

TEMA Nº 8 → ANÁLISIS DE REGRESIÓN LINEAL SIMPLE Y MÚLTIPLE

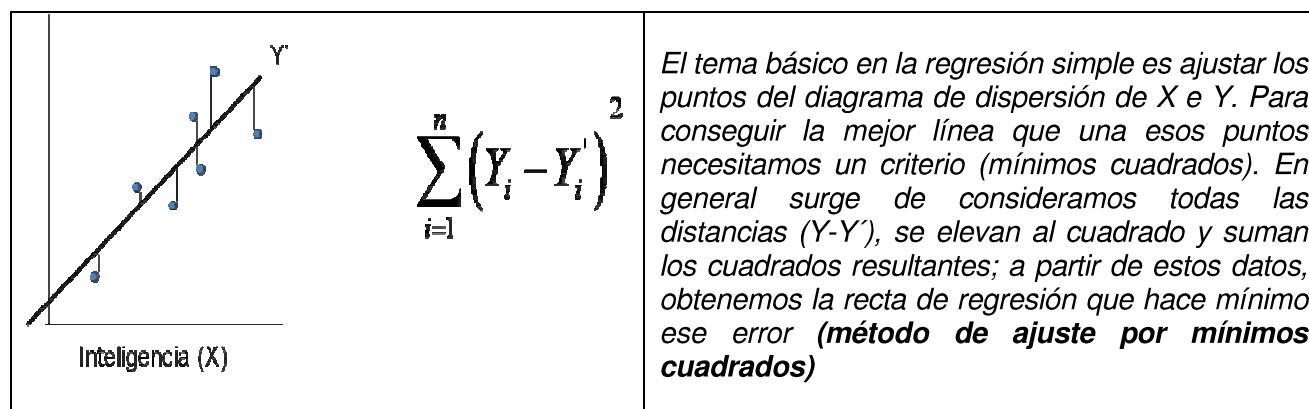
ANÁLISIS DE REGRESIÓN SIMPLE (ARS)

Regresión lineal simple → Cuando una variable (VI) aporta información sobre otra variable (VD) decimos que están relacionadas. El **ARS** estudia la relación entre dos variables (X e Y) para hacer predicciones (sabiendo el comportamiento de la VI, predecimos el de la VD). Ambas variables son cuantitativas (medidas en una escala de intervalo o razón) y mantienen una relación lineal según la expresión de la ecuación de regresión lineal:

$Y' = BX + B_0$	$y' = Bx$	$Z'_y = r_{xy}Z_x$
Puntuaciones directas	Puntuaciones diferenciales	Puntuaciones típicas

La estimación será más precisa y ajustada cuanto mayor sea la relación entre las variables (depende del valor del coeficiente de correlación de Pearson). Entre las puntuaciones que predecimos con la recta de regresión (Y') y los valores reales de la VD (Y) existe una distancia ($Y - Y'$) que se denomina error de estimación o predicción.

Siendo la media de los valores pronosticados ($\bar{Y}'$) = la media de los valores observados ($\bar{Y}$)



COEFICIENTES DE REGRESIÓN LINEAL SIMPLE (B y B_0) →

Normalmente los **coeficientes B y B_0 son desconocidos**. El objetivo será estimarlos por medio de la información contenida en una muestra aleatoria seleccionada de la población.

(B_0) recibe el nombre de “ordenada en el origen” (es el valor de Y pronosticado cuando $X = 0$)

(B) recibe el nombre de “pendiente” de la recta (representa el cambio que experimenta Y cuando X aumenta una unidad)

MODELO → $Y_i = B_0 + Bx + \epsilon$ (Y estimada = ordenada en el origen + pendiente + Error)

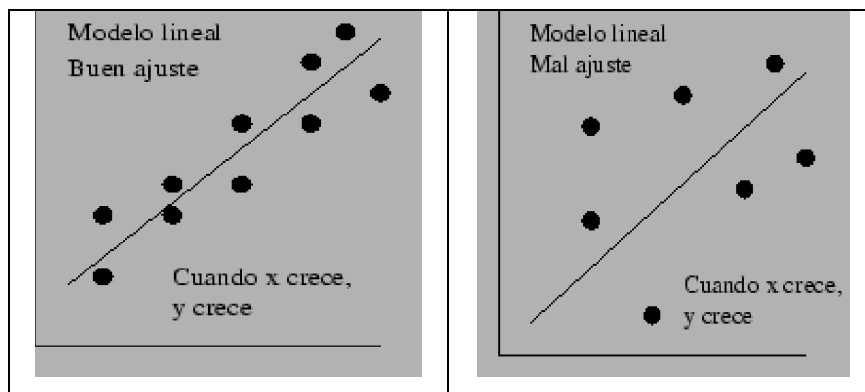
Cálculo de los coeficientes → $B = (n \sum XY) - (\sum X \sum Y) / n \sum X^2 - (\sum X)^2 //$ $B_0 = \bar{Y} - B \bar{X}$
 $B = r_{xy} (S_Y / S_X)$

Para que sean válidas las **inferencias** que se hacen **sobre la VD** utilizando la recta de regresión, se deben cumplir cuatro supuestos básicos (similares a las enunciadas para el ANOVA):

- 1.- **Independencia de las observaciones** (la selección de la muestra debe ser aleatoria)
- 2.- **Homocedasticidad** (las varianzas de las distribuciones de los errores deben ser iguales)
- 3.- **Normalidad de las distribuciones**
- 4.- **Independencia** entre los valores estimados (Y') y los errores de estimación (ϵ) → ($r_{Y' \epsilon} = 0$)

BONDAD DE AJUSTE DE LA RECTA DE REGRESIÓN

Alude a cómo es de explicativa la recta de regresión respecto a los datos sobre los que se ha ajustado. Para analizar la **bondad de ajuste**, se analiza la varianza de la VD antes y después de ajustar la recta de regresión.



