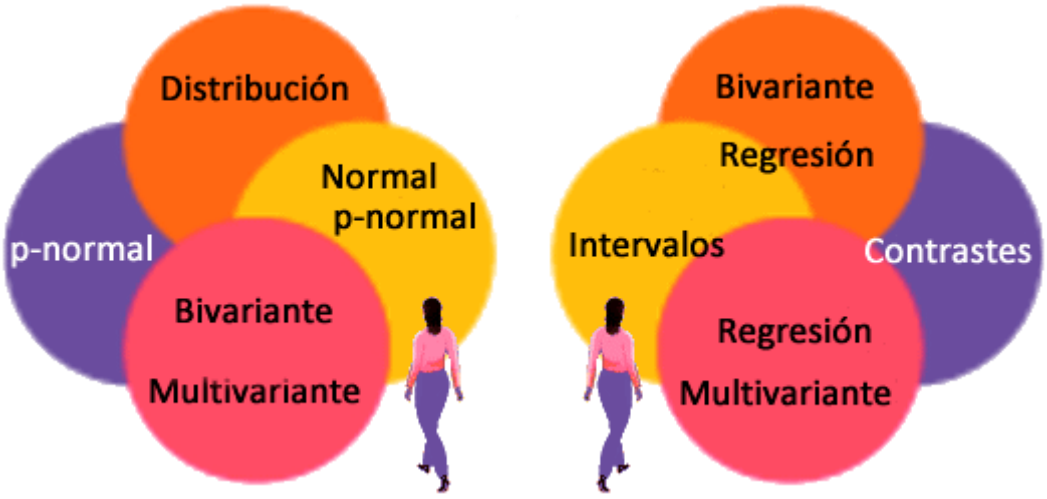
UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....



UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... **Grupo** .....

Apellidos ..... **Nombre** .....

Ejercicio del día .....

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

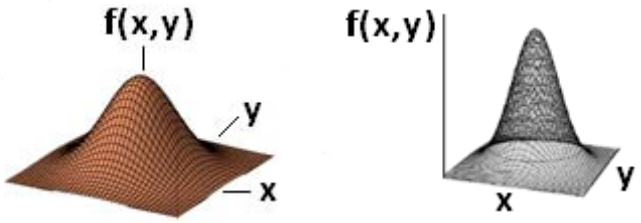
Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

DISTRIBUCIÓN NORMAL BIVARIANTE

(función de densidad conjunta)



Se dice que dos variables aleatorias continuas X e Y se distribuyen según una distribución normal bivariante

$$(X,Y) \approx N\left(\begin{bmatrix}\mu_X\\ \mu_Y\end{bmatrix}, \begin{bmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_2^2\end{bmatrix}\right)$$ si su función de densidad conjunta es:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2\,\sigma_X^2\,\sigma_Y^2\,(1-\rho^2)}\left[\sigma_Y^2\,(x-\mu_X)^2 + \sigma_Y^2\,(y-\mu_Y)^2 - 2\,\sigma_X\sigma_Y\,\rho\,(x-\mu_X)(y-\mu_Y)\right]}$$

Siendo la forma cuadrática Q(x,y):

$$Q(x,y) = \frac{1}{\sigma_X^2\,\sigma_Y^2\,(1-\rho^2)}\left[\sigma_Y^2\,(x-\mu_X)^2 + \sigma_Y^2\,(y-\mu_Y)^2 - 2\,\sigma_X\sigma_Y\,\rho\,(x-\mu_X)(y-\mu_Y)\right] =$$

$$= \frac{1}{(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\,\rho\,\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]$$

También, puede expresarse: $Q(x,y) = (x-\mu_X \quad y-\mu_Y)\,\Sigma^{-1}\begin{pmatrix}x-\mu_X\\ y-\mu_Y\end{pmatrix}$

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}Q(x,y)} = \frac{1}{2\pi\sqrt{|\Sigma|}}\, e^{-\frac{1}{2}Q(x,y)}$$

$E(x) = \mu_X$ $E(y) = \mu_Y$ $Cov(x,y) = \sigma_{XY}$ Coeficiente de correlación: $\rho = \frac{\sigma_{XY}}{\sigma_v\,\sigma_v}$

Matriz de varianzas y covarianzas:

$$\Sigma = Var(x,y) = \begin{pmatrix}Var(x) & Cov(x,y)\\ Cov(x,y) & Var(y)\end{pmatrix} = \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_Y^2\end{pmatrix}$$ $\Sigma \equiv$ matriz simétrica definida positiva

$$|\Sigma| = \sigma_X^2\sigma_Y^2 - \sigma_{XY}^2$$

$$(\Sigma)^{-1} = \frac{1}{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}\begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_Y^2\end{pmatrix}$$

A partir de esta función de densidad conjunta se deduce:

- Las distribuciones marginales son distribuciones normales: $X \sim N(\mu_X,\sigma_X)$ $Y \sim N(\mu_Y,\sigma_Y)$
- Las varianzas condicionadas son constantes:

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$\text{Var}(Y|X=x) = \sigma_2^2\,(1-\rho^2)$ $\quad$ $\text{Var}(X|Y=y) = \sigma_1^2\,(1-\rho^2)$

- Sí $\rho = 0 \;\rightarrow\; f(x,y) = f_1(x)\,.\,f_2(y) \;\rightarrow$ Las variables son independientes (están incorreladas)

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y}\, e^{-\frac{1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]} = \left(\frac{1}{2\pi\sigma_X}\, e^{-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}}\right)\left(\frac{1}{2\pi\sigma_Y}\, e^{-\frac{1}{2}\frac{(x-\mu_Y)^2}{\sigma_Y^2}}\right) =$$

$$= f_1(x)\,.\,f_2(y)$$

Si X e Y son variables aleatorias independientes $\rightarrow$ $\rho = 0$

$\rho = \dfrac{\sigma_{XY}}{\sigma_X\sigma_Y}$ donde $\sigma_{XY} = E(XY) - E(X)\,E(Y)$. Cuando X e Y son variables aleatorias independientes siempre se verifica que $E(XY) = E(X)\,E(Y)$

- Dos variables aleatorias continuas X e Y con una distribución bivariante $(X,Y) \approx N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$ tienen una función de densidad conjunta:

$$f(x,y) = \frac{1}{2\pi}\, e^{-\frac{1}{2}(x^2+y^2)} = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2} \,.\, \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}y^2} = f_1(x)\,.\,f_2(y)$$

- Toda variable aleatoria $X \sim N(\mu,\sigma)$ puede expresarse $X = \mu + \sigma\,.\,z$, donde $z \sim N(0,1)$. La generalización de esta propiedad en dimensión $p > 1$ permite definir la distribución normal multivariante.

Se dice que $X = (x,y)^t$ tiene una ley normal bivariante con parámetros $\mu = (\mu_X, \mu_Y)^t$ y matriz de varianzas-covarianzas $\Sigma = \text{Var}(x,y) = \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_2^2\end{pmatrix}$, y se denotará $X \sim N_2(\mu, \Sigma)$ sí

$$\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix} + \begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{XY} & \sigma_Y^2\end{pmatrix}\begin{pmatrix}z_1\\z_2\end{pmatrix}$$

donde z_1 y z_2 son variables aleatorias independientes con ley $N(0,1)$.

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

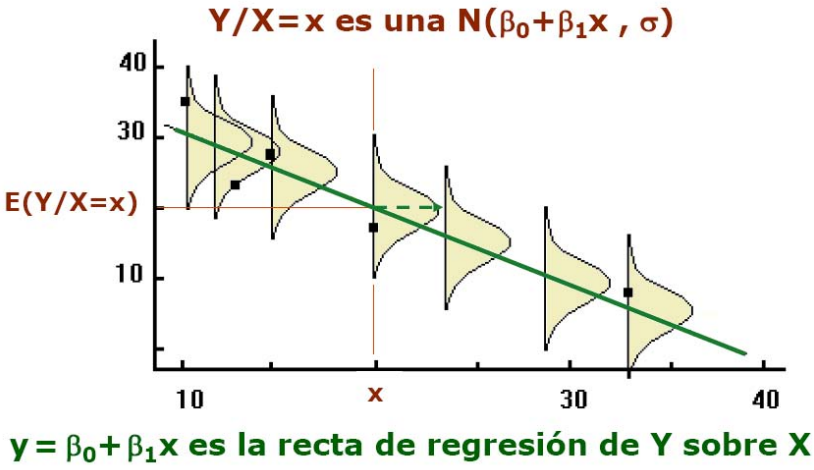
Ejercicio del día .....

REGRESIÓN LINEAL: Las esperanzas condicionadas (líneas de regresión)

Las técnicas de regresión lineal simple parten de dos variables cuantitativas: Una variable explicativa y otra variable de respuesta.

La ecuación de la recta $y = \beta_0 + \beta_1 x$ trata de explicar la variable y mediante valores de la variable x. Para ello, se dispone de un modelo de probabilidad (ley normal) y de n pares de datos (x_i, y_i) que provienen del modelo establecido.

Normal bivariante
Distribuciones condicionadas



La distribución conjunta $(X,Y) \approx N_2(\mu, \Sigma) \equiv N_2\left(\begin{bmatrix}\mu_X \\ \mu_Y\end{bmatrix}, \begin{bmatrix}\sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2\end{bmatrix}\right)$

$$E\left[Y \mid X=x\right] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) = \mu_Y + \frac{\sigma_{XY}}{\sigma_x^2}(x-\mu_X)$$

$$E\left[X \mid Y=y\right] = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y) = \mu_X + \frac{\sigma_{XY}}{\sigma_Y^2}(y-\mu_Y)$$

$$Var\left[Y \mid X=x\right] = \sigma_Y^2(1-\rho^2) \qquad Var\left[X \mid Y=y\right] = \sigma_X^2(1-\rho^2)$$

$$E\left[Y \mid X=x\right] \approx N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X),\ \sigma_Y^2(1-\rho^2)\right)$$

$$E\left[X \mid Y=y\right] \approx N\left(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y),\ \sigma_X^2(1-\rho^2)\right)$$

FUNCIÓN DE REGRESIÓN

$$E\left[E(Y \mid X)\right] = E(Y) \qquad E\left[E(X \mid Y)\right] = E(X)$$

$$Var(Y) = \underbrace{E\left[Var(Y \mid X)\right]}_{\substack{\text{Esperanza} \\ \text{de las varianzas}}} + \underbrace{Var\left[E(Y \mid X)\right]}_{\substack{\text{Varianza} \\ \text{de las esperanzas}}}$$

Sí $Y_1 = AX_1 + B$, $Y_2 = CX_2 + D$ donde $A, B, C, D \in R \rightarrow \rho_{Y_1Y_2} = \frac{AC}{|AC|}\rho_{X_1X_2}$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

TÉCNICAS DE REGRESIÓN LINEAL SIMPLE

La regresión lineal simple parte de dos variables cuantitativas (x, y) : Tratando de explicar la variable de respuesta (y) mediante una función lineal de la variable explicativa (x) representada por la recta

$y = \beta_0 + \beta_1 x$

Para lo que se dispone de un modelo de probabilidad (Ley Normal) y de n pares de datos (x_i, y_i) que provienen del modelo establecido.

Modelo I

Una variable aleatoria Y, cuando el investigador fija los datos (x_i) y obtiene los correspondientes (y_i)

$Y = \beta_0 + \beta_1 x + U$ donde $U \approx N(0, \sigma)$

Modelo II

Dos variables aleatorias (X, Y), el investigador obtiene al azar los valores (x_i, y_i)

$$(X, Y) \approx N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \rightarrow (Y \mid X = x) \approx N\left(\beta_0 + \beta_1 x, \underbrace{\sigma_2\sqrt{1-\rho^2}}_{\sigma}\right)$$

$Y_i = \beta_0 + \beta_1 x_i + u_i$ $Y_i \approx N(\beta_0 + \beta_1 x_i, \sigma)$ independientes $i = 1, 2, \ldots, n$ $u_i \approx N(0, \sigma)$

Muestra aleatoria

Bajo el supuesto
- Normalidad: $u_i \approx$ Normal
- Linealidad: $E(u_i) = 0$
- Homocedasticidad: $V(u_i) = \sigma^2$
- Independencia: u_i independientes

β_0 { Representa el valor medio de la respuesta (y) cuando la variable explicativa (x) es cero.

β_1 { Representa el incremento de la respuesta media (y) cuando la variable explicativa (x) aumenta en una unidad.

ESTIMACIÓN PUNTUAL DE LOS PARÁMETROS DE LA REGRESIÓN

Gauss propuso en 1809 el Método de los Mínimos Cuadrados para obtener los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que mejor se ajustan a los datos: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

El método consiste en minimizar la suma de los cuadrados de las distancias verticales entre los datos y las estimaciones, es decir, minimizar la suma de los residuos al cuadrado.

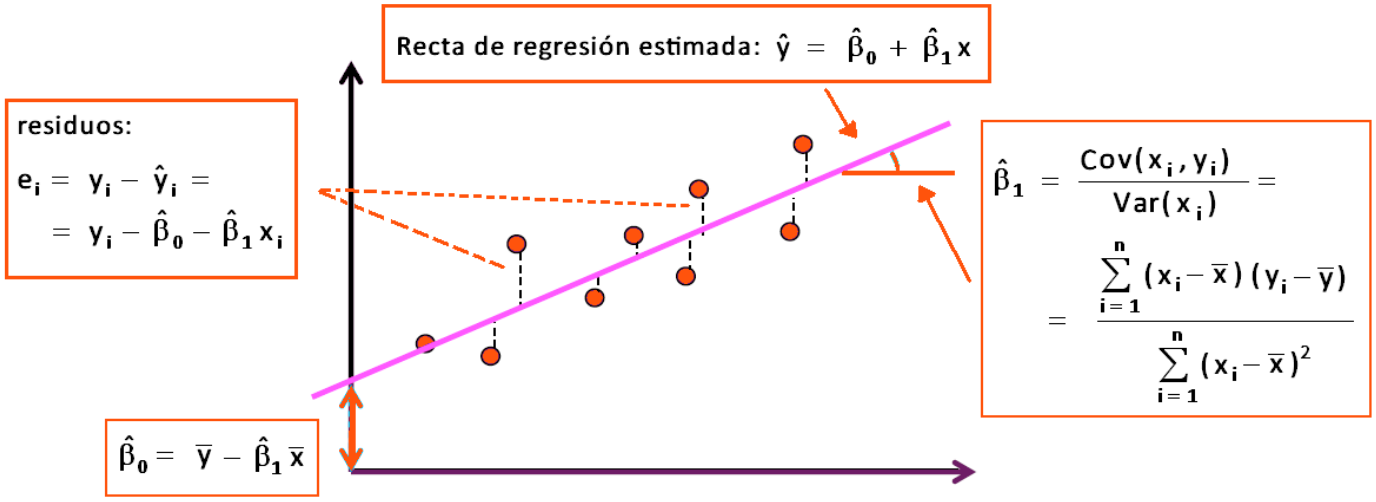
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....



Covarianza : $Cov(x,y) = \sigma_{XY} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ es una medida de la dependencia lineal.

Dependiente de las unidades de media de las variables.

Coeficiente de correlación lineal: $\rho = \dfrac{\sigma_{XY}}{\sigma_x \cdot \sigma_y}$

No depende de las unidades de media de las variables.

Varianza residual: $\sigma_r^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n} = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$ estimador sesgado y máximo verosímil de σ^2

Cuasiarianza residual: $s_r^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2} = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ estimador centrado y máximo verosímil de σ^2

Coeficientes de regresión

Intercepto: $\hat{\beta}_0 = \bar{y} - \dfrac{\sigma_{XY}}{\sigma_x^2}\bar{x}$

Pendiente de la recta: $\hat{\beta}_1 = \dfrac{\sigma_{XY}}{\sigma_x^2}$

HIPÓTESIS DEL MODELO DE REGRESIÓN LINEAL SIMPLE

Normalidad: Los errores siguen una distribución normal $u_i \approx N(0, \sigma)$

Homocedasticidad: La varianza de los errores es constante $Var(u_i) = \sigma^2$

Linealidad: Los residuos se distribuyen sin forma alrededor del cero

Independencia: Las observaciones son independientes $E(u_i\, u_j) = 0$

Homogeneidad: El valor promedio del error es cero $E(u_i) = 0$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

INFERENCIAS SOBRE EL MODELO DE REGRESIÓN

Habiendo obtenido estimaciones puntuales de los coeficientes de regresión, utilizando Intervalos de Confianza (IC) se pueden obtener una medida de precisión de dichas estimaciones.

Utilizando Contrastes de Hipótesis se puede comprobar si un determinado valor puede ser el auténtico valor del parámetro.

$$IC_{1-\alpha}(\beta_0) = \left(\hat{\beta}_0 \pm t_{\alpha/2,\, n-2} \cdot s_r \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n \cdot \sigma_x^2}} \right) \qquad n \cdot \sigma_x = (n-1) \cdot s_x$$

$$IC_{1-\alpha}(\beta_1) = \left(\hat{\beta}_1 \pm t_{\alpha/2,\, n-2} \cdot s_r \cdot \sqrt{\frac{1}{n \cdot \sigma_x^2}} \right)$$

$$IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-2) \cdot s_r^2}{\chi^2_{\alpha/2,\, n-2}} , \frac{(n-2) \cdot s_r^2}{\chi^2_{1-\alpha/2,\, n-2}} \right)$$

CONTRASTE PARA LA PENDIENTE DE LA REGRESIÓN $\begin{cases} H_0 : \beta_1 = 0 & \text{No hay relación lineal} \\ H_1 : \beta_1 \neq 0 \end{cases}$

Con un nivel de significación α se rechaza la hipótesis nula H_0 sí el cero no se encuentra en el intervalo de confianza $IC_{1-\alpha}(\beta_1) = \left(\hat{\beta}_1 \pm t_{\alpha/2,\, n-2} \cdot \underbrace{s_r \cdot \sqrt{\frac{1}{n \cdot \sigma_x^2}}}_{\text{Error típico}} \right)$

Se acepta la hipótesis nula H_0 sí el estadístico de contraste $t_{p/2} = \dfrac{|\hat{\beta}_1|}{s_r \cdot \sqrt{\dfrac{1}{n \cdot \sigma_x^2}}} \leq t_{\alpha/2,\, n-2}$

Se acepta la hipótesis nula H_0 sí $p-\text{valor} = \alpha_p = 2 \cdot P\left(t_{n-2} \leq t_{p/2}\right) > \alpha$

CONTRASTE PARA EL INTERCEPTO DE LA REGRESIÓN $\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$

Se acepta la hipótesis nula H_0 sí el estadístico de contraste $t_{p/2} = \dfrac{|\hat{\beta}_0|}{s_r \cdot \sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{n \cdot \sigma_x^2}}} \leq t_{\alpha/2,\, n-2}$

Se acepta la hipótesis nula H_0 sí $p-\text{valor} = \alpha_p = 2 \cdot P\left(t_{n-2} \leq t_{p/2}\right) > \alpha$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

ESTIMACIÓN DE UNA RESPUESTA PROMEDIO Y PREDICCIÓN DE NUEVA RESPUESTA

- Estimar el valor medio de la variable Y para un cierto valor $X = x_0$

$$IC_{1-\alpha}(\bar{y}) = \left(\hat{y}_0 \pm t_{\alpha/2,\, n-2} \cdot s_r \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n \cdot \sigma_x^2}} \right) \qquad n \cdot \sigma_x = (n-1) \cdot s_x$$

- Predecir el valor que tomará la variable Y para un cierto valor $X = x_0$

$$IC_{1-\alpha}(y \mid x_0) = \left(\hat{y}_0 \pm t_{\alpha/2,\, n-2} \cdot s_r \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n \cdot \sigma_x^2}} \right) \qquad n \cdot \sigma_x = (n-1) \cdot s_x$$

CONTRASTES DE LA REGRESIÓN (ANOVA): Descomposición de la variabilidad en la regresión

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i} + \underbrace{e_i}_{y_i - \hat{y}_i} \quad \rightarrow \quad y_i = \hat{y}_i + (y_i - \hat{y}_i) \quad \rightarrow \quad (y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SCR}$$

SCT ≡ Suma de cuadrados total
SCE ≡ Suma de cuadrados explicada
SCR ≡ Suma de cuadrados residual

TABLA ANOVA

Suma de cuadrados	Grados Libertad	Varianza	Estadístico	p-valor
$SCE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1	$\frac{SCE}{1}$	$F = \frac{SCE/1}{SCR/n-2}$	
$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-2$	$\frac{SCR}{n-2} = s_r^2$		
$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$			

Se establece el contraste:

H_0 : El modelo de regresión no sirve para explicar la respuesta

H_1 : El modelo de regresión sirve para explicar la respuesta

Se rechaza la hipótesis nula H_0 con un nivel de significación α sí $F > F_{\alpha,1,(n-2)}$

Coeficiente de Determinación (proporción de variabilidad explicada por la regresión): $R^2 = \frac{SCE}{SCT}$

En regresión simple: $R = \rho = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$ $\qquad SCR = n\,\sigma_y^2\,(1-\rho^2)$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

DISTRIBUCIÓN NORMAL MULTIVARIANTE (p-dimensional o p-normal)

Sí en lugar de una variable, se consideran un conjunto de variables $(X_1, X_2, \ldots, X_p)$ que siguen conjuntamente una distribución normal se obtiene una distribución normal multivariante.

La función de densidad multivariante tiene la siguiente expresión:

$$f(X_1, X_2, \ldots, X_p) = \frac{1}{\sqrt{(2\pi)^p \left|\Sigma\right|}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

donde μ es el vector teórico de medias y Σ es la matriz teórica de varianzas-covarianzas definida positiva.

$(x-\mu)^t \Sigma^{-1}(x-\mu) \equiv$ Cuadrado de la distancia del punto x al centro de la distribución μ.

La distribución normal multivariante es una generalización de la distribución normal univariante.

Así, el vector x es una generalización del escalar X: $X = \begin{bmatrix} X_1 \\ X_2 \\ .... \\ X_p \end{bmatrix}$

El vector μ es una generalización de la media teórica μ de la distribución univariante.

Media o centro de gravedad poblacional multivariante.

$$\mu = \begin{bmatrix} E(X_1) \\ E(X_2) \\ .... \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ .... \\ \mu_p \end{bmatrix}$$

La generalización de la varianza escalar σ^2 no es un vector sino la matriz de varianzas-covarianzas Σ, esto se debe a que la matriz de varianzas-covarianzas recoge, además de las varianzas de las p variables, las covarianzas entre cada par de variables. Su expresión es: $\Sigma = E(x-\mu)(x-\mu)^t$

$$\Sigma = E(x-\mu)(x-\mu)^t = \begin{bmatrix} E(x_1-\mu_1)^2 & E(x_1-\mu_1)(x_2-\mu_2) & ..... & E(x_1-\mu_1)(x_p-\mu_p) \\ E(x_2-\mu_2)(x_1-\mu_1) & E(x_2-\mu_2)^2 & ..... & E(x_2-\mu_2)(x_p-\mu_p) \\ ..... & ..... & ..... & ..... \\ E(x_p-\mu_p)(x_1-\mu_1) & E(x_p-\mu_p)(x_2-\mu_2) & ..... & E(x_p-\mu_p)^2 \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & ..... & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & ..... & \sigma_{2p} \\ ..... & ..... & ..... & ..... \\ \sigma_{p1} & \sigma_{p2} & ..... & \sigma_p^2 \end{bmatrix}$$

En la diagonal principal de la matriz Σ aparecen las varianzas teóricas de las p variables. Fuera de la diagonal principal aparecen las covarianzas entre cada par de variables.

Señalar que la matriz de covarianzas es simétrica.

Teniendo en cuenta que el coeficiente de correlación lineal poblacional es:

$$\rho_{ij} = \frac{E(X_i-\mu_i)(X_j-\mu_j)}{\sqrt{E(X_i-\mu_i)^2 E(X_j-\mu_j)^2}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \quad \text{se obtiene} \quad \sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Sustituyendo, resulta la matriz de varianzas-covarianzas: $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1p}\sigma_1\sigma_p \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2p}\sigma_2\sigma_p \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{p1}\sigma_p\sigma_1 & \rho_{p2}\sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{bmatrix}$

Sí todas las variables tienen una varianza poblacional igual a la unidad, la matriz de varianzas-covarianzas queda simplificada. La matriz R_p es la matriz de correlación poblacional, y se corresponde con la matriz R de correlación muestral. $R_p = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$

Para indicar que el vector X de dimensión p tiene una distribución normal multivariante con media μ y matriz de varianzas-covarianzas Σ se utiliza la notación: $X \approx N_p(\mu, \Sigma)$

Al considerar una variable aleatoria univariada Y definida como una combinación lineal de $X_1, X_2, \ldots, X_p$, se tiene $Y = a^t X$, donde $a^t = (a_1, a_2, \ldots, a_p)^t$ es un vector de constantes, entonces:

$$E(Y) = a^t \mu \qquad Var(Y) = a^t \Sigma a$$

Estos resultados pueden generalizarse al caso en que A es una matriz de constantes $p \times m$. En este caso, $A^t X$ es un vector $m \times 1$, y tiene un vector de medias y matriz de varianzas-covarianzas:

$$E(A^t X) = A^t \mu \qquad Var(A^t X) = A^t \Sigma A$$

TEOREMA CENTRAL DEL LÍMITE MULTIVARIADO

Si las filas de la matriz de datos X representan una muestra de una variable aleatoria multivariada X con $E(X) = \mu$ y $Cov(X) = \Sigma$, entonces la distribución asintótica de $\bar{X}$ es una normal multivariada con vector de medias μ y matriz de varianzas-covarianzas $\frac{1}{n}\Sigma$.

ESTANDARIZACIÓN MULTIVARIANTE

Sí $X \approx N_p(\mu, \Sigma)$ y se define $Y = \Sigma^{-1/2}(X - \mu)$ entonces $(Y_1, Y_2, \ldots, Y_p)$ son independientes N(0, 1)

Consecuencias de la Estandarización:

- Sí $X \approx N_p(\mu, \Sigma) \rightarrow E(X) = \mu$ y $Var(X) = \Sigma$ puesto que $X = \Sigma^{1/2} Y + \mu$
- La distribución $(X - \mu)^t \Sigma^{-1} (X - \mu) \approx \chi_p^2$ puesto que $(X - \mu)^t \Sigma^{-1} (X - \mu) = Y^t Y$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

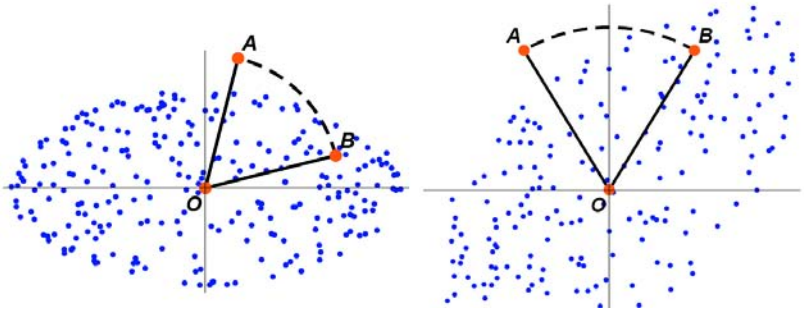
Apellidos ..... Nombre .....

Ejercicio del día .....

DISTANCIA DE MAHALANOBIS

Sí $X = (X_1, X_2, \ldots, X_p)^t$ es un vector con media μ y matriz de varianzas-covarianzas Σ, la distancia de Mahalanobis de X a μ se expresa:

$$d_M(X, \mu) = \sqrt{(X-\mu)^t \, \Sigma^{-1} \, (X-\mu)}$$



- La distancia d_M es adimensional y coincide con la distancia euclídea entre los datos estandarizados de forma multivariante.
- La distancia d_M tiene en cuenta las diferentes variabilidades y correlaciones de las variables.
- La distancia d_M bajo el supuesto de normalidad se distribuye como una χ^2_p

COMBINACIÓN LINEAL MULTIVARIANTE

Sí $X \approx N_p(\mu, \Sigma)$ cualquier combinación lineal $b^t X = b_1 X_1 + b_2 X_2 + \ldots + b_p X_p \approx N(b^t \mu, b^t \Sigma \, b)$

También sí $b^t X \approx N(b^t \mu, b^t \Sigma \, b) \;\; \forall b \;\; \rightarrow \;\; X \approx N_p(\mu, \Sigma)$

Sí $X \approx N_p(\mu, \Sigma)$ con $\pm c\sqrt{\lambda_j e_j}$ $\quad |\Sigma| > 0 \;\; \rightarrow \;\; \begin{cases} (X-\mu)^t \, \Sigma^{-1} \, (X-\mu) \approx \chi^2_p \\ P\left[(X-\mu)^t \, \Sigma^{-1} \, (X-\mu) \leq \chi^2_{\alpha,p}\right] = 1-\alpha \end{cases}$

$\chi^2_{\alpha,p} \equiv$ Cuantil superior de orden α

Sí $X \approx N_p(\mu, \Sigma)$ y $B_{q \times p} \rightarrow$ Las q combinaciones lineales $BX = \begin{pmatrix} b_{11} X_1 + \ldots + b_{1p} X_p \\ ---------- \\ b_{q1} X_1 + \ldots + b_{qp} X_p \end{pmatrix}$ tienen una distribución $N_q(B\mu, B \, \Sigma \, B^t)$

Sí $X \approx N_p(\mu, \Sigma)$ y $d_{p \times 1}$ es un vector de constantes, $(X + d) \approx N_p(\mu + d, \Sigma)$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

INDEPENDENCIA MULTIVARIANTE

Al particionar $X \approx N_p(\mu, \Sigma)$ todos los subconjuntos $X_{(p \times 1)}$ tienen una distribución normal.

Particionando X, μ y Σ $\rightarrow$ $X_1 \approx N_q(\mu_1, \Sigma_{11})$

$$X = \begin{pmatrix} X_1 \\ -- \\ X_2 \end{pmatrix} \begin{matrix} (q \times 1) \\ \\ ((p-q) \times 1) \end{matrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ -- \\ \mu_2 \end{pmatrix} \begin{matrix} (q \times 1) \\ \\ ((p-q) \times 1) \end{matrix} \qquad \Sigma = \begin{matrix} (q \times q) \\ \\ (p-q) \times q \end{matrix} \begin{pmatrix} \Sigma_{11} & | & \Sigma_{12} \\ -- & | & -- \\ \Sigma_{21} & | & \Sigma_{22} \end{pmatrix} \begin{matrix} (q \times (p-q)) \\ \\ ((p-q) \times (p-q)) \end{matrix}$$

donde, $\Sigma_{11} = \text{Var}(X_1)$, $\Sigma_{22} = \text{Var}(X_2)$, $\Sigma_{12} = \text{Covar}(X_1, X_2)$, $\Sigma_{21} = \Sigma_{12}^t = \text{Covar}(X_2, X_1)$

Sí $X_{1(q_1 \times 1)}$ y $X_{2(q_2 \times 1)}$ son independientes $\rightarrow$ $\text{Covar}(X_1, X_2) = 0_{(q_1 \times q_2)}$

Sí $X = \begin{pmatrix} X_1 \\ -- \\ X_2 \end{pmatrix} \approx N_{q_1+q_2} \left[\begin{pmatrix} \mu_1 \\ -- \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & | & \Sigma_{12} \\ -- & | & -- \\ \Sigma_{21} & | & \Sigma_{22} \end{pmatrix} \right]$ $\rightarrow$ X_1 y X_2 son independientes $\Leftrightarrow$ $\Sigma_{12} = 0$

Sí $X_1 \approx N_{q_1}(\mu_1, \Sigma_{11})$ y $X_2 \approx N_{q_2}(\mu_2, \Sigma_{22})$, siendo X_1 y X_2 independientes $\Rightarrow$

$$\Rightarrow \begin{pmatrix} X_1 \\ -- \\ X_2 \end{pmatrix} \approx N_{q_1+q_2} \left[\begin{pmatrix} \mu_1 \\ -- \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & | & 0 \\ -- & | & -- \\ 0 & | & \Sigma_{22} \end{pmatrix} \right]$$

DISTRIBUCIONES CONDICIONADAS MULTIVARIANTES

Sí $X = \begin{pmatrix} X_1 \\ -- \\ X_2 \end{pmatrix} \approx N_p \left[\begin{pmatrix} \mu_1 \\ -- \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & | & \Sigma_{12} \\ -- & | & -- \\ \Sigma_{21} & | & \Sigma_{22} \end{pmatrix} \right]$ con $|\Sigma_{22}| > 0$ $\rightarrow$ La distribución condicionada

$(X_1 | X_2 = x_2)$ es una distribución normal con parámetros: $\begin{cases} E(X_1 | X_2 = x_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \text{Var}(X_1 | X_2 = x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{cases}$

Análogamente,

$(X_2 | X_1 = x_1)$ es una distribución normal con parámetros: $\begin{cases} E(X_2 | X_1 = x_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) \\ \text{Var}(X_2 | X_1 = x_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{cases}$

CONTORNOS DE UNA NORMAL MULTIVARIADA

Siendo $\lambda_1, \lambda_2, \ldots, \lambda_p$ los valores propios de Σ y $e_1, e_2, \ldots, e_p$ los correspondientes vectores propios, donde $\Sigma e_j = \lambda_j e_j$, $j = 1, 2, \ldots, p$, el contorno dado por $\{x : (X-\mu)^t \Sigma^{-1} (X-\mu) = c^2\}$ es una elipsoide centrada en μ y cuyos ejes son $\pm c\sqrt{\lambda_j e_j}$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

La densidad normal multivariada tiene un valor máximo cuando la distancia cuadrática $(x-\mu)^t \Sigma^{-1}(x-\mu) = 0$, es decir, cuando $x = \mu$. Por tanto, el punto μ es el punto de máxima densidad, o la moda, y también es la media.

