FUNDAMENTOS Y HERRAMIENTAS PARA LA MODELIZACIÓN DE PROCESOS TÉCNICOS-CIENTÍFICOS DE INVESTIGACIÓN

Bloque III: Modelización estadística.

Regresión y correlación

Roberto Espejo Mohedano e-mail: ma1esmor@uco.es

Modelización.

- Necesidad del estudio simultaneo de varias variables aleatorias.
  - Analizar las posibles relaciones entre ellas. (Estudios de correlación).
  - Intentar establecer el modelo matemático que las relacione. (Estudios de regresión o ajuste).

- Interés por estimar una magnitud Y (variable dependiente) en función de una o varias $X_1, X_2, \cdots, X_k$ variables explicativas (variables independientes).
  - Imposibilidad de predicción exacta. Perturbación aleatoria ε.
  - Modelos lineales:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Modelos comunes

Lineales

Lineal simple: $\hat{Y} = \beta_0 + \beta_1 X$

Parabólico: $\hat{Y} = \beta_0 + \beta_2 X + \beta_1 X^2$

Cúbico: $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

Polinómico: $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^2 + \cdots + \beta_h X^h$

No Lineales

Exponencial: $\hat{Y} = \beta_0 k^{\beta_1 X}$

Potencial: $\hat{Y} = \beta_0 X^{\beta_1}$

Hiperbólico: $\hat{Y} = 1 / \beta_0 + \beta_1 X$

Logístico: $\hat{Y} = 1 / e^{-\beta_0 - \beta_1 X}$

Estimación de los coeficientes: $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$

Problema de inferencia:

Contrastes sobre los coeficientes y sobre el ajuste del modelo.

Correlación simple.

$$R_{xy} = \frac{S_{xy}}{S_x S_y} \in (-1;1)$$

$$R_{x_j x_i} = \frac{S_{x_j x_i}}{S_{x_j} S_{x_i}} \; \forall \; i \neq j = 1,2,\ldots$$

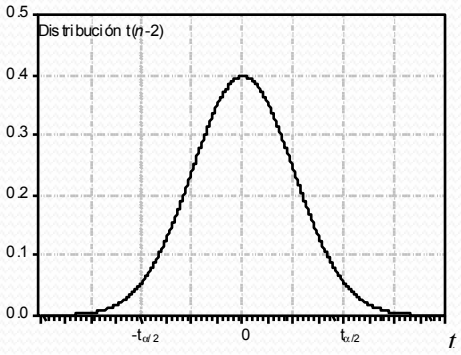
Matriz de Correlación: $\Gamma_{yx_1x_2\ldots} = \begin{pmatrix} 1 & R_{yx_1} & R_{yx_2} & \cdots \\ R_{yx_1} & 1 & R_{x_1x_2} & \cdots \\ R_{yx_2} & R_{x_1x_2} & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$

Contrastes de incorrelación

Poblaciones *X* e *Y* numéricas.

Hipótesis: $\begin{cases} H_0: X \text{ está incorrelada con } Y \\ H_1: X \text{ está correlada con } Y \end{cases} \Rightarrow \begin{cases} H_0: \rho_{xy} = 0 \\ H_1: \rho_{xy} \neq 0 \end{cases}$

Estadístico: $T = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \in t_{(n-2)} \qquad r = r_{xy} = \dfrac{S_{xy}}{S_x S_y}$



Regla de decisión de nivel α: $C_0 = (-t_{\alpha/2}; +t_{\alpha/2})$

Decisión: A nivel α: Si $t \in C_0$ ➔ se acepta H_0

Para todo α ≤ *p* ➔ se acepta H_0
Para todo α > *p* ➔ se rechaza H_0

Probabilidad límite: *p* = *P*(T > |*t*|)

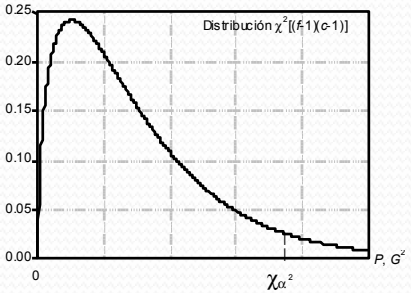
Contrastes de independencia

Poblaciones no numéricas. $X \in S_x = \{x_1^*, \ldots, x_f^*\}$ $Y \in S_y = \{y_1^*, \ldots, y_c^*\}$

