

Tema 2: Métodos de Regresión Estadística

2.1 Regresión simple

El objetivo del Análisis de regresión es determinar una función matemática sencilla que describa el comportamiento de una variable dados los valores de otra u otras variables. En el Análisis de regresión simple, se pretende estudiar y explicar el comportamiento de una variable que notamos y , y que llamaremos variable dependiente o variable de interés, a partir de otra variable, que notamos x , y que llamamos variable explicativa, variable de predicción o variable independiente.

Para cumplir dicho objetivo, el primer paso que debe realizar el investigador, es representar las observaciones de ambas variables en un gráfico llamado diagrama de dispersión o nube de puntos. A partir de esta representación el investigador puede especificar la forma funcional de la función de regresión.

Suponemos un modelo en la forma

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

donde β_0, β_1 son los parámetros desconocidos que vamos a estimar, y donde y_i y ε_i son v.a., ε_i es el error aleatorio o perturbación y la función de regresión está dada por $m(x) = \beta_0 + \beta_1 x$.

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

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i,$$

donde $\hat{\beta}_0$ y $\hat{\beta}_1$ están dados por

$$\begin{cases} \hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

Cada coeficiente $\hat{\beta}_i$ representa el efecto sobre la respuesta cuando la variable aumenta en una unidad.

Ejemplos de modelos de regresión simple:

y	X	\mathcal{E}
Consumo de gasolina de un vehículo	Velocidad media	Efecto conductor, tipo de carretera, condiciones ambientales, etc.
Presupuesto de una universidad	Número de alumnos	Efecto del número de profesores, del número de laboratorios, de los m^2 de instalaciones, del número de personal de administración, etc.

2.1.2 Coeficiente de correlación

Es una medida de la asociación lineal que existe entre las variables x e y .

Se define por

$$r = \frac{Cov(x, y)}{S_x S_y}, \text{ con } -1 \leq r \leq 1$$

A continuación se interpretan algunos valores extremos de este coeficiente. Aunque estos valores son poco habituales en la práctica, ilustran cómo interpretar este coeficiente.

$r = -1 \Rightarrow$ asociación lineal negativa perfecta entre x e y $r = 1 \Rightarrow$ asociación lineal positiva perfecta entre x e y $r = 0 \Rightarrow$ no existe ninguna asociación lineal entre x e y
--

2.1.3 Descomposición de la variabilidad y contraste 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 puede ser notada 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 suele resumirse en la siguiente tabla

Tabla ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Regresión	$S.C.R_{eg} = \sum (\hat{y}_i - \bar{y})^2$	1	$S.C.R_{eg}$	$F_{exp} = \frac{S.C.R_{eg}}{S.C.E.}$ $\frac{1}{n-2}$
Error	$S.C.E. = \sum (y_i - \hat{y}_i)^2$	$n-2$	$\frac{S.C.E.}{n-2}$	
Total	$S.C.T. = \sum (y_i - \bar{y})^2$	$n-1$	$\frac{S.C.T.}{n-1}$	

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

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Fijado un nivel de significación α se rechaza H_0 si $F_{exp} > F_{\alpha,1,n-2}$. En la práctica, SPSS proporciona el *p-valor* o nivel mínimo de significación para el rechazo de H_0 , que permite resolver el contraste de hipótesis fijado un nivel de significación.

Si *p-valor* < α , entonces se rechaza H_0
 Si *p-valor* \geq α , entonces no se rechaza H_0

2.1.4 Coeficiente de determinación

Es una medida relativa del grado de asociación lineal entre x e y . Se define por

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

y 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 $S.C.R. = S.C.T.$, por lo que toda la variación de y es explicada por el modelo de regresión.
- Si $R^2 = 0$ entonces $S.C.T. = S.C.E.$, por lo que 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.

Ejercicio: Se quiere estudiar la posible relación existente entre la tasa de desempleo en cierta región del sur de Europa con la tasa de delitos durante 15 meses, en esa zona en un período de recesión económica.

Tasa de desempleo	Tasa de delitos
15.3	30.7
14.8	32.5
12.6	26.3
18.4	35.9
17.9	34.3
15.6	28.5
18.4	37.4
14.1	29.7
20.2	38.8
21.7	40.4
20.4	40.9
16.5	33.3
15.9	30.1
17.9	35.7
16.7	31.4

Se pide:

- Contrasta la existencia de relación lineal entre la tasa de delitos y la tasa de desempleo.
- Teniendo en cuenta el apartado anterior, realiza un ajuste lineal para predecir la tasa de delitos. Interpreta los coeficientes de la ecuación ajustada y obtén la tasa de delitos estimada para una tasa de paro igual al 20%.
- ¿Qué cantidad de la variación de la tasa de delitos puede explicarse por su relación lineal con la tasa de desempleo?