Las curvas de nivel de una distribución multivariada se utilizan para construir Intervalos de Confianza multivariados:

$$P\left[(X-\mu)^t \Sigma^{-1}(X-\mu) \leq \chi^2_{p,\alpha}\right] = 1-\alpha$$

Sí $Y = \Sigma^{-1/2}(X-\mu)$ y A es una matriz de rango p $\rightarrow$ $Y^t A Y = \chi^2_p$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

MODELO DE REGRESIÓN LINEAL MÚLTIPLE

La regresión múltiple tiene como objetivo analizar un modelo que pretende explicar el comportamiento de una variable (endógena, explicada o dependiente), que se denota como Y, utilizando la información proporcionada por los valores tomados por un conjunto de variables explicativas (exógenas o independientes), que se denotan por $X_1, X_2, \ldots, X_p$.
Las variables exógenas como la endógena son cuantitativas.

El modelo lineal (modelo econométrico) viene dado de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + u$$

Los coeficientes (parámetros) $b_1, b_2, \ldots, b_p$ denotan la magnitud del efecto que las variables explicativas (exógenas o independientes $X_1, X_2, \ldots, X_p$ tienen sobre la variable explicada (endógena o dependiente) del modelo. El término u se denomina término de error del modelo.

Se dispone de un conjunto de T observaciones para cada una de las variables endógenas y exógenas, que permite escribir el modelo de la forma:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_p X_{pt} + u_t \quad t = 1, 2, 3, \ldots, T$$

Que aparezca un término independiente (no necesaria) en el modelo puede interpretarse como la presencia de una primera variable X_0 cuyo valor sea siempre 1.

Se aborda el problema: Suponiendo que la relación entre la variable Y y el conjunto de variables $X_1, X_2, \ldots, X_p$ es como se ha descrito en el modelo, y que se dispone de un conjunto de T observaciones para cada una de las variables (la endógena y las exógenas), ¿cómo se pueden asignar valores numéricos a los parámetros $b_1, b_2, \ldots, b_p$ en base a la información muestral?. Estos valores se conocen como estimaciones de los parámetros.

El modelo lineal se formula bajo las hipótesis:

- Las variables $X_1, X_2, \ldots, X_p$ son deterministas (no son variables aleatorias), ya que su valor es un valor constante que proviene de una muestra tomada.
- La variable u (termino de error) es una variable aleatoria que verifica:

  Hipótesis de homocedasticidad: $u \approx N(0, \sigma)$ con varianza σ^2 no dependiente de t

  Hipótesis de no autocorrelación: $Cov(u_i, u_j) = 0 \ \forall i, j \quad i \neq j$
- La variable Y es aleatoria, ya que depende de la variable aleatoria u
- Se supone que todas las variables X que son relevantes para la explicación de la variable Y, están incluidas en la definición del modelo lineal.
- Hipótesis de independencia: Las variables $X_1, X_2, \ldots, X_p$ son linealmente independientes, es decir, no existe relación lineal exacta entre ellas. Cuando no se cumple, se dice que el modelo presenta multicolinealidad.
- Hipótesis de normalidad: A veces se considera que los residuos sean normales para todo t.

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

ESTIMACIÓN DE LOS PARÁMETROS

Sea la muestra aleatoria: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + u_i \quad (i = 1,2, \cdots ,n)$

En forma matricial: $Y = X\beta + U$ siendo X = 'matriz del diseño'

Datos	Y	X_1	X_2	...	X_k
1	Y_1	X_{11}	X_{21}	...	X_{k1}
2	Y_2	X_{12}	X_{22}	...	X_{k2}
⋮	⋮	⋮	⋮	...	⋮
n	Y_n	X_{1n}	X_{2n}	...	X_{kn}

La nube de puntos está en un espacio de dimensión $(k+1)$.

Es difícil de visualizar para $k > 2$

$\hat{\beta} = [X'X]^{-1} X'Y$ donde X' es la matriz transpuesta del diseño

donde, $X'X = \begin{pmatrix} \sum_{i=1}^n X_{i1}^2 & \sum_{i=1}^n X_{i1}X_{i2} & \cdots & \sum_{i=1}^n X_{i1}X_{ik} \\ \sum_{i=1}^n X_{i2}X_{i1} & \sum_{i=1}^n X_{i2}^2 & \cdots & \sum_{i=1}^n X_{i2}X_{ik} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^n X_{ik}X_{i1} & \sum_{i=1}^n X_{ik}X_{i2} & \cdots & \sum_{i=1}^n X_{ik}^2 \end{pmatrix}$, $X'Y = \begin{pmatrix} \sum_{i=1}^n X_{i1}Y_i \\ \sum_{i=1}^n X_{i2}Y_i \\ \vdots \\ \sum_{i=1}^n X_{ik}Y_i \end{pmatrix}$

Cada uno de los coeficientes β_i representa el efecto de la variable independiente sobre la variable explicada. Es decir, el valor estimado $\hat{\beta}_i$ indica la variación que experimenta la variable dependiente cuando la variable independiente X_i varía en una unidad y todas las demás permanecen constantes.

Cuando el modelo tiene término independiente, las matrices anteriores se simplifican con las expresiones:

$X'X = \begin{pmatrix} n & \sum_{i=1}^n X_{i2} & \cdots & \sum_{i=1}^n X_{ik} \\ \sum_{i=1}^n X_{i2} & \sum_{i=1}^n X_{i2}^2 & \cdots & \sum_{i=1}^n X_{i2}X_{ik} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^n X_{ik} & \sum_{i=1}^n X_{ik}X_{i2} & \cdots & \sum_{i=1}^n X_{ik}^2 \end{pmatrix}$, $X'Y = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{i2}Y_i \\ \vdots \\ \sum_{i=1}^n X_{ik}Y_i \end{pmatrix}$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

$$u_i = Y_i - \hat{Y}_i = Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}\right] \qquad \hat{\sigma} = S_R^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} u_i^2$$

En un principio, para estimar la varianza del error aleatorio U, parece razonable utilizar la varianza de los errores de predicción, también denominados residuos del modelo.

Es decir, parece razonable utilizar $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2$. Sin embargo, este estimador es sesgado $E(\hat{\sigma}^2) \neq \sigma^2$

Por tanto, se utiliza como estimador: $S_R^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} u_i^2$

DEMOSTRACIÓN ($Y = X\beta + U \;\rightarrow\; \hat{\beta} = \left[X'X\right]^{-1}X'Y$)

Sea $Y = X\beta + U$, el correspondiente modelo ajustado será $\hat{Y} = X\hat{\beta}$, con lo cual, $\hat{U} = Y - \hat{Y} = Y - X\hat{\beta}$

Denominando S a la suma de los cuadrados de los residuos:

$$S = \hat{U}'\hat{U} = \left[\hat{u}_1, \hat{u}_2, \bullet\bullet\bullet, \hat{u}_n\right]\begin{bmatrix}\hat{u}_1\\ \hat{u}_2\\ \bullet\\ \bullet\\ \bullet\\ \hat{u}_n\end{bmatrix} = \sum_{i=1}^{n}\hat{u}_i^2 \qquad U' \text{ matriz transpuesta de U}$$

$$S = \left[Y - X\hat{\beta}\right]'\left[Y - X\hat{\beta}\right] = Y'Y - \hat{\beta}'X'Y - Y'X\beta + \hat{\beta}'X'X\hat{B} = \underbrace{Y'Y - \hat{\beta}'X'Y - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{B}}_{\text{un escalar es igual a su transpuesto}} =$$

$$\hat{\beta}'X'Y = X'Y\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{B}$$

Para minimizar S se aplica el criterio mínimo-cuadrático, derivando respecto de $\hat{\beta}$:

$$\frac{S}{\vartheta\hat{\beta}} = -2X'Y + 2X'X\hat{B} = 0 \;\Rightarrow\; X'X\hat{B} = X'Y \;\rightarrow\; \left[X'X\right]^{-1}(X'X)\hat{B} = \left[X'X\right]^{-1}X'Y \;\rightarrow\; \hat{B} = \left[X'X\right]^{-1}X'Y$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

DISTRIBUCIÓN $\hat{\beta} \in N\left(\beta, \sigma^2 \left[X'X\right]^{-1}\right)$

a) Las estimaciones de los parámetros vienen dada por la expresión $\hat{\beta} = \left[X'X\right]^{-1} X'Y$, siendo X' la matriz transpuesta del diseño.

b) El vector de observaciones Y se distribuye según una normal multivariante de media $X\beta$ y de matriz de varianzas-covarianzas $\sigma^2 I$, es decir, $Y \in N(X\beta, \sigma^2 I)$.

c) $\hat{\beta}$ es combinación lineal de las componentes del vector Y, por lo que se distribuye según una variable aleatoria normal, donde su media y matriz de varianzas-covarianzas será:

$$E(\hat{\beta}) = E\left(\left[X'X\right]^{-1} X'Y\right) = \left[X'X\right]^{-1} X'E(Y) = \left[X'X\right]^{-1} X'X\beta = \beta \quad \rightarrow \quad \hat{\beta} \text{ es un estimador insesgado de } \beta$$

$$Var(\hat{\beta}) = Var\left(\left[X'X\right]^{-1} X'Y\right) = \left[X'X\right]^{-1} X' \left[Var(Y)\right] X \left[X'X\right]^{-1} = \left[X'X\right]^{-1} X' \sigma^2 X \left[X'X\right]^{-1} = \sigma^2 \left[X'X\right]^{-1}$$

d) Con el ajuste de mínimos cuadrados: $u_i = Y_i - \hat{Y}_i = Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki}\right]$

$\hat{\beta}_i \in N\left(\beta, \sigma^2 q_{i+1,i+1}\right)$, donde $q_{i+1,i+1}$ son los elementos de la diagonal principal $\left[X'X\right]^{-1}$

Análogamente, la covarianza $Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 q_{i+1,i+1}$

e) La estimación de la varianza residual σ^2 se hace mediante $S_R^2 = \dfrac{1}{n-k-1} \sum_{i=1}^{n} u_i^2$, pudiéndose comprobar que el estimador es insesgado: $E\left[S_R^2\right] = \sigma^2$

f) $\hat{\beta}_i \in N\left(\beta, \sigma^2 q_{i+1,i+1}\right)$, se demuestra que $\dfrac{(n-k-1)\, S_R^2}{\sigma^2} \approx \chi^2_{n-k-1}$

Se obtiene, $\dfrac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{q_{i+1,i+1}}} \approx N(0,1)$

h) Siendo t-Student: $t_k = \dfrac{N(0,1)}{\sqrt{\dfrac{1}{k}\chi_k^2}}$, resulta $t = \dfrac{\dfrac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{q_{i+1,i+1}}}}{\sqrt{\dfrac{1}{(n-k-1)} \dfrac{(n-k-1)\, S_R^2}{\sigma}}} = \dfrac{\hat{\beta}_i - \beta_i}{S_R \sqrt{q_{i+1,i+1}}} \approx t_{n-k-1}$

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

CONTRASTE DE HIPÓTESIS [t - Student]

Se plantea si la variable X_i influye sobre la variable de respuesta Y. En otras palabras, si el valor del parámetro en la población es cero o no.

Se establece la hipótesis nula $H_0: \beta_i = 0$ frente a la hipótesis alternativa $H_1: \beta_i \neq 0$

El estadístico observado $t = \dfrac{\hat{\beta}_i - \beta_i}{S_R \sqrt{q_{i+1,i+1}}}$. bajo la hipótesis nula resulta, $t = \dfrac{\hat{\beta}_i}{S_R \sqrt{q_{i+1,i+1}}}$

Se acepta la hipótesis nula H_0 cuando $\overbrace{\left| \dfrac{\hat{\beta}_i}{S_R \sqrt{q_{i+1,i+1}}} \right|}^{\text{estadístico experimental}} \leq \overbrace{t_{\frac{\alpha}{2};\,(n-k-1)}}^{\text{estadístico teórico}}$. En caso contrario, se rechaza.

Si $n > 30$, se acepta la hipótesis nula H_0 cuando $|t| \leq 2$.

En caso contrario, se acepta la hipótesis alternativa H_1, concluyendo que la variable X_i i-ésima influye en la respuesta.

CÁLCULO DEL COEFICIENTE DE CORRELACIÓN PARCIAL

En un modelo de regresión lineal múltiple, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$, se puede calcular fácilmente el coeficiente de correlación parcial entre la variable de respuesta Y y una variable regresora X, controlado por el resto de variables regresoras.

Para ello se utiliza el contraste individual de la t respecto a la variable X, y que se define como:

$$t_i = \frac{\hat{\beta}_i}{S_R \sqrt{q_{i+1,i+1}}} \qquad i = 1,2,\ldots,k,$$

Obteniéndose la relación: $R^2_{YiC} = \dfrac{t_i^2}{t_i^2 + n - (k+1)}$

donde $C = \{1,2,\ldots,i-1,i+1,\ldots,k\}$ conjunto de índices de todas las variables regresoras excepto el índice i-ésimo.

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

INTERVALOS DE CONFIANZA DE LOS PARÁMETROS $\hat{\beta}_i$

Las estimaciones de los parámetros vienen dada por la expresión $\hat{\beta} = \left[X'X\right]^{-1} X'Y$ (siendo X' la matriz transpuesta del diseño).

Por otra parte, $\hat{\beta}_i \in N\left(\beta, \sigma^2 q_{i+1,i+1}\right)$, donde la varianza residual σ^2 se estima por $S_R^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$, donde $q_{i+1,i+1}$ son los elementos de la diagonal principal $\left[X'X\right]^{-1}$.

$$IC_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i \pm t_{\alpha/2,(n-k-1)} S_R \sqrt{q_{i+1,i+1}}\right]$$

CONTRASTE DE HIPÓTESIS - INTERVALOS DE CONFIANZA

Hipótesis nula: $H_0 : \beta_i = 0$ (X_i no influye en Y) Hipótesis alternativa: $H_1 : \beta_i \neq 0$ (X_i influye en Y)

Se acepta la hipótesis nula H_0 (X_i no influye en Y), con un nivel de significación $(1-\alpha)$, cuando el 0 se encuentra en el intervalo de confianza. En caso contrario, se acepta la hipótesis alternativa.

Este contraste es equivalente al contraste de la t-Student para cada β_i

INTERVALO DE CONFIANZA PARA LA VARIANZA DE LOS RESIDUOS

$$IC_{\sigma^2} = \left(\frac{(n-k-1)S_R^2}{\chi^2_{\frac{\alpha}{2},(n-k-1)}} ; \frac{(n-k-1)S_R^2}{\chi^2_{1-\frac{\alpha}{2},(n-k-1)}}\right) = \left(\frac{SCR}{\chi^2_{\frac{\alpha}{2},(n-k-1)}} ; \frac{SCR}{\chi^2_{1-\frac{\alpha}{2},(n-k-1)}}\right)$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

DESCOMPOSICIÓN DE LA VARIABILIDAD

$$SCT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\right]^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\overbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i).(\hat{Y}_i - \bar{Y})}^{=0}$$

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SCE} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SCR}$$

suma cuadrados total (n−1) grados libertad | suma cuadrados explicada k grados libertad | suma cuadrados residual (n−k−1) grados libertad

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \rightarrow 1 = \underbrace{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}_{SCR/SCT} + \underbrace{\frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}_{R^2 = SCE/SCT}$$

Una vez estimado el modelo es conveniente obtener una medida acerca de la bondad del ajuste realizado. Un estadístico que facilita esta medida es el Coeficiente de Determinación R^2, que se define:

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

El Coeficiente de Determinación permite, además, seleccionar entre modelos clásicos que tengan el mismo número de regresores, ya que la capacidad explicativa de un modelo es mayor cuanto más elevado sea el valor que tome este coeficiente.

Por otra parte, el valor coeficiente de determinación crece con el número de regresores del modelo. Por ello, si los modelos que se comparan tienen distinto número de regresores, no puede establecerse comparación entre sus R^2

En este caso debe emplearse el coeficiente de determinación corregido $\bar{R}^2$, que depura el incremento que experimenta el coeficiente de determinación cuando el número de regresores es mayor.

$$\bar{R}^2 = 1 - \frac{SCR/n-k-1}{SCT/n-1} = 1 - \left(\frac{n-1}{n-k-1}\right)\left[1 - R^2\right]$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

ANÁLISIS DE LA VARIANZA: TABLA ANOVA

Variación	Suma cuadrados	Grados libertad	Media cuadrática	F-Snedecor
Explicada	$SCE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	k	$\frac{SCE}{k}$	$F = \frac{SCE/k}{SCR/(n-k-1)}$
Residual	$SCR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	$n-k-1$	$\frac{SCR}{n-k-1}$	
Total	$SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$n-1$		

CONTRASTE DE HIPÓTESIS

Hipótesis nula: H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ el modelo no es explicativo

Hipótesis alternativa: H_1: Al menos un $\beta_i \neq 0$ el modelo es explicativo

A un nivel de confianza $(1-\alpha)$ se rechaza H_0 sí $F \geq F_{\alpha\,;\,k,(n-k-1)}$

RELACIÓN F-Snedecor - COEFICIENTE DE DETERMINACIÓN

Coeficiente de determinación: $R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$

Distribución F-Snedecor:

$$F = \frac{SCE/k}{SCR/(n-k-1)} = \frac{SCE}{SCT} \cdot \frac{SCT}{SCR} \cdot \frac{n-k-1}{k} = R^2 \cdot \frac{1}{\frac{SCR}{SCT}} \cdot \frac{n-k-1}{k} = R^2 \cdot \frac{1}{\frac{SCT-SCE}{SCT}} \cdot \frac{n-k-1}{k} =$$

$$= R^2 \cdot \frac{1}{1-R^2} \cdot \frac{n-k-1}{k} = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} \quad \rightarrow \quad F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

RESUMEN DE CONTRASTES

Contraste Conjunto F-Snedecor	Contrastes Individuales t-Student	Conclusión
Modelo explicativo	Todas las X_i son explicativas	Se toman todas las X_i
Modelo explicativo	Algunas X_i son explicativas	Se toman las X_i explicativas
Modelo explicativo	Ninguna X_i es explicativa	Posible Multicolinealidad (revisar el Modelo)
Modelo no explicativo	Todas las X_i son explicativas	Posible Multicolinealidad (revisar el Modelo)
Modelo no explicativo	Algunas X_i son explicativas	Posible Multicolinealidad (revisar el Modelo)
Modelo no explicativo	Ninguna X_i es explicativa	El modelo no explica Y

PREDICCIÓN EN EL MODELO DE REGRESIÓN

Una vez estimado y validado el modelo, una de sus aplicaciones más importantes consiste en poder realizar predicciones acerca del valor que tomaría la variable dependiente en el futuro o para una unidad extramuestral.

Esta predicción se puede realizar tanto para un valor individual como para un valor medio, o esperado, de la variable dependiente, siendo posible efectuar una predicción puntual o por intervalos. Su cálculo se realiza mediante las siguientes expresiones:

🕮 Intervalo de confianza para un valor medio de Y para los valores $(X_{10}, X_{20}, \cdots\cdots, X_{k0})$ de las variables explicativas $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_{10} + \hat{\beta}_2 X_{20} + \cdots + \hat{\beta}_K X_{K0}$

$$IC_{E(Y_0)} = \left[\hat{Y}_0 \pm t_{\alpha/2,(n-k-1)} S_R \sqrt{(1 \;\; X_{10} \;\; X_{20} \;\cdots\; X_{k0})\,(X'X)^{-1} \begin{pmatrix} 1 \\ X_{10} \\ X_{20} \\ \bullet \\ \bullet \\ \bullet \\ X_{K0} \end{pmatrix}} \right]$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

📖 Intervalo de confianza para un valor individual de Y para los valores $(X_{10}, X_{20}, \cdots\cdots, X_{k0})$ de las variables explicativas $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_{10} + \hat{\beta}_2 X_{20} + \cdots + \hat{\beta}_K X_{K0}$

$$IC_{E(Y_0)} = \left[\hat{Y}_0 \pm t_{\alpha/2,(n-k-1)} S_R \sqrt{1 + (1 \;\; X_{10} \;\; X_{20} \;\cdots\; X_{k0})\,(X'X)^{-1} \begin{pmatrix} 1 \\ X_{10} \\ X_{20} \\ \bullet \\ \bullet \\ \bullet \\ X_{K0} \end{pmatrix}} \right]$$

MATRIZ DE VARIANZAS-COVARIANZAS: $VC = \begin{bmatrix} \sigma_y^2 & S_{yx_1} & S_{yx_2} \\ S_{x_1 y} & \sigma_{x_1}^2 & S_{x_1 x_2} \\ S_{x_2 y} & S_{x_2 x_1} & \sigma_{x_2}^2 \end{bmatrix}$ $\begin{cases} S_{yx_1} = S_{x_1 y} \\ S_{yx_2} = S_{x_2 y} \\ S_{x_1 x_2} = S_{x_2 x_1} \end{cases}$

Los coeficientes $(\hat{\beta}_1, \hat{\beta}_2)$ vienen dados, respectivamente, con signo negativo $(-)$, por el cociente de los adjuntos (S_{yx_1}, S_{yx_2}) entre el adjunto de σ_y^2:

$$\hat{\beta}_1 = -\frac{VC_{yx_1}}{VC_y} \qquad \hat{\beta}_2 = -\frac{VC_{yx_2}}{VC_y} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

donde, $VC_y = \begin{vmatrix} \sigma_{x_1}^2 & S_{x_1 x_2} \\ S_{x_2 x_1} & \sigma_{x_2}^2 \end{vmatrix}$ $VC_{yx_1} = -\begin{vmatrix} S_{x_1 y} & S_{x_1 x_2} \\ S_{x_2 y} & \sigma_{x_2}^2 \end{vmatrix}$ $VC_{yx_2} = \begin{vmatrix} S_{x_1 y} & \sigma_{x_1}^2 \\ S_{x_2 y} & S_{x_2 x_1}^2 \end{vmatrix}$

Coeficiente de determinación múltiple: $R^2 = r_{y\,x_1 x_2}^2 = 1 - \dfrac{|CV|}{\sigma_y^2\, C_{yy}}$

Coeficientes de correlación parcial: $r_{yx_1 \cdot x_2} = -\dfrac{VC_{yx_1}}{\sqrt{VC_{yy}\, VC_{x_1 x_1}}}$ $r_{yx_2 \cdot x_1} = -\dfrac{VC_{yx_2}}{\sqrt{VC_{yy}\, VC_{x_2 x_2}}}$

siendo, $VC_{\sigma_y^2} = VC_{yy}$, $VC_{\sigma_{x_1}} = VC_{x_1 x_1}$, $VC_{\sigma_{x_2}} = VC_{x_2 x_2}$

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

MATRIZ DE CORRELACIONES

La matriz de correlaciones de las variables explicativas R_x está formada por los coeficientes de correlación lineal simple:

$$R_x = \begin{bmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{x_1y} & 1 & r_{x_1x_2} \\ r_{x_2y} & r_{x_2x_1} & 1 \end{bmatrix} \quad \text{donde} \quad \begin{cases} r_{yx_1} = r_{x_1y} \\ r_{yx_2} = r_{x_2y} \\ r_{x_1x_2} = r_{x_2x_1} \end{cases} \qquad r_{yx_1} = \frac{S_{yx_1}}{\sigma_y \sigma_{x_1}} \qquad r_{yx_2} = \frac{S_{yx_2}}{\sigma_y \sigma_{x_2}}$$

Coeficiente de determinación múltiple: $R^2 = r^2_{yx_1x_2} = \dfrac{r^2_{yx_1} + r^2_{yx_2} - 2\, r_{yx_1} r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}}$

En la regresión lineal múltiple surge el problema de que exista una correlación lineal simple perfecta entre dos (o más variables) explicativas, ya que esto implica que una (o más) columna(s) de la matriz X de observaciones son combinación lineal de otra(s), con lo que el rango de esta matriz X se reduce.

En un principio el rango de $\left[X'X\right]$ es p (número de variables explicativas), pero si existe alguna combinación lineal entre las columnas de X, entonces el rango es menor que p, con lo que el determinante de $\left|X'X\right| = 0$, lo que impide calcular la matriz inversa $\left[X'X\right]^{-1}$, y en consecuencia el vector de coeficientes $\beta = \left[X'X\right]^{-1} X'Y$ queda indeterminado.

Analizando la matriz de las correlaciones R_x se decide si existe o no multicolinealidad:

- Sí $\left|R_x\right| = 0 \rightarrow$ Existe multicolinealidad
- Sí $\left|R_x\right| \approx 0 \rightarrow$ Existe cuasi-multicolinealidad o multicolinealidad imperfecta

En caso de multicolinealidad se requiere modificar el modelo o realizar algún tipo de transformación que la elimine.

El coeficiente de correlación simple está basado en la suposición de la aproximación a la distribución normal bivariante. Si se tiene más de dos variables, el modelo básico para la correlación múltiple, sería una ampliación de esta distribución, denominada distribución normal multivariante (p-Normal).

Si hay tres variables, habrá tres correlaciones simples entre ellas, $r_{yx_1 \cdot x_2}$, $r_{yx_2 \cdot x_1}$, $r_{x_1x_2 \cdot y}$. Estos coeficientes miden la relación lineal que existen entre estas variables, dos a dos, sin tener en cuenta la posible influencia de la tercera.

La correlación parcial se define como la correlación entre dos variables si las demás variable no varían, es decir, el valor de las demás variables son fijos. El coeficiente de correlación parcial $r_{yx_1 \cdot x_2}$, es la correlación entre la variable Y e X_1 siendo constante el valor de la variable X_2. Es decir, mide la correlación entre la variable Y e X_1 que no sea un reflejo de sus relaciones con la variable X_2

Coeficientes de correlación parcial: $r_{yx_1 \cdot x_2} = \dfrac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}} \qquad r_{yx_2 \cdot x_1} = \dfrac{r_{yx_2} - r_{yx_1} \cdot r_{x_2x_1}}{\sqrt{(1 - r^2_{yx_1})(1 - r^2_{x_2x_1})}}$

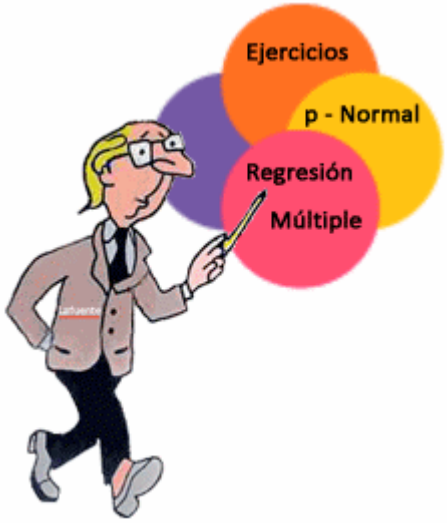
(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... **Grupo** .....

Apellidos ..... **Nombre** .....

Ejercicio del día .....



UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

1 Sea X una variable aleatoria con distribución normal de media cero y varianza 1, e Y una variable aleatoria de forma que la distribución de Y condicionada por $X = x$ es normal de media x y varianza 1.

¿La variable aleatoria bidimensional (X,Y) sigue una distribución normal bidimensional?

$$f(y|x) = \frac{f(x,y)}{f_1(x)} \rightarrow f(x,y) = f(y|x)\,.\,f_1(x)$$

$$f(y|x) = \frac{1}{2\pi}\, e^{-\frac{1}{2}(y-x)^2}$$

$$f_1(x) = \frac{1}{2\pi\sigma_1}\, e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma_1^2}} = \frac{1}{2\pi}\, e^{-\frac{x^2}{2}}$$

$$f(x,y) = \frac{1}{2\pi}\, e^{-\frac{1}{2}(y-x)^2} \,.\, \frac{1}{2\pi}\, e^{-\frac{x^2}{2}} = \frac{1}{2\pi}\, e^{-\frac{1}{2}(y^2+2x^2-2xy)} \qquad (x,y)\in R^2$$

Que se corresponde con la densidad de una normal bidimensional con vector de medias $\big(E(X)=0, E(Y)=0\big)$ y matriz de varianzas-covarianzas

$$\Sigma = Var(x,y) = \begin{pmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

2 Sea (X,Y) una variable aleatoria bidimensional con función de densidad:

$$f(x,y) = \frac{\sqrt{3}}{\pi}\, e^{-x^2-3y^2} \qquad (x,y)\in R^2$$

Calcular el vector de medias, la matriz de varianzas-covarianzas y coeficiente de correlación.

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2\,\sigma_1^2\,\sigma_2^2\,(1-\rho^2)}\left[\sigma_2^2(x-\mu_1)^2 + \sigma_1^2(y-\mu_2)^2 - 2\,\sigma_1\sigma_2\,\rho\,(x-\mu_1)(y-\mu_2)\right]}$$

Igualando coeficientes, se tiene:

$$\frac{1}{2\sigma_1\sigma_2\sqrt{1-\rho^2}} = \sqrt{3} \qquad \mu_1 = 0 \qquad \mu_2 = 0$$

$$2\sigma_1^2(1-\rho^2) = 1$$

$$2\sigma_2^2(1-\rho^2) = \frac{1}{3}$$

$$\frac{2\rho}{\sigma_1\sigma_2}\,.\,\frac{1}{2(1-\rho^2)} = 0$$

Operando se obtiene: $\sigma_1^2 = \frac{1}{2}$ $\sigma_2^2 = \frac{1}{6}$ $\rho = 0$

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

3 Sea (X,Y) una variable aleatoria con distribución normal bidimensional cuyo vector de medias es $(50,45)$ y matriz de varianzas-covarianzas $\Sigma = \begin{pmatrix} 6 & 2 \\ 2 & 4 \end{pmatrix}$

Se consideran las variables aleatorias: $\begin{cases} U = 4X + Y \\ V = X + 4Y \end{cases}$

Calcular: $P(U \geq 250)$ y $P(V \geq 220)$

La matriz de varianzas-covarianzas de la variable aleatoria (U,V) viene dada por:

$$\Sigma^* = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} 6 & 2 \\ 2 & 4 \end{pmatrix}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 116 & 74 \\ 74 & 86 \end{pmatrix}$$

En consecuencia, las medias y varianzas de las variables aleatorias (U,V) son:

$$\begin{cases} U = 4X + Y \\ V = X + 4Y \end{cases} \rightarrow \begin{cases} \mu_U = 4 \times 50 + 45 = 245 & \sigma_u^2 = 116 \\ \mu_V = 50 + 4 \times 45 = 230 & \sigma_V^2 = 86 \end{cases}$$

$$P(U \geq 250) = P\left(\frac{U-245}{\sqrt{116}} \geq \frac{250-245}{\sqrt{116}}\right) = P(z \geq 0{,}464) = 0{,}3228$$

$$P(V \geq 220) = P\left(\frac{V-230}{\sqrt{86}} \geq \frac{220-230}{\sqrt{86}}\right) = P(z \geq -1{,}0787) = 0{,}8599$$

4 Se sabe que las rectas de regresión de una variable aleatoria normal bidimensional (X,Y) son:

$y = 2 + x$, $y = 4 + \frac{3}{10}x$ y la suma de las varianzas es 2.

Calcular la matriz de varianzas-covarianzas de la variable aleatoria (X,Y), así como el vector de medias.