Cuando hay una relación lineal entre dos variables, la varianza de la VD se puede descomponer en dos varianzas: la de los pronósticos y la de los errores $\rightarrow S^2_Y = S^2_{Y'} + S^2_{\epsilon}$

$S^2_{Y'} / S^2_Y = SC_{REGRESIÓN} / SC_Y = R^2_{XY}$ R^2_{XY} (Coeficiente de Determinación)	Proporción de varianza de la VD explicada por la varianza de la VI
$S^2_{\epsilon} / S^2_Y = SC_{RESIDUOS} / SC_Y = 1 - R^2_{XY}$ $1 - R^2_{XY}$ (Coeficiente de Alienación)	Proporción de varianza de la VD no explicada por la varianza de la VI

El coeficiente de Determinación (R^2_{XY}) es susceptible de varias interpretaciones:

- Como medida de la **mejora introducida por la ecuación de regresión** (proporción en que se reduce el error de la VD cuando utilizamos la recta de regresión para estimarla)
- Como medida de la **bondad de ajuste de la recta a los puntos** (cuando no podemos asumir el supuesto de normalidad de los errores). Cuando podemos asumir el supuesto de normalidad de los errores, la **varianza de los errores S^2_{ϵ}** puede interpretarse como una medida de la bondad de ajuste de la regresión (a menor varianza de error, mejor ajuste)

$$S_{\epsilon} = S_Y \sqrt{1 - R^2_{XY}} \quad (\text{error típico})$$

INFERENCIAS SOBRE LA REGRESIÓN

Contraste de hipótesis sobre la correlación $\rightarrow H_0: \rho_{XY} = 0$ y $H_1: \rho_{XY} \neq 0$

La hipótesis nula plantea que no hay relación entre la VD y la VI. Para dilucidar la significación se pueden utilizar dos procedimientos equivalentes (relación entre **t** y **F** $\rightarrow t^2_n = F_{1,n}$):

- ❖ Una **razón F** (cociente entre medias cuadráticas) $\rightarrow$ Coeficiente de determinación dividido entre el Coeficiente de alienación con sus respectivos grados de libertad.

$$F = \frac{R^2}{(1 - R^2) / (N - 2)}$$

- ❖ Mediante el estadístico **t**

$$T = \frac{r_{XY} \sqrt{n - 2}}{\sqrt{1 - r^2_{XY}}}$$

Intervalo de Confianza

$$Z' \pm Z_{\alpha/2} \sqrt{1 / (n - 3)}$$

PROBLEMA EJEMPLO →

En un centro de Psicología clínica se ha encontrado que, en una muestra aleatoria simple de 45 pacientes fumadores, la ingesta media diaria de chicles de nicotina (n) y el consumo medio diario de cigarrillos (c) presentan una correlación de $r_{nc} = (-0,89)$. Se cumplen los supuestos del modelo de correlación lineal ¿Es significativa la relación lineal entre la ingestión media de chicles de nicotina y el consumo medio de cigarrillos? Considere un nivel de confianza del 99 %.

Hipótesis → $H_0: \rho_{NC} = 0$ y $H_1: \rho_{NC} \neq 0$

Estadísticos de contraste (Los resultados confirman la equivalencia) → $[t^2 (-12,8)^2 = F (164)]$

$$F = R^2 / [(1 - R^2) / (N - 2)] \rightarrow F = (0,7921) / (0,00483) = 164$$

$$t = (-0,89) \sqrt{45 - 2} / \sqrt{(1 - (-0,89)^2)} = (-5,84) / 0,456 = (-12,80)$$

Decisión → (se comprueba que

La F teórica = $0,99 F_{1, 43gl} = 7,1 < 164$; por tanto rechazamos H_0 (el coeficiente de correlación es significativo)

La t teórica = $0,99 t_{43gl} = (-2,42) < (-12,80)$; por tanto rechazamos H_0 (el coeficiente de correlación es significativo)

Interpretación → Para un nivel de confianza del 99%, la correlación es significativamente distinta de cero, por lo que debe pensarse que existe relación lineal significativa entre el consumo medio de cigarrillos y la ingesta media de chicles de nicotina.

Intervalo de confianza → $Z' \pm Z_{\alpha/2} \sqrt{1 / (n - 3)}$

La distribución muestral del coeficiente de correlación no es simétrica, lo que supone que los límites del intervalo de confianza no son equidistantes del valor de r_{xy} (por esta razón se aplica la transformación Z' de Fisher) → (las tablas transforman directamente la correlación r en Z')

$$Z' = \frac{1}{2} [\ln(1+r) - \ln(1-r)] \rightarrow Z' = \frac{1}{2} [\ln(1+0,89) - \ln(1-0,89)] = (0,636) - (-2,21) / 2 = 1,42$$

$$\sigma_{Z'} = \sqrt{1 / (n-3)} \rightarrow \sigma_{Z'} = \sqrt{1 / 43} \rightarrow 0,152$$

$$Z' \pm Z_{\alpha/2} \sqrt{1 / (n - 3)} \rightarrow 1,42 \pm (2,42 \cdot 0,152) \rightarrow (1,05 \text{ y } 1,79) \rightarrow \text{transformados en } r \text{ (0,78 y 0,945)}$$

INFERENCIAS SOBRE LA REGRESIÓN

Contraste de hipótesis para los parámetros de la regresión (β y β_0) → $Y' = \beta X + \beta_0$

Se contrasta si los coeficientes de regresión: pendiente (B) e intercepto (B_0) son distintos de cero

Hipótesis estadísticas $H_0: \beta = 0$ (pendiente)
 $H_1: \beta \neq 0$

$H_0: \beta_0 = \beta_0$ (intercepto)
 $H_1: \beta_0 \neq \beta_0$

ESTADÍSTICOS DE CONTRASTE Y ERRORES TÍPICOS	$T = \beta - 0 / \sigma_\beta$	$\sigma_\beta = S_Y / S_X (\sqrt{1 - r^2_{XY} / n-2})$
	$T = \beta_0 - 0 / \sigma_{\beta_0}$	$\sigma_{\beta_0} = \sigma_\epsilon (\sqrt{1/n + \bar{X}^2 / (n-1) S_X^2})$
INTERVALOS DE CONFIANZA: B y B_0 (Coeficientes Regresión) Y' (Valores estimados)	$IC(B) = B \pm T_{n-2; 1-\alpha/2} \sigma_B$	
	$IC(B_0) = B_0 \pm T_{n-2; 1-\alpha/2} \sigma_{B_0}$	
	$IC(Y') = Y' \pm T_{n-2; 1-\alpha/2} \cdot (\sigma_\epsilon \sqrt{1/n + (X - \bar{X})^2 / (n-1) S_X^2})$	

PROBLEMA EJEMPLO →

Consideremos, a modo de ejemplo, los siguientes datos referidos a $n = 25$ sujetos:

ESTADÍSTICOS DESCRIPTIVOS			
	X_1	X_2	Y
Sumatorio (Σ)	882	239	2354
Media	35'28	9'56	94'16
esviación típica	9'7105	2'0833	10'5423
varianza	94'2933	4'34	111'14