Solución:

- Contrasta la existencia de relación lineal entre la tasa de delitos y la tasa de desempleo.**

Se introducen las variables **tasa_desem** y **tasa_paro** como se muestra en la figura

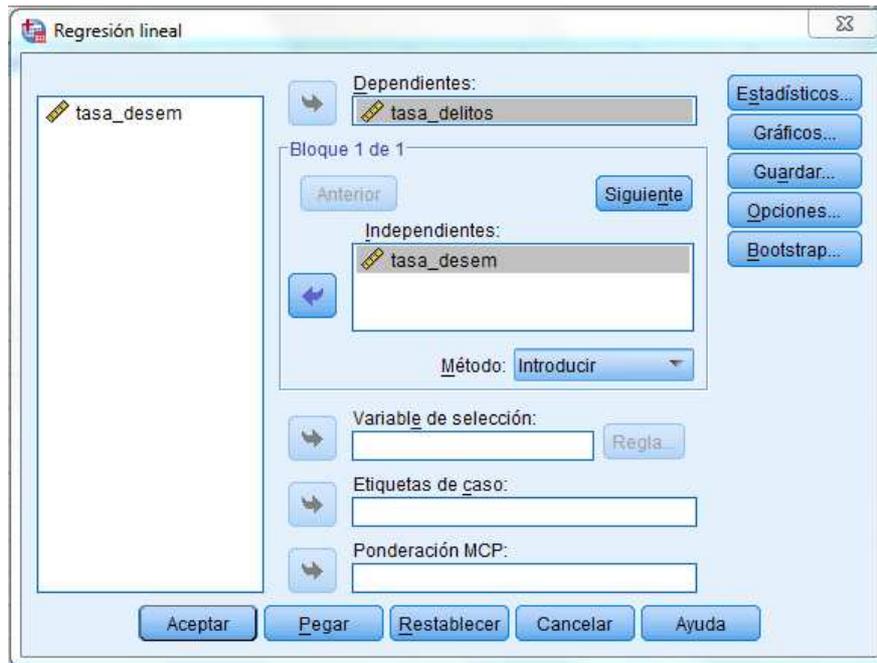
	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
1	tasa_desem	N Numérico	8	1		Ninguna	Ninguna	8	Derecha	Escala
2	tasa_delitos	N Numérico	8	1		Ninguna	Ninguna	8	Derecha	Escala

Una vez introducidas las variables del ejercicio, seleccionamos las opciones **Analizar** -> **Regresión** -> **Lineales**

	tasa_desem	tasa_delitos
1	15,3	30,7
2	14,8	32,5
3	12,6	26,3
4	18,4	35,9
5	17,9	34,3
6	15,6	28,5
7	18,4	37,4
8	14,1	29,7
9	20,2	38,8
10	21,7	40,4
11	20,4	40,9
12	16,5	33,3
13	15,9	30,1
14	17,9	35,7
15	16,7	31,4

The screenshot shows the 'Analyze' menu in SPSS. The 'Regresión' (Regression) option is highlighted, and its sub-menu is open, showing 'Lineales...' (Linear...) as the selected option. Other options in the sub-menu include 'Modelación lineal automatizada...', 'Estimación curvilínea...', 'Mínimos cuadrados parciales...', 'Logística binaria...', 'Logística multinomial...', 'Ordinal...', 'Probit...', 'No lineal...', 'Estimación ponderada...', 'Mínimos cuadrados en dos fases...', and 'Escalamiento óptimo (CATREG)...'.

Seleccionamos las variables del modelo



y obtenemos la siguiente salida de SPSS

ANOVA^a

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	244,306	1	244,306	113,577	,000 ^b
	Residual	27,963	13	2,151		
	Total	272,269	14			

a. Variable dependiente: tasa_delitos

b. Variables predictoras: (Constante), tasa_desem

Consideramos el modelo de regresión lineal simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

donde la variable de interés es la tasa de delitos y la variable de predicción es la tasa de paro. Realizamos el contraste de regresión,

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

El p-valor del contraste (resaltado en un óvalo de color rojo en la figura anterior) es muy pequeño, mucho menor que el nivel de significación habitual de 0.05, lo que conduce al rechazo de la hipótesis nula y a concluir la validez del modelo de regresión lineal.