Recta de regresión de Y sobre X: $y - \mu_2 = \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1)$

Recta de regresión de X sobre Y: $x - \mu_1 = \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2)$

Por otra parte, la pendiente de la recta de regresión de X sobre Y es mayor que la de Y sobre X. En consecuencia, $y = 2 + x$ es la recta de regresión de X sobre Y, con lo que:

$$\rho \frac{\sigma_1}{\sigma_2} = 1 \rightarrow x = \mu_1 + (y - \mu_2) \rightarrow x = y + \mu_1 - \mu_2 \Rightarrow \mu_1 - \mu_2 = -2$$

De forma análoga con la recta de regresión de Y sobre X: $y = 4 + \frac{3}{10}x$

$$\rho \frac{\sigma_2}{\sigma_1} = \frac{3}{10} \rightarrow y = \mu_2 + \frac{3}{10}(x - \mu_1) \rightarrow y = \frac{3}{10}x + \mu_2 - \frac{3}{10}\mu_1 \Rightarrow \mu_2 - \frac{3}{10}\mu_1 = 4$$

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Resolviendo el sistema: $\begin{cases} \mu_1 - \mu_2 = -2 \\ -\frac{3}{10}\mu_1 + \mu_2 = 4 \end{cases} \rightarrow \mu_1 = \frac{20}{7} \ , \ \mu_2 = \frac{34}{7}$

El valor de varianzas y coeficientes de correlación se obtiene resolviendo el sistema:

$$\begin{cases} \rho \frac{\sigma_1}{\sigma_2} = 1 \\ \rho \frac{\sigma_2}{\sigma_1} = \frac{3}{10} \\ \sigma_1^2 + \sigma_2^2 = 2 \end{cases} \rightarrow \begin{vmatrix} \rho^2 = \frac{3}{10} \rightarrow \rho = \sqrt{\frac{3}{10}} \\ \frac{\sigma_2^2}{\sigma_1^2} = \frac{3}{10} \\ \sigma_1^2 + \frac{3}{10}\sigma_1^2 = 2 \rightarrow \sigma_1^2 = \frac{20}{13} \ , \ \sigma_2^2 = \frac{6}{13} \end{vmatrix}$$

La covarianza: $Cov(x,y) = \sigma_{12} = \sigma_{21} \equiv \rho\,\sigma_1\,\sigma_2 = \sqrt{\frac{3}{10} \times \frac{20}{13} \times \frac{6}{13}} = \frac{6}{13}$

La matriz de varianzas-covarianzas: $\Sigma = \begin{pmatrix} \frac{20}{13} & \frac{6}{13} \\ \frac{6}{13} & \frac{6}{13} \end{pmatrix}$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

5 Entre las parejas en las que hombre y mujer trabajan, la distribución conjunta de los salarios mensuales (X,Y), expresados en euros, se ajusta a una distribución normal bivariante con los siguientes parámetros:

X (salario de los hombres) , $X \sim N(1720, 100)$; Y (salario de las mujeres) , $Y \sim N(1420, 80)$; $\rho = 0{,}71$

a) Salario medio de las mujeres cuya pareja tiene un salario de 1800.

b) Probabilidad de que la mujer gane más que su pareja, cuando el salario del hombre es 1700.

c) Probabilidad de que la mujer gane menos que su pareja, cuando el salario del hombre es 1800.

a) La línea de regresión de Y sobre X proporciona los valores medios de la variable Y cuando X toma el valor concreto x. En el caso de la distribución normal bivariante, se tiene:

$$E\left[Y \mid X = x\right] = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1) \rightarrow E\left[Y \mid X = 1800\right] = 1420 + 0{,}71 \times \frac{80}{100}(1800 - 1720) = 1465{,}44$$

b) $P(Y > 1700 \mid X = 1700)$

Las distribuciones condicionadas se distribuyen normalmente:

$$E\left[Y \mid X = x\right] = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1) \qquad Var\left[Y \mid X = x\right] = \sigma_2^2 (1 - \rho^2)$$

$$E\left[Y \mid X = 1700\right] = 1420 + 0{,}71 \times \frac{80}{100}(1700 - 1720) = 1408{,}64$$

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$\text{Var}\left[Y \mid X=1700\right] = 80^2 \text{x}\,(1-0{,}71^2) = 3173{,}76 \;\rightarrow\; \sigma_{Y|X} = 56{,}3361$$

$$(Y \mid X) \approx N(1408{,}64\,,\; 56{,}3361)$$

$$P(Y>1700 \mid X=1700) = P\left(\frac{Y-1408{,}64}{56{,}3361} > \frac{1700-1408{,}64}{56{,}3361}\right) = P(z>5{,}17) \simeq 0$$

c) $P(Y<1800 \mid X=1800)$ donde $(Y \mid X) \approx N(1465{,}44\,,\; 56{,}3361)$

$$\text{Var}\left[Y \mid X=1800\right] = 80^2 \text{x}\,(1-0{,}71^2) = 3173{,}76 \;\rightarrow\; \sigma_{Y|X} = 56{,}3361$$

$$P(Y<1800 \mid X=1800) = P\left(\frac{Y-1465{,}44}{56{,}3361} < \frac{1800-1465{,}44}{56{,}3361}\right) = P(z<5{,}94) \simeq 1$$

Dada una variable aleatoria (X,Y) con distribución normal y función de densidad:

$$f(x,y) = k\, e^{-x^2+xy-2y^2} \quad \forall (x,y) \in R^2$$

a) Cacular el valor de k y matriz de varianzas-covarianzas.

b) $P(-0{,}4 \le Y \le 0{,}4 \mid X=3)$

a) La función de densidad de una variable aleatoria normal con vector de medias $\mu=(\mu_1,\mu_2)$, varianzas σ_1^2 , σ_2^2 y coeficiente de correlación ρ^2 , viene dada por:

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2\,\sigma_1^2\,\sigma_2^2\,(1-\rho^2)}\left[\sigma_2^2\,(x-\mu_1)^2 + \sigma_1^2\,(y-\mu_2)^2 - 2\,\sigma_1\sigma_2\,\rho\,(x-\mu_1)(y-\mu_2)\right]}$$

Igualando coeficientes, se tiene:

$\mu_1 = \mu_2 = 0$

$\rho = 1/\sqrt{8}$ $\qquad \sigma_{12} = \sigma_{21} \equiv \rho\,\sigma_1\sigma_2 = \sqrt{\frac{1}{8}\text{x}\frac{4}{7}\text{x}\frac{2}{7}} = \frac{1}{7}$

$\sigma_1^2 = 4/7$

$\sigma_2^2 = 2/7$ $\qquad X \sim N(0, 2/\sqrt{7})$ $\quad Y \sim N(0, \sqrt{2/7})$

$k = \sqrt{7}/2\pi$

La matriz de varianzas-covarianzas: $\Sigma = \begin{pmatrix} \frac{4}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{2}{7} \end{pmatrix}$

b) $P(-0{,}4 \le Y \le 0{,}4 \mid X=3)$

Las distribuciones condicionadas se distribuyen normalmente:

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$E\left[Y \mid X=x\right] = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x-\mu_1) \qquad Var\left[Y \mid X=x\right] = \sigma_2^2 (1-\rho^2)$$

$$E\left[Y \mid X=3\right] = 0 + \frac{1}{\sqrt{8}} x \frac{\frac{\sqrt{2}}{\sqrt{7}}}{\frac{2}{\sqrt{7}}} (3-0) = \frac{3}{4}$$

$$Var\left[Y \mid X=3\right] = \frac{2}{7} x \left(1-\frac{1}{8}\right) = \frac{1}{4} \rightarrow \sigma_{Y|X} = \frac{1}{2}$$

$$(Y \mid X) \approx N(0{,}75,\ 0{,}5)$$

$$P(-0{,}4 \le Y \le 0{,}4 \mid X=3) = P\left(\frac{-0{,}4-0{,}75}{0{,}5} \le \frac{Y-0{,}75}{0{,}5} \le \frac{0{,}4-0{,}75}{0{,}5}\right) = P(-2{,}3 \le z \le -0{,}7) =$$

$$= P(0{,}7 \le z \le 2{,}3) = 0{,}2420 - 0{,}0107 = 0{,}2313$$

Sean cuatro variables independientes: $X \sim N(0,2)$, $Y \sim N(0,4)$, $Z \sim N(0,6)$, $V \sim N(0,8)$.

Se definen las variables aleatorias: $\begin{cases} T_1 = X+Y+Z+V \\ T_2 = X+Y+Z \\ T_3 = X \end{cases}$

Calcular la función de densidad de la variable aleatoria tridimensional (T_1, T_2, T_3)

Matriz de varianzas-covarianzas de la variable aleatoria (X, Y, Z, V) $\quad \Sigma = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix}$

La matriz de varianzas-covarianzas de la variable aleatoria (T_1, T_2, T_3) es: $\Sigma^* = A\Sigma A^t$

$$\Sigma^* = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 20 & 12 & 2 \\ 12 & 12 & 2 \\ 2 & 2 & 2 \end{pmatrix}$$

En consecuencia, la variable aleatoria (T_1, T_2, T_3) sigue una distribución normal tridimensional $N_3(\mu, \Sigma)$ con vector de medias $\mu = (0, 0, 0)$ y matriz de varianzas-covarianzas $\begin{pmatrix} 20 & 12 & 2 \\ 12 & 12 & 2 \\ 2 & 2 & 2 \end{pmatrix}$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

8 Dada la variable aleatoria normal tridimensional (X_1, X_2, X_3) con función de densidad:

$$f(x_1, x_2, x_3) = \frac{\sqrt{48}}{\sqrt{8\pi^3}} e^{-(x_1^2 + 2x_2^2 + 3x_3^2)} \quad \forall (x_1, x_2, x_3) \in R^3$$

Cacular la matriz de varianzas-covarianzas.

Función de densidad: $f(X_1, X_2, \ldots, X_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)}$

Se debe verificar: $-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu) = -(x_1^2 + 2x_2^2 + 3x_3^2)$

Donde, $\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ y $\frac{1}{2}\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \rightarrow \Sigma = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{pmatrix} \rightarrow |\Sigma| = 48$

$$\Sigma^{-1} = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/6 \end{pmatrix}$$

9 Sea $X = (X_1, X_2, X_3)^t$ un vector aleatorio con ley normal de media $\mu = (-1, 1, 0)^t$ y matriz de varianzas-covarianzas $\Sigma = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix}$

a) Encontrar la ley de la variable aleatoria $Y = X_1 + 2X_2 - 3X_3$

b) Encontrar un vector $\vec{a}(2 \times 1)$ para que las variables X_1 y $X_1 - \vec{a} \begin{pmatrix} X_2 \\ X_3 \end{pmatrix}$ sean independientes

c) Calcular la distribución $X_3 \,|\, (X_1, X_2)^t$

a) La variable aleatoria $Y = X_1 + 2X_2 - 3X_3 = \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix} (X_1 \quad X_2 \quad X_3)$ siendo $b = (1, 2, -3)^t$

Como $X \sim N_3(\mu, \Sigma) \rightarrow$ la variable aleatoria $Y = b^t X \approx N(b^t \mu, b^t \Sigma b)$

$$E(Y) = E(b^t X) = b^t \mu = \begin{pmatrix} 1 & 2 & -3 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = 1$$

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$\text{Var}(Y) = \text{Var}(b^t X) = b^t \Sigma\, b = \begin{pmatrix}1 & 2 & -3\end{pmatrix}\begin{pmatrix}1 & 0 & 1\\ 0 & 3 & 1\\ 1 & 1 & 2\end{pmatrix}\begin{pmatrix}1\\ 2\\ -3\end{pmatrix} = 13$$

b) Sea $a = (a_1, a_2)^t$ tal que $X_1 - \begin{pmatrix}a_1 & a_2\end{pmatrix}\begin{pmatrix}X_2\\ X_3\end{pmatrix} = X_1 - a_1 X_2 - a_2 X_3$ sean independientes

Como independencia, bajo el supuesto de normalidad, equivale a incorrelación $(\rho = 0)$, hay que encontrar un vector $\text{Cov}(X_1,\ X_1 - a_1 X_2 - a_2 X_3) = 0$

Siendo: $X_1 - a_1 X_2 - a_2 X_3 = (1, -a_1, -a_2)X$, $X_1 = (1,0,0)X$

$$0 = \text{Cov}(X_1,\ X_1 - a_1 X_2 - a_2 X_3) = (1,0,0)\,\Sigma\begin{pmatrix}1\\ -a_1\\ -a_2\end{pmatrix} = (1,0,0)\begin{pmatrix}1 & 0 & 1\\ 0 & 3 & 1\\ 1 & 1 & 2\end{pmatrix}\begin{pmatrix}1\\ -a_1\\ -a_2\end{pmatrix} = 1 - a_2 \rightarrow a_2 = 1$$

En consecuencia, todo vector $a = (\lambda, 1)$ $\lambda \in R$ hace que las variables X_1 y $X_1 - a_1 X_2 - a_2 X_3$ sean independientes.

c) La distribución $\left[X_3 \,\middle|\, (X_1, X_2)^t\right]$ es una normal univariante con media y esperanza:

$$X = \begin{pmatrix}X_1\\ X_2\\ X_3\end{pmatrix} \approx N_3(\mu, \Sigma) \text{ donde } \mu = \begin{pmatrix}\mu_1\\ \mu_2\\ \mu_3\end{pmatrix} = \begin{pmatrix}-1\\ 1\\ 0\end{pmatrix} \text{ y } \Sigma = \left(\begin{array}{cc|c}1 & 0 & 1\\ 0 & 3 & 1\\ \hline 1 & 1 & 2\end{array}\right) = \left(\begin{array}{c|c}\Sigma_{(12)(12)} & \Sigma_{(12)3}\\ \hline \Sigma_{3(12)} & \Sigma_{33}\end{array}\right)$$

$$E(X_3 \,|\, (X_1, X_2)^t) = \mu_3 + \Sigma_{3(12)}\,\Sigma^{-1}_{(12)(12)}\begin{pmatrix}x_1 - \mu_1\\ x_2 - \mu_2\end{pmatrix}$$

$$\text{Var}(X_3 \,|\, (X_1, X_2)^t) = \Sigma_{33} - \Sigma_{3(12)}\,\Sigma^{-1}_{(12)(12)}\,\Sigma_{(12)3}$$

$$E\left[X_3 \,|\, (X_1, X_2)^t\right] = 0 + \begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 3\end{pmatrix}^{-1}\begin{pmatrix}x_1 + 1\\ x_2 - 1\end{pmatrix} = \begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 1/3\end{pmatrix}\begin{pmatrix}x_1 + 1\\ x_2 - 1\end{pmatrix} = \begin{pmatrix}1 & 1/3\end{pmatrix}\begin{pmatrix}x_1 + 1\\ x_2 - 1\end{pmatrix} =$$

$$= x_1 + \frac{1}{3}x_2 + \frac{2}{3}$$

$$\text{Var}(X_3 \,|\, (X_1, X_2)^t) = 2 - \begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 3\end{pmatrix}^{-1}\begin{pmatrix}1\\ 1\end{pmatrix} = 2 - \begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 1/3\end{pmatrix}\begin{pmatrix}1\\ 1\end{pmatrix} = 2 - \begin{pmatrix}1 & 1/3\end{pmatrix}\begin{pmatrix}1\\ 1\end{pmatrix} =$$

$$= 2 - \left(1 + \frac{1}{3}\right) = \frac{2}{3}$$

Por tanto, $\left[X_3 \,|\, (X_1, X_2)^t\right] \approx N\left(x_1 + \frac{1}{3}x_2 + \frac{2}{3},\ \frac{2}{3}\right)$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

10 Sean (X_1, X_2, X_3, X_4) vectores aleatorios independientes con ley $N_2(\mu, \Sigma)$, donde $\mu = (1,2)^t$ y $\Sigma = \begin{pmatrix} 1 & 0{,}1 \\ 0{,}1 & 2 \end{pmatrix}$

a) Encontrar la ley del vector aleatorio $Y = \frac{1}{4}X_1 - \frac{1}{4}X_2 + \frac{1}{4}X_3 - \frac{1}{4}X_4$

b) Función de densidad del vector Y

c) Calcular la correlación correspondiente a la matriz de varianzas-covarianzas Σ. Cambios que se producen en la densidad de Y al variar el valor de $\rho = \text{Corr}(Y_1, Y_2)$

a) El vector Y tiene una ley normal bivariante con media y matriz de varianzas-covarianzas:

$$E(Y) = E\left(\frac{1}{4}X_1 - \frac{1}{4}X_2 + \frac{1}{4}X_3 - \frac{1}{4}X_4\right) = \frac{1}{4}E(X_1) - \frac{1}{4}E(X_2) + \frac{1}{4}E(X_3) - \frac{1}{4}E(X_4) = \frac{\mu}{4} - \frac{\mu}{4} + \frac{\mu}{4} - \frac{\mu}{4} = 0$$

$$V(Y) = V\left(\frac{1}{4}X_1 - \frac{1}{4}X_2 + \frac{1}{4}X_3 - \frac{1}{4}X_4\right) = \frac{1}{16}V(X_1) + \frac{1}{16}V(X_2) + \frac{1}{16}V(X_3) + \frac{1}{16}V(X_4) = \frac{1}{4}\Sigma$$

En consecuencia, $Y \approx N_2\left(\vec{0}, \frac{1}{4}\Sigma\right)$

b) Función de densidad del vector Y: $f(Y_1, Y_2) = \frac{1}{\sqrt{(2\pi)^2 \left|\frac{1}{4}\Sigma\right|}} e^{-\frac{1}{2}y^t\left(\frac{1}{4}\Sigma\right)^{-1}y} = \frac{1}{\pi\sqrt{|\Sigma|}} e^{-2y^t\Sigma^{-1}y}$

c) La matriz de varianzas-covarianzas de Y es $\Sigma = \begin{pmatrix} 1 & 0{,}1 \\ 0{,}1 & 2 \end{pmatrix}$

Coeficiente de correlación: $\rho = \frac{\sigma_{12}}{\sigma_1 \cdot \sigma_2} = \frac{0{,}1}{1 \cdot \sqrt{2}} = 0{,}071$

- Análisis del coeficiente de correlación en la densidad:

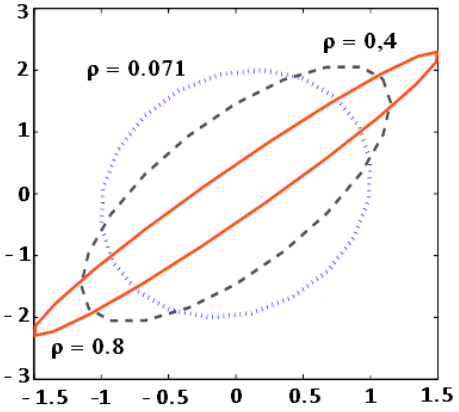
Siendo la covarianza $\sigma_{12} \equiv \rho\,\sigma_1\,\sigma_2$, la matriz Σ puede escribirse $\Sigma = \begin{pmatrix} 1 & \sqrt{2}\,\rho \\ \sqrt{2}\,\rho & 2 \end{pmatrix}$

Variación total: Traza $(\Sigma) = 3$ no depende de ρ

Varianza generalizada:

$$|\Sigma| = 2(1-\rho^2) \equiv \begin{cases} \rho^2 = 0 \rightarrow |\Sigma| = 2 \\ \rho^2 = 1 \rightarrow |\Sigma| = 0 \end{cases}$$

A medida que aumenta la correlación ρ disminuye la dispersión global.



(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

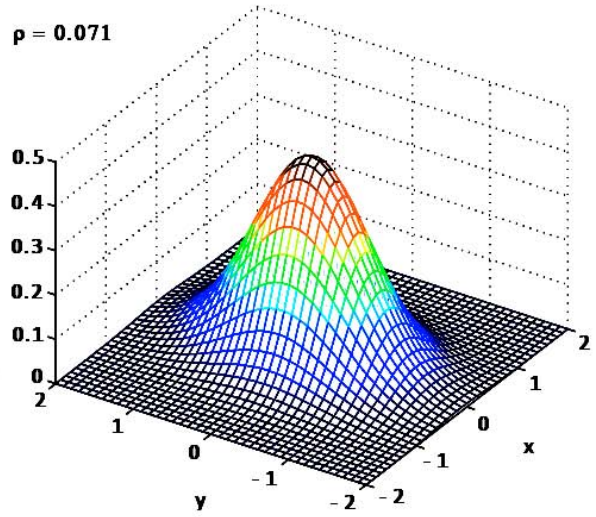
UAM
Universidad Autónoma de Madrid

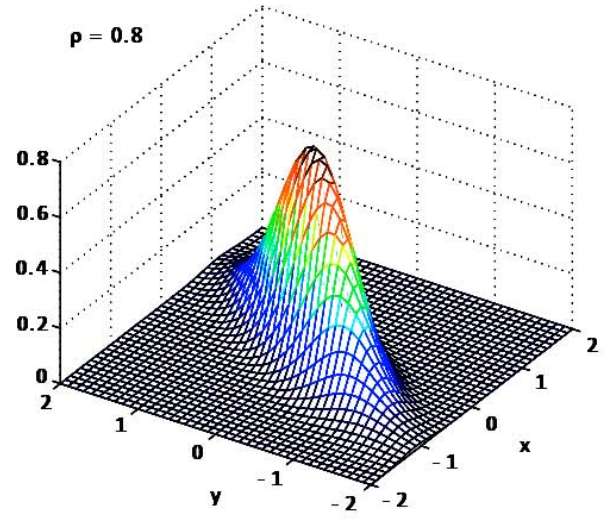
Asignatura ..... Grupo .....

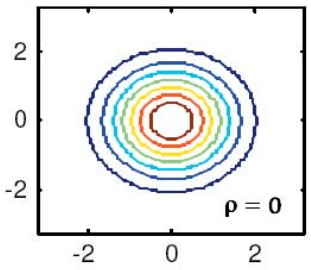
Apellidos ..... Nombre .....

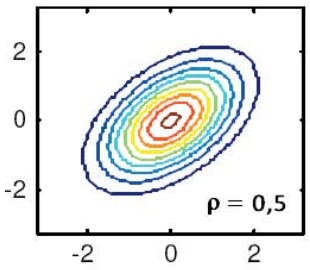
Ejercicio del día .....

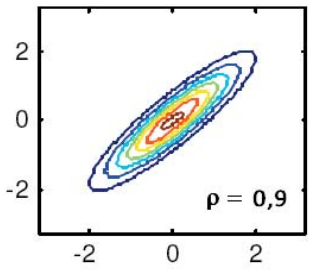
Cuando $\rho = 0$ la superficie es la clásica campana de gauss, en la posición marcada por el vector de medias de la distribución.

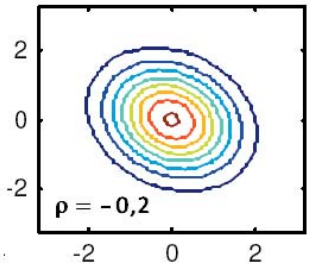


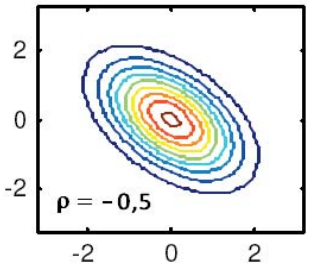


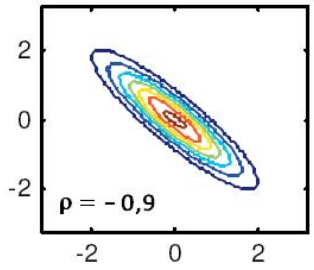












UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

11 Sean Y_1 e Y_2 variables aleatorias independientes con ley $N(0,1)$.

Sean las variables $X_1 = 2Y_1 - Y_2 + 3$ y $X_2 = Y_1 + Y_2 + 5$

a) Hallar la función de densidad de (X_1, X_2)

b) ¿Cómo se distribuye la variable $Z = 3X_1 + 4X_2 - 1$?

a) Como son variables normales independientes, la distribución conjunta de Y_1 e Y_2 es una variable bidimensional estándar:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \approx N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$$

Por otra parte, (X_1, X_2) se puede expresar en función de (Y_1, Y_2):

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} + \begin{pmatrix} 3 \\ 5 \end{pmatrix}, \text{ es decir, } X = B \cdot Y + C$$

Por tanto, X sigue una distribución normal bidimensional: $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \approx N_2\left[\begin{pmatrix} 3 \\ 5 \end{pmatrix}, \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}\right]$

$$\mu_X = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 3 \\ 5 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$\Sigma_X = B \cdot \Sigma_Y \cdot B^t = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix} \text{ matriz definida positiva}$$

b) La variable $Z = 3X_1 + 4X_2 - 1$ se puede expresar en función de (X_1, X_2):

$$Z = 3X_1 + 4X_2 - 1 = \begin{pmatrix} 3 & 4 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} - 1$$

Z sigue una distribución normal unidimensional: $Z \approx N\left(28, \sqrt{101}\right)$

$$\mu_Z = \begin{pmatrix} 3 & 4 \end{pmatrix}\mu_X - 1 = \begin{pmatrix} 3 & 4 \end{pmatrix}\begin{pmatrix} 3 \\ 5 \end{pmatrix} - 1 = 28$$

$$\sigma_Z^2 = \begin{pmatrix} 3 & 4 \end{pmatrix}\Sigma_X \begin{pmatrix} 3 & 4 \end{pmatrix}^t = \begin{pmatrix} 3 & 4 \end{pmatrix}\begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 3 \\ 4 \end{pmatrix} = 101$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

12 Sea (X_1,X_2,X_3) una variable aleatoria con distribución $N_3\left[\begin{pmatrix}2\\1\\3\end{pmatrix},\begin{pmatrix}4&1&2\\1&1&2\\2&2&9\end{pmatrix}\right]$

a) Sea una nueva variable tridimensional tal que $\begin{cases}Z_1 = X_1 - 2X_2 + 3X_3\\ Z_2 = 3X_1 - X_2 + 2X_3\\ Z_3 = -X_1 + 2X_2 + 2X_3\end{cases}$

Demostrar que (Z_1,Z_2,Z_3) sigue una distribución normal tridimensional y obtener sus parámetros.

b) Sea una nueva variable $Z_4 = 3X_1 - X_2 + 7X_3$. Probar que (Z_1,Z_2,Z_3,Z_4) es una nueva variable cuatridimensional y calcular su esperanza y su matriz de varianzas-cuasivarianzas.

c) Sea otra nueva variable $Z_5 = X_1 + 3X_2 - 4X_3$. ¿Es (Z_1,Z_2,Z_5) una normal tridimensional?

a) La variable tridimensional $Z = (Z_1,Z_2,Z_3)^t$ puede expresarse como una transformación lineal de $X = (X_1,X_2,X_3)^t$:

$$\begin{pmatrix}Z_1\\Z_2\\Z_3\end{pmatrix} = \begin{pmatrix}1&-2&3\\3&-1&2\\-1&2&2\end{pmatrix}\begin{pmatrix}X_1\\X_2\\X_3\end{pmatrix} \quad Z = B\,.\,X \quad \text{rango}(B) = \begin{pmatrix}1&-2&3\\3&-1&2\\-1&2&2\end{pmatrix} = 3 \rightarrow$$ Las filas son linealmente independientes

$$\mu_Z = B\,.\,\mu = \begin{pmatrix}1&-2&3\\3&-1&2\\-1&2&2\end{pmatrix}\begin{pmatrix}2\\1\\3\end{pmatrix} = \begin{pmatrix}9\\11\\6\end{pmatrix}$$

$$\Sigma_Z = B\,.\,\Sigma\,.\,B^t = \begin{pmatrix}1&-2&3\\3&-1&2\\-1&2&2\end{pmatrix}\begin{pmatrix}4&1&2\\1&1&2\\2&2&9\end{pmatrix}\begin{pmatrix}1&3&-1\\-2&-1&2\\3&2&2\end{pmatrix} = \begin{pmatrix}73&69&52\\69&83&41\\52&41&48\end{pmatrix}$$ definida positiva

b) La variable cuatridimensional $Z^t = (Z_1,Z_2,Z_3,Z_4)^t$ también es una transformación lineal de X:

$$\begin{pmatrix}Z_1\\Z_2\\Z_3\\Z_4\end{pmatrix} = \begin{pmatrix}1&-2&3\\3&-1&2\\-1&2&2\\3&-1&7\end{pmatrix}\begin{pmatrix}X_1\\X_2\\X_3\end{pmatrix} \quad Z^* = B^*\,.\,X$$

Z^* no sigue una distribución normal porque tiene una componente más que X

Esperanza y matriz de varianzas-covarianzas de Z^*:

$$E(Z^*) = E(B^*\,.\,X) = B^*\,.\,E(X) = \begin{pmatrix}1&-2&3\\3&-1&2\\-1&2&2\\3&-1&7\end{pmatrix}\begin{pmatrix}2\\1\\3\end{pmatrix} = \begin{pmatrix}9\\11\\6\\26\end{pmatrix}$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$V(Z^*) = V(B^*.X) = B^*.V(X).(B^*)^t = \begin{pmatrix} 1 & -2 & 3 \\ 3 & -1 & 2 \\ -1 & 2 & 2 \\ 3 & -1 & 7 \end{pmatrix} \begin{pmatrix} 4 & 1 & 2 \\ 1 & 1 & 2 \\ 2 & 2 & 9 \end{pmatrix} \begin{pmatrix} 1 & 3 & -1 & 3 \\ -2 & -1 & 2 & -1 \\ 3 & 2 & 2 & 7 \end{pmatrix} =$$

$$= \left(\begin{array}{ccc|c} 73 & 69 & 52 & 194 \\ 69 & 83 & 41 & 193 \\ \underline{52} & \underline{41} & \underline{48} & \underline{141} \\ 194 & 193 & 141 & 528 \end{array}\right)$$ ampliación de la matriz calculada $\Sigma_Z = \begin{pmatrix} 73 & 69 & 52 \\ 69 & 83 & 41 \\ 52 & 41 & 48 \end{pmatrix}$

c) ¿Es (Z_1, Z_2, Z_5) una normal tridimensional?

$$Z^0 = \begin{cases} Z_1 = X_1 - 2X_2 + 3X_3 \\ Z_2 = 3X_1 - X_2 + 2X_3 \\ Z_5 = X_1 + 3X_2 - 4X_3 \end{cases} \qquad \begin{pmatrix} Z_1 \\ Z_2 \\ Z_5 \end{pmatrix} = \begin{pmatrix} 1 & -2 & 3 \\ 3 & -1 & 2 \\ 1 & 3 & -4 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \quad Z^0 = B^0.X$$

$$\left|B^0\right| = \begin{vmatrix} 1 & -2 & 3 \\ 3 & -1 & 2 \\ 1 & 3 & -4 \end{vmatrix} = 0 \rightarrow \text{rango}\,(B^0) = 2 \quad Z_5 = Z_2 - 2Z_1$$

En consecuencia, Z^0 no es una normal tridimensional.

Esperanza y matriz de varianzas-covarianzas de Z^0 :

$$\mu_{z^0} = B^0.\mu = \begin{pmatrix} 1 & -2 & 3 \\ 3 & -1 & 2 \\ 1 & 3 & -4 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 9 \\ \underline{11} \\ -7 \end{pmatrix} \qquad E(Z_5) = E(Z_2) - 2E(Z_1) = 11 - 2.9 = -7$$

$$\text{Var}(Z_5) = \text{Var}(Z_2 - 2Z_1) = \text{Var}(Z_2) + 4\,\text{Var}(Z_1) - 4\,\text{Cov}(Z_1, Z_2) = 83 + 4.73 - 4.69 = 99$$

$$\text{Cov}(Z_1, Z_5) = \text{Cov}(Z_1, Z_2 - 2Z_1) = \text{Cov}(Z_1, Z_2) - 2\,\text{Cov}(Z_1, Z_1) = 69 - 2.73 = -77$$

$$\text{Cov}(Z_2, Z_5) = \text{Cov}(Z_2, Z_2 - 2Z_1) = \text{Cov}(Z_2, Z_2) - 2\,\text{Cov}(Z_2, Z_1) = 83 - 2.69 = -55$$

$$\text{Var}(Z^0) = \left(\begin{array}{cc|c} 73 & 69 & -77 \\ \underline{69} & \underline{83} & \underline{-55} \\ -77 & -55 & 99 \end{array}\right)$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Las calificaciones del primer parcial de Instrumentos Estadísticos se distribuyen como una normal de media 4 y desviación típica 1, y las del segundo parcial como una normal de media 6 y desviación típica 2. Suponiendo que las notas de ambos parciales siguen una distribución normal bidimensional, siendo 0,8 el coeficiente de correlación.

Un alumno aprueba la asignatura si la media aritmética de ambas calificaciones es al menos 5. Hallar la probabilidad de que en un grupo de 8 alumnos sólo haya 3 que aprueban la asignatura.

Sea X ≡ "Calificación del primer parcial", $X \approx N(4,1)$ e Y ≡ "Calificación del segundo parcial", $Y \approx N(6,2)$ con $\rho = 0,8$

$$\sigma_{xy} = \rho \times \sigma_x \times \sigma_y = 0,8 \times 1 \times 2 = 1,6$$

La distribución conjunta de X e Y es una normal bidimensional

$$\begin{pmatrix} X \\ Y \end{pmatrix} \approx N_2\left[\begin{pmatrix} 4 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 1,6 \\ 1,6 & 4 \end{pmatrix}\right]$$

$(X + Y)$ es una normal unidimensional:

$$X + Y = \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} \approx N\left[\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 4 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1,6 \\ 1,6 & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] = N(10, 8,2) \qquad \sigma = \sqrt{8,2}$$

Probabilidad de que un alumno apruebe:

$$P\left(\frac{X+Y}{2} \geq 5\right) = P(X + Y \geq 10) = P\left[\frac{X + Y - 10}{\sqrt{8,2}} \geq \frac{10 - 10}{\sqrt{8,2}}\right] = P(z \geq 0) = 0,5$$

Siendo U = "Número de alumnos que aprueban entre ocho", $U \approx B(8, 0,5)$

$$P(U = 3) = \binom{8}{3} 0,5^3\, 0,5^5 = 0,21875$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Cierta enfermedad produce muertes anualmente, las mujeres siguen una ley $N(\mu, 4)$, mientras que los hombres siguen una $N(100, 5)$, el coeficiente de correlación entre ambas variables es de 0,8. Se conoce que la probabilidad de que el número total de enfermos que mueren al año sea superior a 300 es de 0,5. Suponiendo que la distribución conjunta es una normal bidimensional, considerando un periodo de 10 años, calcular la probabilidad de que 8 años el número de mujeres que mueren sea superior a 195.

Sea $X \equiv$ "Número de mujeres que mueren al año", $X \approx N(\mu, 4)$ e $Y \equiv$ "Número de hombres que mueren al año", $Y \approx N(100, 5)$ con $\rho = 0{,}8$

$$\sigma_{xy} = \rho \times \sigma_x \times \sigma_y = 0{,}8 \times 4 \times 5 = 16$$

La distribución conjunta de X e Y es una normal bidimensional

$$\begin{pmatrix} X \\ Y \end{pmatrix} \approx N_2\left[\begin{pmatrix} \mu \\ 100 \end{pmatrix}, \begin{pmatrix} 16 & 16 \\ 16 & 25 \end{pmatrix}\right]$$

$(X + Y)$ es una normal unidimensional:

$$X + Y = \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} \approx N\left[\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} \mu \\ 100 \end{pmatrix}, \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 16 & 16 \\ 16 & 25 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] = N(\mu + 100, 73) \quad \sigma = \sqrt{73}$$

Probabilidad de que el total de enfermos muera al año:

$$P(X + Y \geq 300) = P\left[\frac{X + Y - (\mu + 100)}{\sqrt{73}} \geq \frac{300 - (\mu + 100)}{\sqrt{73}}\right] = P\left[z \geq \frac{200 - \mu}{\sqrt{73}}\right] = 0{,}5$$

$$\frac{200 - \mu}{\sqrt{73}} = 0 \rightarrow \mu = 200 \rightarrow X \approx N(200, 4)$$

$$P(X > 195) = P\left[\frac{X - 200}{4} > \frac{195 - 200}{4}\right] = P(z < 1{,}25) = 1 - 0{,}1056 = 0{,}8944$$

Siendo U = "8 mujeres mueran entre 10", $U \approx B(10, 0{,}8944)$

$$P(U = 8) = \binom{10}{8} 0{,}8944^8\ 0{,}1056^2 = 0{,}2055$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

15 Sea (X, Y) una variable aleatoria bidimensional, donde X sigue una distribución normal $N(0, 1)$ y la distribución $(Y|X = x)$ sigue una normal $N\left(2X - 3, \sqrt{12}\right)$

a) Demostrar que la distribución (X, Y) es una normal bidimensional.

b) Obtener la distribución marginal de la Y y el coeficiente de correlación entre X e Y.

