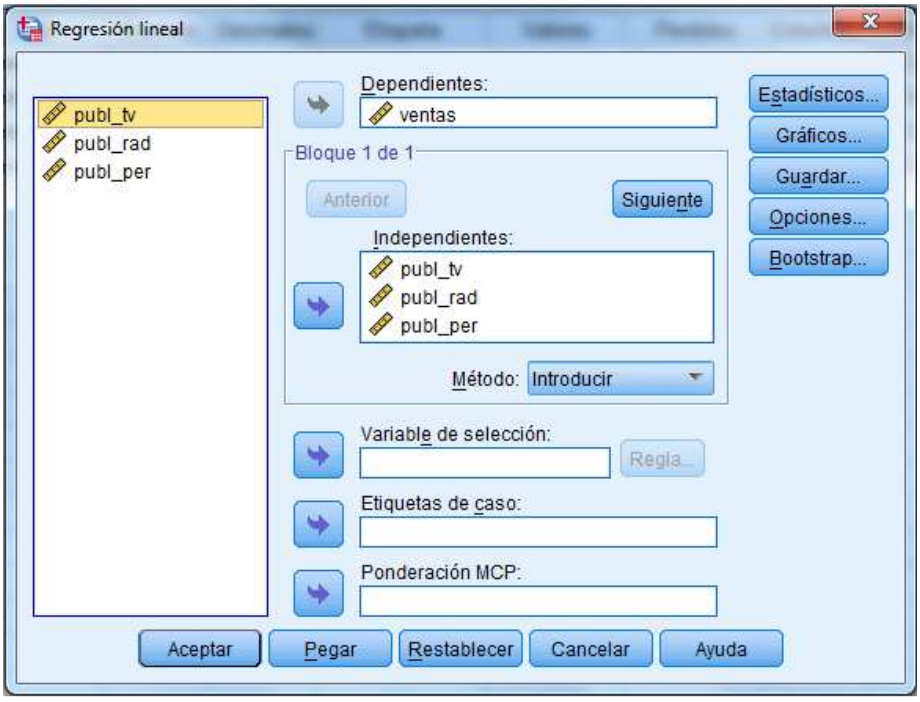
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

El valor $\hat{\beta}_0$ es el valor de la respuesta ajustada cuando todas las variables predictivas tienen un valor igual a cero. Cada $\hat{\beta}_i \ i = 1,2,3$ representa el cambio en la respuesta estimada para un aumento igual a una unidad de la correspondiente variable x_i cuando todas las demás variables independientes se mantienen constantes.

Para obtener dichas estimaciones mediante el paquete SPSS seleccionamos **Analizar -> Regresión -> Lineales.**



Se introducen las tres variables explicativas en el campo Variables Independientes y la variable ventas en el campo Variable Dependiente, como muestra la siguiente figura.



Se pulsa Aceptar y se obtiene como resultado la siguiente salida del programa

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	2,108	2,055		1,026	,320
	publ_tv	3,432	1,121	,358	3,060	,007
	publ_rad	,001	,030	,006	,050	,961
	publ_per	11,347	1,802	,711	6,298	,000

a. Variable dependiente: ventas

En esta figura aparecen los parámetros estimados de regresión $\hat{\beta}_0 = 2.108$, $\hat{\beta}_1 = 3.432$, $\hat{\beta}_2 = 0.001$ y $\hat{\beta}_3 = 11.347$.

La ecuación de regresión ajustada está dada por:

$$\hat{y} = 2.108 + 3.432x_1 + 0.001x_2 + 11.347x_3 \ .$$

Las ventas estimadas son iguales a 2108 euros si no se produce inversión en publicidad (ni en televisión, ni en radio ni en periódicos).

Por cada mil euros invertidos en publicidad en televisión las ventas esperadas aumentan en 3432 euros, supuesto que permanecen constantes las otras variables.

Por cada mil euros invertidos en publicidad en radio, las ventas estimadas aumentan únicamente en 1 euro, suponiendo que se mantienen constantes las otras variables independientes.

Por cada mil euros invertidos en publicidad en periódicos se produce un incremento en las ventas esperadas de 11347 euros, supuestas constantes las restantes variables predictivas.

A la vista de estos resultados parece recomendable la inversión en publicidad en periódicos frente a la publicidad en televisión o en radio.

b) Contrastar la significación del modelo propuesto.

El contraste de significación del modelo de regresión permite verificar si ninguna variable explicativa es válida para la predicción de la variable de interés.

Este contraste puede escribirse

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 : \text{al menos un } \beta_i \neq 0 \quad i = 1,2,3. \end{cases}$$

El p-valor asociado a este contraste aparece en la tabla ANOVA:

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	247,677	3	82,559	26,440	,000[b]
	Residual	49,960	16	3,123		
	Total	297,638	19			

a. Variable dependiente: ventas

b. Variables predictoras: (Constante), publ_per, publ_rad, publ_tv

El *p-valor* asociado al contraste es menor que $\alpha = 0.05$, por lo que rechazamos la hipótesis nula. Esto implica que al menos una de las variables independientes contribuye de forma significativa a la explicación de la variable respuesta.

c) ¿Puede eliminarse alguna variable del modelo? Realiza los contrastes de significación individuales

En la siguiente salida de SPSS aparecen los *p-valores* asociados a los contrastes de regresión individuales