- b) Teniendo en cuenta el apartado anterior, realiza un ajuste lineal para predecir la tasa de delitos. Interpreta los coeficientes de la ecuación ajustada y obtén la tasa de delitos estimada para una tasa de paro igual al 20%**

Una vez hemos estudiado la existencia de relación lineal entre estas dos variables, se obtiene la ecuación ajustada de regresión a partir de la salida de SPSS.

$$\hat{y} = 5.296 + 1.663x$$

Coeficientes^a

Modelo	Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
	B	Error tip.	Beta		
1 (Constante)	5,296	2,694		1,965	,071
tasa_desem	1,663	,156	,947	10,657	,000

a. Variable dependiente: tasa_delitos

El valor 1.663 es el aumento porcentual que se produce en la tasa de delitos estimada por cada punto porcentual que aumenta la tasa de desempleo.

Para obtener la estimación de la tasa de delitos con una tasa de paro igual al 20%, basta evaluar la variable **tasa de paro** igual a 20 en la ecuación ajustada de regresión.

$$\hat{y} = 5.296 + 1.663 \times 20 = 38.529$$

- a) ¿Qué cantidad de la variación de la tasa de delitos puede explicarse por su relación lineal con la tasa de desempleo?**

El coeficiente de determinación representa el porcentaje de variabilidad de la variable dependiente explicada por el modelo de regresión. Toma valores entre 0 y 1, puede ser medido también en tanto por ciento y es una de las medidas de bondad del ajuste más empleadas en la práctica. Este coeficiente (R^2 en la tabla **Resumen del modelo** y resaltado en un óvalo de color rojo), es igual a 0.897, por lo que el 89.7 % de la variación de la tasa de delitos puede explicarse por su relación lineal con la tasa de desempleo.

El valor R que aparece en dicha tabla (R en la tabla **Resumen del modelo** de SPSS) representa el valor absoluto del coeficiente de correlación. Este coeficiente tiene igual signo que $\hat{\beta}_1 = 1.663$ y por tanto R es igual a 0.947, lo que denota una fuerte asociación lineal positiva de las dos variables.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,947 ^a	,897	,889	1,4666

a. Variables predictoras: (Constante), tasa_desem

Ejercicio: Como resultado de una nueva política empresarial, se ha aumentado progresivamente la inversión en formación de los empleados en una multinacional de software. Se sospecha que este incremento en inversión ha tenido gran importancia en los beneficios de la empresa. Se recogen datos correspondientes a los gastos en inversión (en millones de euros) y a los beneficios brutos de la empresa (en millones de euros) en los últimos diez meses

Gastos en inversión	Beneficios
0.2	25.3
0.1	26.7
0.7	31.4
0.8	33.5
1.1	39.7
1.3	40.6
2.4	45.5
2.9	56.8
3.5	75.4
3.9	97.2

- Representa el diagrama de dispersión
- ¿Existe relación lineal entre las variables? Bondad de ajuste
- Realiza un ajuste exponencial. ¿Qué modelo es preferible? Razona la respuesta

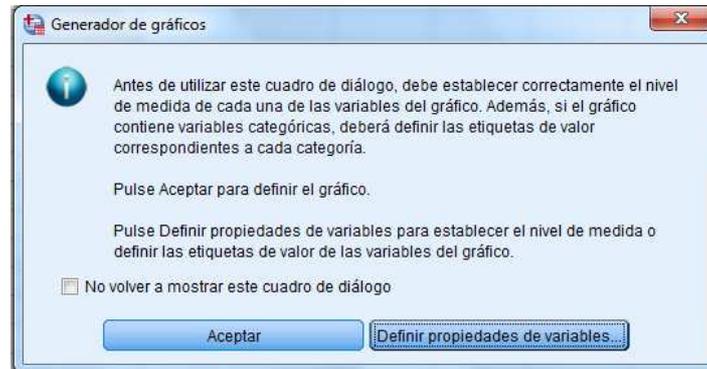
Solución:

a) Representa el diagrama de dispersión

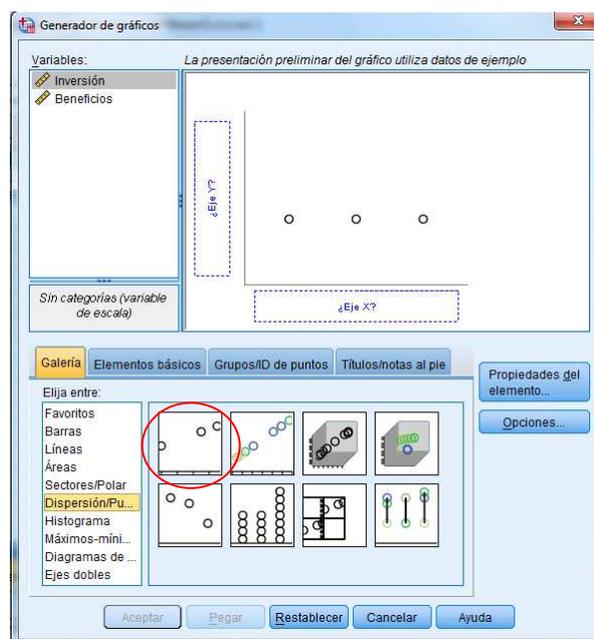
Una vez hemos introducido las variables **Inversión** y **Beneficios** y sus valores en SPSS, seleccionamos **Gráficos -> Generador de gráficos**