Sea X ≡ "Número de mujeres que mueren al año", $X \approx N(\mu, 4)$ e Y ≡ "Número de hombres que mueren al año", $Y \approx N(100, 5)$ con $\rho = 0{,}8$

$$f_{(X,Y)}(x,y) = f_X(x) \cdot f_{(Y|X=x)}(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}\cdot\sqrt{12}} e^{-\frac{1}{2.12}[y-(2x-3)]^2} =$$

$$= \frac{1}{2\pi\sqrt{12}} e^{-\frac{1}{2}\left(\frac{16x^2+y^2-4xy-12x+6y+9}{12}\right)}$$

- Señalar que dos variables aleatorias continuas X e Y se distribuyen según una distribución normal bivariante $(X,Y) \approx N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{bmatrix}\right)$ con función de densidad $f(x,y) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}Q(x,y)}$

donde $Q(x,y) = (x-\mu_1 \quad y-\mu_2)\,\Sigma^{-1}\begin{pmatrix}x-\mu_1\\ y-\mu_2\end{pmatrix}$

$$E(Y) = E\left[E(Y|X=x)\right] = E(2X-3) = 2E(X) - 3 = -3$$

$$Var(Y) = \underbrace{E\left[Var(Y|X)\right]}_{\text{Esperanza de las varianzas}} + \underbrace{Var\left[E(Y|X)\right]}_{\text{Varianza de las esperanzas}} = 12 + Var\left[2X-3\right] = 12 + 4\,Var(X) = 12 + 4\times 1 = 16$$

con lo cual, $(X,Y) \approx N\left(\begin{bmatrix}0\\ -3\end{bmatrix}, \begin{bmatrix}1 & \sigma_{12}\\ \sigma_{21} & 16\end{bmatrix}\right)$ donde $\Sigma = \begin{bmatrix}1 & \sigma_{12}\\ \sigma_{21} & 16\end{bmatrix}$ $\quad |\Sigma| = 12$

$$\begin{vmatrix}1 & \sigma_{12}\\ \sigma_{21} & 16\end{vmatrix} = 12 \rightarrow 16 - \sigma_{12}^2 = 12 \rightarrow \sigma_{12} = 2$$

La distribución de $(X,Y) \approx N_2\left(\begin{bmatrix}0\\ -3\end{bmatrix}, \begin{bmatrix}1 & 2\\ 2 & 16\end{bmatrix}\right)$

Se comprueba que $Q(x,y) = (x \quad y+3)\begin{bmatrix}1 & 2\\ 2 & 16\end{bmatrix}^{-1}\begin{pmatrix}x\\ y+3\end{pmatrix} = \frac{16x^2+y^2-4xy-12x+6y+9}{12}$

b) Coeficiente de correlación: $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} = \frac{2}{1.4} = 0{,}5$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

La función de densidad de una variable aleatoria (X, Y) se expresa:

$$f_{(X,Y)}(x,y) = \begin{cases} k\, f_{(U,V)}(x,y) & -0{,}5 \le x \le 0{,}2 \quad 0{,}5 \le y \le 1 \\ 0 & \text{en el resto} \end{cases}$$

donde $f_{(U,V)}$ es la función de densidad de una variable normal bidimensional de media 0 y matriz de varianzas-covarianzas I.

a) Calcular la función de densidad de (X, Y)

b) Hallar las funciones de densidad marginales de X y de Y

c) Calcular la esperanza de $\left(V \mid U + V = 2\right)$

a) Para hallar k se considera: $\int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f_{(X,Y)}(x,y)\, dx\, dy = 1$, resultando:

$$f(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} = f_1(x) \cdot f_2(y)$$

$$1 = \int_{x=-0{,}5}^{0{,}2} \int_{y=0{,}5}^{1} k \cdot \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dx\, dy = k \cdot \int_{x=-0{,}5}^{0{,}2} \frac{1}{2\pi} e^{-\frac{1}{2}x^2} dx \cdot \int_{y=0{,}5}^{1} \frac{1}{2\pi} e^{-\frac{1}{2}y^2} dy =$$

$$= k \cdot P(-0{,}5 \le z \le 0{,}2) \cdot P(0{,}5 \le z \le 1) = k \cdot (1 - 0{,}3085 - 0{,}4207) \cdot (0{,}3085 - 0{,}1587) = 0{,}04\, k$$

$$k = \frac{1}{0{,}04} = 25$$

con lo que, $$f_{(X,Y)}(x,y) = \begin{cases} \frac{25}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} & -0{,}5 \le x \le 0{,}2 \quad 0{,}5 \le y \le 1 \\ 0 & \text{en el resto} \end{cases}$$

b) Funciones de densidad marginales:

$$f_X(x) = \int_{y=0{,}5}^{1} \frac{25}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} dy = \frac{25}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \cdot \int_{y=0{,}5}^{1} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}y^2} dy =$$

$$= \frac{25}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \cdot P(0{,}5 \le z \le 1) = \frac{25}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \cdot (0{,}3085 - 0{,}1587) = \frac{3{,}745}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

Marginal de X: $$f_X(x) = \begin{cases} \frac{3{,}745}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} & -0{,}5 \le x \le 0{,}2 \\ 0 & \text{en el resto} \end{cases}$$

Análogamente,

$$f_Y(y) = \int_{x=-0{,}5}^{0{,}2} \frac{25}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} dx = \frac{25}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}y^2} \cdot \int_{x=-0{,}5}^{0{,}2} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} dx =$$

$$= \frac{25}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}y^2} \cdot P(0{,}5 \le z \le 1) = \frac{25}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}y^2} \cdot (1 - 0{,}3085 - 0{,}4207) = \frac{6{,}77}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}y^2}$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

Marginal de Y: $f_Y(y) = \begin{cases} \frac{6,77}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} & 0,5 \le y \le 1 \\ 0 & \text{en el resto} \end{cases}$

b) Para calcular $E(V \mid U+V=2)$ se necesita hallar la distribución conjunta de V y de $U+V$

$$\begin{pmatrix} U \\ U+V \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} U \\ V \end{pmatrix} \approx N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right] \text{ dado que } (U,V) \approx N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$$

$E(V \mid U+V=2)$ coincide con la recta de regresión: $E[V \mid U+V=x] = \mu_V + \frac{\sigma_{V(U+V)}}{\sigma^2_{(U+V)}}(x - \mu_{(U+V)})$

$$E(V \mid U+V=2) = 0 + \frac{1}{2}(2-0) = 1$$

En determinado país, la renta y el consumo anual (en cientos de euros) de una familia, (X, Y), siguen una distribución normal bidimensional de medias $\mu_X = 100$ y $\mu_Y = 90$, y de varianzas $\sigma^2_X = 900$ y $\sigma^2_Y = 625$. Se sabe que sólo el 10% de familias se endeuda anualmente.

a) ¿Qué porcentaje de familias ahorra el 20% de su renta anual?
b) ¿Qué porcentaje de familias tiene un consumo anual superior a 15. 000 euros?
c) Hallar el consumo esperado de una familia que tiene una renta anual de 56.000 euros

a) Denotando la renta (X) y el consumo (Y), se tiene:

La diferencia entre la renta (X) y el consumo (Y) es mayor o igual que el 20% de X: $P(X - Y \ge 0,2X)$

El 10% de las familias se adeuda anualmente: $P(X \le Y) = 0,10$

Según los datos: $(X,Y) \approx N_2\left[\begin{pmatrix} 100 \\ 90 \end{pmatrix}, \begin{pmatrix} 900 & \sigma_{XY} \\ \sigma_{XY} & 625 \end{pmatrix}\right]$

- $P(X - Y \ge 0,2X) = P(0,8X - Y \ge 0) = P(T \ge 0)$

Se necesita conocer la distribución de la variable $T = 0,8X - Y$

La variable T en forma matricial: $T = 0,8X - Y = \begin{pmatrix} 0,8 & -1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}$, sigue una distribución normal unidimensional $T \approx N\left(-10, \sqrt{1201 - 1,6\,\sigma_{XY}}\right)$

$$E(T) = \begin{pmatrix} 0,8 & -1 \end{pmatrix}\begin{pmatrix} 100 \\ 90 \end{pmatrix} = 80 - 90 = -10$$

$$V(T) = \begin{pmatrix} 0,8 & -1 \end{pmatrix}\begin{pmatrix} 900 & \sigma_{XY} \\ \sigma_{XY} & 625 \end{pmatrix}\begin{pmatrix} 0,8 \\ -1 \end{pmatrix} = \begin{pmatrix} 720 - \sigma_{XY} & 0,8\,\sigma_{XY} - 625 \end{pmatrix}\begin{pmatrix} 0,8 \\ -1 \end{pmatrix} = 1201 - 1,6\sigma_{XY}$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

- Cálculo de la covarianza σ_{XY}: $P(X \le Y) = 0{,}10 \rightarrow P(X - Y \le 0) = 0{,}10 \equiv P(W \le 0) = 0{,}10$

Denotando por $W = X - Y$, en forma matricial: $W = \begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}$

Se necesita conocer la distribución de W:

$$E(W) = \begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} 100 \\ 90 \end{pmatrix} = 10$$

$$V(W) = \begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} 900 & \sigma_{XY} \\ \sigma_{XY} & 625 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 900 - \sigma_{XY} & \sigma_{XY} - 625 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1525 - 2\sigma_{XY}$$

$$W \approx N\left(10, \sqrt{1525 - 2\sigma_{xy}}\right)$$

$$P(W \le 0) = P\left[\frac{W - 10}{\sqrt{1525 - 2\sigma_{xy}}} \le \frac{0 - 10}{\sqrt{1525 - 2\sigma_{xy}}}\right] = P\left[z \le \frac{-10}{\sqrt{1525 - 2\sigma_{xy}}}\right] = 0{,}10$$

$$P\left[z \ge \frac{10}{\sqrt{1525 - 2\sigma_{xy}}}\right] = 0{,}10 \rightarrow \frac{10}{\sqrt{1525 - 2\sigma_{xy}}} = 1{,}28 \rightarrow \sigma_{xy} = 731{,}98$$

- La distribución $T \approx N\left(-10, \sqrt{1201 - 1{,}6\,\sigma_{xy}}\right) = N\left(-10, \sqrt{1201 - 1{,}6 \cdot 731{,}98}\right) = N(-10, 5{,}46)$

$$P(X - Y \ge 0{,}2X) = P(0{,}8X - Y \ge 0) = P(T \ge 0) = P\left[\frac{T + 10}{5{,}46} \ge \frac{0 - 10}{5{,}46}\right] = P\left[z \le \frac{10}{5{,}46}\right] =$$

$$= P(z \le 1{,}83) = 1 - 0{,}0336 = 0{,}9664$$

En consecuencia, el 96,64% de las familias ahorra más del 20% de sus ingresos.

b) Las distribuciones marginales de una distribución p-dimensional son normales unidimensionales. En este sentido, el consumo anual (Y) sigue una distribución $N\left(90, \sqrt{625}\right) = N(90, 25)$

$$P(Y > 150) = P\left[\frac{Y - 90}{25} > \frac{150 - 90}{25}\right] = P(z > 2{,}4) = 0{,}0082$$

El 0,82% de las familias tiene una renta anual superior a 150.000 euros.

c) $(X, Y) \approx N_2\left[\begin{pmatrix} 100 \\ 90 \end{pmatrix}, \begin{pmatrix} 900 & 731{,}98 \\ 731{,}98 & 625 \end{pmatrix}\right]$

La curva general de la regresión de Y sobre X coincide con la recta:

$$E\left[Y \mid X = x\right] = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X)$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

$$E\left[Y \middle| X=560\right] = 90 + \frac{731{,}98}{900}(560-100) = 464{,}12 \text{ (46.412 euros)}$$

La predicción efectuada es muy fiable al tener un coeficiente de correlación elevado:

$$\rho = \frac{\sigma_{XY}}{\sigma_X . \sigma_Y} = \frac{731{,}98}{\sqrt{900}.\sqrt{625}} = 0{,}976 \text{ (97,6\%)}$$

18 Sea una variable aleatoria $(X, Y) \approx N_2\left[\begin{pmatrix}-1\\2\end{pmatrix}, \begin{pmatrix}1 & -1\\-1 & 4\end{pmatrix}\right]$

a) Hallar el coeficiente de correlación entre X y $X+Y$ y $P(X+Y \le 5 \,|\, X > 4)$

b) Calcular las rectas de regresión de U y V, siendo $U = 2X + Y$ y $V = X - Y$

a) $\mu = \begin{pmatrix}-1\\2\end{pmatrix}$, $\Sigma = \begin{pmatrix}1 & -1\\-1 & 4\end{pmatrix}$

$$\begin{pmatrix}X\\X+Y\end{pmatrix} = \begin{pmatrix}1 & 0\\1 & 1\end{pmatrix}\begin{pmatrix}X\\Y\end{pmatrix}$$

$$\mu_{X+Y} = \begin{pmatrix}1 & 0\\1 & 1\end{pmatrix}\begin{pmatrix}-1\\2\end{pmatrix} = \begin{pmatrix}-1\\1\end{pmatrix}$$

$$\Sigma_{X+Y} = \begin{pmatrix}1 & 0\\1 & 1\end{pmatrix}\begin{pmatrix}1 & -1\\-1 & 4\end{pmatrix}\begin{pmatrix}1 & 0\\1 & 1\end{pmatrix}^t = \begin{pmatrix}1 & 0\\1 & 1\end{pmatrix}\begin{pmatrix}1 & -1\\-1 & 4\end{pmatrix}\begin{pmatrix}1 & 1\\0 & 1\end{pmatrix} = \begin{pmatrix}1 & 0\\0 & 3\end{pmatrix}$$

$$\begin{pmatrix}X\\X+Y\end{pmatrix} \approx N_2\left[\begin{pmatrix}-1\\1\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 3\end{pmatrix}\right] \rightarrow \sigma_{X,X+Y} = 0 \rightarrow X \text{ y } X+Y \text{ son independientes}$$

$$P(X+Y \le 5 \,|\, X > 4) = P(X+Y \le 5) = P\left[\frac{X+Y-1}{\sqrt{3}} \le \frac{5-1}{\sqrt{3}}\right] = P(z \le 2{,}31) = 1 - 0{,}0104) = 0{,}9896$$

b) En forma de matriz: $\begin{pmatrix}U\\V\end{pmatrix} = \begin{pmatrix}2 & 1\\1 & -1\end{pmatrix}\begin{pmatrix}X\\Y\end{pmatrix}$

$$\mu = \begin{pmatrix}2 & 1\\1 & -1\end{pmatrix}\begin{pmatrix}-1\\2\end{pmatrix} = \begin{pmatrix}0\\-3\end{pmatrix}$$

$$\Sigma = \begin{pmatrix}2 & 1\\1 & -1\end{pmatrix}\begin{pmatrix}1 & -1\\-1 & 4\end{pmatrix}\begin{pmatrix}2 & 1\\1 & -1\end{pmatrix}^t = \begin{pmatrix}2 & 1\\1 & -1\end{pmatrix}\begin{pmatrix}1 & -1\\-1 & 4\end{pmatrix}\begin{pmatrix}2 & 1\\1 & -1\end{pmatrix} = \begin{pmatrix}4 & -1\\-1 & 7\end{pmatrix}$$

$$\begin{pmatrix}U\\V\end{pmatrix} \approx N_2\left[\begin{pmatrix}0\\-3\end{pmatrix}, \begin{pmatrix}4 & -1\\-1 & 7\end{pmatrix}\right]$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Regresión de U sobre V: $\tilde{U} = \mu_U + \frac{\sigma_{UV}}{\sigma_V^2}(V - \mu_V) = 0 - \frac{1}{7}(V + 3) = -\frac{1}{7}V - \frac{3}{7}$

Regresión de V sobre U: $\tilde{V} = \mu_V + \frac{\sigma_{UV}}{\sigma_U^2}(U - \mu_U) = -3 - \frac{1}{4}(U - 0) = -3 - \frac{1}{4}U$

Se observa que apenas existe relación lineal: $\rho_{UV} = \frac{\sigma_{UV}}{\sigma_U \cdot \sigma_V} = \frac{-1}{2 \cdot \sqrt{7}} = -0{,}1889$

Se pretenden estimar los gastos en alimentación de una familia en base a la información que proporcionan las variables regresoras "ingresos mensuales" y "número de miembros de la familia". Para ello se recoge una muestra aleatoria simple de 15 familias, cuyos resultados se facilitan en la tabla adjunta. (El gasto e ingreso se expresan en cien mil euros).

Gasto Alimentación	Ingresos	Tamaño
0,43	2,10	3
0,31	1,10	4
0,32	0,90	5
0,46	1,60	4
1,25	6,20	4
0,44	2,30	3
0,52	1,80	6
0,29	1,00	5
1,29	8,90	3
0,35	2,40	2
0,35	1,20	4
0,78	4,70	3
0,43	3,50	2
0,47	2,90	3
0,38	1,40	4

En forma matricial: $Y = X\beta + U$, $\hat{\beta} = \left[X'X\right]^{-1}X'Y$, donde X' matriz transpuesta

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

$$Y = \begin{pmatrix} 0{,}43 \\ 0{,}31 \\ 0{,}32 \\ 0{,}46 \\ 1{,}25 \\ 0{,}44 \\ 0{,}52 \\ 0{,}29 \\ 1{,}29 \\ 0{,}35 \\ 0{,}35 \\ 0{,}78 \\ 0{,}43 \\ 0{,}47 \\ 0{,}38 \end{pmatrix} = X\beta + U = \begin{pmatrix} 1 & 2{,}1 & 3 \\ 1 & 1{,}1 & 4 \\ 1 & 0{,}9 & 5 \\ 1 & 1{,}6 & 4 \\ 1 & 6{,}2 & 4 \\ 1 & 2{,}3 & 3 \\ 1 & 1{,}8 & 6 \\ 1 & 1 & 5 \\ 1 & 8{,}9 & 3 \\ 1 & 2{,}4 & 2 \\ 1 & 1{,}2 & 4 \\ 1 & 4{,}7 & 3 \\ 1 & 3{,}5 & 2 \\ 1 & 2{,}9 & 3 \\ 1 & 1{,}4 & 4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$$

Aplicando el criterio de los mínimos cuadrados ordinarios MCO, la función que mejor se ajusta a los datos es la que minimiza la varianza del error U, lo que conlleva a un sistema de ecuaciones normales:

$$\text{ecuaciones normales MCO} \begin{cases} \sum_{i=1}^{15} Y_i = N\beta_0 + \beta_1 \sum_{i=1}^{15} X_{1i} + \beta_2 \sum_{i=1}^{15} X_{2i} \\ \sum_{i=1}^{15} X_{1i} Y_i = \beta_0 \sum_{i=1}^{15} X_{1i} + \beta_1 \sum_{i=1}^{15} X_{1i}^2 + \beta_2 \sum_{i=1}^{15} X_{1i} X_{2i} \\ \sum_{i=1}^{15} X_{2i} Y_i = \beta_0 \sum_{i=1}^{15} X_{2i} + \beta_1 \sum_{i=1}^{15} X_{1i} X_{2i} + \beta_2 \sum_{i=1}^{15} X_{2i}^2 \end{cases}$$

Con estos datos, se obtiene:

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Y_i	X_{1i}	X_{2i}	X_{1i}^2	X_{2i}^2	$X_{1i}X_{2i}$	$X_{1i}Y_i$	$X_{2i}Y_i$
0,43	2,1	3	4,41	9	6,3	0,903	1,29
0,31	1,1	4	1,21	16	4,4	0,341	1,24
0,32	0,9	5	0,81	25	4,5	0,288	1,6
0,46	1,6	4	2,56	16	6,4	0,736	1,84
1,25	6,2	4	38,44	16	24,8	7,750	5
0,44	2,3	3	5,29	9	6,9	1,012	1,32
0,52	1,8	6	3,24	36	10,8	0,936	3,12
0,29	1	5	1	25	5	0,29	1,45
1,29	8,9	3	79,21	9	26,7	11,481	3,87
0,35	2,4	2	5,76	4	4,8	0,84	0,7
0,35	1,2	4	1,44	16	4,8	0,42	1,4
0,78	4,7	3	22,09	9	14,1	3,666	2,34
0,43	3,5	2	12,25	4	7	1,505	0,86
0,47	2,9	3	8,41	9	8,7	1,363	1,41
0,38	1,4	4	1,96	16	5,6	0,532	1,52
8,07	42	55	188,08	219	140,8	32,063	28,96

$$\begin{cases} \sum_{i=1}^{15} Y_i = N\beta_0 + \beta_1 \sum_{i=1}^{15} X_{1i} + \beta_2 \sum_{i=1}^{15} X_{2i} \\ \sum_{i=1}^{15} X_{1i}Y_i = \beta_0 \sum_{i=1}^{15} X_{1i} + \beta_1 \sum_{i=1}^{15} X_{1i}^2 + \beta_2 \sum_{i=1}^{15} X_{1i}X_{2i} \\ \sum_{i=1}^{15} X_{2i}Y_i = \beta_0 \sum_{i=1}^{15} X_{2i} + \beta_1 \sum_{i=1}^{15} X_{1i}X_{2i} + \beta_2 \sum_{i=1}^{15} X_{2i}^2 \end{cases} \rightarrow \begin{cases} 15\beta_0 + 42\beta_1 + 55\beta_2 = 8,07 \\ 42\beta_0 + 188,08\beta_1 + 140,08\beta_2 = 32,063 \\ 55\beta_0 + 140,08\beta_1 + 219\beta_2 = 28,96 \end{cases}$$

en forma matricial,

$$\overbrace{\begin{pmatrix} 15 & 42 & 55 \\ 42 & 188,08 & 140,8 \\ 55 & 140,8 & 219 \end{pmatrix}}^{X'X} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 8,07 \\ 32,063 \\ 28,96 \end{pmatrix} \Rightarrow \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \overbrace{\begin{pmatrix} 1,36 & -0,092 & -0,282 \\ -0,092 & 0,016 & 0,013 \\ -0,282 & 0,013 & 0,067 \end{pmatrix}}^{[X'X]^{-1}} \overbrace{\begin{pmatrix} 8,07 \\ 32,063 \\ 28,96 \end{pmatrix}}^{X'Y}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} -0,16 \\ 0,149 \\ 0,077 \end{pmatrix} \rightarrow Y = -0,16 + 0,149\ X_1 + 0,077\ X_2 + \text{Residuo} \quad \text{(Modelo regresión lineal)}$$

A partir de la ecuación $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ se obtienen las predicciones y residuos asociados $u_i = Y_i - \hat{Y}_i$ a las observaciones muestrales.

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

De este modo, para la primera observación ($Y_1 = 0,43$; $X_{11} = 2,1$; $X_{21} = 3$), se tiene:

$$\begin{cases} Y = -0,16 + 0,149\ (2,1) + 0,077\ (3) = 0,3839 \\ u_1 = Y_1 - \hat{Y}_1 = 0,43 - 0,3839 = 0,0461 \end{cases}$$

En esta línea, considerando todos los puntos muestrales, se obtiene:

Predicciones: $\hat{Y}_i$	Residuos: $u_i = Y_i - \hat{Y}_i$	$u_i^2 = (Y_i - \hat{Y}_i)^2$
0,3839	0,046	0,0021
0,3119	-0,002	0,0000
0,3591	-0,039	0,0015
0,3864	0,074	0,0054
1,0718	0,178	0,0318
0,4137	0,026	0,0007
0,5702	-0,050	0,0025
0,374	-0,084	0,0071
1,3971	-0,107	0,0115
0,3516	-0,002	0,0000
0,3268	0,023	0,0005
0,7713	0,009	0,0001
0,5155	-0,086	0,0073
0,5031	-0,033	0,0011
0,3566	0,023	0,0005

$$\sum_{i=1}^{15} (Y_1 - \hat{Y}_1)^2 = 0,0721$$

De donde, la suma de cuadrados RESIDUAL, es decir, la variabilidad de Y respecto a la recta ajustada será:

$$SCR = \sum_{i=1}^{15} (Y_1 - \hat{Y}_1)^2 = 0,0721$$

$$S_R^2 = \frac{SCR}{15-2-1} = \frac{0,0721}{12} = 0,006$$

$$S_R = \sqrt{0,006} = 0,0775$$

INTERVALOS DE CONFIANZA PARAMÉTROS DEL MODELO $(1-\alpha) = 0,90$

- Intervalo de confianza para la varianza

$$n-k-1 = 15-2-1 = 12 \quad S_R^2 = 0,006 \quad S_R^2 = 0,0721 \quad \chi^2_{0,05,12} = 21,026 \quad \chi^2_{0,95,12} = 5,226$$

$$IC(\sigma^2) = \left(\frac{(n-k-1)\,S_R^2}{\chi^2_{\frac{\alpha}{2},(n-k-1)}} \; ; \; \frac{(n-k-1)\,S_R^2}{\chi^2_{1-\frac{\alpha}{2},(n-k-1)}} \right) = \left(\frac{SCR}{\chi^2_{\frac{\alpha}{2},(n-k-1)}} \; ; \; \frac{SCR}{\chi^2_{1-\frac{\alpha}{2},(n-k-1)}} \right) = \left[\frac{0,0721}{21,026} \; ; \; \frac{0,0721}{5,226} \right] =$$

$$= \left[0,0034 \; ; \; 0,0138 \right] \rightarrow \quad 0,0034 \leq \sigma^2 \leq 0,0138$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

- Varianza de los estimadores del modelo $\hat{\beta} \in N\left(\beta, \sigma^2 \left[X'X\right]^{-1}\right)$:

$$\left[X'X\right]^{-1} = \begin{pmatrix} 1{,}36 & -0{,}092 & -0{,}282 \\ -0{,}092 & 0{,}016 & 0{,}013 \\ -0{,}282 & 0{,}013 & 0{,}067 \end{pmatrix}$$

$$\mathrm{Var}(\hat{\beta}_i) = \sigma^2 \left[X'X\right]^{-1} \approx S_R^2 \left[X'X\right]^{-1} = (0{,}006) \overbrace{\begin{pmatrix} 1{,}36 & & \\ & 0{,}016 & \\ & & 0{,}067 \end{pmatrix}}^{q_{i+1,i+1} \equiv \text{elemento de} \left[X'X\right]^{-1}} = \overbrace{\begin{pmatrix} 0{,}00816 & & \\ & 0{,}000096 & \\ & & 0{,}0004 \end{pmatrix}}^{S_R^2 q_{i+1,i+1}}$$

de donde se deduce, $\begin{cases} \mathrm{Var}(\beta_0) = 0{,}00816 \Rightarrow \sigma_{\beta_0} = \sqrt{0{,}00816} = 0{,}0903 \\ \mathrm{Var}(\beta_1) = 0{,}000096 \Rightarrow \sigma_{\beta_1} = \sqrt{0{,}000096} = 0{,}0098 \\ \mathrm{Var}(\beta_2) = 0{,}0004 \Rightarrow \sigma_{\beta_1} = \sqrt{0{,}0004} = 0{,}02 \end{cases}$

- Intervalo de confianza para los parámetros: $IC_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i \pm t_{\alpha/2,(n-k-1)} S_R \sqrt{q_{i+1,i+1}}\right]$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} -0{,}16 \\ 0{,}149 \\ 0{,}077 \end{pmatrix}, \quad \sqrt{S_R^2 q_{i+1,i+1}} = \begin{pmatrix} \sqrt{0{,}00816} & & \\ & \sqrt{0{,}000096} & \\ & & \sqrt{0{,}0004} \end{pmatrix}, \quad t_{\alpha/2,n-k-1} = t_{0{,}05,12} = 1{,}782$$

$$IC_{1-\alpha}(\beta_0) = \left[-0{,}160 \pm (1{,}782)\sqrt{0{,}00816}\right] = \left[-0{,}321\,;\, 0{,}001\right]$$

$$IC_{1-\alpha}(\beta_1) = \left[0{,}149 \pm (1{,}782)\sqrt{0{,}000096}\right] = \left[0{,}1315\,;\, 0{,}1665\right] \text{ (Ingreso)}$$

$$IC_{1-\alpha}(\beta_2) = \left[0{,}077 \pm (1{,}782)\sqrt{0{,}0004}\right] = \left[0{,}0414\,;\, 0{,}1126\right] \text{ (Tamaño)}$$

- Contraste de Hipótesis individual para X_2 (tamaño familiar)

Se plantea si la variable X_2 (tamaño familiar) influye sobre la variable de respuesta Y (gastos). En otras palabras, si el valor del parámetro en la población es cero o no.

Para ello, se establece la hipótesis nula $H_0: \beta_2 = 0$ frente a la hipótesis alternativa $H_1: \beta_2 \neq 0$.

El estadístico observado $t = \dfrac{\hat{\beta}_2 - \beta_2}{S_R \sqrt{q_{33}}}$, bajo la hipótesis nula, resulta: $t = \dfrac{\hat{\beta}_2}{S_R \sqrt{q_{33}}}$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

donde, $\hat{\beta}_2 = 0,077$ $S_R q_{33} = (0,0775)\sqrt{0,0004} = 0,00155$ $t_{0,05,12} = 1,782$

El estadístico experimental $t = \dfrac{\hat{\beta}_2}{S_R\sqrt{q_{33}}} = \dfrac{0,077}{0,00155} = 49,67$

Siendo $|t| > t_{0,05,12}$ se rechaza la hipótesis nula, afirmando, con un 90% de fiabilidad, que el número de miembros de la familia influye en los gastos de alimentación.

Se observa que en el Intervalo de Confianza para β_2: $IC_{1-\alpha}(\beta_2) = [0,0414 ; 0,1126]$ no se encuentra el cero, con lo que se rechaza la hipótesis nula $H_0: \beta_2 = 0$, concluyendo que el número de miembros de la familia (tamaño) si influye en los gastos de alimentación (Y).

Partiendo de la Información:

y_j	x_{1j}	x_{2j}
1	2	17
3	1	15
4	3	13
6	4	10
7	5	9
10	8	8
15	7	6
16	9	5
18	11	3
20	12	4

a) Estimar el modelo de regresión
b) Obtener una medida de fiabilidad del ajuste lineal
c) ¿Qué parte de la variabilidad de Y queda explicada a través del plano de regresión?
d) Calcular los coeficientes de correlación lineal simple

a) En forma matricial: $Y = X\beta + U$, $\hat{\beta} = [X'X]^{-1}X'Y$, donde X' matriz transpuesta

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$Y = \begin{pmatrix} 1 \\ 3 \\ 4 \\ 6 \\ 7 \\ 10 \\ 15 \\ 16 \\ 18 \\ 20 \end{pmatrix} = X\beta + U = \begin{pmatrix} 1 & 2 & 17 \\ 1 & 1 & 15 \\ 1 & 3 & 13 \\ 1 & 4 & 10 \\ 1 & 5 & 9 \\ 1 & 8 & 8 \\ 1 & 7 & 6 \\ 1 & 9 & 5 \\ 1 & 11 & 3 \\ 1 & 12 & 4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$$

Los coeficientes β, estimados bajo la condición mínimo cuadrática, vienen dados por: $\hat{\beta} = [X'X]^{-1} X'Y$

$$[X'X] = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 3 & 4 & 5 & 8 & 7 & 9 & 11 & 12 \\ 17 & 15 & 13 & 10 & 9 & 8 & 6 & 5 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 17 \\ 1 & 1 & 15 \\ 1 & 3 & 13 \\ 1 & 4 & 10 \\ 1 & 5 & 9 \\ 1 & 8 & 8 \\ 1 & 7 & 6 \\ 1 & 9 & 5 \\ 1 & 11 & 3 \\ 1 & 12 & 4 \end{bmatrix} = \begin{bmatrix} 10 & 62 & 90 \\ 62 & 514 & 405 \\ 90 & 405 & 1014 \end{bmatrix} \Rightarrow |X'X| = 30294 \neq 0$$

Adviértase que cuando la matriz es singular, es decir, cuando $|X'X| = 0$, no existe matriz inversa y, en consecuencia, las estimaciones de los coeficientes β quedan indeterminadas.