MATRIZ DE CORRELACIONES DE ORDEN CERO			
	X_1	X_2	Y
X_1		(- 0'231)	0'436
X_2			0'504
Y			

1.- Ecuación de regresión (predecir el comportamiento de la variable Y a partir de la variable X_1):

Cálculo de los coeficientes de regresión lineal (consideramos los datos de X_1 e Y):

$$B = r_{XY}(S_Y / S_X) = 0'436 (10'5423 / 9'7105) = 0'473$$

$$B_0 = \bar{Y} - B \bar{X} = 94'16 - (0,473 \cdot 35'28) = 77'465$$

$$\text{Ecuación de regresión} \rightarrow Y' = 77'465 + 0'473 X$$

2.- Contrastar hipótesis de que los parámetros de la regresión son nulos (no son significativos)

Previamente calculamos el **error típico** $\rightarrow \sigma_e = S_e = S_Y \sqrt{1 - R^2_{XY}}$

$$S_e = 10'54 \sqrt{1 - 0'436^2} \rightarrow 10'54 \cdot 0'9 = 9'486$$

El intercepto (ordenada en el origen) es nulo ($\beta_0 = 0$).

Hipótesis estadísticas $\rightarrow H_0: \beta_0 = 0$ y $H_1: \beta_0 \neq 0$

$$\text{Estadístico de Contraste} \rightarrow T = B_0 - 0 / \sigma_{B_0} \rightarrow T = (77'465 - 0) / 7'035 = 11'01$$

$$\sigma_{B_0} = \sigma_e (\sqrt{1/n + \bar{X}^2 / (n-1) S_X^2}) = \sigma_{B_0} = 9'486 (\sqrt{1/25 + 35'28^2 / 24 \cdot 94'29}) = 7'035$$

Distribución $T_{\alpha/2 = 0,025; n-2}$ y $T_{1-\alpha/2 = 0,975; n-2} \rightarrow \pm 2'069 < 11,01$ (Rechazamos H_0)

$$\text{Intervalo de confianza} \rightarrow IC(B_0) = B_0 \pm T_{n-2; 1-\alpha/2} \sigma_{B_0} \rightarrow 77'465 \pm (2'069 \cdot 7'035) = (92'02 \text{ y } 62'915)$$

El intercepto no es nulo ($\beta_0 \neq 0$) $\rightarrow 0$ no está en el intervalo de confianza.

La pendiente de la recta de regresión es nula $\rightarrow \beta = 0$

Hipótesis estadísticas $\rightarrow H_0: \beta = 0$ $H_1: \beta \neq 0$

$$\text{Estadístico de Contraste} \rightarrow T = \beta - 0 / \sigma_\beta \rightarrow T = (0,473 - 0) / 0'197 = 2'401$$

$$\sigma_\beta = S_Y / S_X (\sqrt{1 - r^2_{XY} / n-2}) \rightarrow \sigma_\beta = (10'54 / 9'71) \sqrt{0'035} = 0'197$$

Distribución $T_{\alpha/2=0,025; n-2}$ y $T_{1-\alpha/2=0,975; n-2} \rightarrow \pm 2'069 < 2'401$ (Rechazamos H_0)

Intervalo confianza $\rightarrow IC(B) = B \pm T_{n-2; 1-\alpha/2} \sigma_B \rightarrow 0,473 \pm (2'069 \cdot 0'197) = (0'066 \text{ y } 0'88) \rightarrow$
 Pendiente no nula ($\beta \neq 0$) $\rightarrow 0$ no está en el intervalo de confianza.

3.- Pronóstico mediante la ecuación de regresión de la puntuación que ha obtenido un sujeto en Y, cuando su puntuación en $X_1 = 31$ y calculamos el intervalo de confianza para este pronóstico individual (para $\alpha = 0,05$). Utilizamos la ecuación de regresión simple: $Y' = B_0 + B X \rightarrow Y' = 77'465 + 0'473 X$

Pronóstico $\rightarrow Y' = 77'465 + 0'473 (31) \rightarrow Y' = 92'128$

Para calcular el **intervalo de confianza** de este pronóstico individual $Y' = 92'128$, necesitamos calcular el error típico del pronóstico con σ_e^2 desconocida y estimada mediante S_e^2 .

$$IC(Y') = Y' \pm T_{n-2; 1-\alpha/2} \cdot (\sigma_e \sqrt{1/n + (X - \bar{X})^2 / (n-1) S_x^2})$$

$$Y' \pm T_{n-2; \alpha/2} \sigma_Y \rightarrow 92'128 \pm 2'069 \cdot 9'486 \rightarrow (72,502 \text{ y } 111'754)$$

$$\text{Donde } \rightarrow T_{\alpha/2=0,025; n-2} \text{ y } T_{1-\alpha/2=0,975; n-2} \rightarrow \pm 2'069 \text{ y } \sum (X - \bar{X})^2 = 2262'96$$

$$\sigma_Y = 9'486 \sqrt{(0'04 + 2262'96) / (24 \cdot 94'29)} \rightarrow \sigma_Y = (9'486) \cdot (1) = 9'486$$

Precisión de las estimaciones de los parámetros (ρ , β y β_0) según el tamaño muestral:

La amplitud del intervalo es función inversa del tamaño muestral. El error típico de la distribución muestral de los parámetros (factor que multiplica el valor t en la construcción del intervalo de confianza) está aproximadamente en una proporción inversa a $\sqrt{n}$ (cuando se cuadruplica el tamaño muestral el error típico se divide por la mitad). Así, si se pretende reducir el error típico para un n dado $\rightarrow n^* = n (\sigma_{\beta_0} / \sigma_{\beta_0}^*)^2$

Problema ejemplo: Se desea pasar de un error típico ($\sigma_{\beta_0} = 0'78$) para un $n = 50$ a un error típico menor ($\sigma_{\beta_0}^* = 0'3$) $\rightarrow$ La muestra necesaria sería: $n^* = n (\sigma_{\beta_0} / \sigma_{\beta_0}^*)^2 \rightarrow n^* = 50 (0'78 / 0'3)^2 = 338$.

ANÁLISIS DE REGRESIÓN MÚLTIPLE (ARM)

El modelo de la Regresión Múltiple incorpora dos o más variables independientes o predictoras (X_1, X_2) actuando sobre una variable dependiente o criterio (Y). Estudiaremos el caso (el modelo) que considera fijos los valores de las X_i y aleatorios únicamente los de la variable Y .