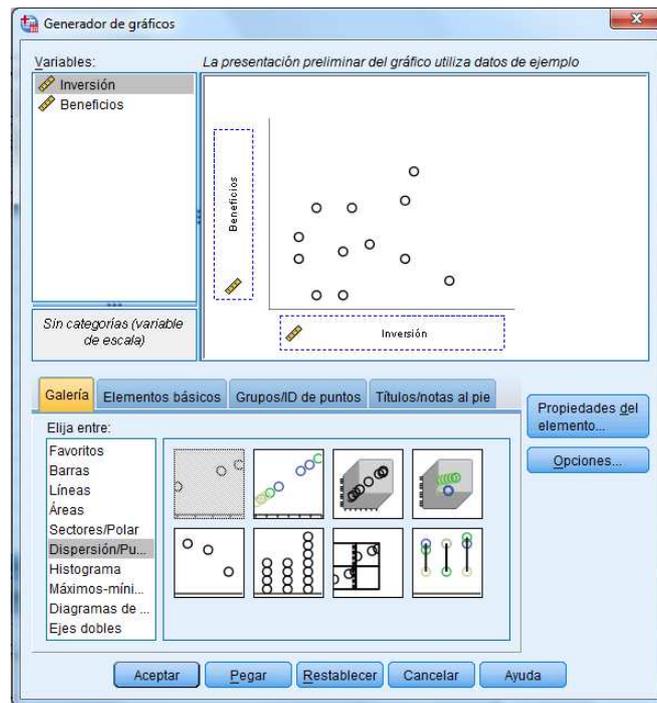
Obtenemos el siguiente cuadro de diálogo, donde pulsamos **Aceptar**



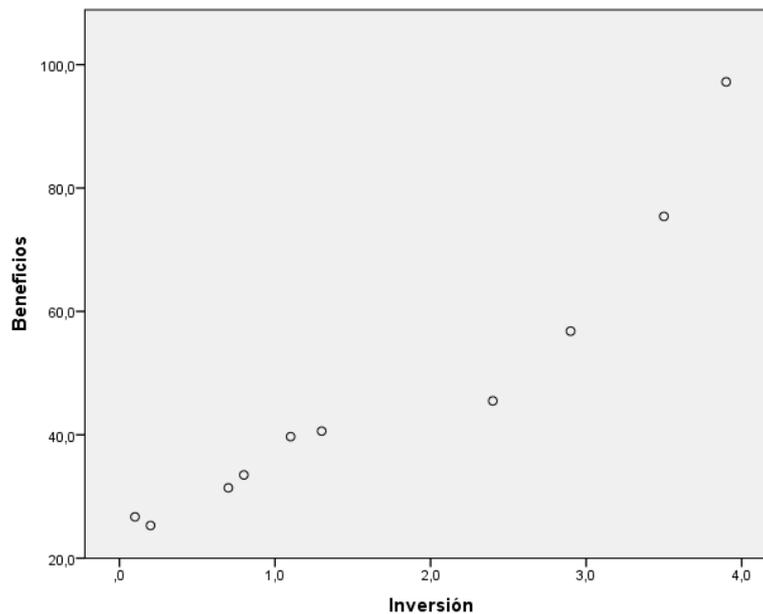
En la ventana resultante, seleccionamos en la pestaña **Galería** la opción **Dispersión/Puntos**, y hacemos doble-click en el primero de los ocho modelos de gráficos disponibles, como aparece resaltado en la figura



A continuación, en el campo **Variables** de dicha ventana, se arrastran las variables **Inversión** y **Beneficios** hacia el **Eje x** y **Eje y** del gráfico central, respectivamente.