Esto ocurre porque existe multicolinealidad entre X_1 y X_2, esto es, que existe una relación lineal entre estas variables. Para ello se calcula la regresión de X_1 sobre X_2 y su correspondiente coeficiente de correlación lineal será 1.

$$[X'X] = \begin{bmatrix} 10 & 62 & 90 \\ 62 & 514 & 405 \\ 90 & 405 & 1014 \end{bmatrix} \Rightarrow \text{Adj}[X'X] = \text{Adj} \begin{bmatrix} 10 & 62 & 90 \\ 62 & 514 & 405 \\ 90 & 405 & 1014 \end{bmatrix} = \begin{bmatrix} 357171 & -26418 & -21150 \\ -26418 & 2040 & 1530 \\ -21150 & 1530 & 1296 \end{bmatrix}$$

$$[X'X]^{-1} = \begin{bmatrix} 10 & 62 & 90 \\ 62 & 514 & 405 \\ 90 & 405 & 1014 \end{bmatrix} = \frac{1}{30294} \begin{bmatrix} 357171 & -26418 & -21150 \\ -26418 & 2040 & 1530 \\ -21150 & 1530 & 1296 \end{bmatrix} = \begin{bmatrix} 11{,}7902 & -0{,}8721 & -0{,}6982 \\ -0{,}8721 & 0{,}0673 & 0{,}0505 \\ -0{,}6982 & 0{,}0505 & 0{,}0428 \end{bmatrix}$$

De otra parte,

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$[X'Y] = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 3 & 4 & 5 & 8 & 7 & 9 & 11 & 12 \\ 17 & 15 & 13 & 10 & 9 & 8 & 6 & 5 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 4 \\ 6 \\ 7 \\ 10 \\ 15 \\ 16 \\ 18 \\ 20 \end{bmatrix} = \begin{bmatrix} 100 \\ 843 \\ 621 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = [X'X]^{-1} X'Y = \frac{1}{30294} \begin{bmatrix} 357171 & -26418 & -21150 \\ -26418 & 2040 & 1530 \\ -21150 & 1530 & 1296 \end{bmatrix} \begin{bmatrix} 100 \\ 843 \\ 621 \end{bmatrix} = \begin{bmatrix} 10{,}32 \\ 0{,}93 \\ -0{,}67 \end{bmatrix}$$

El modelo de regresión será: $Y = 10{,}32 + 0{,}93X_1 - 0{,}67X_2$

- También se podía haber realizado teniendo en cuenta las ecuaciones normales mínimo cuadráticas:

$$\text{ecuaciones normales MCO} \begin{cases} \sum_{i=1}^{10} Y_i = N\beta_0 + \beta_1 \sum_{i=1}^{10} X_{1i} + \beta_2 \sum_{i=1}^{10} X_{2i} \\ \sum_{i=1}^{10} X_{1i} Y_i = \beta_0 \sum_{i=1}^{10} X_{1i} + \beta_1 \sum_{i=1}^{10} X_{1i}^2 + \beta_2 \sum_{i=1}^{10} X_{1i} X_{2i} \\ \sum_{i=1}^{10} X_{2i} Y_i = \beta_0 \sum_{i=1}^{10} X_{2i} + \beta_1 \sum_{i=1}^{10} X_{1i} X_{2i} + \beta_2 \sum_{i=1}^{10} X_{2i}^2 \end{cases}$$

Con estos datos, se obtiene:

Y_i	X_{1i}	X_{2i}	Y_i^2	X_{1i}^2	X_{2i}^2	$X_{1i}X_{2i}$	$X_{1i}Y_i$	$X_{2i}Y_i$
1	2	17	1	4	289	34	2	17
3	1	15	9	1	225	15	3	45
4	3	13	16	9	169	39	12	52
6	4	10	36	16	100	40	24	60
7	5	9	49	25	81	45	35	63
10	8	8	100	64	64	64	80	80
15	7	6	225	49	36	42	105	90
16	9	5	256	81	25	45	144	80
18	11	3	324	121	9	33	198	54
20	12	4	400	144	16	48	240	80
100	62	90	1416	514	1014	405	843	621

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$\begin{cases} \sum_{i=1}^{10} Y_i = N\beta_0 + \beta_1 \sum_{i=1}^{10} X_{1i} + \beta_2 \sum_{i=1}^{10} X_{2i} \\ \sum_{i=1}^{10} X_{1i} Y_i = \beta_0 \sum_{i=1}^{10} X_{1i} + \beta_1 \sum_{i=1}^{10} X_{1i}^2 + \beta_2 \sum_{i=1}^{10} X_{1i} X_{2i} \\ \sum_{i=1}^{10} X_{2i} Y_i = \beta_0 \sum_{i=1}^{10} X_{2i} + \beta_1 \sum_{i=1}^{10} X_{1i} X_{2i} + \beta_2 \sum_{i=1}^{10} X_{2i}^2 \end{cases} \rightarrow \begin{cases} 10\beta_0 + 62\beta_1 + 90\beta_2 = 100 \\ 62\beta_0 + 514\beta_1 + 405\beta_2 = 843 \\ 90\beta_0 + 405\beta_1 + 1014\beta_2 = 621 \end{cases}$$

en forma matricial,

$$\overbrace{\begin{pmatrix} 10 & 62 & 90 \\ 62 & 514 & 405 \\ 90 & 405 & 1014 \end{pmatrix}}^{X'X} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 100 \\ 843 \\ 621 \end{pmatrix} \Rightarrow \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \overbrace{\frac{1}{30294} \begin{pmatrix} 357171 & -26418 & -21150 \\ -26418 & 2040 & 1530 \\ -21150 & 1530 & 1296 \end{pmatrix}}^{[X'X]^{-1}} \overbrace{\begin{pmatrix} 100 \\ 843 \\ 621 \end{pmatrix}}^{X'Y}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 10{,}32 \\ 0{,}93 \\ -0{,}67 \end{pmatrix} \rightarrow Y = 10{,}32 + 0{,}93\ X_1 - 0{,}67\ X_2 + \text{Residuo} \quad \text{(Modelo regresión lineal)}$$

b) La fiabilidad del ajuste se refleja mediante R^2 coeficiente de determinación lineal:

A partir de la ecuación $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ se obtienen las predicciones y residuos asociados $u_i = Y_i - \hat{Y}_i$ a las observaciones muestrales.

De este modo, para la primera observación ($Y_1 = 1$; $X_{11} = 2$; $X_{21} = 17$), se tiene:

$$\begin{cases} \hat{Y}_1 = 10{,}32 + 0{,}93\ (2) - 0{,}67\ (17) = 0{,}79 \\ u_1 = Y_1 - \hat{Y}_1 = 1 - 0{,}79 = 0{,}21 \end{cases}$$

En esta línea, considerando todos los puntos muestrales, se obtiene:

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Y_i	$\hat{Y}_i$	$u_i = Y_i - \hat{Y}_i$	$(Y_i - \bar{Y})^2$	$u_i^2 = (Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \bar{Y})^2$
1	0,79	0,21	81	0,0441	84,8241
3	1,2	1,8	49	3,24	77,44
4	4,4	-0,4	36	0,16	31,36
6	7,34	-1,34	16	1,7956	7,0756
7	8,94	-1,94	9	3,7636	1,1236
10	12,4	-2,4	0	5,76	5,76
15	12,81	2,19	25	4,7961	7,8961
16	15,34	0,66	36	0,4356	28,5156
18	18,54	-0,54	64	0,2916	72,9316
20	18,8	1,2	100	1,44	77,44
100			416	21,7266	394,3666

$$SCT = \sum_{i=1}^{10}(Y_i - \bar{Y})^2 = 416$$

$$SCR = \sum_{i=1}^{10}(Y_1 - \hat{Y}_1)^2 = 21,7266$$

$$SCE = \sum_{i=1}^{10}(\hat{Y}_i - \bar{Y})^2 = 394,3666$$

Coeficiente de determinación lineal: $R^2 = \dfrac{SCE}{SCT} = \dfrac{\sum_{i=1}^{10}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{10}(Y_i - \bar{Y})^2} = \dfrac{394,3666}{416} = 0,948$

Coeficiente de determinación lineal corregido por el número de grados de libertad:

$$\bar{R}^2 = 1 - \frac{\frac{SCR}{n-k-1}}{\frac{SCT}{n-1}} = 1 - \frac{21,7266/7}{416/9} = 0,933$$

c) La parte de variabilidad contenida en Y que queda explicada por el método es precisamente $R^2 = 0,948 = 94,8\%$, que es suficientemente alta.

d) Para calcular los coeficientes de correlación lineal simple, se recurre a la tabla anterior, para calcular medias, varianzas y covarianzas

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{N} = \frac{100}{10} = 10 \qquad \sigma_y^2 = \frac{\sum_{i=1}^{10} y_i^2}{N} - \bar{y}^2 = \frac{1416}{10} - 10^2 = 41,6$$

$$\bar{x}_1 = \frac{\sum_{i=1}^{10} x_{1i}}{N} = \frac{62}{10} = 6,2 \qquad \sigma_{x_1}^2 = \frac{\sum_{i=1}^{10} x_{1i}^2}{N} - \bar{x}_1^2 = \frac{514}{10} - 6,2^2 = 12,96$$

$$\bar{x}_2 = \frac{\sum_{i=1}^{10} x_{2i}}{N} = \frac{90}{10} = 9 \qquad \sigma_{x_2}^2 = \frac{\sum_{i=1}^{10} x_{2i}^2}{N} - \bar{x}_2^2 = \frac{1014}{10} - 9^2 = 20,4$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{N} = \frac{100}{10} = 10 \qquad \sigma_y^2 = \frac{\sum_{i=1}^{10} y_i^2}{N} - \bar{y}^2 = \frac{1416}{10} - 10^2 = 41{,}6$$

$$m_{x_1 y} = \frac{\sum_{i=1}^{10} x_{1i}\, y_i}{N} - \bar{x}_1\, \bar{y} = \frac{843}{10} - (6{,}2).10 = 22{,}3$$

$$m_{x_2 y} = \frac{\sum_{i=1}^{10} x_{2i}\, y_i}{N} - \bar{x}_2\, \bar{y} = \frac{621}{10} - (9).10 = -27{,}9$$

$$m_{x_1 x_2} = \frac{\sum_{i=1}^{10} x_{1i}\, x_{2i}}{N} - \bar{x}_1\, \bar{x}_2 = \frac{405}{10} - (6{,}2).9 = -15{,}3$$

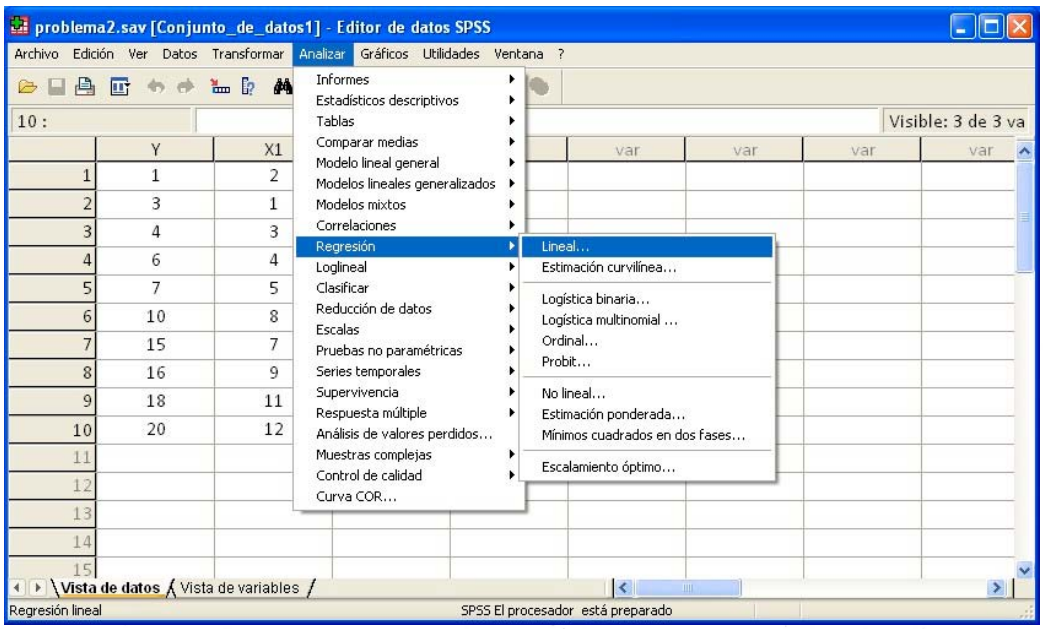
Coeficientes de correlación lineal simple:

$$r_{x_1 y} = \frac{m_{x_1 y}}{\sigma_{x_1} . \sigma_y} = \frac{22{,}3}{\sqrt{12{,}96}\,\sqrt{41{,}6}} = 0{,}96$$

$$r_{x_2 y} = \frac{m_{x_2 y}}{\sigma_{x_2} . \sigma_y} = \frac{-27{,}9}{\sqrt{20{,}4}\,\sqrt{41{,}6}} = -0{,}96$$

$$r_{x_1 x_2} = \frac{m_{x_1 x_2}}{\sigma_{x_1} . \sigma_{x_2}} = \frac{-15{,}3}{\sqrt{12{,}96}\,\sqrt{20{,}4}} = -0{,}94$$

Práctica en SPSS



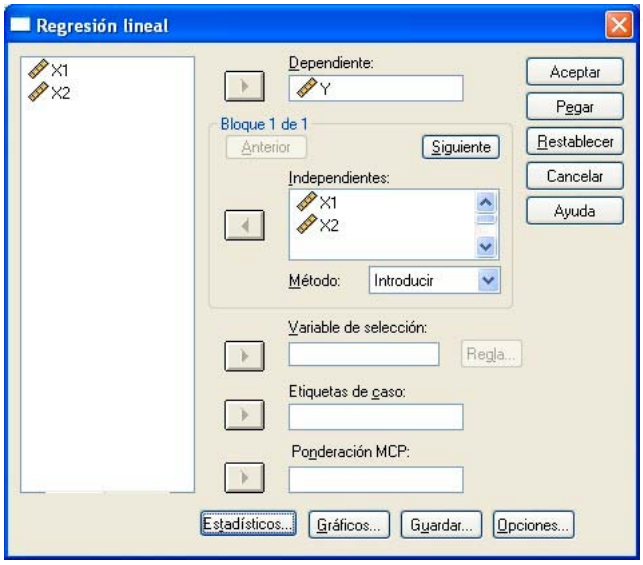
UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

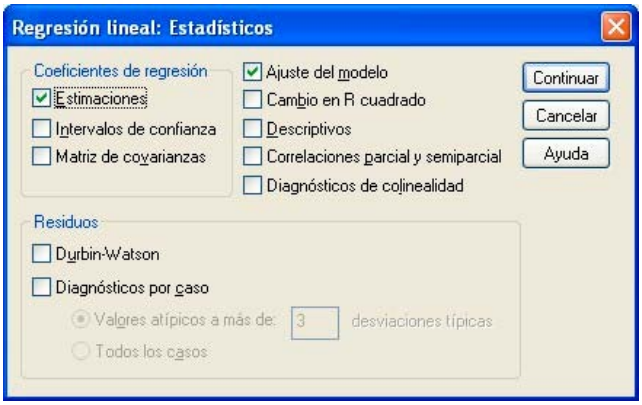
UAM
Universidad Autónoma de Madrid

Asignatura ..... **Grupo** .....

Apellidos ..... **Nombre** .....

Ejercicio del día .....





En el Visor de SPSS, se reflejan los resultados:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,974[a]	,948	,933	1,760

a. Variables predictoras: (Constante), X2, X1

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	394,305	2	197,153	63,612	,000[a]
	Residual	21,695	7	3,099		
	Total	416,000	9			

a. Variables predictoras: (Constante), X2, X1

b. Variable dependiente: Y

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	10,318	6,045		1,707	,132
	X1	,926	,457	,517	2,027	,082
	X2	-,673	,364	-,471	-1,849	,107

a. Variable dependiente: Y

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

En la tabla adjunta se refleja la calificación obtenida por una muestra de 12 estudiantes y su relación con la inteligencia y la motivación como variables predictoras..

Nota (Y)	Inteligencia (X_1)	Motivación (X_2)
4	85	10
5	100	20
8	95	35
7	80	30
10	180	45
6	90	25
6	110	10
7	120	15
4	80	10
4	95	15
6	160	15
9	150	45

Se pide:

1) Coeficientes de correlación simples, coeficiente de determinación múltiple y ecuación del plano de regresión.

2) Tabla ANOVA: Análisis de la Varianza

1) Tabla de cálculo para elaborar la matriz de varianzas–covarianzas

Y_i	X_{1i}	X_{2i}	Y_i^2	X_{1i}^2	X_{2i}^2	$X_{1i}Y_i$	$X_{2i}Y_i$	$X_{1i}X_{2i}$
4	85	10	16	7225	100	340	40	850
5	100	20	25	10000	400	500	100	2000
8	95	35	64	9025	1225	760	280	3325
7	80	30	49	6400	900	560	210	2400
10	180	45	100	32400	2025	1800	450	8100
6	90	25	36	8100	625	540	150	2250
6	110	10	36	12100	100	660	60	1100
7	120	15	49	14400	225	840	105	1800
4	80	10	16	6400	100	320	40	800
4	95	15	16	9025	225	380	60	1425
6	160	15	36	25600	225	960	90	2400
9	150	45	81	22500	2025	1350	405	6750
76	1345	275	524	163175	8175	9010	1990	33200

$$\bar{y} = \frac{\sum_{i=1}^{12} y_i}{N} = \frac{76}{12} = 6{,}3333 \qquad \sigma_y^2 = \frac{\sum_{i=1}^{12} y_i^2}{N} - \bar{y}^2 = \frac{524}{12} - 6{,}3333^2 = 3{,}5560$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$\bar{x}_1 = \frac{\sum_{i=1}^{12} x_{1i}}{N} = \frac{1345}{12} = 112{,}0833 \qquad \sigma^2_{x_1} = \frac{\sum_{i=1}^{12} x^2_{1i}}{N} - \bar{x}_1^2 = \frac{163175}{12} - 112{,}0833^2 = 1035{,}2430$$

$$\bar{x}_2 = \frac{\sum_{i=1}^{12} x_{2i}}{N} = \frac{275}{12} = 22{,}9167 \qquad \sigma^2_{x_2} = \frac{\sum_{i=1}^{12} x^2_{2i}}{N} - \bar{x}_2^2 = \frac{8175}{12} - 22{,}9167^2 = 156{,}0764$$

$$S_{yx_1} = \frac{\sum_{i=1}^{12} y_i\, x_{1i}}{N} - \bar{y}\,\bar{x}_1 = \frac{9010}{12} - (6{,}3333)\,.\,(112{,}0833) = 40{,}9762$$

$$S_{yx_2} = \frac{\sum_{i=1}^{12} y_i\, x_{2i}}{N} - \bar{y}\,\bar{x}_2 = \frac{1990}{12} - (6{,}3333)\,.\,(22{,}9167) = 20{,}6950$$

$$S_{x_1x_2} = \frac{\sum_{i=1}^{12} x_{1i}\, x_{2i}}{N} - \bar{x}_1\,\bar{x}_2 = \frac{33200}{12} - (112{,}0833)\,.\,(22{,}9167) = 198{,}0873$$

Matriz de varianzas–covarianzas: $VC = \begin{bmatrix} \sigma^2_y & S_{yx_1} & S_{yx_2} \\ S_{x_1y} & \sigma^2_{x_1} & S_{x_1x_2} \\ S_{x_2y} & S_{x_2x_1} & \sigma^2_{x_2} \end{bmatrix} = \begin{bmatrix} 3{,}5560 & 40{,}9762 & 20{,}6950 \\ 40{,}9762 & 1035{,}2430 & 198{,}0873 \\ 20{,}6950 & 198{,}0873 & 156{,}0764 \end{bmatrix}$

La matriz de correlaciones de las variables explicativas R_x está formada por los coeficientes de correlación lineal simple:

$$R_x = \begin{bmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{x_1y} & 1 & r_{x_1x_2} \\ r_{x_2y} & r_{x_2x_1} & 1 \end{bmatrix} \quad \text{donde} \quad r_{yx_1} = \frac{S_{yx_1}}{\sigma_y\,\sigma_{x_1}} \qquad r_{yx_2} = \frac{S_{yx_2}}{\sigma_y\,\sigma_{x_2}} \qquad r_{x_1x_2} = \frac{S_{x_1x_2}}{\sigma_{x_1}\,.\,\sigma_{x_2}}$$

$$r_{yx_1} = \frac{S_{yx_1}}{\sigma_y\,.\,\sigma_{x_1}} = \frac{40{,}9762}{\sqrt{3{,}5560}\,\sqrt{1035{,}2430}} = 0{,}6753 \qquad r_{yx_2} = \frac{S_{yx_2}}{\sigma_y\,.\,\sigma_{x_2}} = \frac{20{,}6950}{\sqrt{3{,}5560}\,\sqrt{156{,}0764}} = 0{,}8784$$

$$r_{x_1x_2} = \frac{S_{x_1x_2}}{\sigma_{x_1}\,.\,\sigma_{x_2}} = \frac{198{,}0873}{\sqrt{1035{,}2430}\,\sqrt{156{,}0764}} = 0{,}4928$$

Por tanto, la matriz de las correlaciones: $R_x = \begin{bmatrix} 1 & 0{,}6753 & 0{,}8784 \\ 0{,}6753 & 1 & 0{,}4928 \\ 0{,}8784 & 0{,}4928 & 1 \end{bmatrix}$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Coeficiente de determinación $R^2_{yx_1} = r^2_{yx_1} = 0{,}6753^2 = 0{,}4560$

El 45,6% de la varianza de la nota (Y) queda explicada por el coeficiente de inteligencia (X_1)

Coeficiente de determinación $R^2_{yx_2} = r^2_{yx_2} = 0{,}8784^2 = 0{,}7715$

El 77,15% de la varianza de la nota (Y) queda explicada por la motivación del estudiante (X_2)

El Coeficiente de determinación múltiple viene definido por:

$$\text{a) } r^2_{yx_1x_2} = 1 - \frac{|CV|}{\sigma^2_y\, C_{yy}} \qquad \text{b) } R^2 = r^2_{yx_1x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2\, r_{yx_1}\, r_{yx_2}\, r_{x_1x_2}}{1 - r^2_{x_1x_2}}$$

$$\text{a) } |CV| = \begin{vmatrix} 3{,}5560 & 40{,}9762 & 20{,}6950 \\ 40{,}9762 & 1035{,}2430 & 198{,}0873 \\ 20{,}6950 & 198{,}0873 & 156{,}0764 \end{vmatrix} = 65555{,}5424$$

$$VC_{yy} = \begin{vmatrix} 1035{,}2430 & 198{,}0873 \\ 198{,}0873 & 156{,}0764 \end{vmatrix} = 122338{,}4221$$

$$r^2_{yx_1x_2} = 1 - \frac{|CV|}{\sigma^2_y\, C_{yy}} = 1 - \frac{|CV|}{\sigma^2_y\, C_{yy}} = 1 - \frac{65555{,}5424}{(3{,}5560)(122338{,}4221)} = 0{,}8492$$

$$\text{b) } r^2_{yx_1x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2 r_{yx_1}\, r_{yx_2}\, r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{0{,}6753^2 + 0{,}8784^2 - 2\,.\,0{,}6753\,.\,0{,}8784\,.\,0{,}4928}{1 - 0{,}4928^2} = 0{,}8492$$

Los coeficientes $(\hat\beta_1, \hat\beta_2)$, respectivamente, con signo negativo $(-)$, vienen dados por el cociente de los adjuntos (S_{yx_1}, S_{yx_2}) entre el adjunto de σ^2_y:

$$\hat\beta_1 = -\frac{VC_{yx_1}}{VC_y} \qquad \hat\beta_2 = -\frac{VC_{yx_2}}{VC_y} \qquad \hat\beta_0 = \bar Y - \hat\beta_1 \bar X_1 - \hat\beta_2 \bar X_2$$

$$\hat\beta_1 = -(-)\, \frac{\begin{vmatrix} 40{,}9762 & 198{,}0873 \\ 20{,}6950 & 156{,}0764 \end{vmatrix}}{\begin{vmatrix} 1035{,}2430 & 198{,}0873 \\ 198{,}0873 & 156{,}0764 \end{vmatrix}} = \frac{2296}{122338{,}4221} = 0{,}01877 \approx (0{,}019)$$

$$\hat\beta_2 = (-)\, \frac{\begin{vmatrix} 40{,}9762 & 1035{,}2430 \\ 20{,}6950 & 198{,}0873 \end{vmatrix}}{\begin{vmatrix} 1035{,}2430 & 198{,}0873 \\ 198{,}0873 & 156{,}0764 \end{vmatrix}} = -\frac{-13307{,}489}{122338{,}4221} = 0{,}1088 \approx (0{,}109)$$

$$\hat\beta_0 = \bar Y - \hat\beta_1 \bar X_1 - \hat\beta_2 \bar X_2 = 6{,}3333 - (0{,}01877)\,(112{,}0833) - (0{,}1088)\,(22{,}9167) = 1{,}737$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Ecuación del plano de regresión: $\hat{Y} = 1{,}737 + 0{,}019\,X_1 + 0{,}109\,X_2$

2) Tabla ANOVA: Análisis de la Varianza

A partir de la ecuación de l plano de regresión $\hat{Y}_i = 1{,}737 + 0{,}019\,X_{1i} + 0{,}109\,X_{2i}$ se obtienen las predicciones y residuos asociados $u_i = Y_i - \hat{Y}_i$ a las observaciones muestrales.

De este modo, para la primera observación $(Y_1 = 4\ ,\ X_{11} = 85\ ,\ X_{21} = 10)$, se tiene:

$$\begin{cases} \hat{Y}_1 = 1{,}737 + 0{,}019\ (85) + 0{,}109\ (10) = 4{,}420 \\ u_1 = Y_1 - \hat{Y}_1 = 4 - 4{,}420 = -0{,}420 \end{cases}$$

En esta línea, considerando todos los puntos muestrales, se obtiene:

Y_i	X_{1i}	X_{2i}	$\hat{Y}_i$	$u_i = Y_i - \hat{Y}_i$	$(Y_i - \bar{Y})^2$	$u_i^2 = (Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \bar{Y})^2$
4	85	10	4,420	-0,420	5,444	0,177	3,658
5	100	20	5,790	-0,790	1,778	0,624	0,295
8	95	35	7,328	0,672	2,778	0,451	0,990
7	80	30	6,503	0,497	0,444	0,247	0,029
10	180	45	10,012	-0,012	13,445	0,000	13,532
6	90	25	6,146	-0,146	0,111	0,021	0,035
6	110	10	4,890	1,110	0,111	1,233	2,083
7	120	15	5,621	1,379	0,444	1,901	0,506
4	80	10	4,327	-0,327	5,444	0,107	4,026
4	95	15	5,152	-1,152	5,444	1,327	1,394
6	160	15	6,372	-0,372	0,111	0,139	0,002
9	150	45	9,449	-0,448	7,111	0,201	9,706
76	1345	275			42,667	6,428	36,256

$$SCT = \sum_{i=1}^{12} (Y_i - \bar{Y})^2 = 42{,}667$$

$$SCR = \sum_{i=1}^{12} (Y_1 - \hat{Y}_1)^2 = 6{,}428$$

$$SCE = \sum_{i=1}^{12} (\hat{Y}_i - \bar{Y})^2 = 36{,}256$$

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{12} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{12} (Y_i - \bar{Y})^2} = \frac{36{,}256}{42{,}667} = 0{,}849 \quad \text{coeficiente de determinación múltiple}$$

$$S_R^2 = \frac{SCR}{n-k-1} = \frac{6{,}428}{9} = 0{,}714 \quad \text{varianza residual} \qquad S_R = \sqrt{0{,}714} = 0{,}845 \quad \text{error típico estimación}$$

$$\bar{R}^2 = 1 - \frac{SCR\,/\,(n-k-1)}{SCT\,/\,(n-1)} = 1 - \frac{6{,}428\,/\,9}{42{,}667\,/\,11} = 0{,}816 \quad \text{coeficiente de determinación múltiple corregido}$$

$$F = \frac{SCE\,/\,k}{SCR\,/\,(n-k-1)} = \frac{36{,}256\,/\,2}{6{,}428\,/\,9} = 25{,}389 \quad \text{estadístico observado F de Fisher-Snedecor}$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

ANÁLISIS DE LA VARIANZA: TABLA ANOVA

Variación	Suma cuadrados	Grados libertad	Media cuadrática	F-Snedecor
Regresión	$SCE = 36,256$	$k = 2$	$\frac{36,256}{2} = 18,128$	$F = \frac{18,128}{0,714} = 25,389$
Residual	$SCR = 6,428$	$n-k-1 = 9$	$\frac{6,428}{9} = 0,714$	
Total	$SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$n-1 = 12-1 = 11$		

CONTRASTE DE HIPÓTESIS

Hipótesis nula: H_0: $\beta_1 = \beta_2 = 0$ el modelo no es explicativo

Se rechaza H_0 : $F = 25,389 > 4,2565 = F_{0,05\,;\,2,9}$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

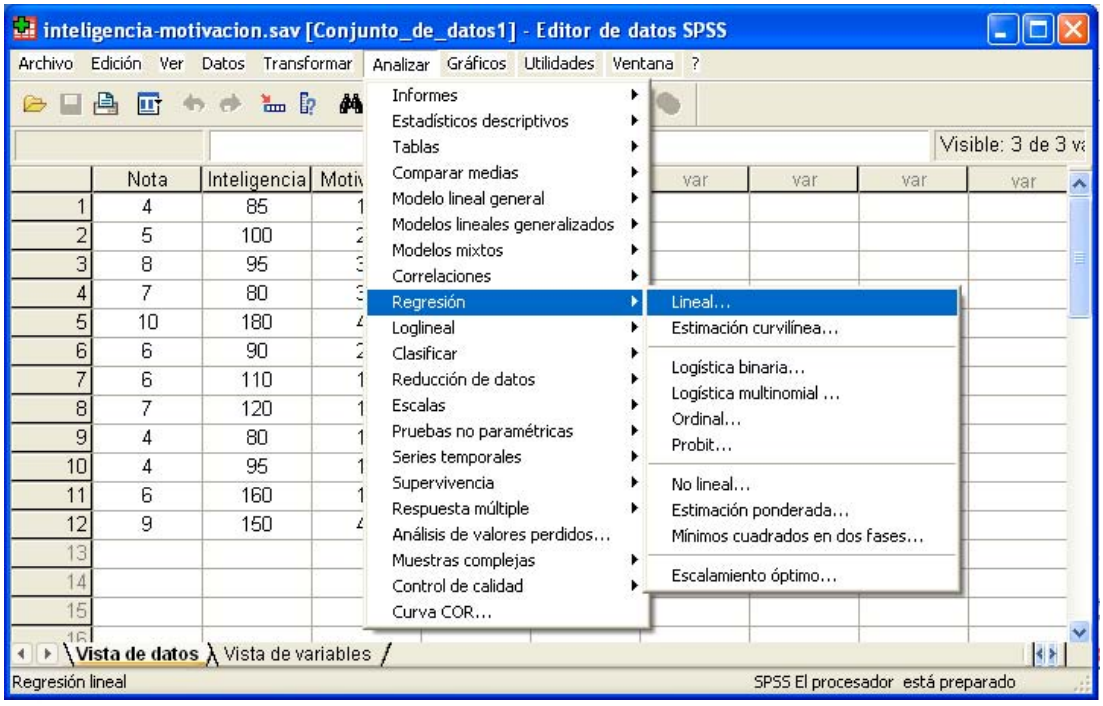
Universidad Autónoma de Madrid

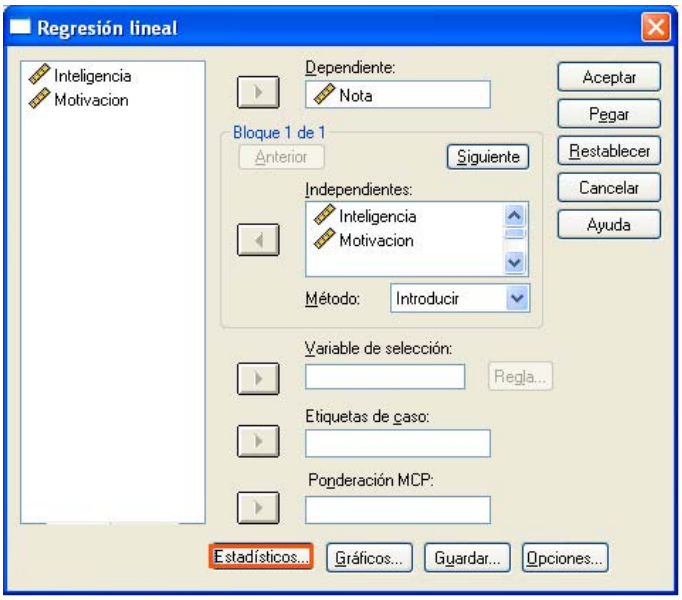
Asignatura ..... **Grupo** .....

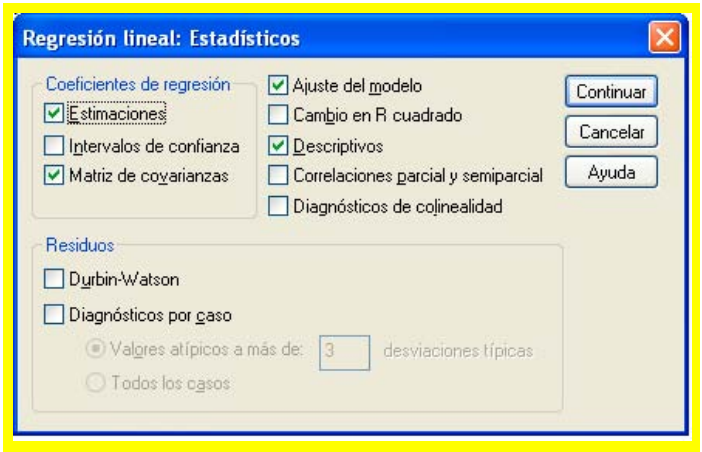
Apellidos ..... **Nombre** .....

Ejercicio del día .....

Guía Práctica en SPSS _____ Analizar /Regresión / Lineal _____







UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Estadísticos descriptivos

	Media	Desviación típ.	N
Nota	6,33	1,969	12
Inteligencia	112,08	33,606	12
Motivacion	22,92	13,049	12

Correlaciones

		Nota	Inteligencia	Motivacion
Correlación de Pearson	Nota	1,000	,675	,878
	Inteligencia	,675	1,000	,493
	Motivacion	,878	,493	1,000
Sig. (unilateral)	Nota	.	,008	,000
	Inteligencia	,008	.	,052
	Motivacion	,000	,052	.
N	Nota	12	12	12
	Inteligencia	12	12	12
	Motivacion	12	12	12

Correlaciones bivariadas:

$r_{yx_1} = r_{x_1y} = 0,6753$

$r_{yx_2} = r_{x_2y} = 0,8784$

$r_{x_1x_2} = r_{x_2x_1} = 0,4928$

Correlaciones parciales:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1-r^2_{yx_2})(1-r^2_{x_1x_2})}} = \frac{0,6753 - 0,8784 \cdot 0,4928}{\sqrt{(1-0,8784^2)(1-0,4928^2)}} = 0,583$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_2x_1}}{\sqrt{(1-r^2_{yx_1})(1-r^2_{x_2x_1})}} = \frac{0,8784 - 0,6753 \cdot 0,4928}{\sqrt{(1-0,6753^2)(1-0,4928^2)}} = 0,850$$

Correlaciones semiparciales:

La correlación semiparcial entre variables permite conocer el poder explicativo de cada predictor en el modelo de regresión cuando éste se incorpora después de otro u otros regresores ya contemplados. De este modo, se decide si ampliar o no el número de variables explicativas al modelo en función de que los coeficientes de correlación semiparcial de éstas (incorporadas de nuevo) sean significativos o no lo sean.

Coeficiente de determinación semiparcial incorporando variable X_2 :

$$r^2_{y.(x_1.x_2)} = r^2_{y.x_1x_2} - r^2_{y.x_2} = 0,8492 - 0,8784^2 = 0,0776 \rightarrow r_{y.(x_1.x_2)} = \sqrt{0,0776} = 0,279$$

Coeficiente de determinación semiparcial incorporando variable X_1 :

$$r^2_{y.(x_2.x_1)} = r^2_{y.x_1x_2} - r^2_{y.x_1} = 0,8492 - 0,6753^2 = 0,3931 \rightarrow r_{y.(x_2.x_1)} = \sqrt{0,3931} = 0,627$$

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,922[a]	,849	,816	,845

a. Variables predictoras: (Constante), Motivacion, Inteligencia

El coeficiente de correlación múltiple al cuadrado corregido es una forma de minimizar la suma de todas las correlaciones que mantienen las variables predictoras con el criterio.

Es decir, una forma de aminorar lo inflado que puede resultar dicho coeficiente.

