Tema 2. Descripción Conjunta de Varias Variables

Cuestiones de Verdadero/Falso

1. La covarianza mide la relación lineal entre dos variables, pero depende de las unidades de medida utilizadas.
2. El análisis de regresión simple se usa cuando varias variables independientes contribuyen a la variación de una variable dependiente.
3. La variable que se va a predecir en la regresión se denomina variable independiente.
4. Si en el análisis de regresión simple la pendiente de la recta es negativa, entonces hay correlación lineal negativa entre las variables.
5. Si el coeficiente de correlación lineal es cero, las rectas de regresión de Y sobre X y X sobre Y son perpendiculares.
6. La nube de puntos es una representación gráfica que nos permite visualizar la posible relación entre dos variables.
7. El coeficiente de correlación lineal r_{xy} siempre toma valores entre -2 y 2.
8. El coeficiente de determinación puede tomar valores negativos.
9. Un coeficiente de correlación lineal casi cero indica que la relación lineal entre las variables dependiente e independiente es muy débil.
10. Si el coeficiente de correlación lineal vale 1 se dice que hay correlación lineal perfecta positiva.
11. La variable que se predice en el análisis de regresión es la variable dependiente.
12. Un coeficiente de correlación negativo entre la variable dependiente Y y la variable independiente X indica que valores grandes de X están asociados a valores pequeños de Y.
13. Para encontrar las frecuencias relativas marginales de la variable columna en una tabla de contingencia, se debe dividir el total de la columna por el tamaño muestral.
14. La recta de regresión de X sobre Y se obtiene despejando de la recta de regresión de Y sobre X

Cuestiones a completar

1. En el análisis de regresión, la variable que va a ser predicha se denomina variable (dependiente, independiente) _____
2. ¿Cuál de los siguientes valores $(-1, 0'65, 0, 1'06, -0'48, 1)$ no puede ser un coeficiente de correlación?

   _____
3. En general, un coeficiente de correlación lineal de 0'97 indica que las observaciones están (muy próximas a, muy alejadas de) _____ la recta de regresión.

4. El valor del coeficiente de correlación lineal está entre _____ y el del coeficiente de determinación está entre _____ (-1 y $+1$, -1 y 0, 0 y $+1$, $-0'5$ y $+0'5$).
5. Un valor del coeficiente de correlación igual a (0, -1, $+1$) _____ indica que hay relación perfecta negativa entre la variable independiente X y la variable dependiente Y.
6. La relación entre el tiempo de funcionamiento de una máquina y el gasto en reparación es (positiva, negativa) _____
7. El coeficiente de determinación puede obtenerse elevando al cuadrado (la varianza de X, la desviación típica de Y, el coeficiente de correlación lineal) _____
8. Las tablas de frecuencias cruzadas que se usan para describir la asociación entre variables cualitativas se llaman tablas de (contingencia, distribución de probabilidad) _____
9. Al calcular las frecuencias relativas condicionales de la variable columna dada la variable fila de una tabla de contingencia, las celdas se dividen por el (total de la fila, total de la columna, tamaño muestral) _____
10. El coeficiente de correlación entre la variable X = "número de cigüeñas que anidan en diferentes poblaciones" e Y ="número de nacimientos anuales en cada población" será cercano a (1, -1, 0) _____
11. Pon un ejemplo donde se muestre que una elevada correlación (positiva o negativa) no implica causalidad _____

Cuestiones de Elección Múltiple

En los años 20, al famoso mafioso de Chicago *Don Pito Corleone*, que tenía como hobby la estadística, lo acribillaron a tiros justo cuando iba a resolver las preguntas de la 1 a la 8. Aquí tenemos los datos que iba a utilizar, "rematemos su faena":

x_i	-3	-2	1	3	5
y_i	8	4	2	-1	-3

1. El coeficiente de correlación vale: a) 0'56 b) $-0'48$ c) $-1'5$ d) $-0'97$
2. La ecuación de la recta de regresión de Y sobre X es:
   a) $y^* = 1'1 - 0'56x$ b) $y^* = 0'54 + 2x$ c) $y^* = 3 - 1'25x$ d) $y^* = 3'5 - 0'5x$
3. La ecuación de la recta de regresión de X sobre Y es:
   a) $x^* = 2'3 - 0'5y$ b) $x^* = 2'31 - 0'76y$ c) $x^* = 2 - 0'8y$ d) $x^* = 0'67 + 0'34y$
4. La predicción y^* cuando $x = 2$ es: a) 0'5 b) $-7'4$ c) 0'2 d) 1'23
5. La predicción x^* cuando $y = 3$ es: a) 0'55 b) $0'46$ c) 0'04 d) $-0'2$
6. El coeficiente de determinación vale: a) 0'9 b) 0'97 c) 1'5 d) 0'946
7. La calidad del ajuste se puede calificar de:
   a)excelente porque R^2 es casi 1 b) pésima porque los parámetros de la recta son pequeños
   c)pésima porque r es cercano a -1 d) regular porque las rectas Y sobre X y X sobre Y son diferentes

