SITUACIÓN 1. En un colegio se estudia la relación que puede existir entre la calificación global de los estudiantes al final del curso (Y) y la obtenida en el curso anterior (X). Efectuado un análisis de regresión, con una muestra de estudiantes, la prueba para determinar la significación del modelo de regresión ajustado es la siguiente:

	S.C.	g.l.	M.C.	F	Prob.
Regresión	28,097	1	28,097	V	W
Residual	23,778	6	T		
Total	U				

En primer lugar deducimos los datos que faltan en la tabla de ANOVA.

$$U = 28{,}097 + 23{,}778 = 51{,}875 \qquad T = \frac{23{,}778}{6} = 3{,}963$$

$$V = \frac{28{,}097}{3{,}963} = 7{,}09$$

Completamos la tabla de ANOVA.

	S.C.	g.l.	M.C.	F
Regresión	28,097	1	28,097	7,09
Residual	23,778	6	3,963	
Total	51,875	7	7,411	

1- ¿Cuál fue el tamaño de la muestra elegida?: A) 6; B) 7; **C) 8**.

Sabiendo los grados de libertad de la suma de cuadrados total: $n - 1 = 7 \rightarrow n = 8$

2- ¿Cuál es la varianza muestral insesgada de la calificación de los estudiantes (Y) al final del curso?: A) 51,875; **B) 7,411**; C) 3,397.

La varianza muestral insesgada (cuasivarianza) es igual a:

$$\hat{S}_y^2 = \frac{\sum (Y - \bar{Y})^2}{n - 1}$$

Que podemos calcular con la tabla de ANOVA, dado que es igual a la media cuadrática total.

3- ¿Qué proporción de la varianza de Y es explicada por la varianza de X?: A) 0,736; **B) 0,542**; C) No hay datos suficientes para calcularlo.

La proporción de varianza de Y explicada por la varianza de X es igual al coeficiente de determinación (coeficiente de correlación de Pearson al cuadrado), que podemos deducir de la tabla de ANOVA:

$$r_{xy}^2 = \frac{SC_{reg}}{SC_Y} = \frac{28{,}097}{51{,}875} = 0{,}542$$

4- Con un nivel de confianza del 95%, ¿cuál es el valor de la F teórica, por encima del cual se debería rechazar la hipótesis de que no hay relación lineal significativa entre X e Y?: **A) 5,987**; B) 5,59; C) 6,608.

Consultando las tablas F para 1 y 6 grados de libertad comprobamos que la puntuación 5,987 supera una proporción igual a 5,987.

5- Uno de los supuestos básicos del análisis de regresión simple es: A) los pronósticos y los errores están correlacionados; **B) las distribuciones condicionadas de los errores deben de tener una distribución normal**; C) No deben estar relacionadas la variables predictora y la variable dependiente o criterio.

SITUACIÓN 2. En un colegio se estudia la relación que puede existir entre la calificación global de los estudiantes al final de curso (Y) y la obtenida el curso anterior (X). Efectuado un análisis de regresión, con una muestra de estudiantes, la prueba para determinar la significación del modelo de regresión ajustado es la siguiente:

	S.C.	g.l.	M.C.	F
Regresión	42,734	1	42,734	17,345
Residual	U	V	W	
Total	64,909	10		

Deducimos fácilmente los datos que nos faltan en la tabla:

$$U = 64{,}909 - 42{,}734 = 22{,}175;\ V = 10 - 1 = 9;\ \ W = \frac{U}{V} = \frac{22{,}175}{9} = 2{,}4639$$

La tabla completa queda:

	SC	gl	MC	F
Regresión	42,734	1	42,734	17,345
Residual	22,175	9	2,4639	
Total	64,909	10		

1- ¿Aproximadamente, cuál es la varianza muestral insesgada de los errores del modelo de regresión? A) 2,217; B) 4,273; **C) 2,464**.

La varianza muestral insesgada de los errores del modelo de regresión coincide con la media cuadrática residual.

2- ¿Con un nivel de confianza del 95%, cuál es el valor de la F teórica, por encima del cual se debería rechazar la hipótesis de que hay no relación lineal significativa entre Y y X?; **A) 5,117**; B) 5,59; C) 6,608.