El error típico de la estimación $S^2_R = \sqrt{\frac{SCR}{n-k-1}} = \sqrt{\frac{6,428}{9}} = 0,845$ representa una medida de la parte

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

de variabilidad de la variable dependiente (Nota) que no es explicada por la recta de regresión. En general, cuando mejor es el ajuste, más pequeño es este error típico.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	1,737	,885		1,964	,081
	Inteligencia	,019	,009	,320	2,153	,060
	Motivacion	,109	,022	,721	4,847	,001

a. Variable dependiente: Nota

La tabla de coeficientes proporciona información sobre la ecuación de regresión y sus coeficientes.

La ecuación representa la regresión de la inteligencia (X_1) y la motivación (X_2) sobre la nota:

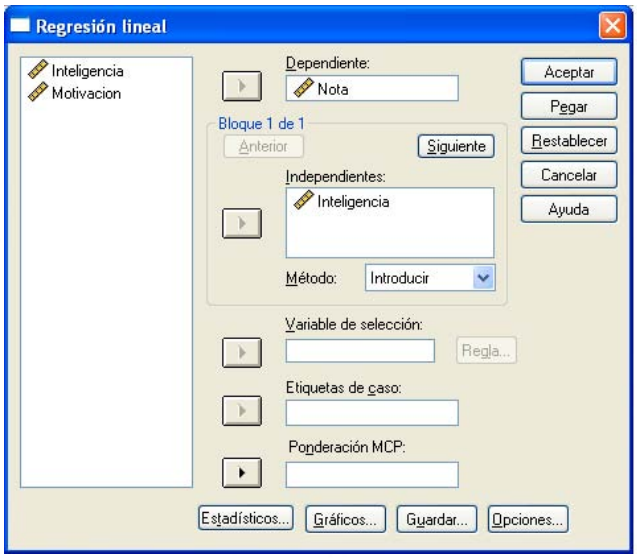
$\hat{Y} = 1{,}737 + 0{,}019\ X_1 + 0{,}109\ X_2$

La nota prevista en ausencia de motivación y de inteligencia alguna es de 1,737

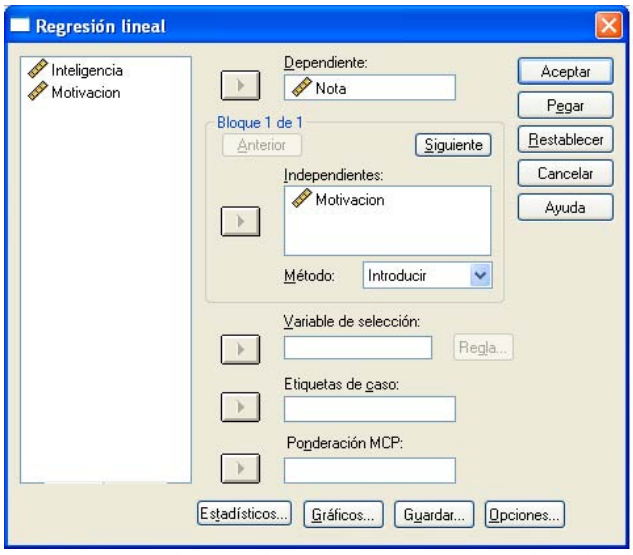
Señalar que estos coeficientes no son independientes entre sí. De hecho, reciben el nombre de coeficientes de regresión parcial porque el valor estimado para coeficiente se ajusta teniendo en cuenta la presencia de variables independientes. Conviene, por tanto, interpretarlos con cautela.

Adviértase que si se hubiera probado en una ecuación de regresión simple el efecto de la inteligencia sobre la nota o bien el efecto de la motivación sobre la nota, los resultados hubieran sido otros:

a)



b)



UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

a)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	1,897	1,594		1,190	,261
	Inteligencia	,040	,014	,675	2,896	,016

a. Variable dependiente: Nota

Un efecto estadísticamente significativo, que no lo es en la ecuación múltiple. Obsérvese la coincidencia del coeficiente estandarizado (Beta) con el valor de la correlación bivariada calculado en el cuadro de correlaciones.

b)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	3,295	,595		5,536	,000
	Motivacion	,133	,023	,878	5,814	,000

a. Variable dependiente: Nota

La interpretación de los coeficientes de la ecuación de regresión múltiple, pues, hay que hacerla con mucha cautela. Un coeficiente pequeño no es reflejo de una baja correlación entre la variable que lo acompaña y el criterio sino que puede significar que la información compartida entre dicha variable y otra (u otras) predictoras del modelo es muy alta.

La recomendación es, teniendo en cuenta que el modelo se elabora con estos condicionantes, que las variables predictoras que explican Y sean lo más independientes entre sí, es decir, no mantengan relación o colinealidad . Sin embargo, la mayoría de las veces la ausencia total de colinealidad entre ellas es imposible de satisfacer y por tanto hay que jugar con dicha información compartida entre predictores a la hora de aportar la interpretación última de la ecuación de regresión.

El signo del coeficiente de correlación parcial de una variable puede no ser el mismo que el del coeficiente de correlación simple entre esa variable y la dependiente. Esto es debido a los ajustes que se llevan a cabo para poder obtener la mejor ecuación posible. Aunque existen diferentes explicaciones para justificar el cambio del signo de un coeficiente de regresión, una de las que más seriamente deben de ser consideradas es la que se refiere a un alto grado de asociación entre algunas de las variables independientes (colinealidad).

Los coeficientes Beta estandarizados están basados en las puntuaciones típicas y, por tanto, son directamente comparables entre sí. Indican la cantidad de cambio, en puntuaciones típicas, que se produce en la variable dependiente por cada cambio de una unidad en la correspondiente variable independiente (manteniendo constantes el resto de variables independientes).
Estos coeficientes proporcionan una pista muy útil sobre la importancia relativa de cada variable independiente en la recta de regresión. En general, una variable tiene tanto más peso (importancia) en la ecuación de regresión cuanto mayor (en valor absoluto) es su coeficiente de regresión estandarizado. En la tabla se observa que la variable Motivación es la más importante.

En función de los coeficientes Beta estandarizados, la ecuación de regresión sería:

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

$\hat{Y} = 0{,}320\ X_1 + 0{,}721\ X_2$

Los coeficientes son directamente comparables. En este caso, el efecto esperado (en desviaciones típicas) de la nota a cargo de la motivación (X_2) es de $0{,}721 - 0{,}320 = 0{,}401$ desviaciones típicas más que el cambio esperado por el efecto exclusivo de la variable inteligencia (X_1).

La pruebas de significación: Pruebas t y sus niveles críticos sirven para contrastar la hipótesis nula de que un coeficiente de regresión vale 0 en la población.

Hipótesis nula: H_0: $\beta_1 = \beta_2 = 0$ el modelo no es explicativo

Hipótesis alternativa: H_1: Al menos un $\beta_i \neq 0$ el modelo es explicativo

Cuando el nivel crítico (p-valor ≡ Sig.) $> \alpha \rightarrow$ se acepta la hipótesis nula.

La variable Motivación presenta un p-valor = 0,001 < 0,05 → Con una significación del 5% no contribuye a explicar lo que ocurre con la variable dependiente.

VALIDEZ DEL MODELO: El ANOVA y su correspondiente índice F es otra información proporcionada por el comando regresión en el SPSS para el modelo múltiple.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	36,238	2	18,119	25,368	,000[a]
	Residual	6,428	9	,714		
	Total	42,667	11			

a. Variables predictoras: (Constante), Motivacion, Inteligencia

b. Variable dependiente: Nota

$H_0: R_{yx_1x_2} = 0$

Sig. = 0,000 < 0,05

→ Se rechaza H_0

REGRESIÓN POR PASOS: STEPWISE

Hasta ahora el control de las variables utilizadas para construir el modelo de regresión recae sobre el propio analista. El analista decide qué variables independientes desea incluir en la ecuación de regresión seleccionándolas en la lista de variables independientes.

Sin embargo, se encuentran situaciones en las que, existiendo un elevado número de variables independientes, no existe una teoría o un trabajo previo que oriente al analista en la elección de variables relevantes. Este tipo de situaciones pueden afrontarse utilizando procedimientos diseñados para seleccionar, entre una gran cantidad de variables, solo un conjunto reducido de las mismas, aquellas que permiten obtener el mejor ajuste posible.

Un análisis pormenorizado sobre los predictores que se trabajan en el ejemplo conduce a afirmar que no todos son igualmente relevantes. Quizás sea conveniente depurar el modelo de regresión múltiple, eliminado regresores que no explican mucho sobre la variable Y al presentar una predicción no significativa.

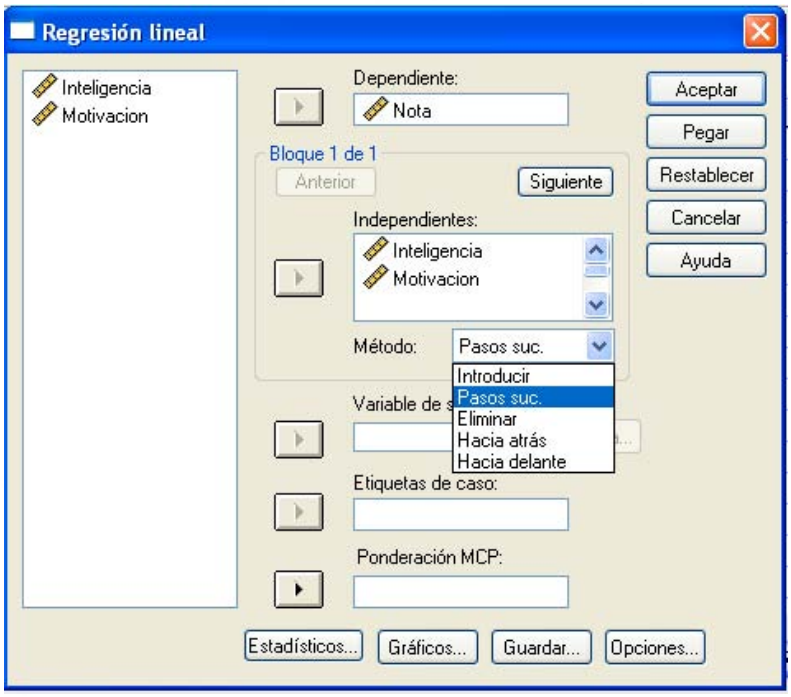
UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....



Calculando la ecuación de regresión múltiple mediante el método por pasos (stepwise) en vez de mediante el método "introducir", la salida obtenida informa desde el principio sobre la significación o no de cada variable predictora en el modelo, calculando el poder de predicción del modelo una vez eliminada aquella variable no relevante en el mismo.

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Motivacion		Por pasos (criterio: Prob. de F para entrar <= ,050, Prob. de F para salir >= ,100).

a. Variable dependiente: Nota

La única variable contemplada en la ecuación es la Motivación, dado que únicamente su coeficiente satisface la condición de tener asociado un p-valor $< 0,05$.

El coeficiente de correlación simple (sólo existe dicha variable en la ecuación) con la nota es de 0,8784 (obtenido en el análisis de correlaciones bivariadas) y su coeficiente de determinación es de 0,772.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,878[a]	,772	,749	,987

a. Variables predictoras: (Constante), Motivacion

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	32,927	1	32,927	33,807	,000[a]
	Residual	9,740	10	,974		
	Total	42,667	11			

a. Variables predictoras: (Constante), Motivacion

b. Variable dependiente: Nota

El modelo resulta significativo con un p-valor = 0,000 (mismo del ANOVA de la ecuación de regresión simple Motivación - Nota)

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	3,295	,595		5,536	,000
	Motivacion	,133	,023	,878	5,814	,000

a. Variable dependiente: Nota

La ecuación de regresión es, tal y como se estimaba antes a través del modelo de regresión simple al eliminar la inteligencia:

$$\hat{Y} = 0{,}3295 + 0{,}133\, X_2$$

Variables excluidas[b]

Modelo		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad
						Tolerancia
1	Inteligencia	,320[a]	2,153	,060	,583	,757

a. Variables predictoras en el modelo: (Constante), Motivacion

b. Variable dependiente: Nota

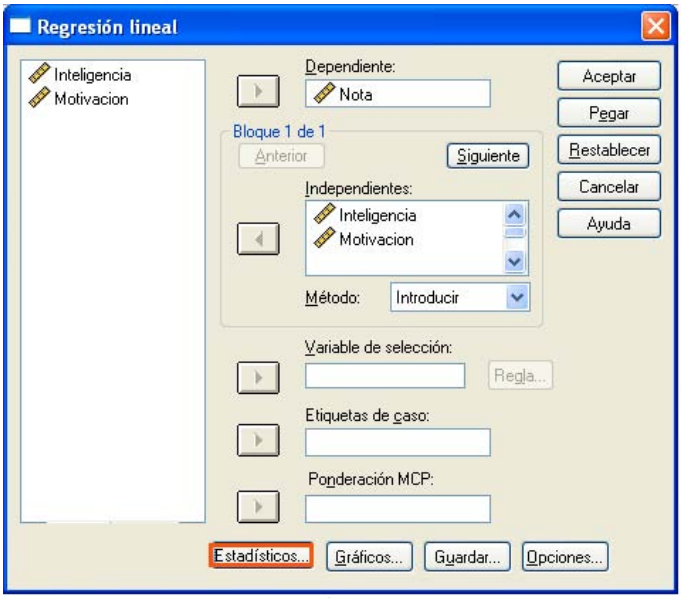
El cuadro presenta información sobre la variable Inteligencia excluida del modelo.

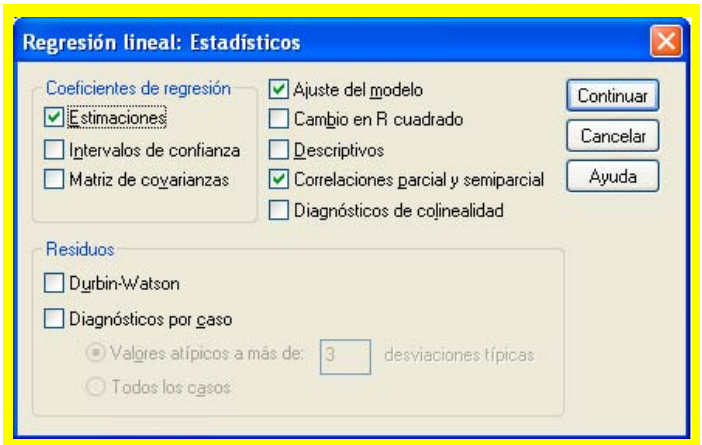
Contiene la correlación parcial de la inteligencia con la nota (0,583) que se refiere a la correlación entre ambas, cuando tanto de la Inteligencia como de la Nota se elimina o no se tiene en cuenta, lo que comparten con la Motivación.

El índice de tolerancia es el complementario del coeficiente de determinación entre inteligencia y motivación (variables predictoras del modelo), de tal manera que valores bajos de este estadístico representan correlaciones altas entre ellas y valores altos correlaciones bajas entre ellas.

Índice de tolerancia: $I.T = 1 - r^2_{x_1 x_2} = 1 - 0{,}4928^2 = 0{,}7571$

Para conocer el poder explicativo de la variable Inteligencia en el modelo de regresión cuando ésta se incorpora después de la variable regresora Motivación, se analizar la correlación semiparcial.





UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones		
		B	Error típ.	Beta			Orden cero	Parcial	Semiparcial
1	(Constante)	1,737	,885		1,964	,081			
	Inteligencia	,019	,009	,320	2,153	,060	,675	,583	,279
	Motivacion	,109	,022	,721	4,847	,001	,878	,850	,627

a. Variable dependiente: Nota

El incremento en la variabilidad explicada de Y al introducir como nueva la variable Inteligencia X_1 es de 0,279, que indica su correlación con las notas (Y) una vez eliminada aquella correlación que éstas mantienen con la Motivación.

Por los análisis realizados se sabe que este incremento en variabilidad explicada por parte de la Inteligencia sobre la Motivación no resulta estadísticamente significativa, de ahí que se considera aconsejable eliminar dicha variable del modelo..

No obstante, con el estadístico F se puede validar la significación de dicho coeficiente de correlación semiparcial:

$$r^2_{y.(x_1.x_2)} = r^2_{y.x_1x_2} - r^2_{y.x_2} = 0{,}8492 - 0{,}8784^2 = 0{,}0776$$

$$F = \frac{\frac{r^2_{y.x_1x_2} - r^2_{y.(x_1.x_2)}}{p-k}}{\frac{1-r^2_{y.x_1x_2}}{n-p-1}} = \frac{\frac{0{,}8492 - 0{,}8784^2}{2-1}}{\frac{1-0{,}8492}{12-2-1}} = 4{,}631 < 5{,}1174 = F_{0{,}05;\,1,9}$$

Concluyendo que la correlación semiparcial de la Inteligencia con la Nota no es estadísticamente significativa. Es decir, la incorporación de la variable Inteligencia al modelo en el que ya se contemplaba la Motivación no aumenta significativamente la explicación de Y.

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

El gerente de una empresa estudia las posibles relaciones entre beneficios anuales, gastos en publicidad anuales y horas extraordinarias anuales de los empleados. Para ello utiliza datos, de estas tres variables, proporcionadas por algunas empresas del sector.

Beneficios (millones)	Gastos Publicidad (millones)	Horas extras (100 horas)
1,3	0,3	4
3,5	1,5	9
2,8	0,7	6
3	1,1	7,5
3,3	1,2	8
4	2	7
3,7	2	8

Se desea saber:

1) Matriz de varianzas–covarianzas.
2) Matriz de correlación.
3) ¿Qué porcentaje de la varianza de los beneficios explicaría una función lineal de los gastos en publicidad?
4) ¿Qué porcentaje de la varianza de los beneficios explicaría una función lineal de las horas extraordinarias anuales de los empleados?
5) Establecer una relación lineal que explique anualmente los beneficios mediante los gastos en publicidad y horas extras.
6) Hallar el coeficiente de correlación múltiple. ¿Qué porcentaje de la varianza de beneficios queda explicado por el modelo lineal obtenido en el apartado anterior?
7) Coeficientes de correlación parcial de beneficios con gastos en publicidad y de beneficios con horas extras de los empleados.
8) Coeficiente de determinación múltiple.

1) Matriz de varianzas–covarianzas: $VC = \begin{bmatrix} \sigma_y^2 & S_{yx_1} & S_{yx_2} \\ S_{x_1y} & \sigma_{x_1}^2 & S_{x_1x_2} \\ S_{x_2y} & S_{x_2x_1} & \sigma_{x_2}^2 \end{bmatrix}$ $\quad S_{yx_1} = S_{x_1y}$ $\quad S_{yx_2} = S_{x_2y}$ $\quad S_{x_1x_2} = S_{x_2x_1}$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

Tabla Operaciones:

Y_i	X_{1i}	X_{2i}	Y_i^2	X_{1i}^2	X_{2i}^2	$X_{1i}Y_i$	$X_{2i}Y_i$	$X_{1i}X_{2i}$
1,3	0,3	4	1,69	0,09	16	0,39	5,2	1,2
3,5	1,5	9	12,25	2,25	81	5,25	31,5	13,5
2,8	0,7	6	7,84	0,49	36	1,96	16,8	4,2
3	1,1	7,5	9	1,21	56,25	3,3	22,5	8,25
3,3	1,2	8	10,89	1,44	64	3,96	26,4	9,6
4	2	7	16	4	49	8	28	14
3,7	2	8	13,69	4	64	7,4	29,6	16
21,6	8,8	49,5	71,36	13,48	366,25	30,26	160	66,75

$$\bar{y} = \frac{\sum_{i=1}^{7} y_i}{N} = \frac{21,6}{7} = 3,0857 \qquad \sigma_y^2 = \frac{\sum_{i=1}^{7} y_i^2}{N} - \bar{y}^2 = \frac{71,36}{7} - 3,0857^2 = 0,6727$$

$$\bar{x}_1 = \frac{\sum_{i=1}^{7} x_{1i}}{N} = \frac{8,8}{7} = 1,2571 \qquad \sigma_{x_1}^2 = \frac{\sum_{i=1}^{7} x_{1i}^2}{N} - \bar{x}_1^2 = \frac{13,48}{7} - 1,2571^2 = 0,3454$$

$$\bar{x}_2 = \frac{\sum_{i=1}^{7} x_{2i}}{N} = \frac{49,5}{7} = 7,0714 \qquad \sigma_{x_2}^2 = \frac{\sum_{i=1}^{7} x_{2i}^2}{N} - \bar{x}_2^2 = \frac{366,25}{7} - 7,0714^2 = 2,3167$$

$$S_{yx_1} = \frac{\sum_{i=1}^{7} y_i\, x_{1i}}{N} - \bar{y}\,\bar{x}_1 = \frac{30,26}{7} - (3,0857)\,.\,(1,2571) = 0,4438$$

$$S_{yx_2} = \frac{\sum_{i=1}^{7} y_i\, x_{2i}}{N} - \bar{y}\,\bar{x}_2 = \frac{160}{7} - (3,0857)\,.\,(7.0714) = 1,0369$$

$$S_{x_1x_2} = \frac{\sum_{i=1}^{7} x_{1i}\, x_{2i}}{N} - \bar{x}_1\,\bar{x}_2 = \frac{66,75}{7} - (1,2571)\,.\,(7,0714) = 0,6462$$

Matriz de varianzas-covarianzas: $VC = \begin{bmatrix} 0,6727 & 0,4438 & 1,0369 \\ 0,4438 & 0,3454 & 0,6462 \\ 1,0369 & 0,6462 & 2,3167 \end{bmatrix}$

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

2) La matriz de correlaciones de las variables explicativas R_x está formada por los coeficientes de correlación lineal simple:

$$R_x = \begin{bmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{x_1y} & 1 & r_{x_1x_2} \\ r_{x_2y} & r_{x_2x_1} & 1 \end{bmatrix} \quad \text{donde} \quad \begin{cases} r_{yx_1} = r_{x_1y} \\ r_{yx_2} = r_{x_2y} \\ r_{x_1x_2} = r_{x_2x_1} \end{cases} \qquad r_{yx_1} = \frac{S_{yx_1}}{\sigma_y \sigma_{x_1}} \qquad r_{yx_2} = \frac{S_{yx_2}}{\sigma_y \sigma_{x_2}}$$

$$r_{yx_1} = \frac{S_{yx_1}}{\sigma_y . \sigma_{x_1}} = \frac{0,4438}{\sqrt{0,6727}\sqrt{0,3454}} = 0,9207 \qquad r_{yx_2} = \frac{S_{yx_2}}{\sigma_y . \sigma_{x_2}} = \frac{1,0369}{\sqrt{0,6727}\sqrt{2,3167}} = 0,83067$$

$$r_{x_1x_2} = \frac{S_{x_1x_2}}{\sigma_{x_1} . \sigma_{x_2}} = \frac{0,6462}{\sqrt{0,3454}\sqrt{2,3167}} = 0,7224$$

Por tanto, la matriz de las correlaciones: $R_x = \begin{bmatrix} 1 & 0,9207 & 0,8306 \\ 0,9207 & 1 & 0,7224 \\ 0,8306 & 0,7224 & 1 \end{bmatrix}$

3) Se requiere hallar el coeficiente de determinación entre (Y, X_1): $R^2_{yx_1} = r^2_{yx_1} = 0,9207^2 = 0,8477$

Es decir, el 84,77% de la varianza de los beneficios (Y) queda explicado por una función lineal del gasto en publicidad (X_1)

4) Se requiere hallar el coeficiente de determinación entre (Y, X_2): $R^2_{yx_2} = r^2_{yx_2} = 0,8306^2 = 0,6899$

Es decir, el 68,99% de la varianza de los beneficios (Y) queda explicado por una función lineal de las horas extras de los empleados (X_2)

5) Hay que determinar el plano de regresión de los beneficios (Y) sobre el gasto en publicidad (X_1) y las horas extras de los empleados (X_2)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

Se tiene como referencia la matriz de las varianzas-covarianzas:

$$VC = \begin{bmatrix} \sigma^2_y & S_{yx_1} & S_{yx_2} \\ S_{x_1y} & \sigma^2_{x_1} & S_{x_1x_2} \\ S_{x_2y} & S_{x_2x_1} & \sigma^2_{x_2} \end{bmatrix} = \begin{bmatrix} 0,6727 & 0,4438 & 1,0369 \\ 0,4438 & 0,3454 & 0,6462 \\ 1,0369 & 0,6462 & 2,3167 \end{bmatrix}$$

Los coeficientes $(\hat{\beta}_1, \hat{\beta}_2)$, respectivamente, con signo negativo $(-)$, vienen dados por el cociente de los adjuntos (S_{yx_1}, S_{yx_2}) entre el adjunto de σ^2_y:

$$\hat{\beta}_1 = -\frac{VC_{yx_1}}{VC_y} \qquad \hat{\beta}_2 = -\frac{VC_{yx_2}}{VC_y} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$$\hat{\beta}_1 = -(-)\frac{\begin{vmatrix} 0{,}4438 & 0{,}6462 \\ 1{,}0369 & 2{,}3167 \end{vmatrix}}{\begin{vmatrix} 0{,}3454 & 0{,}6462 \\ 0{,}6462 & 2{,}3167 \end{vmatrix}} = \frac{0{,}3581}{0{,}3826} = 0{,}9360 \quad \hat{\beta}_2 = (-)\frac{\begin{vmatrix} 0{,}4438 & 0{,}3454 \\ 1{,}0369 & 0{,}6462 \end{vmatrix}}{\begin{vmatrix} 0{,}3454 & 0{,}6462 \\ 0{,}6462 & 2{,}3167 \end{vmatrix}} = -\frac{-0{,}0714}{0{,}3826} = 0{,}1866$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2 = 3{,}0857 - (0{,}9360)\,(1{,}2571) - (0{,}1866)\,(7{,}0714) = 0{,}5895$$

Ecuación del plano de regresión: $\hat{Y} = 0{,}5895 + 0{,}936\,X_1 + 0{,}1866\,X_2$

5• Otra forma de enfocar la situación, desde la Tabla de operaciones, con las ecuaciones MCO:

$$\begin{cases} N\beta_0 + \beta_1 \sum_{i=1}^{7} X_{1i} + \beta_2 \sum_{i=1}^{7} X_{2i} = \sum_{i=1}^{7} Y_i \\ \beta_0 \sum_{i=1}^{7} X_{1i} + \beta_1 \sum_{i=1}^{7} X_{1i}^2 + \beta_2 \sum_{i=1}^{7} X_{1i} X_{2i} = \sum_{i=1}^{7} X_{1i} Y_i \\ \beta_0 \sum_{i=1}^{7} X_{2i} + \beta_1 \sum_{i=1}^{7} X_{1i} X_{2i} + \beta_2 \sum_{i=1}^{7} X_{2i}^2 = \sum_{i=1}^{7} X_{2i} Y_i \end{cases} \rightarrow \begin{cases} 7\beta_0 + 8{,}8\beta_1 + 49{,}5\beta_2 = 21{,}6 \\ 8{,}8\beta_0 + 13{,}48\beta_1 + 66{,}75\beta_2 = 30{,}26 \\ 49{,}5\beta_0 + 66{,}75\beta_1 + 366{,}25\beta_2 = 160 \end{cases}$$

en forma matricial,

$$\overbrace{\begin{pmatrix} 7 & 8{,}8 & 49{,}5 \\ 8{,}8 & 13{,}48 & 66{,}75 \\ 49{,}5 & 66{,}75 & 366{,}25 \end{pmatrix}}^{X'X} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 21{,}6 \\ 30{,}26 \\ 160 \end{pmatrix} \Rightarrow \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \overbrace{\begin{bmatrix} 3{,}6687 & 0{,}6181 & -0{,}6085 \\ 0{,}6181 & 0{,}8648 & -0{,}2412 \\ -0{,}6085 & -0{,}2412 & 0{,}1289 \end{bmatrix}}^{[X'X]^{-1}} \overbrace{\begin{pmatrix} 21{,}6 \\ 30{,}26 \\ 160 \end{pmatrix}}^{X'Y}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{bmatrix} 0{,}5895 \\ 0{,}9360 \\ 0{,}1866 \end{bmatrix} \Rightarrow Y = 0{,}5895 + 0{,}936\,X_1 + 0{,}1866\,X_2 \quad \text{Modelo regresión lineal}$$

6) Coeficiente de correlación múltiple:

A partir de la ecuación de regresión $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ se obtienen las predicciones y residuos asociados $u_i = Y_i - \hat{Y}_i$ a las observaciones muestrales.

De este modo, para la primera observación ($Y_1 = 1{,}3$, $X_{11} = 0{,}3$, $X_{21} = 4$), se tiene:

$$\begin{cases} \hat{Y}_1 = 0{,}5895 + 0{,}936\,(0{,}3) + 0{,}1866\,(4) = 1{,}6167 \\ u_1 = Y_1 - \hat{Y}_1 = 1{,}3 - 1{,}6167 = -0{,}3167 \end{cases}$$

En esta línea, considerando todos los puntos muestrales, se obtiene:

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Y_i	X_{1i}	X_{2i}	$\hat{Y}_i$	$u_i = Y_i - \hat{Y}_i$	$(Y_i - \overline{Y})^2$	$u_i^2 = (Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \overline{Y})^2$
1,3	0,3	4	1,6167	-0,3167	3,1887	0,1003	2,1580
3,5	1,5	9	3,6729	-0,1729	0,1716	0,0299	0,3448
2,8	0,7	6	2,3643	0,4357	0,0816	0,1898	0,5204
3	1,1	7,5	3,0186	-0,0186	0,0073	0,0003	0,0045
3,3	1,2	8	3,2055	0,0945	0,0459	0,0089	0,0144
4	2	7	3,7677	0,2323	0,8359	0,0540	0,4651
3,7	2	8	3,9543	-0,2543	0,3774	0,0647	0,7545
21,6	8,8	49,5			4,7086	0,4479	4,2616

$$SCT = \sum_{i=1}^{7}(Y_i - \overline{Y})^2 = 4,7086$$

$$SCR = \sum_{i=1}^{7}(Y_1 - \hat{Y}_1)^2 = 0,4479$$

$$SCE = \sum_{i=1}^{7}(\hat{Y}_i - \overline{Y})^2 = 4,2616$$

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{7}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{7}(Y_i - \overline{Y})^2} = \frac{4,2616}{4,7086} = 0,9051$$ coeficiente de determinación lineal múltiple

$$S_R^2 = \frac{SCR}{n-k-1} = \frac{0,4479}{4} = 0,112$$ varianza residual

Coeficiente de determinación lineal múltiple corregido por el número de grados de libertad:

$$\overline{R}^2 = 1 - \frac{\frac{SCR}{n-k-1}}{\frac{SCT}{n-1}} = 1 - \frac{0,4479/4}{4,7086/6} = 0,8573$$

7) El coeficiente de correlación parcial entre los Beneficios (Y) y el Gasto en Publicidad (X_1) se puede obtener mediante la expresión:

$$r_{yx_1 \cdot x_2} = -\frac{VC_{yx_1}}{\sqrt{VC_{yy}\,VC_{x_1x_1}}}$$ $VC_{ij} \equiv$ Adjuntos de la matriz de las varianzas-covarianzas

$$VC = \begin{bmatrix} \sigma_y^2 & S_{yx_1} & S_{yx_2} \\ S_{x_1y} & \sigma_{x_1}^2 & S_{x_1x_2} \\ S_{x_2y} & S_{x_2x_1} & \sigma_{x_2}^2 \end{bmatrix} = \begin{bmatrix} 0,6727 & 0,4438 & 1,0369 \\ 0,4438 & 0,3454 & 0,6462 \\ 1,0369 & 0,6462 & 2,3167 \end{bmatrix}$$

$$VC_{yx_1} = -\begin{vmatrix} 0,4438 & 0,6462 \\ 1,0369 & 2,3167 \end{vmatrix} = -0,3581$$

$$VC_{yy} = \begin{vmatrix} 0,3454 & 0,6462 \\ 0,6462 & 2,3167 \end{vmatrix} = 0,3826 \qquad VC_{x_1x_1} = \begin{vmatrix} 0,6727 & 1,0369 \\ 1,0369 & 2,3167 \end{vmatrix} = 0,4833$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....
Apellidos ..... Nombre .....
Ejercicio del día .....