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	2,108	2,055		1,026	,320
	publ_tv	3,432	1,121	,358	3,060	,007
	publ_rad	,001	,030	,006	,050	,961
	publ_per	11,347	1,802	,711	6,298	,000

a. Variable dependiente: ventas

Realizamos tres contrastes de hipótesis, uno para cada coeficiente que acompaña a cada variable explicativa ($i = 1,2,3$)

$$\begin{cases} \mathrm{H}_0 : \beta_i = 0 \\ \mathrm{H}_1 : \beta_i \neq 0 \end{cases} \qquad i = 1,2,3.$$

Para la variable **publ_radio**, *p-valor* = 0.961 > $\alpha = 0.05$, por lo que no rechazamos la hipótesis nula de significación de la variable **publ_radio**. Esta variable no es válida para predecir las ventas de cereales y por tanto puede ser eliminada del modelo.

d) Coeficiente de determinación y de determinación corregido

El coeficiente de determinación es igual a 0.832 y el coeficiente de determinación corregido es igual a 0,801. En este caso no se aprecian grandes diferencias entre los dos coeficientes R^2 y $\overline{R}^2$. El 83.2 % de la variación en las ventas de cereales se explican por su relación lineal con el modelo propuesto. El valor del coeficiente de determinación es satisfactorio.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,912[a]	,832	,801	1,7671

a. Variables predictoras: (Constante), publ_per, publ_rad, publ_tv

Ejercicios propuestos

Ejercicio 1: La siguiente tabla muestra la cantidad de gasolina *y* (en porcentaje con respecto a la cantidad del petróleo en crudo). Se quiere expresar como combinación lineal de cuatro variables: x_1, gravedad del crudo, x_2, presión del vapor del crudo, x_3, temperatura para la cual se ha evaporado un 10% y x_4, temperatura para la cual se ha evaporado el 100%, a partir de los siguientes datos:

Y	X_1	X_2	X_3	X_4
6.9	38.4	6.1	220	235
14.4	40.3	4.8	231	307
7.4	40.0	6.1	217	212
8.5	31.8	0.2	316	365
8.0	40.8	3.5	210	218
2.8	41.3	1.8	267	235
5.0	38.1	1.2	274	285
12.2	50.8	8.6	190	205
10.0	32.2	5.2	236	267
15.2	38.4	6.1	220	300
26.8	40.3	4.8	231	367
14.0	32.2	2.4	284	351
14.7	31.8	0.2	316	379
6.4	41.3	1.8	267	275
17.6	38.1	1.2	274	365
22.3	50.8	8.6	190	275
24.8	32.2	5.2	236	360
26.0	38.4	6.1	220	365
34.9	40.3	4.8	231	395
18.2	40.0	6.1	217	272
23.2	32.2	2.4	284	424
18.0	31.8	0.2	316	428
13.1	40.8	3.5	210	273
16.1	41.3	1.8	267	358
32.1	38.1	1.2	274	444
34.7	50.8	8.6	190	345
31.7	32.2	5.2	236	402
33.6	38.4	6.1	220	410
30.4	40.0	6.1	217	340
26.6	40.8	3.5	210	347
27.8	41.3	1.8	267	416
45.7	50.8	8.6	190	407

Se pide:

a) Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación
b) Contrastar la significación del modelo propuesto.
c) ¿Puede eliminarse alguna variable del modelo? Razona la respuesta
d) Coeficiente de determinación y de determinación corregido

Ejercicio 2:

Se pretende estudiar la posible relación lineal entre el precio de pisos en miles de euros, en una conocida ciudad española y variables como la superficie en m^2 y la antigüedad del inmueble en años. Para ello, se realiza un estudio, en el que se selecciona de forma aleatoria una muestra estratificada representativa de los distintos barrios de la ciudad. Los datos aparecen en la siguiente tabla.

Precio	Superficie	Antigüedad
200	100	20
120	70	15
155	120	30
310	150	20
320	90	12
400	227	7
100	75	22
80	65	28
75	80	30
169	150	43
110	120	49
210	100	21
200	125	15
180	137	28
140	90	30
95	110	33

Se pide:

a) Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación
b) Contrastar la significación del modelo propuesto.
c) ¿Puede eliminarse alguna variable del modelo? Razona la respuesta
d) Coeficiente de determinación y de determinación corregido. Interpretación.

Ejercicio 3:

Salsberry Reality vende casas en la costa este de Estados Unidos. Una de las preguntas más habituales de los potenciales compradores es: "si compramos esta casa, ¿cuánto gastaremos en calefacción durante el invierno?". Para contestar esa pregunta de forma satisfactoria, el departamento de investigación de dicha compañía realizó un estudio en el que se pretende relacionar linealmente el coste de la calefacción en dólares, con las variables temperatura media externa en grados Fahrenheit, el aislamiento del ático en pulgadas y la antigüedad del calentador en años. Los datos se muestran en la siguiente tabla.

Casa	Coste calefacción	Temperatura	Aislamiento	Antigüedad
1	250	35	3	6
2	360	29	4	10
3	165	36	7	3
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	355	10	6	7
8	290	7	10	10
9	230	21	9	11
10	120	55	2	5
11	73	54	12	4
12	205	48	5	1
13	400	20	5	15
14	320	39	4	7
15	72	60	8	6
16	272	20	5	8
17	94	58	7	3
18	190	40	8	11
19	235	27	9	8
20	139	30	7	5

Se pide:

a) Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación
b) Contrastar la significación del modelo propuesto.
c) ¿Cuánto será el coste estimado de la calefacción para una casa con temperatura media externa de 40ºF, 6 pulgadas de aislamiento y 5 años de antigüedad?
d) ¿Puede eliminarse alguna variable del modelo? Razona la respuesta
e) Coeficiente de determinación y de determinación corregido. Interpretación.