8. La predicción para x = 20 la consideraremos:

   a) muy buena, porque R^2 es cercano a 1 b) incierta, porque estamos extrapolando
   c) muy mala, porque r es casi -1 d) debemos hablar con Rappel

9. El coeficiente de correlación lineal entre dos variables mide:

   a) la dispersión de los valores a la media b) la simetría de las dos variables
   c) la relación lineal entre las dos variables d) la forma de la distribución

La siguiente tabla de contingencia recoge la información sobre la preferencia de los jóvenes sobre conocidas marcas de refrescos según edades e implica las preguntas 10 a 14

COLA/EDAD	Menor de 15	De 15 a 25	De 26 a 35
Coco colo	150	100	200
Pepso	300	125	200
Fanto	300	200	300

10. La frecuencia marginal relativa para Edad *menor de 15* es:

    a) 40% b) 24% c) 20% d) 33.33%

11. La frecuencia relativa para Edad *de 15 a 25* y Cola *Fanto* es:

    a) 10.67% b) 15% c) 5 % d) 10%

12. La frecuencia marginal relativa para Cola *Fanto* es:

    a) 40% b) 37.5% c) 42.67% d) 93.75%

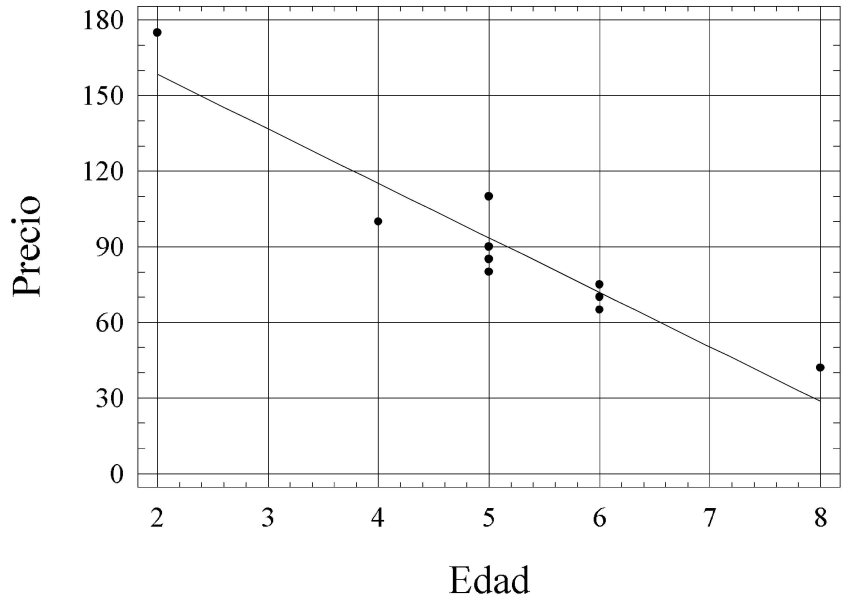
13. La frecuencia condicional relativa de Cola *Pepso* dado que la persona tiene una edad *de 26 a 35 años* es:

    a) 42.67% b) 28.57% c) 32% d) 42.86%

14. La frecuencia condicional relativa de Edad *menor de 15* dado que la persona prefiere la Cola *Coco colo* es:

    a) 20% b) 24% c) 50% d) 33.33%

A continuación se muestra un diagrama de dispersión, para los datos de dos variables: Edad (en años) y Precio (en cientos de euros) para cierta marca de coches usados. Con ella, responderemos a las preguntas de la 15 a la 16:



15. La relación entre ambas variables es:

    a) lineal positiva b) lineal negativa c) no hay relación d) simétrica

16. ¿Cuál de las siguientes ecuaciones sería la que mejor describiría la recta de regresión?

    a) $y^* = 198.9 - 20.5x$ b) $y^* = 198.9 + 20.5x$ c) $y^* = -198.9 - 20.5x$ d) $y^* = -198.9 + 20.5x$

Cuestiones abiertas

1. Desea estudiarse la relación entre la variable Y="Concentración de formaldehido" (medida como la concentración media en ppb) y la variable X="Nivel de aislamiento" (calculada a partir de diversas medidas). Los datos obtenidos fueron:

X	8	1	7	2	3	4	5	8
Y	54	35	50	36	38	38	42	56

   Calcula:

   (a) La recta de regresión de la variable Y sobre la X

   (b) ¿Cómo calificarías la calidad del ajuste? Basa tu respuesta en alguna medida estadística.