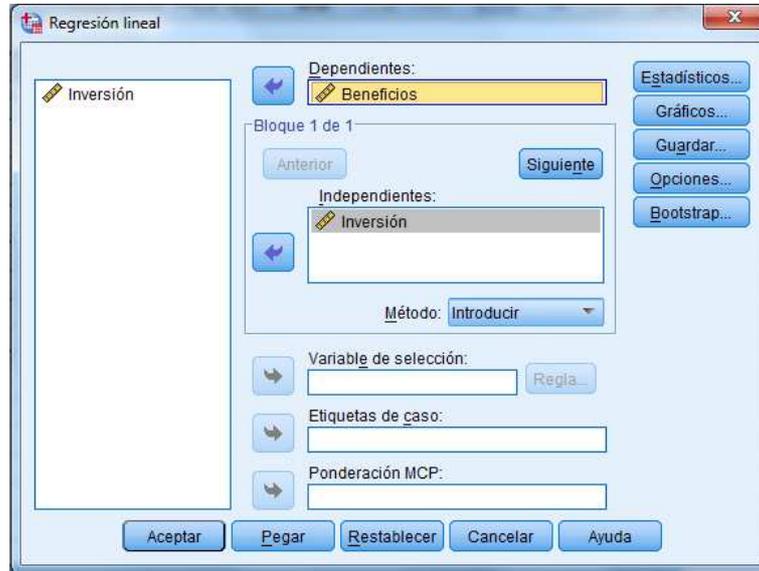
Se pulsa **Aceptary** se obtiene el diagrama de dispersión



A partir de esta gráfico el investigador puede especificar la forma funcional de la función de regresión. Consideramos un ajuste lineal y exponencial.

b) ¿Existe relación lineal entre las variables? Bondad de ajuste

Se eligen las opciones del menú de SPSS **Anализar -> Regresión -> Lineales** e introducimos las variables en los campos correspondientes, como se muestra en la figura



Obtenemos la siguiente salida de SPSS

ANOVA^a

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4314,536	1	4314,536	67,316	,000 ^b
	Residual	512,753	8	64,094		
	Total	4827,289	9			

a. Variable dependiente: Beneficios

b. Variables predictoras: (Constante), Inversión

Se considera el siguiente contraste

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

El p-valor asociado al contraste es menor que el nivel de significación habitual del 5%, por lo que rechazamos la hipótesis nula de no linealidad. El valor del coeficiente de determinación $R^2 = 0.894$ indica que el 89.4 % del aumento de los beneficios de la empresa es explicada por su relación lineal con la inversión en formación de sus empleados.

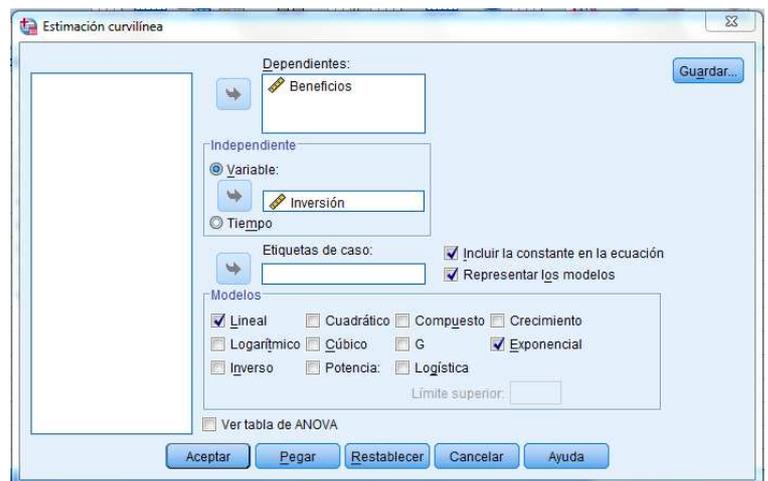
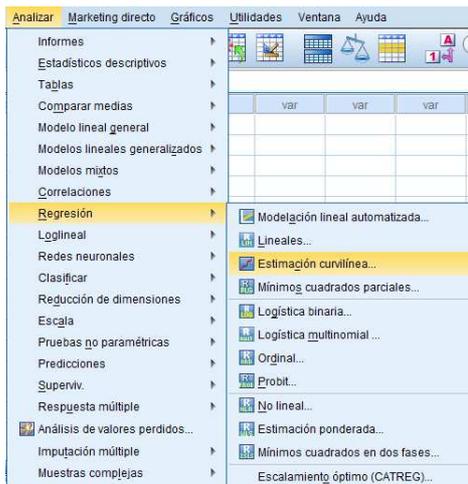
Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,945 ^a	,894	,881	8,0059