Ecuación de regresión múltiple $\rightarrow$ (Plano de Regresión)

$$Y' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \text{ (Modelo)} \quad Y' = B_0 + B_1 X_1 + B_2 X_2 + \dots \text{ (Caso concreto)}$$

Ecuación de regresión en puntuaciones típicas $\rightarrow z'_Y = \beta_1 z_1 + \beta_2 z_2$

Procedimientos de cálculo $\rightarrow B_0$ (intercepto con el eje de la Y cuando X_1 y X_2 valen 0; altura del plano de la regresión) y B_1, B_2 (coeficientes de regresión parcial o pesos de cada variable)

$$\text{Coeficiente } B_0 \rightarrow B_0 = \bar{Y} - B_1 \bar{X}_1 - B_2 \bar{X}_2$$

Coeficientes B $\rightarrow$ A partir de los coeficientes de regresión parcial estandarizados (β)

$B_1 = \beta_1 (S_Y / S_1)$	$\beta_1 = r_{Y1} - r_{Y2} \cdot r_{12} / 1 - r_{12}^2$
$B_2 = \beta_2 (S_Y / S_2)$	$\beta_2 = r_{Y2} - r_{Y1} \cdot r_{12} / 1 - r_{12}^2$
$S_1 \text{ y } S_2 = \text{Desviaciones típicas de las variables } X_1 \text{ y } X_2$	

BONDAD DE AJUSTE DEL PLANO DE REGRESIÓN

Correlación y determinación lineal múltiple →

El coeficiente de correlación múltiple ($R_{Y,12}$) se interpreta como la relación de Y con X_1 y X_2 consideradas conjuntamente.

El coeficiente de determinación múltiple ($R^2_{Y,12}$) se interpreta como → % de la varianza de Y que se debe a la variación conjunta de X_1 y X_2 y como índice de bondad de ajuste al plano de regresión.

Procedimientos de cálculo →

A PARTIR DE LAS CORRELACIONES	$R_{Y,12} = \sqrt{(r^2_{Y1} + r^2_{Y2} - 2 r_{12} r_{1Y} r_{2Y}) / (1 - r^2_{12})}$
A PARTIR DE LOS COEFICIENTES DE REGRESIÓN ESTANDARIZADOS	$R_{Y,12} = \sqrt{\beta_1 r_{Y1} + \beta_2 r_{Y2}}$
A PARTIR DE LAS VARIANZAS	$R^2_{Y,12} = S^2_{Y,12} / S^2_Y = (R_{Y,12})^2$ Coeficiente determinación = (Var pronósticos / Varianza de la VD)
COEFICIENTE DE DETERMINACIÓN MÚLTIPLE CORREGIDO (R^2 AJUSTADO)	$R^2_{Y,12} = 1 - [(1 - R^2_{Y,12}) \cdot (n - 1 / n - p - 1)]$ ($p = n^o$ de VI o predictoras)
VARIANZA DE LOS ERRORES Y ERROR TÍPICO DE ESTIMACIÓN (Al aumentar el coeficiente de correlación múltiple disminuye error típico)	$S^2_{ERROR} = (1 - R^2_{Y,12}) S^2_Y$ $\sigma_e = \sqrt{\sum (Y - \hat{Y})^2 / n - p - 1}$

INFERENCIAS SOBRE LA REGRESIÓN MÚLTIPLE

Significación del coeficiente de correlación múltiple → $H_0: R^2 = 0$ y $H_1: R^2 \neq 0$

$$F = R^2 (n - p - 1) / (1 - R^2) p \quad \text{Distribución } F_{p, Y(n-p-1) gl}$$

Intervalo de confianza → $IC (R^2 \pm t_{n-2; 1-\alpha/2} \cdot \sigma_{R^2} = \text{Límites superior e inferior})$

Donde $\sigma_{R^2} = \sqrt{4R^2(1-R^2)^2(n-p-1)^2 / (n^2 - 1)(n + 3)}$

PROBLEMA EJEMPLO →

Aludimos, a modo de ejemplo, al planteamiento utilizado en el análisis de regresión simple. Disponemos de los siguientes datos referidos a $n = 25$ sujetos:

ESTADÍSTICOS DESCRIPTIVOS			
	X_1	X_2	Y
Sumatorio (Σ)	882	239	2354
Media	35'28	9'56	94'16
Desviación típica	9'7105	2'0833	10'5423
varianza	94'2933	4'34	111'14

MATRIZ DE CORRELACIONES DE ORDEN CERO			
	X_1	X_2	Y
X_1		(- 0'231)	0'436
X_2			0'504
Y			

1.- Pronosticar la puntuación en la variable **Y (VD)** de un sujeto con ($X_1=31$) y ($X_2=9$). Para ello, construimos la ecuación de regresión múltiple:

Coeficientes de regresión parcial →

$B_1 = \beta_1 (S_Y / S_1)$	$\beta_1 = r_{Y1} - r_{Y2} \cdot r_{12} / 1 - r_{12}^2$
$B_1 = 0'583 (10'54 / 9'71) = 0'6328$	$\beta_1 = 0'436 - [(0'504) \cdot (-0'231)] / 0'9466 = 0'583$
$B_2 = \beta_2 (S_Y / S_2)$	$\beta_2 = r_{Y2} - r_{Y1} \cdot r_{12} / 1 - r_{12}^2$
$B_2 = 0'639 (10'54 / 2'08) = 3'238$	$\beta_2 = (0'504) - (0'436) \cdot (-0'231) / 0'9466 = 0'639$
$B_0 = \bar{Y} - B_1 \bar{X}_1 - B_2 \bar{X}_2 \rightarrow B_0 = 94'16 - 0'6328 (35'28) - 3'238 (9'56) = 40'88$	

Ecuación de regresión múltiple → $Y' = 40'88 + (0'6328) X_1 + (3'238) X_2$

Pronóstico → $Y' = 40'88 + (0'6328) (31) + (3'238) (9) \rightarrow 89'64$

2.- Obtener el coeficiente de determinación múltiple y el coeficiente de correlación múltiple:

Cálculo a partir de las correlaciones →

$$R_{Y, 12} = \sqrt{(r_{Y1}^2 + r_{Y2}^2 - 2 r_{12} r_{Y1} r_{Y2}) / (1 - r_{12}^2)} = \sqrt{(0'436^2 + 0'594^2) - 2 (0'436)(0'504)(-0'231)} / 0,986$$

$R_{Y, 12} = 0'759 \rightarrow$ Coeficiente de determinación múltiple → $R^2_{Y, 12} = (0,759)^2 = 0,576$ (Un 57'6% de la varianza de Y se debe a la variación conjunta de X_1 y X_2).