Hipótesis: $\begin{cases} H_0 : X \text{ es independiente de } Y \\ H_1 : X \text{ esta relacionada con } Y \end{cases}$

Muestra: $(x_t, y_t) \quad t = 1, \cdots, n$ tabulada en forma de tabla de contingencia

$$\|n_{ij}\| \quad i = 1 \cdots f; j = 1 \cdots c$$



Estadísticos: $G^2 = 2\sum_{i=1}^{f}\sum_{j=1}^{c} n_{ij} \ln \frac{n_{ij}}{E_{ij}}$

$$P = \sum_{i=1}^{f}\sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \qquad P_c = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(|n_{ij} - E_{ij}| - 0.5\right)^2}{E_{ij}}$$

Regla de decisión de nivel α:

$$C_0 = \left(0, \chi_\alpha^2\right)$$

Decisión: A nivel α: Si $P, P_c, G^2 \in C_0$ ➔ se acepta H_0

Para todo α ≤ *p* ➔ se acepta H_0
Para todo α > *p* ➔ se rechaza H_0

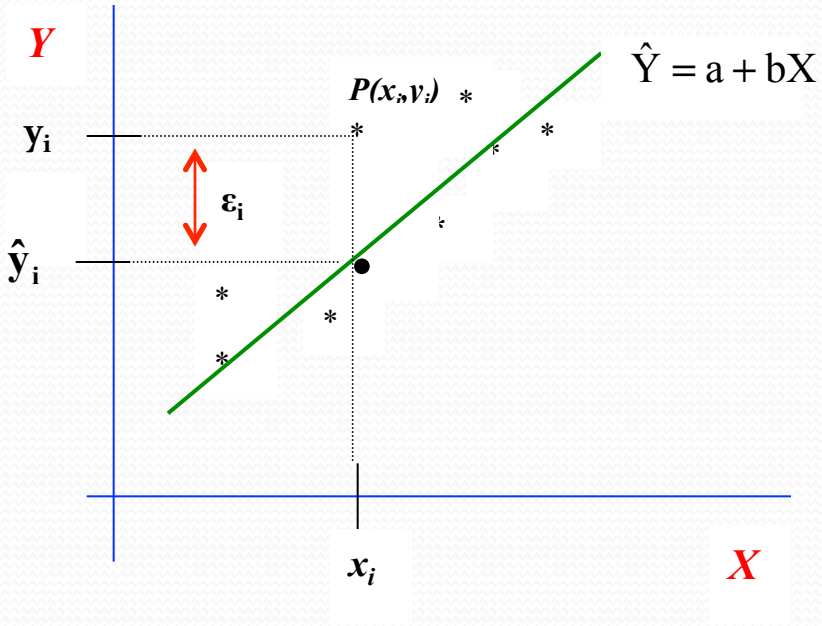
Distribución muestral (asintótica):

$$G^2 / H_0, P / H_0, P_C / H_0 \in \chi^2\left((f-1),(c-1)\right)$$

Condiciones asintóticas: $E_{ij} = \frac{n_{i.}\, n_{.j}}{n} \geq 1 \qquad \forall i, j$

Regresión Simple: Línea de Regresión

Recta de regresión de Y sobre X



$\hat{y}_i$ valor estimado de *Y* para $X=x_i$

ε_i residuos: $\varepsilon_i = y_i - \hat{y}_i$

Mínimos cuadrados $H = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2$

$$b = \frac{S_{XY}}{S_X^2} \qquad a = \overline{Y} - b\overline{X}$$

Los modelos de Regresión requieren Normalidad en la variable Dependiente

Regresión Múltiple: $\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n$

Matriz de Covarianzas

$$\sum_{yx_1x_2\ldots} = \begin{pmatrix} S_y^2 & S_{yx_1} & S_{yx_2} & \ldots \\ S_{x_1y} & S_{x_1}^2 & S_{x_1x_2} & \ldots \\ S_{x_2y} & S_{x_1x_2} & S_{x_2}^2 & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

Vector de Medias

$$\vec{\mu}_{yx_1x_2\ldots} = \begin{pmatrix} \overline{Y} \\ \overline{X}_1 \\ \overline{X}_2 \\ \ldots \end{pmatrix}$$

$$b_i = -\frac{A_{1,i+1}}{A_{1,1}} \quad \forall \quad i = 1,2,\ldots$$

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 - \ldots \qquad \left(\overline{Y} = a + b_1\overline{X}_1 + b_2\overline{X}_2 + \ldots\right)$$