   (c) Predice el valor de Y si X=4.5

SOLUCIONES de las cuestiones de autoevaluación del tema 2

Cuestiones V/F

1. V	2. V	3. F	4. V	5. V	6. V
7. F	8. F	9. V	10. V	11. V	12. V
13. V	14. F				

Cuestiones a completar

1. dependiente	2. 1.06	3. muy próximas a	4. -1 y 1; 0 y 1	5. -1
6. positiva	7. el coeficiente de correlación lineal	8. contingencia	9. total de la fila	10. 0

11. X = "temperatura media mensual" Y= "n^o de matrimonios mensuales"; X = "n^o de ventas de chocolate caliente semanales" Y = "n^o de accidentes de esquí semanales"

Cuestiones de elección múltiple

1. d)	2. c)	3. b)	4. a)	5. c)	6. d)
7. a)	8. b)	9. c)	10. a)	11. a)	12. c)
13. b)	14. d)	15. b)	16. a)		

Cuestiones abiertas

1. (a)

La recta de regresión de la variable Y sobre la X es:

$$Y - \overline{y} = r_{xy}\frac{s_y}{s_x}(X - \overline{x}) = \frac{s_{xy}}{s_x^2}(X - \overline{x})$$

El primer paso será realizar los cálculos pertinentes, ayudándose de la calculadora (lo mejor para evitar errores y ganar tiempo, es saber usar las funciones estadísticas de la calculadora). También es buena idea escribir todos los pasos y cálculos realizados.
[**Consejo**: para evitar equivocaciones debidas a un mal redondeo, lo mejor sería trabajar con todos los decimales (¡es la calculadora la que trabaja!), aunque luego sólo se escriban los primeros decimales. Recuerda también comprobar si los resultados obtenidos parecen razonables.]

$\overline{x} = \frac{8+1+7+2+3+4+5+8}{8} = 4.75$,

$s_x^2 = \frac{8^2+1^2+7^2+2^2+3^2+4^2+5^2+8^2- \quad 8\cdot 4.75^2}{7} = 7.357$, $s_x = \sqrt{7.357} = 2.7124$

$\overline{y} = \frac{54+35+50+36+38+38+42+56}{8} = 43.625$,

$s_y^2 = \frac{54^2+35^2+50^2+36^2+38^2+38^2+42^2+56^2- \quad 8\cdot 43.625^2}{7} = 71.41$ $s_y = \sqrt{71.41} = 8.4505$

$s_{xy} = \frac{8\cdot 54+1\cdot 35+7\cdot 50+2\cdot 36+3\cdot 38+4\cdot 38+5\cdot 42+8\cdot 56- \quad 8\cdot 4.75\cdot 43.625}{7} = 22.1786$

$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{22.1786}{2.7124\cdot 8.4505} = 0.9676$

La recta de regresión de la variable Y sobre la X es:

$$Y - 43.625 = 0.9676\frac{8.45}{2.71}(X - 4.75) \quad \text{, o sea,} \qquad Y = 29.3058 + 3.0146X$$

(b) Para determinar la calidad del ajuste, nos basaremos en el coeficiente de determinación:
$R^2 = r_{xy}^2 = 0.9676^2 = 0.936$, como es bastante cercano a 1, consideraremos que el ajuste es bueno.

(c) Para x= 4.5, el valor predicho se obtiene sustituyendo en la recta de regresión:
$29.3058 + 3.0146 \cdot 4.5 = 42.871$

	_____	Nº aciertos de cuestiones Verdadero/Falso
	_____	Nº aciertos de cuestiones a completar
	_____	2 x Nº aciertos de cuestiones elección múltiple
	_____	18 puntos, si la cuestión abierta es correcta
Suma =	_____	**Puntuación final**

Si tu puntuación final está entre:

0 y 30: estás en peligro, acude urgentemente a tutorías
31 y 45: estás en el filo, te puedes cortar si no vas con cuidado
46 y 63: estás por el buen camino, sigue así
64 y 75: muy bien, eres un hacha