Cálculo a partir de los coeficientes de regresión estandarizados →

$$\beta_1 = r_{Y1} - r_{Y2} \cdot r_{12} / 1 - r_{12}^2 \rightarrow \beta_1 = 0'436 - [(0'504) \cdot (-0'231)] / 0'9466 = 0'583$$

$$\beta_2 = r_{Y2} - r_{Y1} \cdot r_{12} / 1 - r_{12}^2 \rightarrow \beta_2 = (0'504) - (0'436) \cdot (-0'231) / 0'9466 = 0'639$$

$$R_{Y, 12} = \sqrt{\beta_1 r_{Y1} + \beta_2 r_{Y2}} \rightarrow R_{Y, 12} = \sqrt{(0'583 \cdot 0'436) + (0'639 \cdot 0'504)} = 0'759$$

Coeficiente de determinación ajustado →

$$R^2_{Y, 12} = 1 - [(1 - R^2_{Y, 12}) \cdot (n - 1 / n - p - 1)] \rightarrow R^2_{Y, 12} = 1 - [(0'4239) \cdot (24 / (25-2-1))] = 0'538$$

3.- Obtener la Varianza de los errores y Error típico →

$$\text{Varianza de los errores} \rightarrow S^2_{\text{ERROR}} = (1 - R^2_{Y, 12}) S^2_Y \rightarrow S^2_{\text{ERROR}} = (1 - 0'759^2) 111'14 = 47'11$$

$$\text{Error Típico} \rightarrow \sigma_e = \sqrt{\sum (Y - Y')^2 / n - p - 1}; \sigma_e = S_{\text{ERROR}} \sqrt{1130'6 / 22} = 7'17$$

$$\text{Donde: } S^2_{\text{ERROR}} = \sum (Y - Y')^2 / n - 1 \rightarrow \sum (Y - Y')^2 = S^2_{\text{ERROR}} (n - 1) \rightarrow$$

$$\sum (Y - Y')^2 = 47'11 \cdot 24 = 1130'6$$

4.- Averiguar la significación de R^2 con ($\alpha = 0'05$) → $F = R^2 (n - p - 1) / (1 - R^2) p$

$$F = 0'759^2 (25-2-1) / (1-0'759^2) 2 = 14'95 \rightarrow \text{Como } 14'95 > F_{0'05; 2 \text{ y } 22} = 3'44 \text{ (Aceptamos } H_1: R^2 \neq 0)$$

Intervalo de confianza $\rightarrow (R^2 \pm t_{n-2; 1-\alpha/2} \cdot \sigma_{R^2} \rightarrow 0'576 \pm (2'07 \cdot 0'107) = (0'356 \text{ y } 0'796)$

$$\sigma_{R^2} = \sqrt{4R^2(1-R^2)^2(n-p-1)^2 / (n^2 - 1)(n + 3)} \rightarrow \sigma_{R^2} = \sqrt{(200,72 / 17472)} = 0'107$$

5.- Averiguar la significación de las **pendientes de la ecuación de regresión** (se averigua el intervalo de confianza (IC) para cada pendiente y se comprueba si dentro del intervalo de confianza está el valor cero (si el **0 no está** se rechaza la H_0)

Hipótesis estadísticas $\rightarrow H_0: \beta_1 = 0 \text{ y } H_1: \beta_1 \neq 0 \quad H_0: \beta_2 = 0 \text{ y } H_1: \beta_2 \neq 0$

Estadístico de Contraste $\rightarrow IC(B) = B \pm T_{n-2; 1-\alpha/2} \sigma_B$

Donde $\rightarrow \sigma_B = S_Y / S_X \sqrt{(1 / (1 - R^2_i)) \sqrt{(1 - R^2_{Y.12}) / (n-p-1)}}$

Donde $\rightarrow R^2_i$ es la correlación entre las variables predictoras al cuadrado $r^2_{12} = (-0'231)^2$

$$\sigma_{\beta_1} = (10'54 / 9'71) \sqrt{1'056} \sqrt{0'019} = 0'155 \quad // \quad \sigma_{\beta_2} = (10'54 / 2'08) \sqrt{1'056} \sqrt{0'019} = 0'724$$

Distribución $T_{\alpha/2} = 0,025; n-p-1$ y $T_{1-\alpha/2} = 0,975; n-p-1 \rightarrow \pm 2'074 < \text{(Rechazamos } H_0)$

Intervalo confianza $\rightarrow IC(B) = B \pm T_{n-2; 1-\alpha/2} \sigma_B \rightarrow 0,473 \pm (2'069 \cdot 0'197) = (0'066 \text{ y } 0'88) \rightarrow$
Pendiente no nula ($\beta \neq 0$) $\rightarrow 0$ no está en el intervalo de confianza.

CORRELACIÓN PARCIAL Y SEMIPARCIAL $\rightarrow$

Para determinar la contribución de cada VI a la explicación de la VD se utilizan la correlación semiparcial (sr) y su cuadrado (sr^2). Cuando en un modelo intervienen más de dos variables las correlaciones dos a dos no son correlaciones puras (no miden la relación entre esas dos variables al margen del influjo de las otras variables del modelo). **Las correlaciones entre dos variables de orden cero son** correlaciones calculadas sin tener en cuenta la presencia de terceras variables.

Correlación semiparcial (sr): relación entre cada VI y la VD, eliminando el influjo del resto de las VII. Para calcularlo, utilizando el modelo de dos variables predictivas $\rightarrow X_1$ y X_2 , (ajusto una regresión de la 1ª sobre la 2ª, extraigo los residuos y los correlaciono con la VD $\rightarrow$ Coeficiente de correlación semiparcial entre X_1 y la VD)

$$sr_1 = (r_{Y1} - r_{Y2} r_{12}) / \sqrt{1 - r^2_{12}} \quad // \quad sr_2 = (r_{Y2} - r_{Y1} r_{12}) / \sqrt{1 - r^2_{12}}$$

Cuando elevamos al cuadrado sr_1 y sr_2 obtenemos la contribución que cada VI tiene sobre la VD, habiendo eliminado el influjo de las otras VII (contribución exclusiva que cada variable hace a la explicación de la VD)

$$sr^2_1 = R^2_{Y.12} - r^2_{Y2} \quad // \quad sr^2_2 = R^2_{Y.12} - r^2_{Y1}$$

Correlación parcial (pr): se elimina el influjo de los predictores, tanto de la VI como de la VD (correlación entre residuos). Se trata de la correlación pura de dos variables, eliminando el influjo de terceras variables.

$$pr_1 = (r_{Y1} - r_{Y2} r_{12}) / \sqrt{1 - r^2_{Y2}} \sqrt{1 - r^2_{12}} \quad // \quad pr_2 = (r_{Y2} - r_{Y1} r_{12}) / \sqrt{1 - r^2_{Y1}} \sqrt{1 - r^2_{12}}$$

Cuando elevamos al cuadrado pr_1 y pr_2 se interpreta como la proporción de la varianza de la VD (Y) asociada a X_1 y no asociada a X_2 (y viceversa)

$$pr^2_1 = R^2_{Y.12} - r^2_{Y2} / 1 - r^2_{Y2} \quad // \quad pr^2_2 = R^2_{Y.12} - r^2_{Y1} / 1 - r^2_{Y1}$$

Modelo Stepwise (pasos sucesivos): Modelo para la introducción sucesiva de variables en el análisis de regresión múltiple. Si realizamos una regresión con el modelo (**stepwise**) introduciríamos en primer lugar la VI con mayor correlación con la VD (posteriormente la que mayor correlación parcial tenga con la VD y así sucesivamente hasta que la nueva variable no aporte un incremento significativo en el R^2). Resulta un ajuste óptimo del modelo con el menor número de variables.

$$F = R^2_{\text{Cambio}} (N - p - 1) / q (1 - R^2)$$

Donde: R^2_{Cambio} (el cuadrado del coeficiente de correlación semiparcial $\rightarrow sr^2 \rightarrow$ proporción de varianza explicada cuando se incorpora esta variable al modelo). El cambio debe ser significativo.

Siguiendo nuestro **problema ejemplo** $\rightarrow$ si quisiéramos determinar el cambio que se produce en el modelo al incorporar la segunda variable (modelo stepwise):

La primera variable que entraría en el modelo sería la X_2 (es la que más alto correlaciona con Y)

La segunda variable introduce el cambio $\rightarrow sr^2_2 = R^2_{Y.12} - r^2_{Y1} \rightarrow sr^2_2 = 0'759^2 - 0'504^2 = 0'322$

Para averiguar si es significativo $\rightarrow q$ (n° de nuevas variables que entran en el modelo)

$$F = 0'322 (25-2-1) / 1 (1 - 0'576) = 16'71 \quad // \quad {}_{0'95}F_{1 \text{ y } 22} = 4'35 < 16'71 \text{ (cambio significativo)}$$

PROBLEMA EJEMPLO $\rightarrow$

Consideremos, a modo de ejemplo, las variables X_1 , X_2 e Y, cuyas correlaciones son las siguientes:

$$r_{Y1} = 0'7 \quad r_{Y2} = 0'6 \quad R_{Y.12} = 0'8$$

Una primera ojeada puede hacernos pensar que la variable X_1 contribuye a la variabilidad de Y en una proporción de $0'7^2 = 0'49$ y que la variable X_2 contribuye en una proporción de $0'6^2 = 0'36$. No obstante, se sabe por la correlación múltiple que la proporción de variación explicada es de $0'8^2 = 0'64$. El total de ambas contribuciones no es igual a la suma, luego está claro que ambas variables explicativas no son fuentes independientes de variabilidad, sino que comparten una cierta cantidad de la misma (existe redundancia entre ambas variables)

Cálculos (correlaciones semiparciales) $\rightarrow$ Como entre ambas variables explican una proporción de 0'64, es evidente que la contribución adicional de X_1 sobre la que explica X_2 será:

$$sr^2_1 = R^2_{Y.12} - r^2_{Y2} \rightarrow sr^2_1 = 0'64 - 0'36 = 0'28 \rightarrow sr_1 = \sqrt{0'28} = 0'529$$

Esto es, lo que añade X_1 a X_2 es una proporción de variación explicada de 0'28. La raíz cuadrada de este valor se expresa como sr_1 y se define como coeficiente de correlación semiparcial.

De igual modo lo que añade X_2 a X_1 será:

$$sr^2_2 = R^2_{Y.12} - r^2_{Y1} \rightarrow sr^2_2 = 0'64 - 0'49 = 0'15 \rightarrow sr_2 = \sqrt{0'15} = 0'387$$

Es decir, la inclusión de X_2 supone un incremento sobre la proporción de variación explicada por X_1 de 0'15 puntos. Su coeficiente de correlación semiparcial es = 0'387.

Cálculos (correlaciones parciales) $\rightarrow$ Un investigador desea estudiar la relación entre actividad física y sensación de bienestar pero cree que el tiempo de sueño puede afectar la relación entre ambas variables. Fija el nivel de significación en $\alpha = 0,01$, extrae aleatoria e independientemente 30 sujetos y les mide el tiempo que dedican diariamente al deporte (X_1), el tiempo diario de sueño (Y) y la sensación de bienestar (X_2). Se cumplen los supuestos del modelo de correlación lineal y el supuesto de normalidad. El investigador obtiene los siguientes coeficientes de correlación: $r_{12} = 0,80$;

$$r_{1Y} = 0,50 \text{ y } r_{2Y} = 0,6$$

$$pr_1 = (r_{Y1} - r_{Y2} r_{12}) / \sqrt{1 - r_{Y2}^2} \sqrt{1 - r_{12}^2} \rightarrow pr_1 = [0'5 - (0'6 \cdot 0'8)] / \sqrt{0'64} \sqrt{0'36} = 0'416; pr_1^2 = 0'174$$

La correlación entre tiempo dedicado al deporte y sensación de bienestar, habiendo controlado la influencia del tiempo de sueño = 0'416 (17'4 % es el porcentaje de varianza asociado a X_1)

$$pr_2 = (r_{Y2} - r_{Y1} r_{12}) / \sqrt{1 - r_{Y1}^2} \sqrt{1 - r_{12}^2} \rightarrow pr_2 = [0'6 - (0'5 \cdot 0'8)] / \sqrt{0'75} \sqrt{0'36} = 0'385; pr_2^2 = 0'148$$

Interpretación → Para un nivel de confianza del 99%, los resultados indican que es significativa la relación lineal entre el tiempo dedicado al deporte y la sensación de bienestar, habiendo controlado la posible influencia del tiempo de sueño.