Buscando en las tablas F de Fisher para 1 y 9 grados de libertad, comprobamos que la respuesta correcta es A.

3- ¿Aproximadamente cuál es el coeficiente de correlación de Pearson entre X e Y? A) 0,787; B) 0,97; **C) 0,811**.

El cociente entre la suma de cuadrados de la regresión y la suma de cuadrados total es el coeficiente de determinación, luego su raíz cuadrada será el coeficiente de correlación de Pearson.

$$r_{xy} = \sqrt{\frac{42{,}734}{64{,}909}} = \pm 0{,}811$$

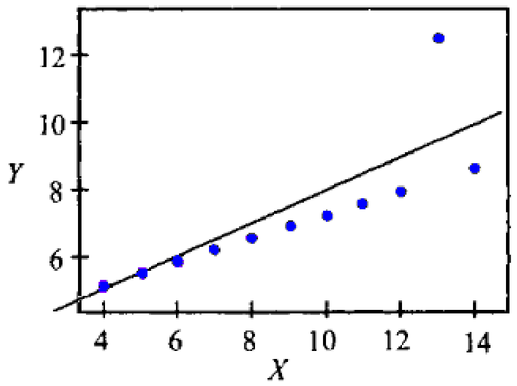
Suponemos que la relación entre la calificación obtenida a final del curso y la obtenida en el curso anterior es directa, por lo que: $r_{xy} = 0{,}811$

4- ¿Cuál es el error típico del modelo de regresión? A) 2,464; **B) 1,57**; C) 1,665.

$$\hat{S}_e = \sqrt{\frac{\sum(Y - Y')^2}{n-2}} = \sqrt{\frac{22{,}175}{9}} = 1{,}57$$

5- Uno de los supuestos básicos del análisis de regresión simple es: **A) los pronósticos y los errores son independientes**; B) las distribuciones condicionadas de los errores deben de tener una distribución uniforme; C) no deben estar relacionadas las variables predictora y la variable dependiente o criterio.

SITUACIÓN 3. Anscombe (1973) presentó cuatro conjuntos de datos artificiales que tenían la misma recta de regresión pero eran claramente diferentes. En la Figura 1 se puede ver uno de estos conjuntos de datos (11 pares de datos) junto con la recta de regresión que mejor ajusta por el criterio de mínimos cuadrados.

	Estadísticos de este conjunto de datos: $S_x = 3{,}316$ $S_y = 2{,}030$ $r_{xy} = 0{,}816$ $\alpha = 0{,}05$

1. El cumplimiento del supuesto de homocedasticidad en regresión no puede evaluarse en este conjunto de datos debido a que: A) el proceso de selección de los datos no ha sido aleatorio; **B) para cada valor de X solo tenemos un valor de Y**; C) no hay normalidad de las distribuciones condicionadas de Y con respecto a X_i.

   La homocedasticidad se refiere a la igualdad de varianzas en Y para los diversos valores de la variable predictora X. No podemos determinar esta igualdad de varianza si solo tenemos un valor de Y para cada valor de X ya que no tenemos una distribución de valores de Y. Luego la opción correcta es la B.
   La opción A es incorrecta ya que la forma en que han sido seleccionados los datos es irrelevante para la determinación de la homocedasticidad. La opción C es incorrecta porque no se refiere a la homocedasticidad sino a la normalidad.

2. La recta de regresión no pasa exactamente por ningún punto porque: **A) existe un valor atípico o extremo**; B) aunque la variabilidad es pequeña en este conjunto de datos, siempre existe una cierta cantidad de ella que impide que el ajuste sea perfecto; C) el ajuste por mínimos cuadrados está mal realizado.

   Debemos observar que todos los puntos dados se sitúan exactamente de forma lineal entre sí. La recta de regresión dibujada ha sido "alejada" debido a la influencia del valor "outlier" o extremo en X = 13. Si no existiera este valor, la recta calculada habría pasado exactamente por el resto de puntos. Luego la opción correcta es la A. La opción B es irrelevante en relación a la pregunta planteada. La opción C es falsa ya que el ajuste por mínimos cuadrados se realizó correctamente por Anscombe.