Coeficiente de correlación parcial entre Y e X_1 :

$$r_{yx_1.x_2} = -\frac{VC_{yx_1}}{\sqrt{VC_{yy}\,VC_{x_1x_1}}} = \frac{0,3581}{\sqrt{(0,3826)(0,4833)}} = 0,833$$

7• El coeficiente de correlación parcial entre los Beneficios (Y) y el Gasto en Publicidad (X_1) se puede obtener mediante la expresión:

$$r_{yx_1.x_2} = \frac{r_{yx_1} - r_{yx_2}\,r_{x_1x_2}}{\sqrt{(1-r^2_{yx_2})(1-r^2_{x_1x_2})}}$$

$r_{ij} \equiv$ Elementos de la matriz de correlaciones simples R_x de la variable explicativa

$$R_x = \begin{bmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{x_1y} & 1 & r_{x_1x_2} \\ r_{x_2y} & r_{x_2x_1} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0,9207 & 0,8306 \\ 0,9207 & 1 & 0,7224 \\ 0,8306 & 0,7224 & 1 \end{bmatrix}$$, con lo cual,

$$r_{yx_1.x_2} = \frac{r_{yx_1} - r_{yx_2}\,r_{x_1x_2}}{\sqrt{(1-r^2_{yx_2})(1-r^2_{x_1x_2})}} = \frac{0,9207 - (0,8306)(0,7224)}{\sqrt{(1-0,8306^2)(1-0,7224^2)}} = 0,833$$

Análogamente, se tiene: $r_{yx_2.x_1} = \frac{r_{yx_2} - r_{yx_1}\,r_{x_2x_1}}{\sqrt{(1-r^2_{yx_1})(1-r^2_{x_2x_1})}} = \frac{0,8306 - (0,9207)(0,7224)}{\sqrt{(1-0,9207^2)(1-0,7224^2)}} = 0,613$

7•• Los coeficientes de correlación parcial se calculan también mediante la expresión:

$$t_i = \frac{\hat{\beta}_i}{\sqrt{S^2_R \cdot q_{i+1,\,i+1}}} \qquad R^2(Y,X_1\,;\,X_2) = \frac{t_i^2}{t_i^2 + n - (k+1)}$$ coeficiente de determinación parcial

se tenía, $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \underbrace{\overbrace{\begin{bmatrix} 3,6687 & 0,6181 & -0,6085 \\ 0,6181 & 0,8648 & -0,2412 \\ -0,6085 & -0,2412 & 0,1289 \end{bmatrix}}^{[X'X]^{-1}}}_{\text{elementos } q_{i+1,i+1}} \overbrace{\begin{pmatrix} 21,6 \\ 30,26 \\ 160 \end{pmatrix}}^{X'Y} = \begin{bmatrix} 0,5895 \\ 0,9360 \\ 0,1866 \end{bmatrix}$

$$t_{gastos_publi} = \frac{\hat{\beta}_1}{\sqrt{S^2_R \cdot q_{1+1,1+1}}} = \frac{0,936}{\sqrt{(0,112)\,(0,8648)}} = 3,0075$$

Coeficiente de determinación parcial:

$$r^2_{yx_1.x_2} = R^2(\text{Beneficios, Gastos_Publi}\,;\,\text{Horas extras}) = \frac{t^2_{gasto_publi}}{t^2_{gasto_publi} + n - (k+1)} = \frac{3,0075^2}{3,0075^2 + 4} = 0,6933$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Coeficiente de correlación parcial:

$$r_{yx_1.x_2} = R(\text{Beneficios, Gastos_Publi ; Horas extras}) = \sqrt{0,6933} = 0,833$$

Análogamente, $t_{horas_extras} = \dfrac{\hat{\beta}_2}{\sqrt{S_R^2 \cdot q_{2+1,2+1}}} = \dfrac{0,1866}{\sqrt{(0,112)\,(0,1289)}} = 1,5530$

Coeficiente de determinación parcial:

$$r^2_{yx_2.x_1} = R^2(\text{Beneficios, Horas extras ; Gastos_Publi}) = \frac{t^2_{horas_extras}}{t^2_{horas_extras} + n - (k+1)} = \frac{1,553^2}{1,533^2 + 4} = 0,37615$$

Coeficiente de correlación parcial:

$$r_{yx_2.x_1} = R(\text{Beneficios, Horas extras ; Gastos_Publi}) = \sqrt{0,37615} = 0,6133$$

8) El Coeficiente de determinación múltiple viene definido por:

$$r^2_{yx_1x_2} = 1 - \frac{|CV|}{\sigma_y^2\, C_{yy}} \qquad R^2 = r^2_{yx_1x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2\, r_{yx_1}\, r_{yx_2}\, r_{x_1x_2}}{1 - r^2_{x_1x_2}}$$

$$VC = \begin{bmatrix} 0,6727 & 0,4438 & 1,0369 \\ 0,4438 & 0,3454 & 0,6462 \\ 1,0369 & 0,6462 & 2,3167 \end{bmatrix} \qquad R_x = \begin{bmatrix} 1 & 0,9207 & 0,8306 \\ 0,9207 & 1 & 0,7224 \\ 0,8306 & 0,7224 & 1 \end{bmatrix}$$

a) $$|CV| = \begin{vmatrix} 0,6727 & 0,4438 & 1,0369 \\ 0,4438 & 0,3454 & 0,6462 \\ 1,0369 & 0,6462 & 2,3167 \end{vmatrix} = 0,0245 \qquad VC_{yy} = \begin{vmatrix} 0,3454 & 0,6462 \\ 0,6462 & 2,3167 \end{vmatrix} = 0,3826$$

$$r^2_{yx_1x_2} = 1 - \frac{|CV|}{\sigma_y^2\, C_{yy}} = 1 - \frac{|CV|}{\sigma_y^2\, C_{yy}} = 1 - \frac{0,0245}{(0,6727)(0,3826)} = 0,905$$

b) $$r^2_{yx_1x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2 r_{yx_1}\, r_{yx_2}\, r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{0,9207^2 + 0,8306^2 - 2\,.\,0,9207\,.\,0,8306\,.\,0,7224}{1 - 0,7224^2} = 0,905$$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

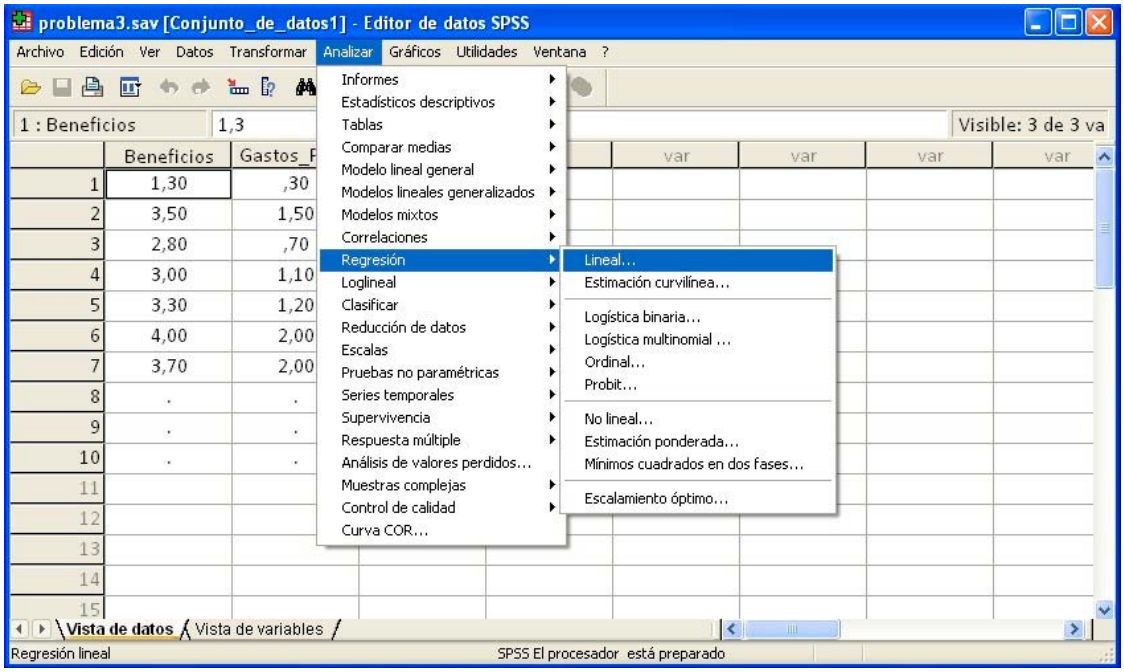
Universidad Autónoma de Madrid

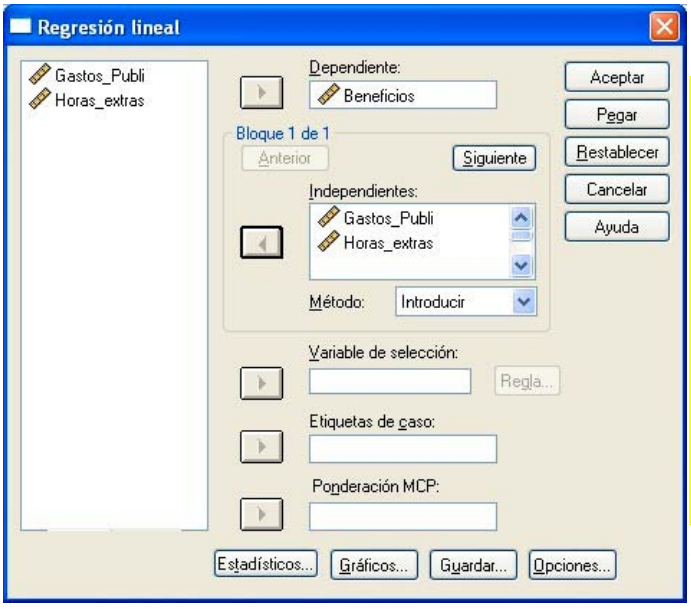
Asignatura ..... **Grupo** .....

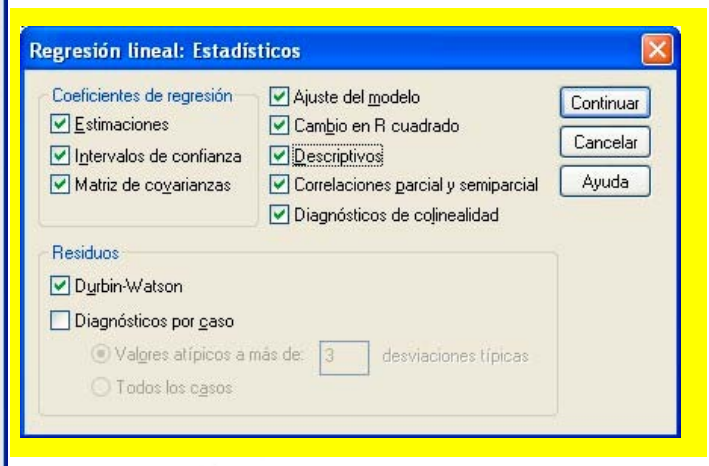
Apellidos ..... **Nombre** .....

Ejercicio del día .....

Guía Práctica en SPSS







UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Estimaciones

Ofrece las estimaciones de los coeficientes de regresión parcial no estandarizados (B) y estandarizados (Beta), junto con las pruebas de significación individuales para contrastar las hipótesis de que el valor poblacional de esos coeficientes es cero.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	,590	,641		,920	,410
	Gastos_Publi	,936	,311	,670	3,007	,040
	Horas_extras	,187	,120	,346	1,553	,195

a. Variable dependiente: Beneficios

En la columna encabezada por [Coeficientes no estandarizados] se encuentran los coeficientes β_i que forman parte de la ecuación en puntuaciones directas:

$$\text{Beneficios} = 0{,}590 + 0{,}936\ (\text{Gastos_Publicidad}) + 0{,}187\ (\text{Horas_extras})$$

Estos coeficientes no estandarizados se interpretan en los términos ya conocidos. Señalar que estos coeficientes no son independientes entre sí. De hecho, reciben el nombre de coeficientes en regresión parcial porque el valor concreto estimado para coeficiente se ajusta teniendo en cuenta la presencia del resto de variables independientes. Conviene, por tanto, interpretarlos con cautela.

El signo del coeficiente de regresión parcial de una variable puede no ser el mismo que el del coeficiente de correlación simple entre esa variable y la dependiente. Esto se produce a los ajustes que se llevan a cabo para obtener la mejor ecuación posible. Aunque existen diferentes explicaciones para justificar el cambio de signo de un coeficiente de regresión, una de las que deben de ser más seriamente consideradas es la que se refiere a la presencia de un alto grado de asociación entre alguna de las variables independientes (Colinealidad).

- Los coeficientes Beta están basados en las puntuaciones típicas y, por tanto, son directamente comparables entre sí. Indican la cantidad de cambio, en puntuaciones típicas, que se producirá en la variable dependiente por cada cambio de una unidad en la correspondiente variable independiente (manteniendo constantes el resto de variables independientes).

Estos coeficientes proporcionan una pista muy útil sobre la importancia relativa de cada variable independiente en la ecuación de regresión. En general, una variable tiene tanto más peso (importancia) en la ecuación de regresión cuanto mayor (en valor absoluto) es su coeficiente de regresión estandarizado.

Observando los coeficientes Beta del ejercicio, la variable Gastos_Publicidad es la más importante.

- Pruebas de significación: Las pruebas t y sus niveles críticos (últimas dos columnas de la tabla) sirven para contrastar la hipótesis nula de que un coeficiente de regresión vale 0 en la población. Niveles críticos (Sig) muy pequeños (generalmente menores que 0,05) indican que se debe rechazar la hipótesis nula.

Un coeficiente de cero indica ausencia de relación lineal, de modo que los coeficientes significativamente distintos de cero informan sobre qué variables son relevantes en la ecuación de regresión.

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Observando el nivel crítico asociado a cada prueba t, las dos variables utilizadas (Gastos_Publicidad, Horas_extras) tienen coeficientes significativamente distintos de cero (en todas, Sig < 0,05).
Por tanto, las dos variables independientes contribuyen significativamente a explicar lo que ocurre con la variable dependiente (Beneficios).

Ajuste del modelo

Muestra el coeficiente de correlación múltiple, su cuadrado corregido y no corregido, y el error típico de los residuos. También incluye la tabla resumen de ANOVA, que contiene al estadístico F de Fisher-Snedecor para contrastar la hipótesis nula de que el coeficiente de correlación múltiple $R_{yx_1x_2} = 0$

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,951[a]	,905	,857	,33464

a. Variables predictoras: (Constante), Horas_extras, Gastos_ Publi

$R^2 = 0,905$ coeficiente determinación múltiple

$\overline{R}^2 = 0,857$ coeficiente determinación múltiple corregido

El estadístico F contrasta la hipótesis nula

$H_0 : R_{yx_1x_2} = 0$

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4,261	2	2,130	19,023	,009[a]
	Residual	,448	4	,112		
	Total	4,709	6			

a. Variables predictoras: (Constante), Horas_extras, Gastos_Publi

b. Variable dependiente: Beneficios

En consecuencia, permite decidir si existe relación lineal significativa entre la variable dependiente y el conjunto de variables independientes tomadas juntas.

El valor de un nivel crítico (Sig < 0,05) indica que existe relación lineal significativa, pudiendo afirmar que el hiperplano definido por la ecuación de regresión ofrece un buen ajuste a la nube de puntos.

$$\begin{cases} SCE = \sum_{i=1}^{7}(\hat{Y}_i - \overline{Y})^2 = 4,261 \quad gl = k = 2 \\ SCR = \sum_{i=1}^{7}(Y_1 - \hat{Y}_1)^2 = 0,448 \quad gl = n-k-1 = 7-2-1 = 4 \\ SCT = \sum_{i=1}^{7}(Y_i - \overline{Y})^2 = 4,709 \quad gl = n-1 = 7-1 = 6 \end{cases}$$

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{7}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{7}(Y_i - \overline{Y})^2} = \frac{4,261}{4,709} = 0,905 \quad \text{coeficiente de determinación múltiple}$$

$$S_R^2 = \frac{SCR}{n-k-1} = \frac{0,448}{4} = 0,112 \quad \text{varianza residual}$$

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

$\bar{R}^2 = 1 - \frac{SCR/(n-k-1)}{SCT/(n-1)} = 1 - \frac{0,448/4}{4,709/6} = 0,857$ coeficiente de determinación múltiple corregido

$F = \frac{SCE/k}{SCR/(n-k-1)} = \frac{4,261/2}{0,448/4} = 19,023$ estadístico observado F de Fisher-Snedecor

Relación entre R^2 y F: $F = \left[\frac{n-k-1}{k}\right]\left[\frac{R^2}{1-R^2}\right] = 2.\frac{0,905}{(1-0,905)} = 19,023$

Se rechaza la hipótesis nula $H_0: \beta_1 = \beta_2 = 0$ cuando $F \geq F_{\alpha;\,k,\,(n-k-1)}$

Intervalos de confianza

Situados en la tabla (Coeficientes de regresión), permitiendo que además de obtener una estimación puntual de los coeficientes de regresión parcial, se pueda obtener el intervalo de confianza para estos coeficientes.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	,590	,641		,920	,410	-1,190	2,369
	Gastos_Publi	,936	,311	,670	3,007	,040	,072	1,800
	Horas_extras	,187	,120	,346	1,553	,195	-,147	,520

a. Variable dependiente: Beneficios

Estos intervalos informan sobre los límites en que se encuentra el valor poblacional de cada coeficiente. Los límites se obtienen sumando y restando 1,96 (SPSS trabaja por defecto con un nivel de significación 0,95) errores típicos al valor del correspondiente coeficiente de regresión.

Una amplitud grande en los intervalos de confianza indica que las estimaciones obtenidas son poco precisas y, probablemente, inestables (coas que puede ocurrir, por ejemplo, cuando existen problemas de colinealidad).

Matriz de covarianzas

Muestra una matriz con las covarianzas y correlaciones existentes entre los coeficientes de regresión parcial.

Correlaciones de los coeficientes[a]

Modelo			Horas_extras	Gastos_Publi
1	Correlaciones	Horas_extras	1,000	-,722
		Gastos_Publi	-,722	1,000
	Covarianzas	Horas_extras	,014	-,027
		Gastos_Publi	-,027	,097

a. Variable dependiente: Beneficios

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

Descriptivos

Ofrece la media y la desviación típica de cada variable y el número de casos utilizados en el análisis.

Además, ofrece la matriz de correlaciones entre el conjunto de variables utilizadas en el análisis.
En la matriz de correlaciones, cada coeficiente de correlación aparece acompañado de su correspondiente nivel crítico (que permite decidir sobre la hipótesis de que el coeficiente de correlación vale 0 en la población) y del número de casos sobre el que se ha calculado cada coeficiente.

En la diagonal de la matriz de correlaciones aparecen unos, pues la relación entre una variable y ella misma es perfecta.

Estadísticos descriptivos

	Media	Desviación típ.	N
Beneficios	3,0857	,88587	7
Gastos_Publi	1,2571	,63471	7
Horas_extras	7,0714	1,64389	7

Correlaciones

		Beneficios	Gastos_Publi	Horas_extras
Correlación de Pearson	Beneficios	1,000	,921	,831
	Gastos_Publi	,921	1,000	,722
	Horas_extras	,831	,722	1,000
Sig. (unilateral)	Beneficios	.	,002	,010
	Gastos_Publi	,002	.	,033
	Horas_extras	,010	,033	.
N	Beneficios	7	7	7
	Gastos_Publi	7	7	7
	Horas_extras	7	7	7

Correlaciones parcial y semiparcial

Esta opción permite obtener los coeficientes de correlación parcial y semiparcial entre la variable dependiente y cada variable independiente.

Un coeficiente de correlación parcial expresa el grado de relación existente entre dos variables tras eliminar de ambas el efecto debido a terceras variables. Es decir, los coeficientes de correlación parcial expresan el grado de relación existente entre cada variable independiente y la variable dependiente tras eliminar de ambas el efecto debido al resto de variables independientes incluidas en la ecuación.

Un coeficiente de correlación semiparcial expresa el grado de relación existente entre dos variables tras eliminar de una de ellas el efecto debido a terceras variables. Es decir, estos coeficientes expresan el grado de relación existente entre la variable dependiente y la parte de cada variable independiente que no está explicada por el resto de variables independientes.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones		
		B	Error típ.	Beta			Orden cero	Parcial	Semiparcial
1	(Constante)	,590	,641		,920	,410			
	Gastos_Publi	,936	,311	,670	3,007	,040	,921	,833	,464
	Horas_extras	,187	,120	,346	1,553	,195	,831	,613	,240

a. Variable dependiente: Beneficios

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Con los coeficientes de correlación parcial y semiparcial, aparecen las correlaciones de orden cero, es decir, los coeficientes de correlación calculados sin tener en cuenta la presencia de terceras variables (se trata de los mismos coeficientes que aparecen en la tabla anterior de correlaciones).

Comparando entre sí estos coeficientes (de orden cero, parcial y semiparcial), pueden encontrarse pautas de relación interesantes: En los datos de la tabla se observa, por ejemplo:

- La relación entre la variable dependiente Beneficios y la variable independiente Gastos_Publicitarios vale 0,921.
- Al eliminar de las variables (Beneficios, Gastos_Publicitarios) el efecto atribuible a las Horas_extras, la relación baja a 0,833 (parcial).
- Cuando el efecto atribuible a Horas_extras se elimina sólo de la variable Beneficios, la relación baja a 0,464 (semiparcial).

Análisis que indica que la relación entre las variables (Beneficios, Horas_extras) tiene mucho menor peso en la relación.

Colinealidad

Existe una colinealidad perfecta cuando una de las variables independientes se relaciona de forma perfectamente lineal con una o más del resto de las variables independientes de la ecuación.

Se dice que existe una colinealidad parcial, o simplemente, colinealidad, cuando entre las variables independientes de una ecuación existen correlaciones altas.

La colinealidad es un problema, porque en el caso de colinealidad perfecta, no es posible estimar los coeficientes de la ecuación de regresión; y en el caso de colinealidad parcial, aumenta el tamaño de los residuos tipificados y esto produce coeficientes de regresión muy inestables (pequeños cambios en los datos, como quitar o añadir un caso, produce cambios muy grandes en los coeficientes de regresión). Esta es una de las razones de encontrarse con coeficientes con signo cambiado: < correlaciones positivas pueden transformarse en coeficientes de regresión negativos (incluso significativamente negativos) >.

Curiosamente, la medida de ajuste R^2 no se altera por la presencia de colinealidad, pero los efectos atribuidos a las variables independientes pueden ser engañosos.

Al evaluar la existencia o no de colinealidad, la dificultad estriba en determinar cuál es el grado máximo de relación permisible entre las variables independientes. Sobre esta cuestión no existe un consenso generalizado, aunque puede servir de guía la presencia de indicios que se pueden encontrar en los resultados de un análisis de regresión (aunque estos indicios pueden tener su origen en otras causas):

- El estadístico F que evalúa el ajuste general de la ecuación de regresión es significativo, pero no lo es ninguno de los coeficientes de regresión parcial.
- Los coeficientes de regresión parcial estandarizados (coeficientes Beta) están inflados tanto en positivo como en negativo (al mismo tiempo, adoptan valores mayores que 1 y menores que –1)
- Existen valores de tolerancia pequeños (próximos a 0,01). La tolerancia de una variable independiente es la proporción de varianza de esa variable que no está asociada (que no depende) del resto de variables independientes incluidas en la ecuación. Por ejemplo, una variable con una tolerancia de 0,01 es una variable que comparte el 99% de su varianza con el resto de variables independientes, lo que significa que se trata de una variable redundante casi por completo.
- Los coeficientes de correlación estimados son muy grandes (por encima de 0,90 en valor absoluto).

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

SPSS ofrece la posibilidad de obtener algunos estadísticos que pueden ayudar a diagnosticar la presencia de colinealidad. Se trata de estadísticos orientativos que, aunque pueden servir de ayuda para determinar si existe mayor o menor grado de colinealidad, no permiten tomar una decisión clara sobre la presencia o no de colinealidad.

Los estadísticos de colinealidad se recogen en la tabla de coeficientes de regresión parcial ya analizada anteriormente, pero ahora contienen información adicional sobre los niveles de tolerancia y sus inversos (FIV).

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	,590	,641		,920	,410		
	Gastos_Publi	,936	,311	,670	3,007	,040	,478	2,090
	Horas_extras	,187	,120	,346	1,553	,195	,478	2,090

a. Variable dependiente: Beneficios

El nivel de tolerancia de una variable se obtiene restando a 1 el coeficiente de determinación múltiple $(1-R^2)$ que resulta al regresar esa variable sobre el resto de variables independientes.

Valores de tolerancia muy pequeños indican que esa variable puede ser explicada por una combinación lineal del resto de variables, lo que significa que existe colinealidad.

Los factores de inflación de la varianza (FIV) son los inversos de los niveles de tolerancia. Reciben este nombre porque son utilizados en el cálculo de las varianzas de los coeficientes de regresión.

Cuanto mayor es el FIV de una variable, mayor es la varianza del correspondiente coeficiente de regresión. De ahí, que uno de los problemas de la presencia de colinealidad (tolerancias pequeñas, FIVs grandes) sea la inestabilidad de las estimaciones de los coeficientes de regresión.

La siguiente tabla del Visor de SPSS muestra la solución resultante de aplicar un análisis de componentes principales a la matriz estandarizada no centrada de productos cruzados de las variables independientes:

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza		
				(Constante)	Gastos_Publi	Horas_extras
1	1	2,889	1,000	,00	,01	,00
	2	,097	5,453	,15	,53	,01
	3	,014	14,614	,85	,46	,99

a. Variable dependiente: Beneficios

Los autovalores informan sobre cuántas dimensiones o factores diferentes subyacen en el conjunto de variables independientes utilizadas.

La presencia de varios autovalores próximos a cero indica que las variables independientes están muy relacionadas entre sí (colinealidad). En este caso, no existe el problema.

Los Índices de condición son la raíz cuadrada del cociente entre el autovalor más grande (2,889) y cada uno del resto de los autovalores (por ejemplo, $\sqrt{2,889 / 0,097} = 5,453$).

En condiciones de no-colinealidad, estos índices no deben superar el valor de 15. Índices mayores que 15 indican un posible problema, índices mayores que 30 informan de un serio problema de colinealidad.

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

La Colinealidad es un problema cuando una dimensión o factor con un Índice de condición alto, contribuye a explicar gran parte de la varianza de los coeficientes de dos o más variables.

Cuando en un conjunto de datos se detecta la presencia de colinealidad, hay que aplicar algún tipo de actuación:

a) Aumentar el tamaño de la muestra (es útil cuando existen pocos casos en relación con el número de variables).

b) Crear indicadores múltiples combinando variables (promediando variables, efectuando un análisis de componentes principales para reducir las variables a un conjunto de componentes independientes y aplicar después el análisis de regresión sobre esos componentes.

c) Excluir variables redundantes (variables que correlacionan muy alto con otras), quedando con las que se consideran más importantes.

(d) Utilizar una técnica de estimación sesgada, como la regresión ridge.

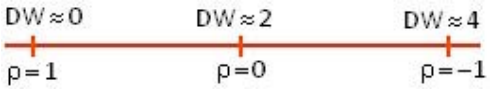
Residuos: Durbin-Watson

Esta opción permite obtener los coeficientes de correlación parcial y semiparcial entre la variable.

El análisis de los residuos proporciona información crucial sobre el cumplimiento de varios supuestos del modelo de regresión lineal: independencia, homocedasticidad, normalidad y linealidad.

El estadístico de Durbin-Watson (1951) proporciona información sobre el grado de independencia existente entre ellos:

$$DW = \frac{\sum_{i=2}^{n}(u_i - u_{i-1})^2}{\sum_{i=1}^{n}u_i^2} \approx 2(1-\rho) \rightarrow \begin{cases} DW \approx 0 & si \quad \rho = 1 \\ DW \approx 2 & si \quad \rho = 0 \\ DW \approx 4 & si \quad \rho = -1 \end{cases}$$



El estadístico de Durbin-Watson oscila entre 0 y 4, toma el valor 2 cuando los residuos son independientes. Los valores menores que 2 indican autocorrelación positiva y los mayores que 2 autocorrelación negativa. Se puede asumir independencia entre los residuos cuando $1,5 \leq DW \leq 2,5$

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,951[a]	,905	,857	,33464	1,933

a. Variables predictoras: (Constante), Horas_extras, Gastos_Publi

b. Variable dependiente: Beneficios

DW = 1,933, valor que se encuentra entre 1,5 y 2,5, se puede asumir que los residuos son independientes.

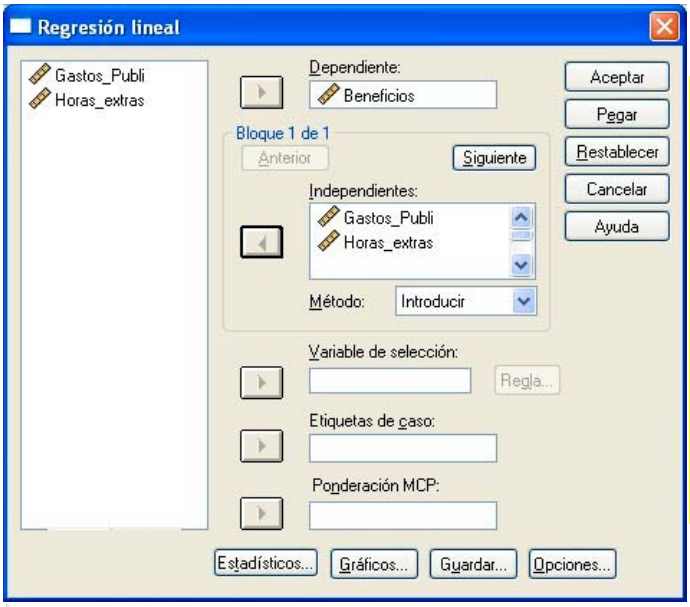
UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

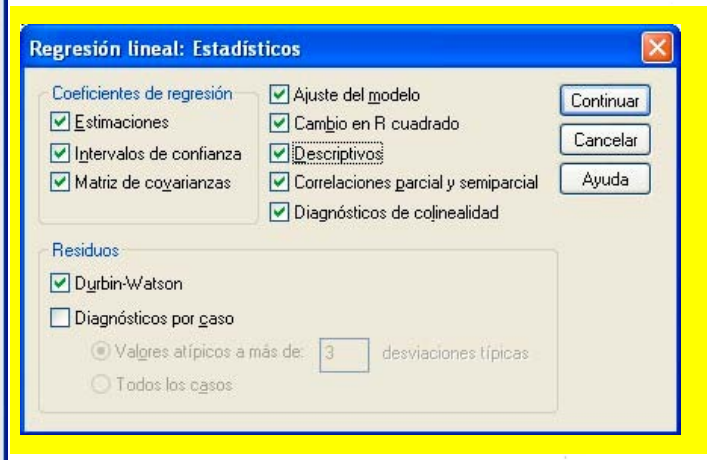
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....





Diagnósticos por caso

Valores atípicos a más de ... Con esta opción, SPSS indica los valores que producen un error grande, concretamente a más de n veces la desviación típica de la variable residuos.

En este caso, si se introducen 2 o 3 desviaciones típicas no se obtiene ningún valor atípico (pudiera ocurrir que al poner 1,5 desviaciones típicas, sí existiera).

La forma de proceder es seleccionar [Cambio en R^2] y [Valores atípicos a más de ...], el Modelo indica el número de pasos dados para construir el modelo de regresión (pasos que sean). También indica si en alguno de los pasos se ha eliminado alguna variable previamente seleccionada; en el ejemplo que ocupa no se elimina ninguna variable.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,951[a]	,905	,857	,33464	,905	19,023	2	4	,009

a. Variables predictoras: (Constante), Horas_extras, Gastos_Publi

b. Variable dependiente: Beneficios

La tabla recoge el valor de R^2 en cada paso, el cambio experimentado por R^2 en cada paso, y el estadístico F y su significación. El estadístico F permite contrastar la hipótesis de que el cambio en R^2 vale cero en la población.

La tabla recoge el valor de R^2 en cada paso, el cambio experimentado por R^2 en cada paso, y el estadístico F y su significación. El estadístico F permite contrastar la hipótesis de que el cambio en R^2 vale cero en la población.

Al seleccionar la primera variable (Modelo 1), el valor de R^2 es 0,905.

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

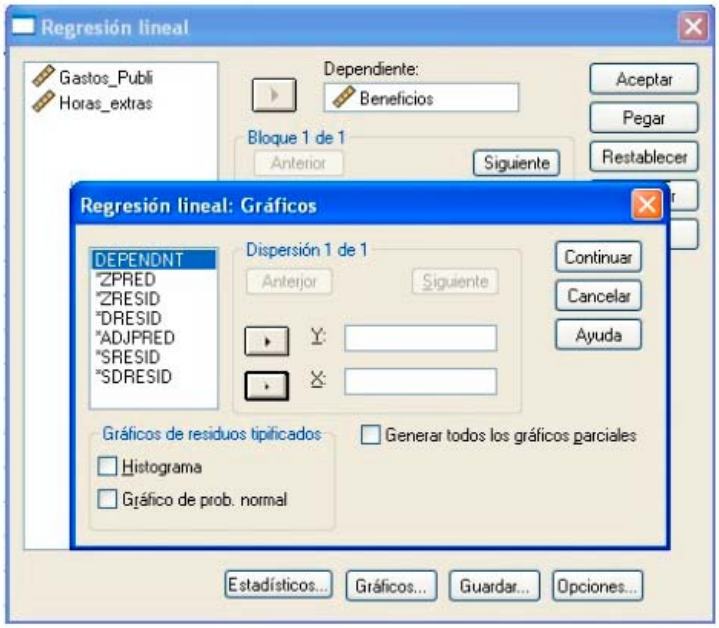
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

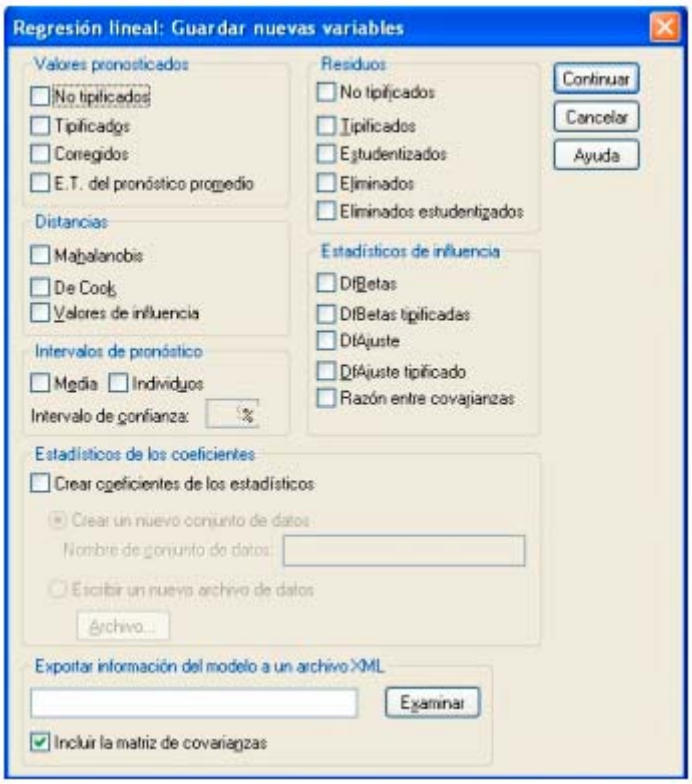
Lógicamente, en el primer paso, $R^2_{cambio} = R^2$. Al contrastar la hipótesis de que el valor poblacional de $R^2_{cambio} = 0$ se obtiene un estadístico F de 19,023 que, con 2 y 4 grados de libertad, tiene una probabilidad asociada p-valor = 0,009 < 0,05 = α 0,009 con lo que se rechaza la hipótesis nula, pudiendo afirmar que la proporción de varianza explicada por la variable Gastos_Publicidad (variable seleccionada en el primer paso) es significativamente distinta de cero.



Las variables que aparecen permiten obtener distintos gráficos de dispersión.

Las variables precedidas por un asterisco son variables creadas por SPSS.

Todas las variables pueden crearse en la opción [Guardar] marcando las opciones pertinentes del recuadro



DEPENDNT: Variable dependiente de la ecuación de regresión.