PROBLEMA EJEMPLO →

Disponemos de las puntuaciones de 5 sujetos en tres variables: promedio de horas de ejercicio al día (X_1), promedio diario de kilocalorías consumidas (X_2) y kilos perdidos (Y). Con los datos expuestos en la tabla, planteamos los análisis más típicos sobre la Regresión lineal múltiple.

X_1	X_2	Y	X_1^2	X_2^2	Y^2	$X_1 X_2$	$X_1 Y$	$X_2 Y$
1	4	4	1	16	16	4	4	16
1,5	3,75	6	2,25	14,0625	36	5,625	9	22,5
2	3	7	4	9	49	6	14	21
2,5	2	9	6,25	4	81	5	22,5	18
3	1,5	10	9	2,25	100	4,5	30	15
$\Sigma = 10$	$\Sigma = 14,25$	$\Sigma = 36$	$\Sigma = 22,5$	$\Sigma = 45,31$	$\Sigma = 282$	$\Sigma = 25,12$	$\Sigma = 79,5$	$\Sigma = 92,5$
$\bar{X}_1 = 2 \quad \bar{X}_2 = 2,85 \quad \bar{Y} = 7,2$ $S^2_{X_1} = 0'5 \quad S^2_{X_2} = 0'94 \quad S^2_Y = 4'56$ $r_{12} = (-0'985) // r_{Y1} = 0'993 // r_{Y2} = (-0'976)$ $Y' = 0,234 + (3,235) X_1 + (0,174) X_2$								

1.- Pronosticar el peso que perderá en dos meses una mujer que hace un promedio de 1,75 horas de ejercicio diarias (X_1) y consume 2,2 Kcal / día (X_2). Construimos la ecuación de regresión:

Coeficientes de regresión parcial →

$B_1 = \beta_1 (S_Y / S_1)$ $B_1 = 1'071 (2'13 / 0'7) = 3'23$	$\beta_1 = r_{Y1} - r_{Y2} \cdot r_{12} / 1 - r_{12}^2$ $\beta_1 = 0'993 - [(-0'976) \cdot (-0'985)] / 0'0298 = 1'071$
$B_2 = \beta_2 (S_Y / S_2)$ $B_2 = 0'079 (2'13 / 0'97) = 0'17$	$\beta_2 = r_{Y2} - r_{Y1} \cdot r_{12} / 1 - r_{12}^2$ $\beta_2 = (-0'976) - (0'993) \cdot (-0'985) / 0'0298 = 0'079$
$B_0 = \bar{Y} - B_1 \bar{X}_1 - B_2 \bar{X}_2 \rightarrow B_0 = 7'2 - 3'235 (2) - 0'174 (2'85) = 0'234$	

Ecuación de regresión múltiple → $Y' = 0'234 + (3'235) X_1 + (0'174) X_2$

Pronóstico $\rightarrow Y' = 0,234 + (3,235) (1,75) + (0,174) (2,2) \rightarrow 6,278$

(La pérdida en peso pronosticada para esa persona será de 6, 278 Kilos)

2.- Obtener el coeficiente de determinación múltiple y el coeficiente de correlación múltiple:

A partir de las correlaciones $\rightarrow$

$$R_{Y, 12} = \sqrt{(r_{Y1}^2 + r_{Y2}^2 - 2 r_{12} r_{1Y} r_{2Y}) / (1 - r_{12}^2)} = \sqrt{(0,986 + 0,952 - 1'91) / (0,0298)} = 0,97$$

Coeficiente de determinación múltiple $\rightarrow R^2_{Y, 12} = (0,97)^2 = 0,94$ (Un 94% de la varianza de Y se debe a la variación conjunta de X_1 y X_2).

A partir de los coeficientes de regresión estandarizados $\rightarrow$

$$\beta_1 = r_{Y1} - r_{Y2} \cdot r_{12} / 1 - r_{12}^2 \rightarrow \beta_1 = (0'993) - [(-0'976) \cdot (-0'985)] / (0,0298) = 1'071$$

$$\beta_2 = r_{Y2} - r_{Y1} \cdot r_{12} / 1 - r_{12}^2 \rightarrow \beta_2 = (-0'976) - (0'993) \cdot (-0'985) / 0'0298 = 0'079$$

$$R_{Y, 12} = \sqrt{\beta_1 r_{Y1} + \beta_2 r_{Y2}} \rightarrow R_{Y, 12} = \sqrt{(1'071 \cdot 0'993) + (0'079 \cdot -0'976)} = 0'97$$

Coeficiente de determinación ajustado $\rightarrow$

$$R^2_{Y, 12} = 1 - [(1 - R^2_{Y, 12}) \cdot (n - 1) / n - p - 1] \rightarrow R^2_{Y, 12} = 1 - [(1 - 0'94) \cdot (4 / (5 - 2 - 1))] = 0'88$$

$$\text{Significación} \rightarrow F = R^2 (n - p - 1) / (1 - R^2) p \rightarrow F = 0'94 (2) / 0'06 (2) = 15'67$$

Como $15'67 > F_{0'05; 2, 2} = 19$ +++++++

Intervalo de confianza $\rightarrow (R^2 \pm t_{n-2; 1-\alpha/2} \cdot \sigma_{R^2} = \text{Limites superior e inferior})$

PROBLEMAS EJEMPLO $\rightarrow$

Se ha llevado a cabo un estudio con objeto de pronosticar las calificaciones que obtendrán los alumnos de A. Datos II (ADII), a partir de las puntuaciones que obtuvieron en la asignatura A. Datos I (ADI). Para ello se ha seleccionado una muestra aleatoria de 5 sujetos, recogiendo sus calificaciones finales en las dos asignaturas:

SUJETOS	AD II $\rightarrow Y$	AD I $\rightarrow X$	XY	Y'	$(Y - Y')^2$	Y'^2
1	6	4	24	5,29	0,5041	0
2	4	3	12	4,895	0,8010	4
3	6	7	42	6,475	0,2256	0
4	7	6	42	6,08	0,8464	1
5	7	9	63	7,265	0,0702	1
	$\Sigma = 30$	$\Sigma = 29$	$\Sigma = 183$	$\Sigma = 30,005$	$\Sigma = 2,45$	$\Sigma = 6$

$$\text{Datos} \rightarrow \Sigma (X - \bar{X})^2 = 22,8 \quad S^2_e = \Sigma (Y - Y')^2 / n - 2 \rightarrow 2,45 / 3 = 0,82 \rightarrow (S_e = 0,904)$$

$$\bar{X} = 29 / 5 = 5,8 \quad \bar{Y} = 30 / 5 = 6 \quad S^2_x = 5'7 \quad S^2_y = 1'5 \quad r_{xy} = 0'673$$

$$Y' = B_0 + B X \rightarrow Y' = 3,71 + 0,395 X$$

Donde, los coeficientes de la regresión son:

$$B = (n \Sigma XY - \Sigma X \Sigma Y) / [n \Sigma X^2 - (\Sigma X)^2] \rightarrow (45 / 114) = 0,395$$

$$B_0 = \bar{Y} - b \bar{X} = 6 - (0,395 \cdot 5,8) = 3,71$$

1.- Contrastar las hipótesis de que los parámetros son nulos (no son significativos)

El intercepto (ordenada en el origen) es nulo ($\beta_0 = 0$).

Hipótesis estadísticas $\rightarrow H_0: \beta_0 = 0$ y $H_1: \beta_0 \neq 0$

Estadístico de Contraste $\rightarrow T = B_0 - 0 / \sigma_{\beta_0} \rightarrow T = (3,71 - 0) / 1,1 = 3'37$

$\sigma_{\beta_0} = \sigma_e (\sqrt{1/n + X^2 / (n-1) S^2_X}) = \sigma_{\beta_0} = 0'904 (\sqrt{1/5 + 5'8^2 / 4 \cdot 5'7} = 1'1$

Distribución $T_{\alpha/2 = 0,025; n-2}$ y $T_{1-\alpha/2 = 0,975; n-2} \rightarrow \pm 3,18 < 3'37$ (Rechazamos H_0).

Intervalo de confianza $\rightarrow IC(B_0) = B_0 \pm T_{n-2; 1-\alpha/2} \sigma_{B_0} \rightarrow 3,71 \pm (3,18 \cdot 1,1) = (7'208 \text{ y } 0'212)$

El intercepto no es nulo ($\beta_0 \neq 0$) $\rightarrow 0$ no está en el intervalo de confianza.

La pendiente de la recta de regresión es nula $\rightarrow \beta = 0$

Hipótesis estadísticas $\rightarrow H_0: \beta = 0$ $H_1: \beta \neq 0$

Estadístico de Contraste $\rightarrow T = \beta - 0 / \sigma_\beta \rightarrow T = (0,395 - 0) / 1'19 = 0'33$

$\sigma_\beta = S_Y / S_X (\sqrt{1 - r^2_{XY} / n-2}) \rightarrow \sigma_\beta = (1'22/2'39) \sqrt{0'182} = 1'19$

Distribución $T_{\alpha/2 = 0,025; n-2}$ y $T_{1-\alpha/2 = 0,975; n-2} \rightarrow \pm 3,18 > 1'19$ (Mantenemos H_0)

Intervalo confianza $\rightarrow IC(B) = B \pm T_{n-2; 1-\alpha/2} \sigma_B$ (como ejemplo, puesto que no hemos aceptado H_1)

$0,395 \pm (3,18 \cdot 1'19) = (4'18 \text{ y } -3'38) \rightarrow$ Pendiente nula ($\beta = 0$) $\rightarrow 0$ está en el intervalo de confianza.

También podemos contrastar la hipótesis ($\beta = 0$) mediante el análisis de varianza de la ecuación de regresión: Hipótesis estadísticas $\rightarrow H_0: \beta = 0$ y $H_1: \beta \neq 0$

TABLA DEL ANOVA

FUENTE DE VARIACIÓN	SUMAS CUADRÁTICAS	GRADOS DE LIBERTAD	MEDIAS CUADRÁTICAS	ESTADÍSTICO DE CONTRASTE
DEBIDA A LA REGRESIÓN Desviación Total	SC REGRESIÓN 3,55	1	MC REGRESIÓN 3,55 / 1 = 3,55	F = (3,55 / 0,82) = 4,353
RESIDUAL O ERROR Varianza explicada por la Regresión	SC RESIDUAL 2,45	3	MC RESIDUAL = S ² _e 2,45 / 3 = 0,82	
TOTAL Desviación respecto a la recta de Regresión	SC TOTAL 6	4	F se distribuye con 1 y 3 gl → 10,13	

$SC \text{ Total } (Y - \bar{Y})^2 = \sum y^2 = 6$ $SC_{\text{REGRESIÓN}} = (6 - 2,45) = 3,55$ $SC_{\text{RESIDUAL}} = \sum (Y - Y') = 2,45$

Decisión $\rightarrow$ No podemos rechazar $H_0 \rightarrow 10,13 > 4,353$. La regresión no es significativa ($\beta = 0$) no hay pendiente. Ambos análisis indican lo mismo; por tanto, la puntuación en A. Datos I no predice adecuadamente la calificación en A. Datos II.

Por último, averiguamos (pronosticamos) la puntuación que obtendrá un alumno en AD II, sabiendo que ha obtenido un 5 en AD I y calculamos el intervalo de confianza para este pronóstico individual (para $\alpha = 0,05$). Utilizamos la recta de regresión simple: $Y' = B_0 + B X \rightarrow 3,71 + 0,395 X$

Pronóstico en AD II, para un alumno que obtuvo un 5 en AD I $\rightarrow Y' = 3,71 + 0,395 (5) \rightarrow 5,685$

Para calcular el **intervalo de confianza** de este pronóstico individual $Y' = 5,685$, necesitamos calcular el error típico del pronóstico con σ_e^2 desconocida y estimada mediante S_e^2 .

$$IC(Y') = Y' \pm T_{n-2; 1-\alpha/2} \cdot (\sigma_e \sqrt{1/n + (X - \bar{X})^2 / (n-1) S_x^2})$$

$$Y' \pm T_{n-2; \alpha/2} \sigma_Y \rightarrow 5,685 \pm 3,18 \cdot 0,82 \rightarrow (3,071 \text{ y } 8,29)$$

$$T_{\alpha/2 = 0,025; n-2} \text{ y } T_{1-\alpha/2 = 0,975; n-2} \rightarrow \pm 3,18$$

$$\sigma_Y = 0,82 \sqrt{(0,2 + 22,8) / (4 \cdot 5,7)} \rightarrow \sigma_Y = (0,82) \cdot (1) = 0,82$$