3. Con los datos proporcionados, la pendiente de la recta de regresión vale aproximadamente: **A) 0'499**; B) 0; C) en la gráfica se ve claramente que la recta corta al eje Y en el punto 5, luego sin hacer cálculos podemos afirmar que esta es la pendiente.

$$B = r_{xy}\frac{S_y}{S_x} = 0{,}816 \cdot \frac{2{,}030}{3{,}316} = 0{,}499542$$

4. Si eliminamos el dato atípico que aparece en la gráfica (X_{10}=13, Y_{10}=12,74), el valor de r_{XY} sería: A) $0{,}499^2$ = 0,249; B) igual al coeficiente de determinación; **C) la unidad porque el ajuste sería perfecto**.

   En este caso, el ajuste sería perfecto ya que todos los puntos se sitúan sobre una línea recta. Si este es el caso, el ajuste sería perfecto (no habría error).

5. Si el valor de R^2 es 0,666 esto significa que: A) también el error típico (σ_ε) es de 0,666; B) la parte de la variabilidad en Y atribuible a otros factores no relacionados linealmente con Y es de 0,666; **C) el porcentaje en que se reduce el error de Y cuando empleamos la recta de regresión para su estimación es del 66,6%**.

   La definición de R^2 (o coeficiente de determinación) es "es la proporción de la variabilidad de la VD que es imputada (o explicada por) la variabilidad de la VI" . Luego la opción correcta es la C.

6. En el modelo de regresión ($Y = B_0 + BX + \varepsilon$), ε representa: A) el efecto del factor manipulado; B) la interacción entre el error y el factor manipulado; **C) el efecto de todas aquellas variables que no hemos manipulado ni controlado.**

   El componente ε del modelo representa el error que todavía queda cuando aplicamos la pendiente y el punto de corte con la ordenada. Por consiguiente, representa la aportación de las variables extrañas al experimento. Opción correcta C.

7. El análisis de la regresión para los datos de la Figura nos indicó que la regresión fue significativa [F(1, 9) = 17,972, MC_{res} = 1,528, p = 0,002]. Luego sabemos que el valor de t con 9 grados de libertad fue aproximadamente: **A) 4,24**; B) 321,12; C) 0.

   Una F y una t se relacionan según la fórmula: $T_n^2 = F_{1,n}$. Luego como sabemos que $F_{1,9}$ = 17,972, el valor de t será: $\sqrt{17{,}972} = 4{,}239$

8. En un análisis de regresión múltiple tendremos tantas pendientes: A) como queramos; B) como nos indiquen los grados de libertad de la fuente de variación "Residual" en la tabla del ANOVA**; C) como variables independientes hayamos introducido en el modelo.**

SITUACIÓN 4. Los siguientes datos relacionan el número de cigarrillos diarios (X) consumidos con el número de radicales libres (Y) localizados en los pulmones de 8 individuos:

Sujeto	X	Y	X Y	X^2	Y^2
1	0	94	0	0	8836
2	10	144	1440	100	20736
3	14	182	2548	196	33124
4	5	120	600	25	14400
5	18	240	4320	324	57600
6	20	234	4680	400	54756
7	30	321	9630	900	103041
8	40	400	16000	1600	160000
Sumatorio	137	1735	39218	3545	452493
Des. Típica	12,24	97,60			
r_{xy} = 0,99					

1- La ecuación de regresión en puntuaciones tipificadas que relaciona el número de cigarrillos con el número de radicales libres viene dada por la expresión: A) $Z_y^{'} = -2{,}1 \cdot Z_X$; B) $Z_y^{'} = 1{,}5 \cdot Z_X$; C) $Z_y^{'} = 0{,}99 \cdot Z_X$

La respuesta correcta es C. El propio enunciado nos proporciona el valor del coeficiente de correlación de Pearson.

2- La ecuación de regresión en puntuaciones tipificadas que relaciona X e Y: **A) tiene un único parámetro**; B) tiene como parámetro el mismo valor que la ecuación de regresión en puntuaciones diferenciales; C) tiene dos parámetros: la pendiente y el punto de corte con la ordenada.

