EJERCICIOS REGRESIÓN

Los datos correspondientes a los ejercicios se encuentran en el fichero Excel "datos ejercicios regresión", salvo el ejercicio 10 cuyos datos están en el fichero "empleados. sav"

1. Los datos de la hoja "***ropa***" muestran, la superficie, en metros cuadrados, y el volumen de ventas, en miles de $, de 14 tiendas de una cadena de ropa para mujer. (*ejemplo* ***adaptado*** *del libro: Berenson, Levine, Krehbiel (2001): Estadística para Administración. Editorial Prentice Hall).*

   Se pide:

   a. Obtener el diagrama de dispersión considerando como variable independiente la superficie de la tienda, e indicar el carácter de la relación entre las dos variables
   b. Obtener la recta de regresión que explica las ventas como una función de la superficie de la tienda y concluir si ésta es significativa en la explicación de las ventas
   c. ¿Puede considerarse adecuado el modelo estimado en el apartado anterior para recoger la relación entre las dos variables?
   d. Estimar el volumen de ventas de una tienda con 300 m^2 de superficie
   e. ¿Cuánto se estima que variarán las ventas si incrementamos la superficie 50 m^2?

2. Los datos de la hoja "***refrescos***" muestran las ventas (en millones) de cajas y los gastos de publicidad (en millones de $) para 7 marcas principales de refrescos (*ejemplo extraído del libro: Anderson, Sweeney, Williams (1999): Estadística para Administración y Economía. Editorial Thomson)*

   Se pide:

   a. Hacer un diagrama de dispersión.
   b. ¿Qué parece indicar el diagrama acerca de la relación entre las variables?
   c. Obtener la ecuación de la recta que explica como las ventas dependen del gasto en publicidad.
   d. Interpretación y significatividad de la pendiente de la recta obtenida
   e. Predecir las ventas para una marca que gaste 70 millones de $ en publicidad

3. Los datos de la hoja "***seguridad social***" recogen información sobre las siguientes variables:

   - Nombre de la Comunidad Autónoma
   - Número de trabajadores con alta en la Seguridad Social en Servicios de Alojamiento
   - Número de trabajadores con alta en la Seguridad Social en Servicios de Comidas y Bebidas
   - Gasto, en millones de euros, de los turistas en cada Comunidad Autónoma
   - PIB, en millones de euros, de cada Comunidad Autónoma

   Se pide

   a. Indicar el tipo/ carácter de la relación entre las variables
   b. Diagrama de dispersión de las variables gasto de los turistas y trabajadores con alta en la SS en servicios de alojamiento
   c. Obtener el modelo lineal que explica el nº de trabajadores con alta en la Seguridad Social en servicios de alojamiento como una función del gasto de los turistas, y hacer un pronóstico para un valor del gasto de 10.000 millones de euros
   d. Porcentaje de variación de la variable nº de trabajadores con alta en la Seguridad Social en servicios de alojamiento que se explica linealmente con el gasto de los turistas
   e. Estimar la repercusión que un aumento de 1 millón de euros del gasto de los turistas, tendrá en el número de trabajadores con alta en la Seguridad Social, en servicios de alojamiento, de una Comunidad Autónoma, ¿y si el aumento fuera de 5 millones?

4. Los datos de la hoja "***absentismo***" muestran, para 50 empleados de una empresa, la siguiente información, (*ejemplo* ***adaptado*** *del libro: Ezequiel Uriel y Joaquín Aldás (2005): Análisis Multivariante Aplicado. Editorial Thomson).*

- Edad en años (edad)
- Años de antigüedad en la empresa (antigüedad)
- Salario mensual en euros (salario)
- Nº de días que cada empleado ha faltado a trabajar en el último año (absentismo)

Se pide:

a. Indicar el carácter de la relación entre las variables absentismo y edad e indicar si la correlación entre las variables es significativa
b. Obtener la recta de regresión que explica el absentismo como una función de la edad
c. Pronosticar el absentismo de un empleado con 60 años
d. Interpretar el significado de la pendiente de la recta. ¿Puede considerarse la edad una variable significativa en la explicación del absentismo?
e. Estimar los parámetros del modelo lineal que explica el absentismo como una función de la edad, el salario y los años de antigüedad
f. Para el modelo obtenido en el apartado anterior, ¿qué variable tiene más influencia sobre el absentismo?
g. Determine si pueden considerarse significativas cada una de las varibles incluidas en el modelo del apartado e y concluya sobre la significativadad global del modelo
h. Haga un diagnóstico sobre la multicolinealidad de las variables incluidas en el modelo del apartado e
i. Restimar el modelo eliminando, una a una, las variables con multicolinealidad

5. Una gran compañía de productos para el consumidor desea medir la efectividad de la publicidad en radio/televisión y en periódicos sobre las ventas de sus productos. Para ello selecciona una muestra de 22 ciudades similares en población. Los datos de la hoja "***publicidad***" muestran el gasto en radio/ televisión y en periódicos y el volumen de ventas en un mes determinado. (*ejemplo* ***adaptado*** *del libro: Berenson, Levine, Krehbiel (2001): Estadística para Administración. Editorial Prentice Hall).*

Se pide

a. Estimar los parámetros del modelo de regresión que explica las ventas en función del resto de variables
b. Interpretar el significado de los coeficientes de cada una de las variables independientes
c. Pronosticar las ventas promedio para una ciudad donde la publicidad en radio/tv es de 20.000 dólares y en periódicos de 20.000 dólares
d. Para un nivel de significación del 5% determinar si cada variable hace una contribución significativa a la explicación de la variable ventas
e. Determinar si existe una relación significativa ente las ventas y las dos variables explicativas para un nivel de significación del 5%
f. Determinar el VIF de cada variable explicativa ¿existe alguna razón para sospechar que existe colinealidad'
g. Obtener el coeficiente de determinación ajustado e interpretar su significado

6. Los datos de la hoja "***turismo***" recogen información sobre el PIB, el nº de establecimientos hoteleros abiertos, el número de plazas hoteleras, nº de viajeros y el personal empleado en el sector turístico para las provincias españolas (se excluyen Ceuta y Melilla)

Se pide:

a. El modelo lineal que explica el PIB en función del nº de viajeros. ¿Es fiable el modelo estimado?
b. Si para un próximo periodo se estima un aumento medio de 5.000 viajeros por provincia ¿Cuál será la repercusión en el PIB?
c. Valor estimado del PIB para un periodo en el que se reciben 100.000 viajeros
d. Una estimación del nº de personas empleadas para 700 establecimientos abiertos.
e. El modelo lineal que explica el PIB en función del nº de viajeros, el nº de establecimientos abiertos, el nº de plazas disponibles y el personal empleado.
f. La fiabilidad del modelo
g. Repercusión en el PIB de un aumento de 1000 viajeros

h. Estimar el PIB de una provincia con 150000 viajeros y 37000 plazas hoteleras

7. Indicar si son verdaderas o falsas las siguientes afirmaciones:

   a. Si dos variables son independientes el coeficiente de correlación vale 0
   b. Dos variables incorrelacionadas linealmente son independientes
   c. Un coeficiente de correlación lineal igual a 1 indica más asociación lineal que un coeficiente igual a -1
   d. El coeficiente de determinación de un modelo de regresión lineal es una medida del porcentaje de error que se comete en la estimación de los parámetros del mismo
   e. La pendiente de una recta de regresión indica la variación que se produce en la variable dependiente cuando la independiente aumenta una unidad
   f. Si dos variables tienen una relación inversa, su coeficiente de correlación es mayor que 0

8. Se dispone de datos de consumo (litros de combustible consumidos por cada 100 kms recorridos), cilindrada (en cc.), peso (en Kg.) y aceleración (segundos en pasar de 0 a 100km/h) para una muestra de 398 coches. Se pretende ajustar un modelo de regresión lineal múltiple para intentar predecir el consumo en función del resto de variables. A continuación se presentan las salidas producidas por SPSS al ajustar dos posibles modelos:

Modelo 1:

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Aceleración 0 a 100 km/h (segundos), Peso total (kg), Cilindrada en cc[a]	.	Introducir

a. Todas las variables solicitadas introducidas

b. Variable dependiente: Consumo (l/100Km)

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,858[a]	,736	,734	2,034

a. Variables predictoras: (Constante), Aceleración 0 a 100 km/h (segundos), Peso total (kg), Cilindrada en cc

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4551,909	3	1517,303	366,695	,000[a]
	Residual	1630,284	394	4,138		
	Total	6182,193	397			

a. Variables predictoras: (Constante), Aceleración 0 a 100 km/h (segundos), Peso total (kg), Cilindrada en cc

b. Variable dependiente: Consumo (l/100Km)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	4,889	,864		5,656	,000		
	Cilindrada en cc	,001	,000	,273	3,327	,001	,100	10,024
	Peso total (kg)	,007	,001	,530	6,981	,000	,116	8,600
	Aceleración 0 a 100 km/h (segundos)	-,188	,046	-,132	-4,129	,000	,652	1,533

a. Variable dependiente: Consumo (l/100Km)

Modelo 2:

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Aceleración 0 a 100 km/h (segundos), Peso total (kg)[a]	.	Introducir

a. Todas las variables solicitadas introducidas

b. Variable dependiente: Consumo (l/100Km)

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,854[a]	,729	,728	2,060

a. Variables predictoras: (Constante), Aceleración 0 a 100 km/h (segundos), Peso total (kg)

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4506,112	2	2253,056	530,975	,000[a]
	Residual	1676,082	395	4,243		
	Total	6182,193	397			

a. Variables predictoras: (Constante), Aceleración 0 a 100 km/h (segundos), Peso total (kg)

b. Variable dependiente: Consumo (l/100Km)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	4,761	,874		5,444	,000		
	Peso total (kg)	,011	,000	,764	26,697	,000	,838	1,193
	Aceleración 0 a 100 km/h (segundos)	-,259	,041	-,182	-6,378	,000	,838	1,193

a. Variable dependiente: Consumo (l/100Km)

Asumiendo que se cumplen el resto de hipótesis del modelo que no aparecen en estas tablas, se pide:

a. Discutir de manera razonada cuál de los dos modelos es el más adecuado.
b. A partir del modelo elegido en el apartado anterior:
    i. discutir de manera razonada su bondad,
    ii. escribir el modelo,
    iii. interpretar los coeficientes,
    iv. estimar el consumo de un coche que tenga una cilindrada de 5000 cc. un peso de 1100 Kg. y una aceleración de 10 segundos

9. Hoja "**eficiencia energética**" Para explicar el gasto en calefacción en edificios residenciales se han observado seis características :
    - Superficie Total
    - Superficie de fachada
    - Superficie de tejado
    - Altura media
    - Gasto en calefaccion
    - Superficie acristalada
    - Número de ventanas

Se pide:

a. Obtenerla matriz de correlaciones en SPSS e incluir en la hoja de respuestas las correlaciones que son significativas y por qué.
b. ¿Cuál es el significado del signo negativo del coeficiente de correlación entre el gasto en calefacción y la superficie total?
c. El modelo de regresión múltiple que explica el gasto en calefacción en función de todas las variables restantes
d. En el modelo anterior ¿Hay alguna variable que no sea significativa para explicar el gasto en calefacción? Justifica la respuesta
e. ¿Cuáles son la hipótesis nula y alternativa del contraste de hipótesis que permite concluir si una variable es significativa o no?
f. Justificar si el modelo estimado presenta multicolinealidad o no. En caso afirmativo proponga un modelo alternativo.
g. El significado del coeficiente de la variable que indica el número de ventanas en el modelo propuesto en el apartado anterior
h. El significado del coeficiente de determinación.

10. Archivo "**datos de empleados**" Los datos del archivo "empleados" recogen información relativa a las siguientes variables para un total de 400 empleados de una gran entidad bancaria
    - sexo
    - fechnac
    - educ (nivel educativo)
    - catlab (categoría laboral. 1: administrativo; 2: seguridad; 3: directivo)
    - salario (salario actual)
    - salini (salario inicial)
    - tiempemp (meses desde el contrato)
    - expprev (experiencia previa)
    - minoría (pertenencia a una minoría étnica. 0: no; 1. si)
    - edad

    a. Obtener la matriz de correlaciones en SPSS e indicar las correlaciones que son significativas a nivel 0,05 y por qué.
    b. ¿Cuál es el significado del signo negativo del coeficiente de correlación entre las variables niveleducativo y experiencia?
    c. El modelo de regresión múltiple que explica el salario actual en función de todas las variables restantes
    d. En el modelo anterior ¿Hay alguna variable que no sea significativa para explicar el salario actual? Justifica la respuesta (nivel de significación 0,05)
    e. ¿Cuáles son la hipótesis nula y alternativa del contraste de hipótesis que permite concluir si una variable es significativa o no?

f. Justificar si el modelo estimado presenta multicolinealidad o no. En caso afirmativo proponga un modelo alternativo.
g. El significado del coeficiente de determinación en este caso.

11. Hoja "**fabricantes de coches**" El archivo contiene información sobre las siguientes variables relativas a una muestra de 153 coches:

Fabricante	
Modelo	
Ventas	Ventas en miles de vehículos
Precio	Precio en miles de euros
Motor	Tamaño del motor
Potencia	Potencia en caballos
Depósito	Capacidad del depósito

a. ¿Puede concluirse la existencia de correlación entre los siguientes pares de variables?:
   - precio de venta y ventas
   - tamaño del motor y ventas

   Indica la hipótesis nula y alternativa del contraste que permite concluir si la correlación entre dos variables es significativa
b. Obtener el diagrama de dispersión de las variables precio de venta y ventas, indicando si la relación es directa o inversa
c. El modelo de regresión múltiple que explica la variable ventas en función de todas las variables numéricas restantes. Para este modelo:
   1. escribe la ecuación del modelo de regresión
   2. indica que variables son significativas y cuales no para explicar las ventas de coches. Justifica la respuesta incluyendo la hipótesis nula y alternativa del contraste de hipótesis que permite concluir si una variable es significativa
   3. ¿es significativo el modelo anterior para explicar las ventas de coches? Justifica la respuesta incluyendo la hipótesis nula y alternativa del contraste de hipótesis que permite concluir si el modelo es significativo o no.
   4. Justificar si el modelo estimado presenta multicolinealidad o no. En caso afirmativo proponga un modelo alternativo

d. Para el modelo alternativo propuesto en el punto c.4, indica
   1. El significado del coeficiente de la variable precio del coche
   2. Bondad de ajuste del modelo

12. "**Hoja cine**" Una nueva distribuidora de cine ha hecho un estudio sobre la relación entre la frecuencia de asistencia al cine y el consumo de otros medios de comunicación para diseñar una campaña publicitaria.

    Se pide:
    a. Diagrama de dispersión para las variables cine y TV
    b. Indique las parejas de variables con correlación significativa
    c. Estimar los parámetros del modelo de regresión de la variable cine sobre el resto de variables
    d. Indique, justificando la respuesta, qué variables son significativas en el modelo estimado
    e. Indique, justificando la respuesta, si el modelo estimado en el apartado anterior es significativo
    f. ¿cuál es el significado del coeficiente de la variable TV?

13. "**Hoja inmobiliaria**" Una inmobiliaria dispone de información sobre las siguientes variables referidas a un total de 25 edificios de oficinas:
    - Metros (superficie en metros cuadrados)
    - Antigüedad (años de antigüedad del edificio)
    - Valor de tasación en euros
    - Número de oficinas

Se pide:

a. Diagrama de dispersión para las variables valor de tasación y metros cuadrados

b. Obtener la matriz de correlaciones e indicar qué correlaciones son significativas y por qué

c. Estimar los parámetros del modelo de regresión de la variable valor de tasación sobre el resto de variables

d. Indique, justificando la respuesta, qué variables son significativas en el modelo estimado

e. Indique, justificando la respuesta, si el modelo estimado en el apartado anterior es significativo

f. Calcular el FIV de cada variable y comentar el resultado

g. Estime de nuevo el modelo utilizando el método de estimación por pasos. Para este modelo:

h. ¿Cómo repercute en el valor de tasación un aumento de la superficie de 50 metros cuadrados?

i. Bondad del modelo estimado