Correlación múltiple: $$R_{yx_1x_2\ldots} = R_y = \frac{S_{y\hat{y}}}{S_y S_{\hat{y}}} \in \left[-1,1\right]$$

Se verifica que: $$R_y^2 \geq R_{yx_i}^2 \;\forall\, i = 1,2,\ldots$$

Coeficiente de determinación: $$R^2 = R_{xy}^2 = \left(\frac{S_{xy}}{S_x S_y}\right)^2 = 1 - \frac{S_\varepsilon^2}{S_y^2} \in \left[\,0\,;1\,\right]$$

"Tanto por 1 (ó %) de la varianza de *Y* explicada por el modelo."

Nos informa de hasta que punto el modelo se ajusta a los datos.

Test de hipótesis sobre el modelo de regresión: Test F

Se basa en el **Teorema de descomposición de la varianza** $S_y^2 = S_{\hat{y}}^2 + S_e^2$

o multiplicando por el número de datos (n), la suma de cuadrados totales se descompone en: $S_y = S_{\hat{y}} + S_e$

Siendo: $S_y = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$ La Suma de cuadrados total

$S_{\hat{y}} = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$ La Suma de cuadrados asociada a las variables explicativas (la regresión)

$S_e = \sum_{i=1}^{n} e_i^2$ La Suma de cuadrados residual, la no explicada por la regresión

Hipótesis a contrastar: $\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ H_1 : \text{algún o algunos } \beta_i \neq 0 \end{cases} \Rightarrow \begin{cases} H_0 : \text{Rechazo Gobalmente el modelo} \\ H_1 : \text{Acepto Gobalmente el modelo} \end{cases}$

Estadístico del contraste: $F = \dfrac{S_{\hat{y}} / (k-1)}{S_e / (n-k-1)}$ Siendo su distribución muestral es: $F / H_0 \in F(k-1; n-k-1)$

Siendo k el número de variables explicativas (independientes) del modelo

Regla de decisión de nivel *α*: $C_0 = (0 ; F_\alpha)$

Para todo $\alpha \leq p$ ➔ se acepta H_0

Para todo $\alpha > p$ ➔ se rechaza H_0

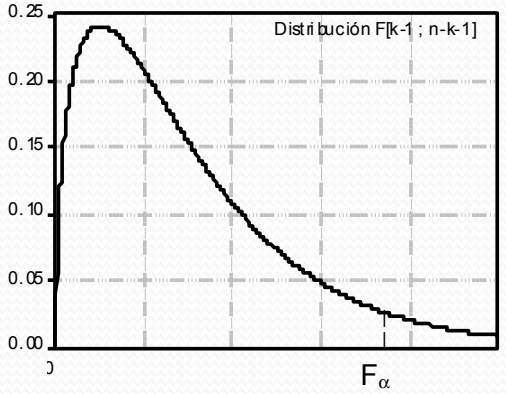
Test de hipótesis sobre el modelo de regresión: Test F

Los cálculos anteriores se resumen en la siguiente tabla de análisis de la Varianza

Fuente de Variación	Grados de Libertad	Sumas de Cuadrados	Medias de Cuadrados	Estadístico F
Regresión	k-1	$S_{\hat{y}} = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$	$M_{\hat{y}} = \frac{S_{\hat{y}}}{k-1}$	$F = \frac{M_{\hat{y}}}{M_e}$
Residuo	n-k-1	$S_e = \sum_{i=1}^{n} e_i^2$	$M_e = \frac{S_e}{n-k-1}$	
Variación Total	n-1	$S_y = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = S_{\hat{y}} + S_e$		

Decisión: A nivel α: Si $F \in C_0$ ➔ se acepta H_0

Para todo α ≤ *p* ➔ se acepta H_0
Para todo α > *p* ➔ se rechaza H_0



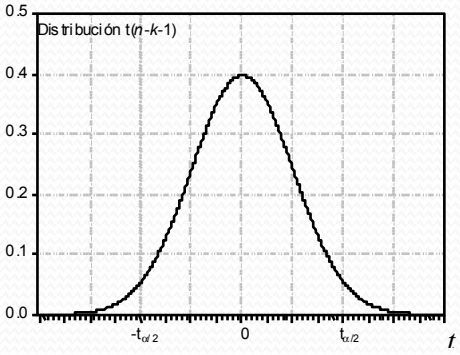
Contrastes sobre los Coeficientes de Regresión