3- La pendiente de la ecuación de regresión para predecir Y a partir de X en puntuaciones directas vale: A) -3,32; **B) 7,92**; C) 0,99.

$$B = r_{xy}\frac{S_y}{S_x} = 0{,}99\frac{97{,}60}{12{,}24} = 7{,}929$$

4- En regresión lineal por mínimos cuadrados se dice que la recta de regresión es una estimación insesgada debido a que: A) se calculan los valores de pendiente y punto de corte con la ordenada tales que hacen mínimo el error al cuadrado; B) cuantifica el incremento que se produce en la estimación de la variable dependiente (Y') cuando la independiente (X) aumenta en una unidad; **C) los valores esperados de los estimadores coinciden con los parámetros que estiman**.

5- Sabiendo que el punto de corte con la ordenada (B_0) vale 81,09, a un sujeto que fume 0 cigarrillos le pronosticaremos un número de radicales libres en pulmones igual a: A) 7,93; **B) 81,09**; C) 0.

6- Conociendo la recta de regresión de mínimos cuadrados de Y sobre X podemos calcular Y'. La correlación de estas puntuaciones predichas (Y') con los errores (Y-Y') vale: **A) 0**; B) lo mismo que la pendiente; C) lo mismo que el coeficiente de alienación.

7- Si queremos poner a prueba la hipótesis $H_0{:}\,\rho = 0$ mediante el estadístico de contraste F obtendríamos un valor aproximado de: A) 594,00; **B) 295,51**; C) 1,00.

$$F = \frac{R^2}{(1-R^2)/(N-2)} = \frac{0{,}99^2}{(1-0{,}99^2)/(8-2)} = 295{,}51$$

8- El error típico de la distribución muestral de la pendiente vale, aproximadamente: A) 0,12; B) 7,97; **C) 0,46**.

$$\sigma_B = \frac{S_y}{S_x}\sqrt{\frac{1-r_{xy}^2}{n-2}} = \frac{97,60}{12,24}\sqrt{\frac{1-0,99^2}{6}} = 0,46$$

9- Si en la Situación 4 incluyéramos también la cantidad de grasas saturadas ingeridas a la semana para predecir el número de radicales libres, el modelo de estimación lineal sería:

A)$\mathbf{Radicales\_Libres' = B_1 Nº\_cigarrillos + B_2 Grasas\_Saturadas + B_0}$;
B)$\mathrm{Radicales\_Libres' = B_1 Nº\_cigarrillos + B_2 Grasas\_Saturadas}$;
C) $\mathrm{Radicales\_Libres' = Nº\_cigarrillos + Grasas\_Saturadas + B_0}$.

10- En un modelo de regresión múltiple, el cálculo de las correlaciones de orden cero: A) implica el cálculo de la regresión entre las variables independientes y, posteriormente, la extracción de los residuos; B) **no nos permite saber qué parte de la varianza de la VD es capaz de explicar independientemente cada una de las VI's**; C) nos permite hallar relaciones bivariadas puras.

SITUACIÓN 5. Un psicólogo escolar considera que la comprensión lectora es un elemento que juega un papel fundamental en el éxito académico de los estudiantes. Para comprobar su hipótesis, selecciona aleatoriamente un grupo de 16 estudiantes de 4º de la ESO a los que aplica un test de comprensión lectora (X). A continuación, y en coordinación con el profesor de ciencias sociales, se les aplica una prueba sobre un tema de historia (Y) que acaban de estudiar, consistiendo dicha prueba en una serie de preguntas sobre un texto básico que tienen que leer y que está relacionado con el tema estudiado.

Sujetos	X	Y
1	6	8
2	4	3
3	7	8
4	3	4
5	5	4
6	9	10
7	7	7
8	4	5
9	2	4
10	5	2
11	8	7
12	9	9
13	10	8
14	6	6
15	3	5
16	1	3

En primer lugar, vamos a calcular varianzas de X e Y, coeficiente de correlación de Pearson y los coeficientes de la recta de regresión.