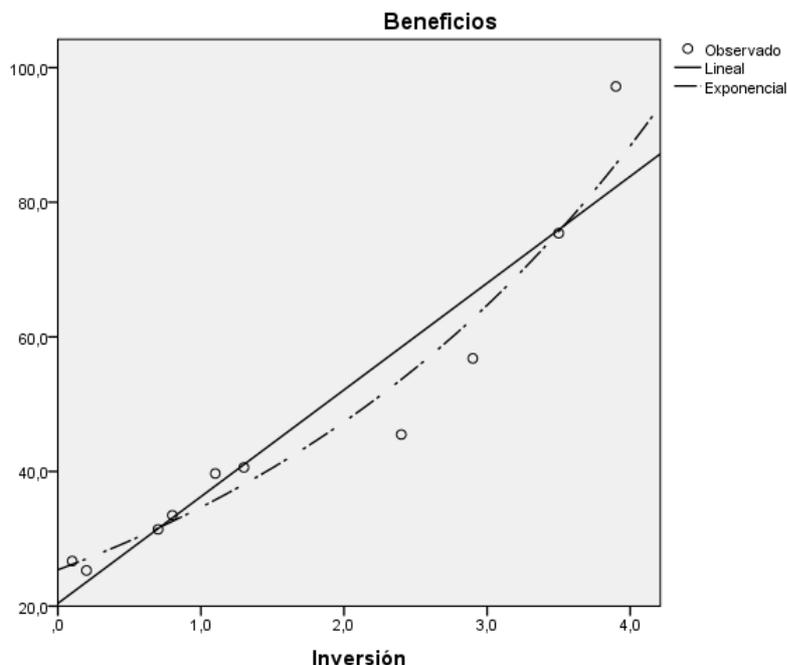
a. Variables predictoras: (Constante), Inversión

c) Realiza un ajuste exponencial. ¿Qué modelo es preferible? Razona la respuesta

Se seleccionan las opciones de los menús **Analizar -> Regresión -> Estimación curvilínea**, e insertamos las variables en sus respectivos campos



Obtenemos el siguiente gráfico:



Las observaciones aparecen representadas por círculos, el ajuste lineal mediante una línea continua y el ajuste exponencial mediante una línea discontinua.

Tanto el modelo lineal como el exponencial parecen aproximarse de forma razonable a las observaciones. SPSS también proporciona la tabla **Resumen del modelo y estimaciones de los parámetros** que permite comparar los dos modelos de forma más precisa

Resumen del modelo y estimaciones de los parámetros

Variable dependiente: Beneficios

Ecuación	Resumen del modelo					Estimaciones de los parámetros	
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	,894	67,316	1	8	,000	20,404	15,862
Exponencial	,958	181,586	1	8	,000	25,398	,312

La variable independiente es Inversión.

Los dos modelos son válidos y pueden ser empleados para obtener predicciones de la variable **Beneficios**. Sin embargo el coeficiente de determinación correspondiente al modelo exponencial es igual a 0,958, superior a 0,894, correspondiente al modelo lineal. Por tanto, es preferible explicar el aumento de beneficios de la empresa con el modelo exponencial que con el modelo lineal.

Ejercicios propuestos Regresión simple

Ejercicio 1:

Dada la siguiente distribución de datos correspondientes al precio de la vivienda y a los metros construidos.

Superficie	Precio
90	9
95	9,5
96,5	10
100,2	12
105,4	11,4
107	9,3
111,7	12,5
124,1	14

Se pide:

- Calcule la ecuación de la recta de regresión del precio sobre la superficie.
- ¿Es válido el modelo de regresión considerado?
- Proporcione una predicción del valor del precio cuando la superficie es 99 m²
- ¿Es buena la predicción que se ha obtenido en el apartado anterior? Razone la respuesta

Ejercicio 2:

Se realiza un estudio con la intención de comparar la capacidad memorística y el cociente intelectual de estudiantes de una universidad española. Se realizan dos pruebas a 6 alumnos, una de capacidad memorística y otra de cociente intelectual:

Capacidad memorística	18,4	22,4	28,3	33,6	36,6	11,5
Cociente intelectual	95	80	104	120	135	91

Se pide:

- Halle la ecuación de la recta de regresión de la capacidad memorística sobre el cociente intelectual.
- Obtenga e interprete los coeficientes de correlación lineal y de determinación.

Ejercicio 3:

En una explotación ganadera, se ha realizado una inversión destinada a mejorar el equipamiento tecnológico de la explotación, durante un periodo de 9 años. Para estudiar futuras inversiones, se realiza un estudio para conocer si la inversión realizada, en millones de euros, ha influido en el beneficio obtenido, medido también en millones de euros. Los datos se recogen en la siguiente tabla:

Inversión	10	12	17	19	21	22	25	20	18
Beneficio	8	9	8	8,5	9	10	12	7,4	8,2

Calcule:

- La recta de regresión del beneficio en dos años respecto de la inversión.
- La previsión de inversión necesaria para obtener un beneficio de 11 millones de euros.
- ¿Se podrían considerar otros modelos? Razona la respuesta.