Tests T

Hipótesis: $\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$

Estadístico: $T = \dfrac{b_j}{\overline{s}_{b_j}} \qquad \overline{s}_{b_j}^2 = \overline{s}_e^2 a_{jj} \qquad \overline{s}_e^2 = \dfrac{1}{n-k-1}\sum e_i^2 \qquad A = \left\| a_{ij} \right\| = (X'X)^{-1}$

Distribución muestral: Si $e_i \in N\left(0_n, \sigma_\varepsilon^2 I_n\right) \rightarrow T \in t(n-k-1)$



Regla de decisión de nivel *α*: $C_0 = (-t_{\alpha/2}; t_{\alpha/2})$

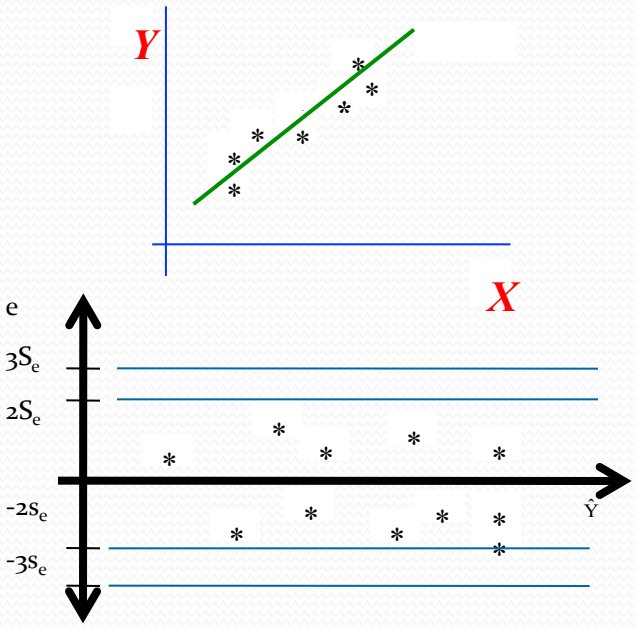
Decisión: A nivel α: Si $t \in C_0$ ➔ se acepta H_0

Para todo α ≤ *p* ➔ se acepta H_0
Para todo α > *p* ➔ se rechaza H_0

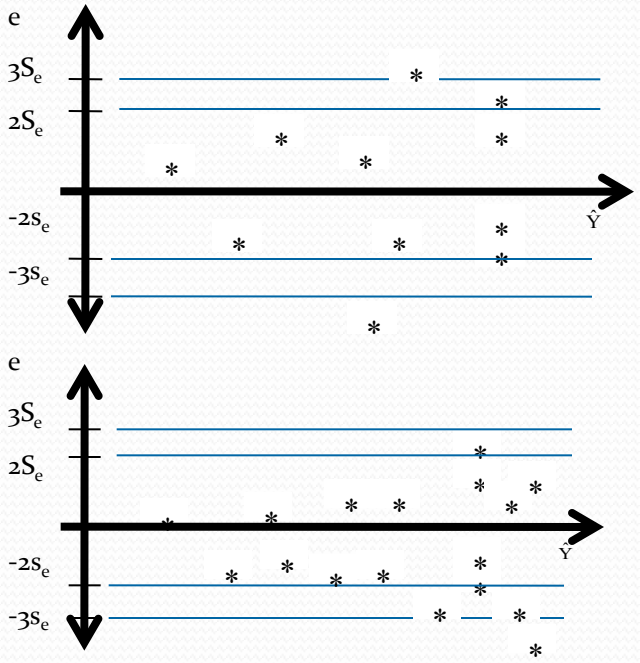
Propiedades de los Residuos

1ª.- $\bar{e} = 0$

2ª.- Var(ε_i) = cte Varianza constante. (homocedásticos)



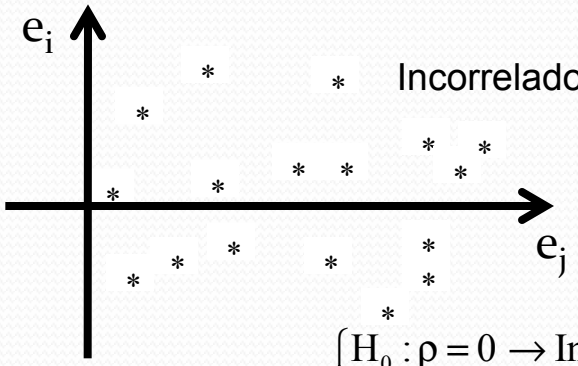
Homocedasticidad



Heterocedasticidad

Propiedades de los Residuos

3ª.- Los residuos han de estar incorrelados entre si.