Sujetos	X	Y	X^2	Y^2	XY
1	6	8	36	64	48
2	4	3	16	9	12
3	7	8	49	64	56
4	3	4	9	16	12
5	5	4	25	16	20
6	9	10	81	100	90
7	7	7	49	49	49
8	4	5	16	25	20
9	2	4	4	16	8
10	5	2	25	4	10
11	8	7	64	49	56
12	9	9	81	81	81
13	10	8	100	64	80
14	6	6	36	36	36
15	3	5	9	25	15
16	1	3	1	9	3
Sumas	89	93	601	627	596

$$S_x^2 = \frac{\sum X^2}{n} - \bar{X}^2 = 6{,}621; \qquad S_y^2 = \frac{\sum Y^2}{n} - \bar{Y}^2 = 5{,}402$$

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} = 0{,}822$$

$$\boldsymbol{r_{xy}^2 = 0{,}822^2 = 0{,}676}$$

$$B = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = r_{xy}\frac{S_y}{S_x} = 0{,}743$$

$$B_0 = \bar{Y} - B\bar{X} = 1{,}681$$

1- Con el fin de analizar la relación entre la comprensión lectora y el rendimiento en la prueba de historia, ¿qué tipo de análisis es el más adecuado para una investigación de este tipo?: A) análisis de varianza con dos factores completamente aletorizados; **B) análisis de correlación/regresión**; C) análisis de la varianza de un factor.

2- ¿Cuál es el coeficiente de correlación de Pearson entre X e Y? A) 0,676; B) 0,324; **C) 0,822**.

3- La recta de regresión que mejor se ajusta a los datos es: Y´= 0,256 + 0,482X; B) Y´= 2,354 + 0,529X; **C) Y´= 1,681 + 0,743X**

4- El valor de la suma de cuadrados de la regresión es: A) 27,991; B) 86,438; **C) 58,447**.

La suma de cuadrados total es: $SC_{Total} = \sum(Y - \bar{Y})^2$

Conociendo S_y^2 *podemos calcularla:* $SC_{Total} = n \cdot S_y^2 = 16 \cdot 5{,}402 = 86{,}4375$

También podemos calcular SC_{Total} *con la fórmula utilizada en los temas de ANOVA:*

$$SC_{Total} = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = [Y] - [T] = 627 - \frac{93^2}{16} = 86{,}4375$$

Dado que el coeficiente de determinación es igual a: $r_{xy}^2 = 0{,}676$*, la suma de cuadrados debida a la regresión es:* $SC_{Reg} = r_{xy}^2 \cdot SC_{Total} = 0{,}676 \cdot 86{,}4375 = 58{,}447$

Completamos la tabla de ANOVA para la regresión lineal:

FV	SC	gl	MC	F
Regres.	58,447	1	58,447	29,233
Residual	27,991	14	1,999	
Total	86,438	15		

5- La proporción de variabilidad de Y explicada por la variabilidad en X, es: **A) 0,676**; B) 0,822; C) 0,324.

6- El valor del estadístico F para el contraste de la regresión es: A) 31,578; **B) 29,233**; C) 25,821.

7- La probabilidad asociada al estadístico F para el contraste de la regresión es 0,0001, de ello puede deducirse que: A) para un nivel de significación de 0,01 NO existe relación entre X e Y; **B) para un nivel de significación de 0,01 existe relación entre X e Y**; C) para un nivel de significación de 0,05 existe relación entre X e Y, pero no para un nivel de significación de 0,01.

8- El valor más aproximado del estadístico de contraste para la pendiente de regresión es: **A) 5,407**; B) 1,996; C) 0,743.

$$F = T^2 \rightarrow T = \sqrt{F} = \sqrt{29{,}233} = 5{,}407$$

9- Para un nivel de significación de 0,05, los límites del intervalo de confianza para la pendiente de la regresión son: A) (-0,125; 3,487); B) (-0,125; 1,037); **C) (0,448; 1,037)**

$$E_{max} = t_{n-2;1-\alpha/2}\sigma_B = t_{n-2;1-\alpha/2}\frac{S_y}{S_x}\sqrt{\frac{1-r_{xy}^2}{n-2}} = 2{,}145\frac{2{,}324}{2{,}573}\sqrt{\frac{1-0{,}676}{16-2}} = 0{,}2947$$

$$0{,}743 \pm 0{,}2947 \rightarrow \begin{cases} 1{,}037 \\ 0{,}448 \end{cases}$$