Ejercicio 4:

Se ha realizado un estudio para conocer si existe relación lineal entre la renta mensual de una familia, en euros y el gasto en actividades de tipo cultural, en euros

Renta	Gasto
2400	120
2100	110
1300	60
1750	55
1255	120
2100	270
3100	110
2200	220
1640	250
1770	140
1853	120
1994	400
2025	85
3200	100

Se pide:

- Representa el diagrama de dispersión. ¿Qué modelo crees conveniente emplear?
- Realiza un ajuste lineal, interpreta los coeficientes de regresión y obtén el gasto en actividades culturales de una familia con una renta mensual disponible de 2500 euros.
- ¿Existe relación lineal? Razona la respuesta.
- A partir del resultado obtenido en el apartado c), ¿qué puedes afirmar ahora sobre la predicción obtenida en el apartado b)?
- ¿Tiene sentido considerar otros modelos? Razona la respuesta.

Ejercicio 5:

Los siguientes datos recogen el precio en dólares de 12 modelos de portátiles, seleccionados al azar, junto con las velocidades del procesador correspondiente, en gigahercios.

Portátil	Velocidad	Precio
1	2.0	2197
2	1.6	1200
3	1.6	950
4	1.8	1947
5	2.0	2058
6	1.2	1012
7	1.6	1252
8	2.0	2017
9	2.0	2137
10	3.2	2455
11	1.4	969
12	2.8	2273

Se pide:

- Calcule la ecuación de la recta de regresión del precio del portátil sobre la velocidad del procesador.
- ¿Es válido el modelo de regresión considerado?
- A la vista de la ecuación de regresión, ¿hay algún portátil que parezca tener, un precio mayor o menor?

Ejercicio 6:

En un estudio sociológico, se postula que la actitud racista viene determinada fundamentalmente por el nivel de paro que existe en esa sociedad. Para avalar esa afirmación, el estudio proporciona los datos relativos a una muestra de diez localidades, cuyos niveles de paro en %, y racismo, medidos según cierto índice, son los siguientes:

Paro	Racismo
7	22
13	29
5	15
33	37
23	31
21	33
18	32
30	40
15	30
27	38

Se pide:

- a) ¿Es válido el modelo de regresión considerado?
- b) ¿Qué valor del índice de racismo predecirías para una localidad con una tasa de paro del 20%? ¿Cómo de buena te parece esta predicción?

Ejercicio 7:

Una empresa basa las predicciones de sus ventas anuales en las estimaciones oficiales de la demanda total de la industria. Los datos siguientes corresponden a las estimaciones oficiales de la demanda total en miles de euros y las ventas efectuadas por la firma en miles de euros en los últimos años.

Demanda	Ventas
230	5
220	7
400	12
330	8
210	5
390	10
280	9
140	8
280	7
290	12

Explica todo lo que puedas sobre la posible relación lineal o no lineal entre las variables **Demanda** y **Ventas**.

2.2 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, \dots, x_k y no sólo de una única variable 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 variable de 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

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

donde $\beta_0, \beta_1, \dots, \beta_k$ 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, \dots, 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 el tipo 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 ser explicada 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

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

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

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

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.

2.2.1 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
<i>Regresión</i>	$B^t X^t Y^t - \frac{1}{n} (\sum y_i)^2$	$k = m - 1$	$\frac{S.C.R.}{m - 1}$	$F_{\text{exp}} = \frac{\frac{S.C.R.}{m - 1}}{\frac{S.C.E.}{n - m}}$
<i>Error</i>	$Y^t Y - B^t X^t Y$	$n - m$	$\frac{S.C.E.}{n - m}$	
<i>Total</i>	$Y^t Y - \frac{1}{n} (\sum y_i)^2$	$n - 1$		

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

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

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

Si *p-valor* < α , 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

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

En este caso el estadístico de contraste sigue una F de Snedecor con 1 y $n - k - 1$ grados de libertad. Este contraste es equivalente al 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 .