Incorrelados

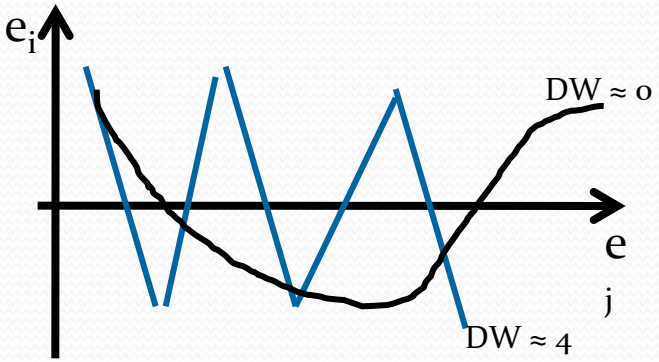
Estadístico de Durbin-Watson

$$DW = \frac{\sum_{i=2}^{n}\left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2} \in \left[0\,;\,4\right]$$

DW ≈ 2, Incorrelados
DW ≈ 4 Autocorrelación negativa
DW ≈ 0 Autocorrelación positiva

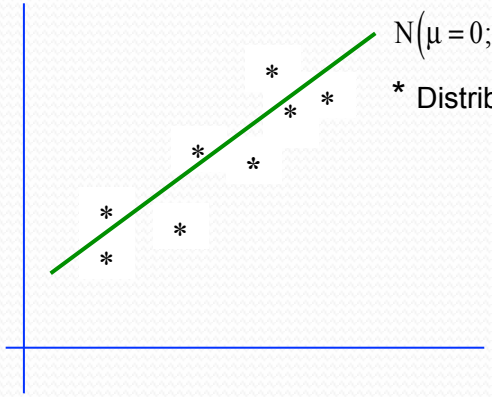
Test de Hipótesis:

$$\begin{cases} H_0 : \rho = 0 \rightarrow \text{Incorrelados} \\ H_1 : \rho \begin{cases} > 0 \rightarrow \text{Autocorrelacion Positiva} \\ \neq 0 \rightarrow \text{Autocorrelados} \\ < 0 \rightarrow \text{Autocorrelacion Negativa} \end{cases} \end{cases}$$



4ª.- Normales

$$e \in N\left(\mu = 0; \sigma^2 = S_e^2\right)$$



$N\left(\mu = 0; \sigma^2 = S_e^2\right)$

* Distribución Residual

Transformaciones no lineales.

Tratan de corregir la falta de simetría de algunas distribuciones de datos. En ocasiones, estas transformaciones corrigen las desviaciones de normalidad.

La transformación mas utilizada es el logaritmo aunque existen otras mas simples como la raíz cuadrada o la inversa.

La transformación de Box-Cox (perteneciente a las transformaciones logarítmicas) también es muy usada y suele corregir las desviaciones de los supuestos necesarios para los modelos de regresión.

Dada una variable *X*, su transformada de Box-Cox se define como:

$$X^* = \left\{ \begin{array}{lll} \dfrac{(X+m)^{\lambda}-1}{\lambda} & si & \lambda \neq 0 \\ \ln(X+m) & si & \lambda = 0 \end{array} \right\}$$

m se escoge de tal forma que (*X* + *m*) sea positivo.

Si λ > 1 se corrige la asimetría a la izquierda y si λ < 1 la asimetría a la derecha.

FUNDAMENTOS Y HERRAMIENTAS PARA LA MODELIZACIÓN DE PROCESOS TÉCNICOS-CIENTÍFICOS DE INVESTIGACIÓN

Bloque III: Modelización estadística.

Regresión y correlación con R (Rcommander)

Roberto Espejo Mohedano e-mail: ma1esmor@uco.es