10- Teniendo en cuenta los límites del intervalo de confianza obtenido para la pendiente de la regresión, podemos decir que, a un nivel de significación de 0,05: **A) se acepta la hipótesis alternativa según la cual existe relación entre X e Y, ya que el intervalo de confianza no incluye el valor cero**; B) se acepta la hipótesis nula según la cual no existe relación entre X e Y, ya que el intervalo de confianza incluye el valor cero; C) se acepta la hipótesis nula según la cual no existe relación entre X e Y, ya que el intervalo de confianza no incluye el valor cero.

SITUACIÓN 6. El artículo "Efecto de la temperatura sobre el pH de la leche" (*Journal of Dairy Research*, 1988, 277-280) informó de un estudio en el que se utilizaron 5 temperaturas diferentes (medidas en grados centígrados) en condiciones experimentales controladas en una granja y se midió el pH de la leche obtenida para cada temperatura. La ecuación de regresión calculada fue $Y' = 0{,}65 + 0{,}22X$. El objetivo era predecir el pH de la leche (Y) a partir de la temperatura (X). Utilice un $\alpha = 0{,}05$. Además se sabe que: $S_Y = 1{,}87$ y $S_X = 7{,}90$.

1- La ecuación genérica de regresión de Y sobre X, $Y = a + b\,X + \epsilon$, es una relación: **A) probabilística ya que depende de la cantidad aleatoria ϵ**; B) funcional cúbica ya que existen tres componentes aditivos en la ecuación; C) determinista ya que el valor de Y está completamente determinado por el valor de X.

2- La distribución del error en la ecuación de regresión lineal simple $Y = a + b\,X + \epsilon$ tiene una media igual a: A) $\bar{Y}$; B) $\bar{X}$; **C) 0**.

3- En regresión lineal simple, se asume que la desviación típica del error en la ecuación $Y = a + b\,X + \epsilon$: A) es igual a la unidad; **B) es la misma para cada valor particular de X**; C) es igual a 0.

4- El estadístico para poner a prueba la significación del coeficiente de correlación entre el pH y la temperatura vale aproximadamente: A) 3,91; **B) 4,36**; C) 1,32.

$$B = r_{xy}\frac{S_y}{S_x} \rightarrow 0{,}22 = r_{xy}\frac{1{,}87}{7{,}90} \rightarrow r_{xy} = 0{,}929$$

$$t_{n-2} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0{,}929\sqrt{5-2}}{\sqrt{1-0{,}929^2}} = 4{,}36$$

5- El estadístico de la Tabla del Anova para poner a prueba si la Regresión como fuente de variación es significativa vale aproximadamente: **A) 19,0**; B) 1,2; C) 2,1.

$$t^2 = F \rightarrow F = 4{,}36^2 = 19{,}01$$

6- El contraste de hipótesis para determinar si la pendiente de la recta de regresión es significativa es: **A) bilateral**; B) unilateral derecho; C) unilateral izquierdo.

7- El intervalo de confianza para contrastar la pendiente vale: **A) (0’38, 0’06)**; B) (-0’15, +0’15); C) (-0’32, -0’15).

$$B \pm t_{1-\alpha/2,n-2}\left(\frac{S_y}{S_x}\sqrt{\frac{1-r_{xy}^2}{n-2}}\right) \rightarrow 0{,}22 \pm 3{,}182 \cdot \frac{1{,}87}{7{,}90}\sqrt{\frac{1-0{,}929^2}{5-2}} \rightarrow \begin{cases} 0{,}38 \\ 0{,}06 \end{cases}$$

8- El coeficiente de correlación semiparcial entre el pH y la temperatura: A) coincide con el coeficiente de regresión estandarizado; **B) no tiene sentido calcularlo ya que estamos en una situación de regresión simple**; C) coincide con el coeficiente de determinación.