ZPRED (pronósticos tipificados): Pronósticos divididos por su desviación típica. Son pronósticos transformados en puntuaciones z (media cero y desviación típica 1)

ZRESID (residuos tipificados): Residuos divididos por su desviación típica. El tamaño de cada residuo tipificado indica el número de desviaciones típicas que se aleja de su media, de modo que, si están normalmente distribuidos (cosa que se asume en el análisis de regresión). El 95% de estos residuos se encontrará en el rango [-1,96, 1,96], lo que permite identificar fácilmente casos con residuos grandes.

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

DRESID (residuos eliminados o corregidos): Residuos obtenidos al efectuar los pronósticos eliminando de la ecuación de regresión el caso sobre el que se efectúa el pronóstico. El residuo correspondiente a cada caso se obtiene a partir del pronóstico efectuado con una ecuación de regresión en la que no se ha incluido ese caso. Son muy útiles para detectar puntos de influencia (casos con gran peso en la ecuación de regresión).

ADJPRED (pronósticos corregidos): Pronósticos efectuados con una ecuación de regresión en la que no se incluye el caso pronosticado (ver residuos eliminados o corregidos). Diferencias importantes entre PRED y ADJPRED delatan la presencia de puntos de influencia (casos con gran peso en la ecuación de regresión).

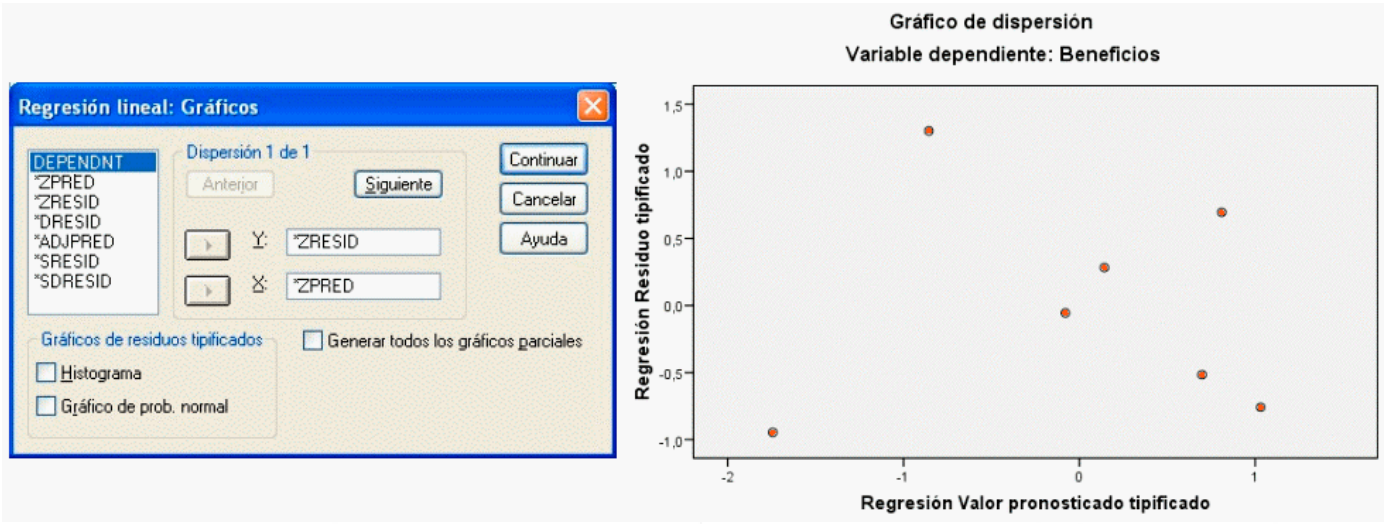
SRESID (residuos estudentizados): Residuos divididos por su desviación típica, basada ésta en cómo de próximo se encuentra un caso a su(s) medias(s) en la(s) variable(s) independiente(s). Al igual que ocurre en los residuos estandarizados (a los que se parecen mucho), los estudentizados están escalados en unidades de desviación típica. Se distribuyen según el modelo de probabilidad t-Student con (n – p - 1) grados de libertad (p se refiere al número de variables independientes). Con muestras grandes, aproximadamente el 95% de estos residuos debería encontrarse en el rango [-2, 2].

SDRESID (residuos corregidos estudentizados): Residuos corregidos divididos por su desviación típica. Útiles también para detectar puntos de influencia.

Algunas de estas variables permiten detectar puntos de influencia, pero, entre todas, hay dos variables (ZPRED, ZRESID) cuyo diagrama de dispersión informa sobre el supuesto de homocedasticidad o igualdad de varianzas.

El supuesto de igualdad de varianzas implica que la variación de los residuos debe de ser uniforme en todo el rango de valores pronosticados. O, lo que es lo mismo, que el tamaño de los residuos es independiente del tamaño de los pronósticos, de donde se desprende que el diagrama de dispersión no debe mostrar ninguna pauta de asociación entre los residuos y los pronósticos.

Para obtener un diagrama de dispersión con las variables (ZPRED, ZRESID):



En el diagrama de dispersión se observa que aunque los residuos y los pronósticos parecen ser independientes (la nube de puntos no sigue ninguna pauta de asociación clara, ni lineal ni de ningún otro tipo), no está claro que las varianzas sean homogéneas. Más bien, parece que a medida que van aumentando el valor de los pronósticos va disminuyendo la dispersión de los residuos.

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Los pronósticos menores que la media (con puntuación típica por debajo de cero) están más concentrados que los pronósticos mayores que la media (con puntuación típica mayor que cero).

Cuando un diagrama de dispersión delata la presencia de varianzas heterogéneas, puede utilizarse una transformación de la variable dependiente para resolver el problema (tal como una transformación logarítmica o una transformación raíz cuadrada). No obstante, al utilizar una transformación de la variable dependiente, debe cuidarse el problema de interpretación que añade el cambio de escala.

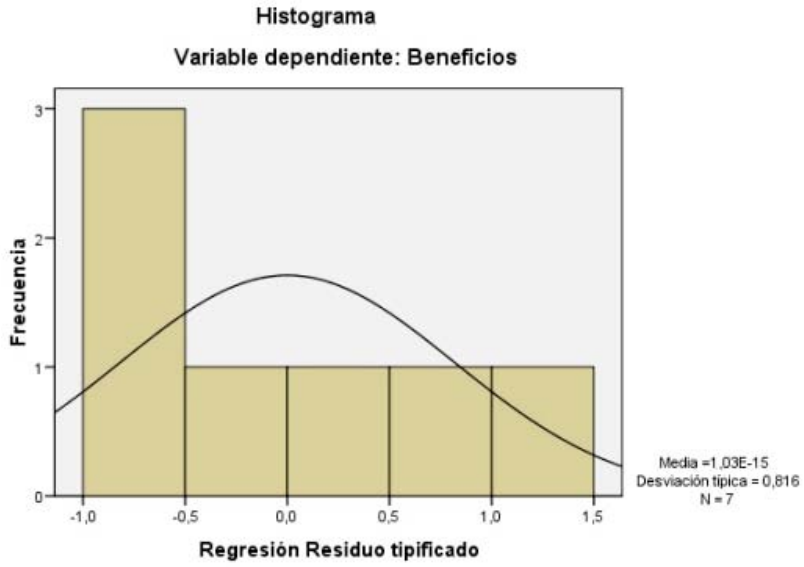
El diagrama de dispersión de las variables (ZPRED, ZRESID) posee la utilidad adicional de permitir detectar relaciones de tipo no lineal entre las variables. Cuando la relación es no lineal, el diagrama puede contener indicios sobre otro tipo de función de ajuste (los residuos estandarizados podrían en lugar de estar homogéneamente dispersos seguir un trazado curvilíneo).

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

Normal

El recuadro de Gráficos de los residuos tipificados contiene dos opciones que informan sobre el grado en que los residuos tipificados se aproximan a una distribución normal: Histograma y Gráfico de probabilidad normal.

HISTOGRAMA: Ofrece un histograma de los residuos tipificados con una curva normal superpuesta. La curva se construye tomando una media de cero y una desviación típica de uno. Es decir, la misma media y la misma desviación típica que los residuos típicos tipificados.



En el histograma se observa que la parte central acumula más casos de los que existen en una curva normal.

La distribución es algo asimétrica a la derecha.

La distribución de los residuos no parece seguir el modelo de probabilidad normal, de modo que los resultados del análisis deben de interpretarse con cautela.

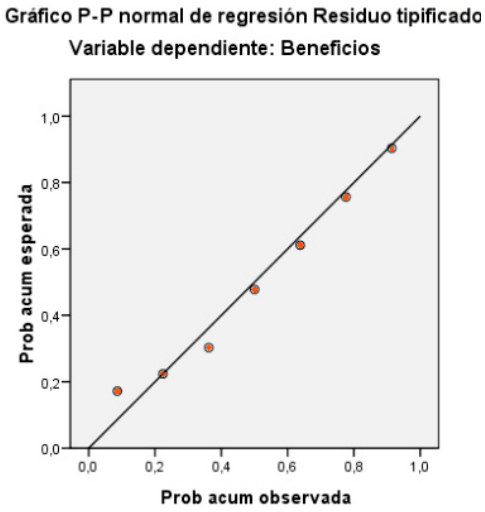
GRÁFICOS DE LOS RESIDUOS TIPIFICADOS. - Permite obtener un diagrama de probabilidad normal. En el eje de abscisas esta representada la probabilidad acumulada que corresponde a cada residuo tipificado. El de ordenadas representa la probabilidad acumulada teórica que corresponde a cada desviación típica en una curva normal N(0, 1).

Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

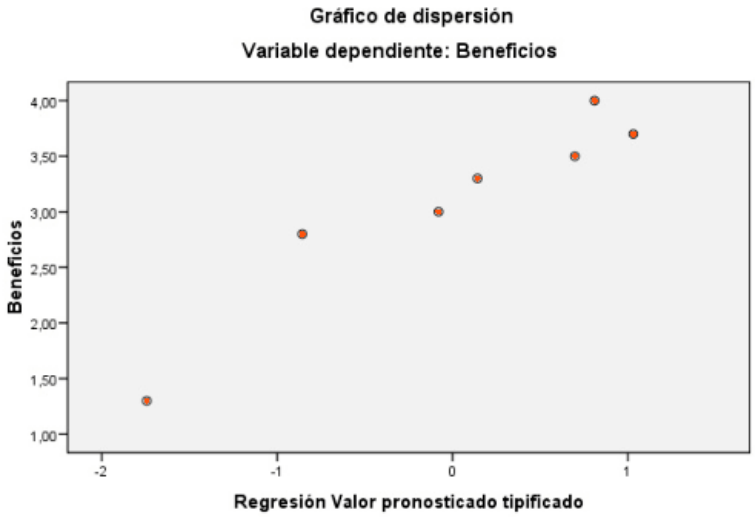
Ejercicio del día .....



Los puntos no se encuentran alineados sobre la diagonal del gráfico, indicando el posible incumplimiento del supuesto de normalidad.

En el Gráfico de valores observados frente a los predichos (DEPENDNT, ZPRED), los valores se deben alinear en la diagonal del cuadrante, si hubiera mucha dispersión, implicaría que no se verifican las hipótesis de homocedasticidad.

En este caso existe igualdad de varianzas



Seleccionando la opción Generar todos los gráficos parciales, SPSS muestra la gráfica de la variable dependiente frente a todas las variables independientes, comprobando si existe linealidad entre las variables.

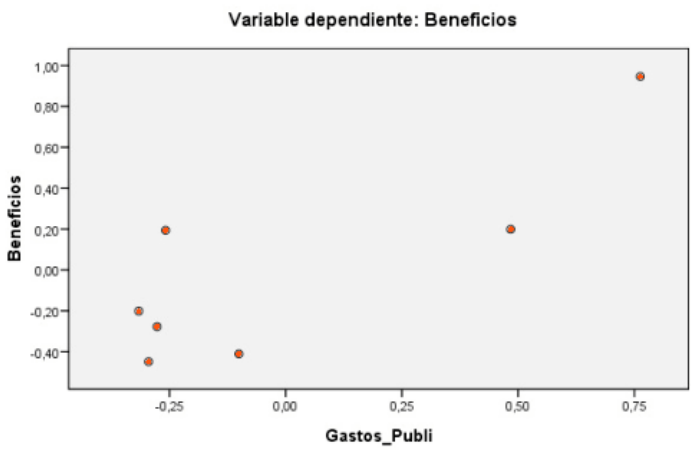
UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

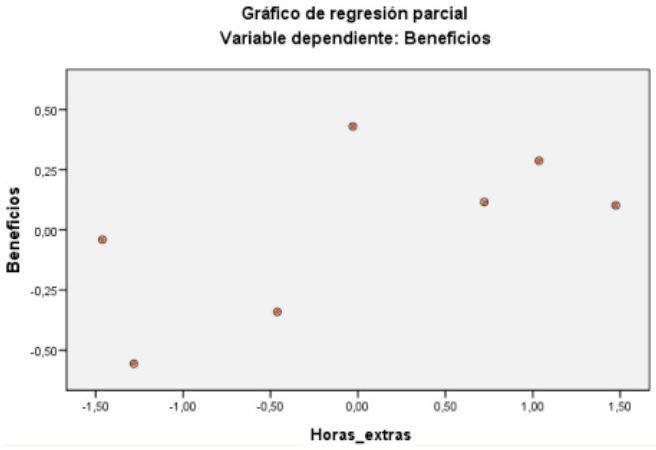
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

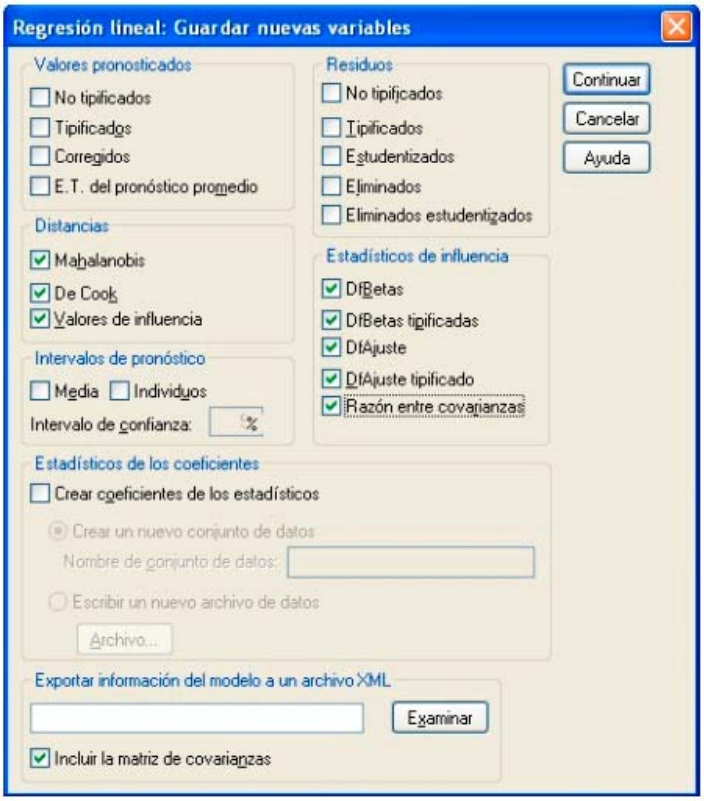
Apellidos ..... Nombre .....

Ejercicio del día .....





Observando los gráficos, se podría imaginar un comportamiento lineal.



Pulsando el botón [Guardar] se abre un abanico de opciones.

Todos los casos contribuyen a la obtención de la recta de regresión, pero no todos lo hacen con la misma fuerza. Los puntos de influencia son casos que afectan de forma importante al valor de la ecuación de regresión.

La presencia de puntos de influencia no tiene por qué constituir un problema en regresión, de hecho lo normal es que en un análisis de regresión no todos los casos tengan la misma importancia (desde el punto de vista estadístico). No obstante, el analista debe de ser consciente de tales puntos, porque, entre otras cosas, podría tratarse de casos con valores erróneos. Siendo conscientes de si existen o no puntos de influencia es posible corregir el análisis

Se marcan todas las opciones de los recuadros Distancias y Estadísticos de influencia (todas estas opciones crean variables nuevas en el archivo de datos).

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Distancias

Este recuadro recoge tres medidas que expresan el grado en que cada caso se aleja de los demás.

⇨ Mahalanobis.- Mide el grado de distanciamiento de cada caso respecto de los promedios del conjunto de variables independientes. En regresión simple, esta distancia se obtiene elevando al cuadrado la puntuación típica de cada caso en la variable independiente.

En regresión múltiple se obtiene multiplicando por (n – 1) el valor de influencia de cada caso.

⇨ Cook.- Mide el cambio que se produce en las estimaciones de los coeficientes de regresión al ir eliminando cada caso de la ecuación de regresión. Una distancia de Cook grande indica que ese caso tiene un peso considerable en la estimación de los coeficientes de regresión.

Para evaluar estas distancias puede utilizarse la distribución F con (p + 1) y (n - p - 1) grados de libertad, donde p es el número de variables independientes y n el tamaño de la muestra.

En general, un caso con una distancia de Cook superior a 1 debe de ser revisado.

⇨ Valores de influencia.- Representan una medida de la influencia potencial de cada caso. Respecto a las variables independientes, un valor de influencia es una medida normalizada del grado de distanciamiento de un punto del centro de su distribución. Los puntos muy alejados pueden influir de forma muy importante en la ecuación de regresión, pero no tienen por qué hacerlo necesariamente.

Con más de 6 variables y al menos 20 casos, se considera que un valor de influencia debe de ser revisados si es mayor que (3p / n). Los valores de influencia tienen un máximo de (n - 1)/n.

Como regla general, para orientar decisiones, los valores menores que 0,2 se consideran poco problemáticos; los valores comprendidos entre 0,2 y 0,5 se consideran arriesgados; y los valores mayores que 0,5 debieran evitarse.

Estadísticos de influencia

Este recuadro contiene varios estadísticos que contribuyen a precisar la posible presencia de puntos de influencia.

⇨ DfBetas (diferencia en las betas).- Mide el cambio que se produce en los coeficientes de regresión estandarizados (betas) como consecuencia de ir eliminando cada caso de la ecuación de regresión. SPSS crea en el Editor de datos tantas variables nuevas como coeficientes beta tiene la ecuación de regresión, es decir, tantos como variables independientes más uno (el correspondiente a la constante de la ecuación).

⇨ DfBetas tipificadas.- Es el cociente entre DfBetas y su error típico. Generalmente, un valor mayor que $2/\sqrt{n}$ delata la presencia de un posible punto de influencia. El SPSS crea en el Editor de datos tantas variables nuevas como coeficientes Beta tiene la ecuación de regresión.

⇨ Df Ajuste (diferencia en el ajuste).- Mide el cambio que se produce en el pronóstico de un caso cuando ese caso es eliminado de la ecuación de regresión.

⇨ Df Ajuste tipificado.- Es el cociente entre DfAjuste y su error típico. Generalmente, se consideran puntos de influencia los casos en los que DfAjuste tipificado es mayor que $2/\sqrt{(p/n)}$, siendo p el número de variables independientes y n el tamaño de la muestra.

⇨ Razón entre las covarianzas (RV).- Indica en qué medida la matriz de productos cruzados (base del análisis de regresión) cambia con la eliminación de cada caso. Se considera que un caso es un punto de influencia si $|RV| > (3+p)/n$

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Crear coeficientes de los estadísticos

SPSS ofrece una tabla resumen que incluye, para todos los estadísticos del recuadro Distancias, el valor mínimo, el máximo la media, la desviación típica y el número de casos. La tabla también recoge información sobre los pronósticos y los residuos

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	1,6168	3,9541	3,0857	,84268	7
Valor pronosticado tip.	-1,743	1,031	,000	1,000	7
Error típico del valor pronosticado	,157	,287	,214	,049	7
Valor pronosticado corregido	2,2021	4,1226	3,1659	,67982	7
Residuo bruto	-,31682	,43560	,00000	,27323	7
Residuo tip.	-,947	1,302	,000	,816	7
Residuo estud.	-1,834	1,525	-,078	1,195	7
Residuo eliminado	-1,18852	,65738	-,08016	,63921	7
Residuo eliminado estud.	-3,978	2,041	-,295	1,931	7
Dist. de Mahalanobis	,465	3,543	1,714	1,177	7
Distancia de Cook	,000	3,084	,652	1,109	7
Valor de influencia centrado	,078	,591	,286	,196	7

a. Variable dependiente: Beneficios

Señalar que los puntos de influencia no tienen por qué tener residuos especialmente grandes, el problema que presentan no es precisamente

A pesar de ello, conviene examinarlos por su desproporcionada influencia sobre la ecuación de regresión.

Como éstos puntos son distintos de los demás, conviene precisar en qué son distintos.

Una vez identificados y examinados, se pueden eliminar del análisis simplemente porque entorpecen el ajuste, o porque su presencia produce medidas de ajuste infladas.

Valores pronosticados

El objetivo principal del análisis es el de poder efectuar pronósticos en casos nuevos. Se han utilizado los coeficientes de regresión parcial (β) para construir la recta de regresión:

$$Y = 0{,}5895 + 0{,}936\ X_1 + 0{,}1866\ X_2$$

Conocidos los pesos de la ecuación de regresión, se puede utilizar la opción del menú Transformar/Calcular variable para obtener los pronósticos que la ecuación asigna a cada caso. Pero esto no es necesario porque el subcuadro Guardar nuevas variables contiene opciones relacionadas con los pronósticos.

Las opciones de este recuadro generan, en el Editor de datos, cuatro nuevas variables, que reciben automáticamente un nombre seguido de un número de serie (nombre_#). Por ejemplo, la primera vez que se solicitan durante una sesión los pronósticos tipificados, la nueva variable con los pronósticos tipificados recibe el nombre de 'zpr_'. Si se vuelven a solicitar pronósticos tipificados durante la misma sesión, la nueva variable recibe el nombre de 'zpr_2', y así sucesivamente.

No tipificados.- Pronósticos que se derivan de la ecuación de regresión en puntuaciones directas, reciben el nombre: pre_#.

Tipificados.- Pronósticos convertidos en puntuaciones típicas (restando a cada pronóstico la media de los pronósticos y dividiendo la diferencia por la desviación típica de los pronósticos), reciben el nombre: zpr_#.

Corregidos.- Pronóstico que corresponde a cada caso cuando la ecuación de regresión se obtiene sin incluir ese caso, nombre: adj_#.

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

E.T. del pronóstico promedio.- Error típico de los pronósticos correspondientes a los casos que tienen el mismo valor en las variables independientes, nombre: sep_#.

Al efectuar un pronóstico hay dos situaciones diferentes:

① Efectuar un pronóstico individual Y_i^* para un caso concreto X_i

② Pronosticar para cada caso la media de los pronósticos Y_0^* correspondientes a todos los casos con el mismo valor X_0 en las(s) variable(s) independiente(s). A esta media se llama pronóstico promedio.

① Al efectuar un pronóstico individual para un determinado valor de X_i, el error de estimación o variación residual $(Y - Y_i^*)$ puede contener dos fuentes de error:

a) La diferencia entre el valor observado en la variable dependiente Y_i y la media poblacional correspondiente a X_0 (μ_{Y/X_0}).

b) La diferencia entre el pronóstico para ese caso (Y_i^* o Y_0^*) y la media poblacional correspondiente a X_0 (μ_{Y/X_0}).

En un pronóstico individual entran en juego las dos fuentes de error, mientras que en un pronóstico promedio sólo entra la segunda fuente de error. En consecuencia, para un valor dado de X_0, el error típico del pronóstico promedio será menor o igual que el error típico del valor individual.

Por tanto, al construir intervalos de confianza para los pronósticos, la amplitud del intervalo cambiará dependiendo del error típico que se tome como referencia.

Intervalos de pronóstico.- Las opciones del recuadro permiten obtener dos tipos de intervalos:

a) **Media:** Intervalo de confianza basado en los errores típicos de los pronósticos promedio.

b) **Individuos:** Intervalo de confianza basado en los errores típicos de los pronósticos individuales.

La opción Intervalo de confianza k % permite establecer el nivel de confianza con el que se construyen los intervalos de confianza.

Cada una de las opciones (media e individuos) genera en el Editor de datos dos nuevas variables con el límite inferior y superior del intervalo. Estas nuevas variables reciben los siguientes nombres:

lmci_#: Límite inferior IC pronóstico medio
lici_#: Límite inferior IC pronóstico individual
umci_#: Límite superior IC pronóstico medio
uici_#: Límite superior IC pronóstico individual

	Beneficios	Gastos_Publi	Horas_extras	PRE_1	ZPR_1	LICI_1	UICI_1	LMCI_1	UMCI_1
1	1,30	,30	4,00	1,6168	-1,743	,39356	2,8401	,8211	2,4125
2	3,50	1,50	9,00	3,6729	,69683	2,5551	4,7907	3,051	4,2944
3	2,80	,70	6,00	2,3644	-,8560	1,3168	3,4120	1,880	2,8484
4	3,00	1,10	7,50	3,0187	-,0796	1,9923	4,0450	2,582	3,4548
5	3,30	1,20	8,00	3,2056	,14221	2,1534	4,2577	2,712	3,6993
6	4,00	2,00	7,00	3,7675	,80908	2,5754	4,9596	3,021	4,5145
7	3,70	2,00	8,00	3,9541	1,0306	2,8554	5,0529	3,368	4,5407

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

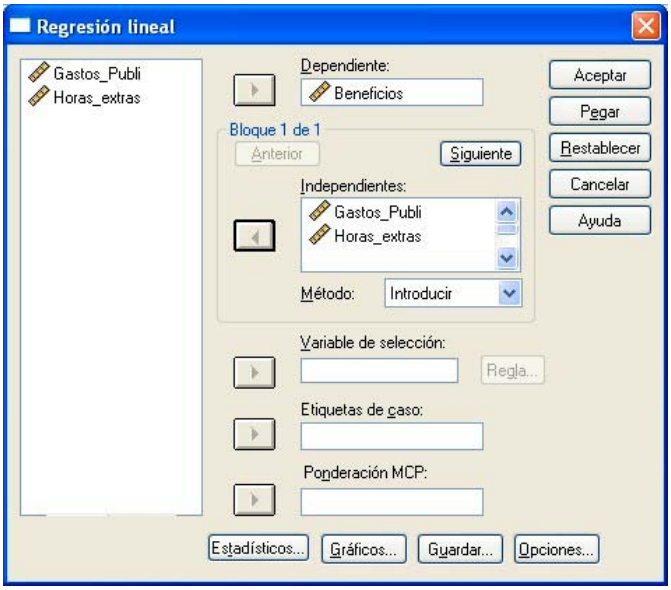
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Criterios de Selección de Variables



Los métodos por pasos que incluye el SPSS para la selección de variables son dos criterios estadísticos:

- Criterio de significación (Probabilidad de F, valor de F)
- Criterio de tolerancia

- Criterio de significación.- Sólo incorpora al modelo de regresión aquellas variables que contribuyen de forma significativa al ajuste del modelo.

La contribución individual de una variable al ajuste del modelo se establece contrastando, a partir del coeficiente de correlación parcial, la hipótesis de independencia entre esa variable y la variable dependiente. Para decidir si se mantiene o se rechaza esa hipótesis de independencia, el SPSS incluye dos criterios de selección:

a) Probabilidad de F.- Una variable pasa a formar parte del modelo de regresión si el nivel crítico asociado a su coeficiente de correlación parcial al contrastar la hipótesis de independencia es menor que 0,05 (probabilidad de entrada). Y queda fuera del modelo de regresión lineal si el nivel crítico es mayor que 0,10 (probabilidad de salida).

b) Valor de F.- Una variable pasa a formar parte del modelo de regresión lineal si el valor del estadístico F utilizado para contrastar la hipótesis de independencia es mayor que 3,84 (valor de entrada). Y queda fuera del modelo de regresión lineal si el valor del estadístico F es menor que 2,71 (valor de salida).

Las opciones del recuadro Criterios del método por pasos permite seleccionar uno de los dos criterios de significación disponibles, así como modificar las probabilidades de entrada y salida.

- Criterio de tolerancia.- Superado el nivel de significación, una variable solo pasa a formar parte del modelo si su nivel de tolerancia es mayor que el nivel establecido por defecto (este nivel es mayor que 0,0001, pero puede cambiarse mediante sintaxis) y, si además, aún correspondiéndole un coeficiente de correlación parcial significativamente distinto de cero, su incorporación al modelo hace que alguna de las variables previamente seleccionadas pase a tener un nivel de tolerancia por debajo del establecido por defecto.

Una forma intuitiva de comprender y valorar el efecto resultante de aplicar estos criterios de selección consiste en observar el cambio que se va produciendo en el coeficiente de determinación R^2 a medida que se van incorporando (o eliminando) variables al modelo.

(IMPRESO EN PAPEL RECICLADO)
UNICAMENTE PARA USO ESCOLAR

Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

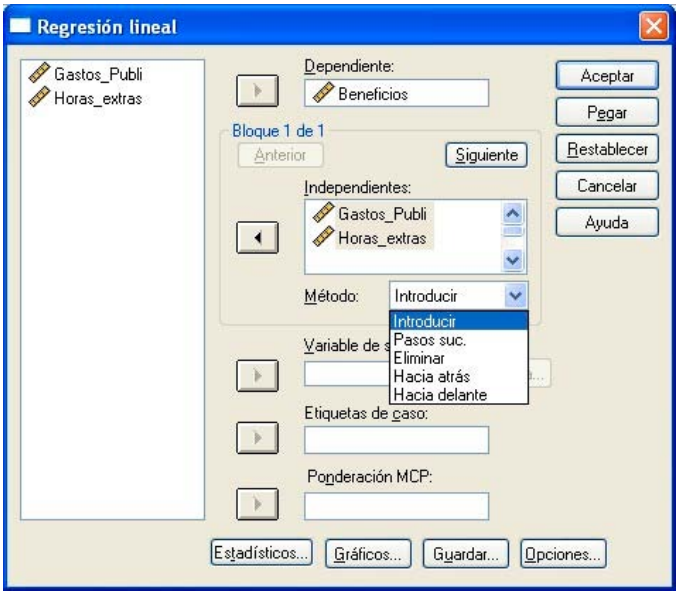
Apellidos ..... Nombre .....

Ejercicio del día .....

Este cambio se puede definir como $R^2_{cambio} = R^2 - R^2_i$, donde R^2_i se refiere al coeficiente de determinación obtenido con todas las variables independientes excepto la i-ésima.

Un cambio grande en R^2 indica que esa variable contribuye de forma importante a explicar lo que ocurre con la variable dependiente.

Para obtener los valores de R^2_{cambio} y su significación (el grado en que el cambio observado en R^2 difiere de cero) hay que marcar la opción Cambio en R cuadrado del botón [Estadísticos].



MÉTODOS DE SELECCIÓN DE VARIABLES

Existen diferentes métodos para seleccionar las variables independientes que debe incluir un modelo de regresión, pero los que mayor aceptación son los métodos de selección por pasos (stepwise). Con estos métodos, se selecciona en primer lugar la mejor variable (con algún criterio estadístico); a continuación, la mejor de las restantes; y así sucesivamente hasta que no queden variables.

Todas las opciones se encuentran disponibles en el menú del botón despegable Método.

Dos de los métodos permiten incluir o excluir, en un solo paso, todas las variables independientes seleccionadas:

Introducir: Construye la ecuación de regresión utilizando todas las variables seleccionadas en la lista de Independientes. Es el método utilizado por defecto.

Eliminar: Elimina en un solo paso todas las variables de la lista de Independientes y ofrece los coeficientes de regresión que corresponderían a cada variable en el caso de que pasaran a formar parte de la ecuación de regresión.

El resto de selección de variables son métodos por pasos, esto es, métodos que van incorporando o eliminando variables paso a paso dependiendo que éstas cumplan o no los criterios de selección:

- Hacia delante: Las variables se incorporan al modelo de regresión una a una.

En el primer paso se selecciona la variable independiente, que además de superar los criterios de entrada, más alto correlaciona (positiva o negativamente) con la dependiente.

En los siguientes pasos se utiliza como criterio de selección el coeficiente de correlación parcial:

<< Van siendo seleccionadas una a una las variables que, además de superar los criterios de entrada, poseen el coeficiente de correlación más alto en valor absoluto (la relación se parcializa controlando el efecto de las variables independientes previamente seleccionadas).

La selección de variables se detiene cuando no quedan variables que superen el criterio de entrada (utilizar como criterio de entrada el tamaño, en valor absoluto, del coeficiente de correlación parcial, es equivalente a seleccionar la variable con menor probabilidad de F o mayor valor de F >>.

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

- Hacia atrás: Comienza incluyendo en el modelo todas las variables seleccionadas en la lista Independientes y luego procede a eliminarlas una a una.

La primera variable eliminada es aquella que, además de cumplir los criterios de salida, pose el coeficiente de regresión más bajo en valor absoluto.

En cada paso sucesivo se van eliminando las variables con coeficientes de regresión no significativos, siempre en orden inverso al tamaño de su nivel crítico.

La eliminación de variables se detiene cuando no quedan variables en el modelo que cumplan los criterios de salida.

- Pasos sucesivos: Es un método mezcla de los métodos Hacia delante y Hacia atrás.

Como el método Hacia delante, en el primer paso comienza seleccionando la variable independiente que, además de superar los criterios de entrada, más alto correlaciona (en valor absoluto) con la variable dependiente.

A continuación, selecciona la variable independiente que, además de superar los criterios de entrada, posee el coeficiente de correlación parcial más alto (en valor absoluto).

Cada vez que se incorpora una nueva variable al modelo, las variables previamente seleccionadas son, al igual que en el método Hacia atrás, evaluadas nuevamente para determinar si siguen cumpliendo o no los criterios de salida. Si alguna variable seleccionada cumple los criterios de salida, es eliminada del modelo.

El proceso se detiene cuando no queden variables que superen el criterio de entrada y las variables seleccionadas no verifiquen los criterios de salida.

(IMPRESO EN PAPEL RECICLADO)

UNICAMENTE PARA USO ESCOLAR

Hoja nº .....

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

UNICAMENTE PARA USO ESCOLAR (IMPRESO EN PAPEL RECICLADO)

Instrumentos Estadísticos Avanzados
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández