Modelo Lineal General

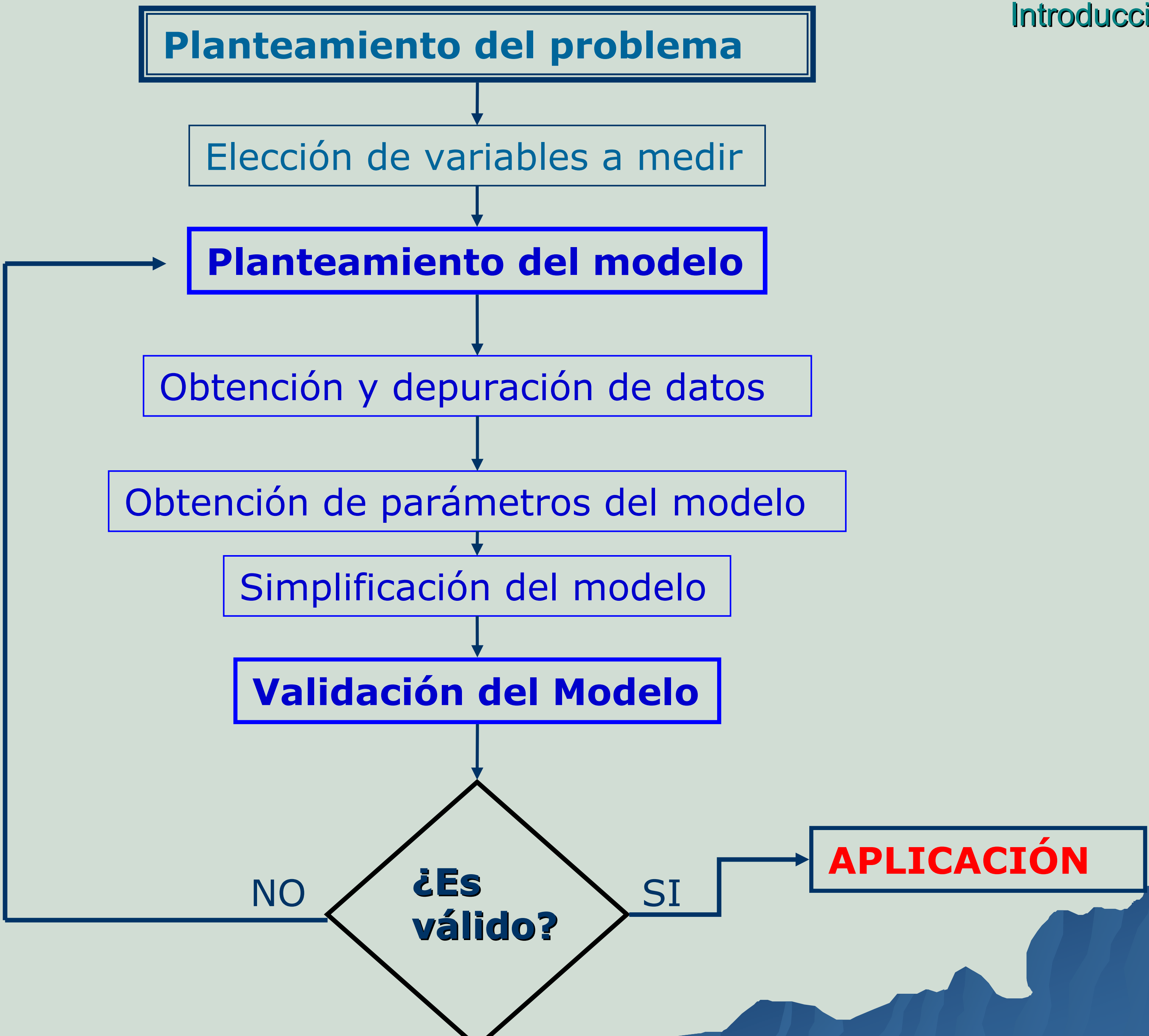
Prof. Susana Martín Fernández

Índice

- Introducción
- Modelo Lineal General
- Análisis de la Varianza
- Regresión Lineal

Introducción

- Un modelo lineal es una relación entre variables matemáticas cuantitativas y/o cualitativas (explicativas) y un vector aleatorio de interés.

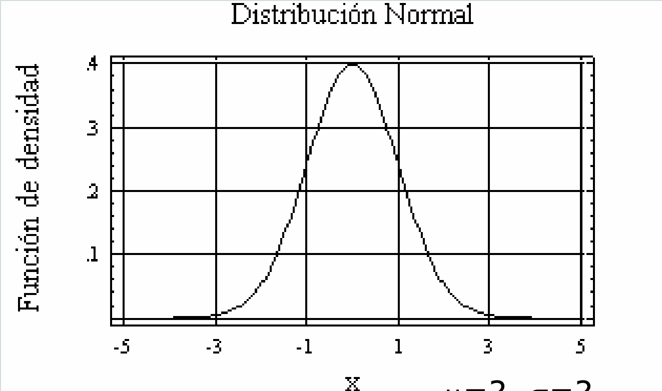

Planteamiento del problema
Elección de variables a medir
Planteamiento del modelo
Obtención y depuración de datos
Obtención de parámetros del modelo
Simplificación del modelo
Validación del Modelo
¿Es válido?
NO
SI
APLICACIÓN

Métodos Estadísticos de Inferencia

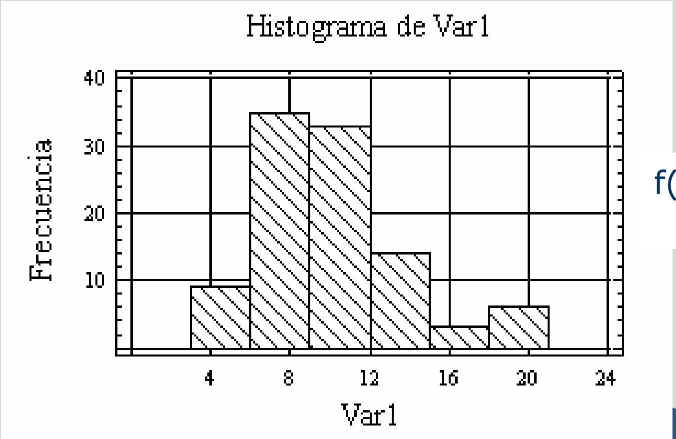
Dada la V.A. X ¿Se conoce su Función de Distribución excepto un nº de parámetros?

SI — Inferencia Paramétrica

NO — Inferencia no Paramétrica



$\mu=?$, $\sigma=?$



f(x)=?

Modelo Lineal General

Sea $X = (X_1, X_2, ..., X_n)$ un vector aleatorio, sea A una matriz *nxk / k<n* de constantes conocidas a_{ij} (valores de las variables explicativas), *i=1,..n; j=1,...,k. Y* sea β un vector **desconocido** de escalares, el vector $X = (X_1, X_2, ..., X_n)$ sigue un modelo lineal si se puede escribir de la siguiente forma:

$$X = \beta A' + \varepsilon$$

Donde $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ es un vector de variables aleatorias no medibles y además cumplen que $E[\varepsilon_i] = 0$ (media cero).

Luego **otra definición** de modelo lineal sería la siguiente:

Sea $X = (X_1, X_2, \ldots, X_n)$ un vector aleatorio, sea A una matriz *nxk / k<n* de constantes conocidas a_{ij} *i=1,..n; j=1,...,k. Y* sea β un vector **desconocido** de escalares, el vector $X = (X_1, X_2, \ldots, X_n)$ sigue un modelo lineal si cumple:

$$E[X] = \beta A'$$

Modelo lineal de forma matricial:

$$(X_1, X_2, \ldots, X_n) = (\beta_1, \beta_2, \ldots, \beta_k) \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1k} & a_{2k} & \cdots & a_{nk} \end{pmatrix} + (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$$

D

Por tanto:

$$X_1 = \beta_1 a_{11} + \ldots + \beta_k a_{1k} + \varepsilon_1$$

$$X_2 = \beta_1 a_{21} + \ldots + \beta_k a_{2k} + \varepsilon_2$$

$$\ldots$$

$$X_i = \beta_1 a_{i1} + \ldots + \beta_j a_{ij} + \ldots + \beta_k a_{ik} + \varepsilon_i$$

$$\ldots$$

$$X_n = \beta_1 a_{n1} + \ldots + \beta_k a_{nk} + \varepsilon_n$$

¿Qué representan estas expresiones?

$$X_1 = \beta_1 a_{11} + \ldots + \beta_k a_{1k} + \varepsilon_1$$

$$X_2 = \beta_1 a_{21} + \ldots + \beta_k a_{2k} + \varepsilon_2$$

$$\ldots$$

$$X_i = \beta_1 a_{i1} + \ldots + \beta_j a_{ij} + \ldots + \beta_k a_{ik} + \varepsilon_i$$

$$\ldots$$

$$X_n = \beta_1 a_{n1} + \ldots + \beta_k a_{nk} + \varepsilon_n$$

Muestra de tamaño n de una variable X

Datos de las k variables numéricas o explicativas medidas en las n unidades muestrales

EJEMPLO:

Queremos modelizar los sólidos en suspensión (**ss**) de una depuradora en función del caudal, Q, y del pH.

Datos:

ss=(376,364,360)

Q=(28.1, 28.9, 30.1)

pH=(7.75, 7.53, 7.9)

Por tanto n=3;

El vector X es:

$X=(X_1, X_2, X_3)=(376,364,360)$

K=2, hay 2 variables numéricas

Q=(28.1, 28.9, 30.1)

pH=(7.75, 7.53, 7.9)

El modelo de forma matricial:

$$(X_1, X_2, X_3) = (\beta_1, \beta_2)\begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{pmatrix} + (\varepsilon_1, \varepsilon_2, \varepsilon_3)$$

Caudal

pH

D

$$(376{,}364{,}360) = (\beta_1, \beta_2)\begin{pmatrix} 28.1 & 28.9 & 30.1 \\ 7.75 & 7.53 & 7.9 \end{pmatrix} + (\varepsilon_1, \varepsilon_2, \varepsilon_3)$$

Se **asume** que las variables del vector $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ cumplen:

- Son independientes

- Siguen un distribución normal

-Todas tienen la misma varianza σ^2 (homocedasticidad) .

- $E[\varepsilon_i] = 0$ (media cero). Condición que ya cumplían por definición.

Bajo estas condiciones se deduce que las variables del vector X=(X1, X2,..., Xn) siguen una distribución normal, son independientes y con varianza constante σ2 .

Objetivo:

El objetivo es encontrar el "mejor" vector de estimadores de los parámetros:

$$\beta = (\beta_1, \beta_2, \ldots, \beta_k)$$

Consistente. Al aumentar el tamaño de la muestra el estimador converge en probabilidad en el parámetro estimado.

D

Invariante. Sea G un grupo de transformaciones que deja a las funciones de distribución $\{F\theta : \theta \in \Theta\}$ invariantes. Un estimador U se dice que es invariante bajo G, si: U(g(x1), g(x2), ..., g(xn)) = U(x1, x2, ..., xn), $\forall$ g$\in$G.

Con varianza mínima. Se dice que U es el estimador de varianza mínima de un parámetro θ, si para cualquier otro estimador Ui de dicho parámetro se cumple que:

var(U) < var(Ui)

Insesgado. E[U]=θ

Obtención del vector de parámetros β:

Para la estimación de los parámetros se pueden utilizar dos métodos:

Error cuadrático mínimo.

$$\min \quad \varepsilon\varepsilon' = \left(X - \beta A'\right)\left(X - \beta A'\right)' = \sum \varepsilon_i^2$$

Método de máxima verosimilitud.

$$f_{\beta,\sigma}(x_1,\ldots,x_n) = \frac{1}{\left(\sqrt{2\pi}\,\sigma\right)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \beta_1 a_{i1} - \ldots - \beta_k a_{ik}\right)^2\right.$$

- **Simplificación del modelo-Hipótesis lineal general**

  El modelo se simplificará si se puede aceptar que alguno de los coeficientes β_i es 0.

Forma teórica y compleja de plantear la simplificación del modelo

$$H_0 : \beta H' = 0$$

$$(\beta_1, \beta_2, \ldots, \beta_k) \begin{pmatrix} h_{11} & h_{21} & \ldots & h_{r1} \\ h_{12} & h_{22} & \ldots & h_{r2} \\ \ldots & \ldots & \ldots & \ldots \\ h_{1k} & h_{2k} & \ldots & h_{rk} \end{pmatrix} = (0, 0, \ldots, 0)$$

Teorema

Sea el modelo lineal

$$X=\beta A'+\varepsilon$$

donde *A* es una matriz *nxk* conocida y de rango *k<n*, sea β un vector desconocido de escalares, y sea $\varepsilon=(\varepsilon_1,\varepsilon_2,...\varepsilon_n)$ un vector de variables aleatorias no observables independientes de forma que todas ellas siguen una distribución Normal $N(0,\sigma^2)$. El coeficiente obtenido por máxima verosimilitud *F* (estadístico) para contrastar la hipótesis lineal $H_0:\beta H'=0$, donde *H* es una matriz *rxk* con rango *r<=k,* hará que se rechace la hipótesis nula para un nivel de significación α si $F\geq F_0$, donde

$P_{H0}(F\geq F_0)$, es decir la probabilidad de rechazar la hipótesis nula cuando ésta es cierta en la realidad es α. Y se demuestra que *F* es una variable aleatoria cuya expresión es:

$$F=\frac{\left(X-\breve{\beta}_0\,A'\right)\left(X-\breve{\beta}_0\,A'\right)'-\left(X-\widehat{\beta}\,A'\right)\left(X-\widehat{\beta}\,A'\right)'}{\left(X-\widehat{\beta}\,A'\right)\left(X-\widehat{\beta}\,A'\right)'}$$

Donde $\hat{\beta}$ es el estimador de máxima verosimilitud de β y $\hat{\beta}_0$ es el estimador máximo verosímil bajo la hipótesis nula.

La variable aleatoria *[(n-k)/r]F* tiene una distribución F-snedecor con *(r,n-k)* grados de libertad bajo .