2.2.3 Coeficiente de determinación

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

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}.$$

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 lo que 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 de determinación presenta el inconveniente de que su valor aumenta al añadir nuevas variables al modelo de regresión, independientemente de que éstas contribuyan de forma significativa a la explicación de la variable respuesta. Para evitar un aumento injustificado de este coeficiente, se introduce el coeficiente de determinación 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 añaden nuevas variables, sino que en caso de añadir variables superfluas al 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:

- Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación
- Contrastar la significación del modelo propuesto.
- ¿Puede eliminarse alguna variable del modelo? Realiza los contrastes de significación individuales
- Coefficiente 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 del tipo:

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

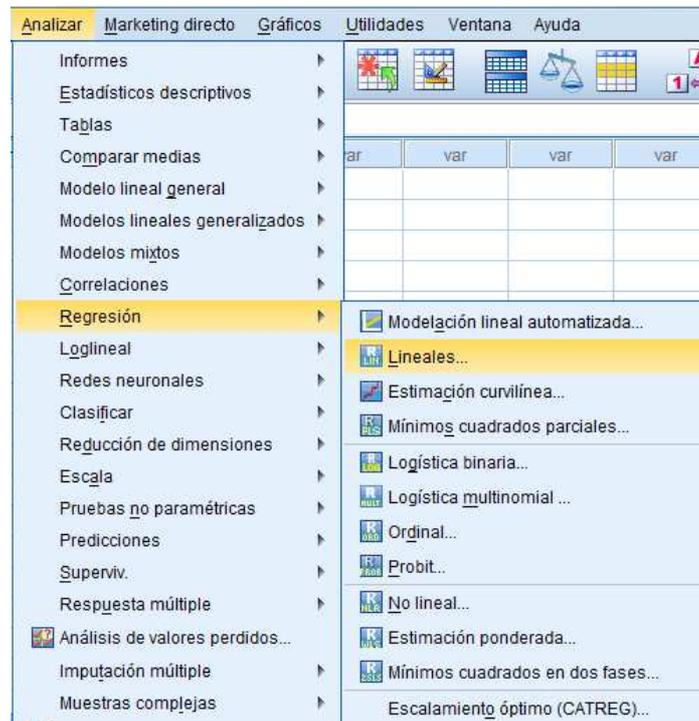
donde y representa las ventas de cereales (en miles de euros), x_1 es la publicidad en televisión (en miles de euros), x_2 es el coste de la publicidad en radio (en miles de euros) y x_3 es la publicidad en periódicos (en miles de euros).

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:

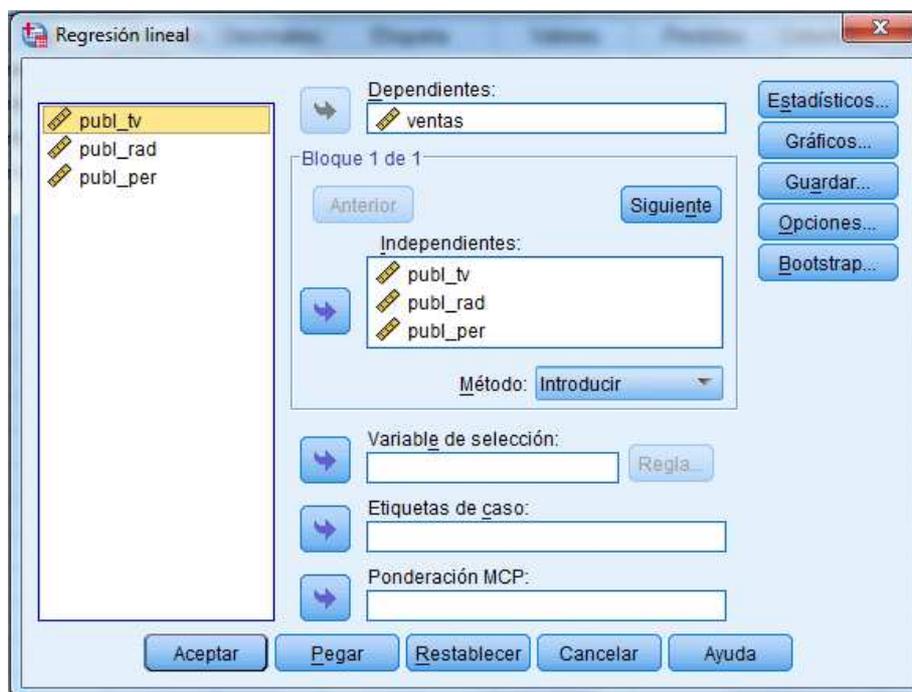
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

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

Coefficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error tip.	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 ($i = 1, 2, 3$)

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases} \quad i = 1, 2, 3.$$

Para la variable **publ_radio**, $p\text{-valor} = 0.961 > \alpha = 0.05$, 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 \bar{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:

- Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación
- Contrastar la significación del modelo propuesto.
- ¿Puede eliminarse alguna variable del modelo? Razona la respuesta
- Coefficiente 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:

- Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación
- Contrastar la significación del modelo propuesto.
- ¿Puede eliminarse alguna variable del modelo? Razona la respuesta
- Coefficiente 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:

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