Modelos Lineales de Regresión

Emilio Picasso

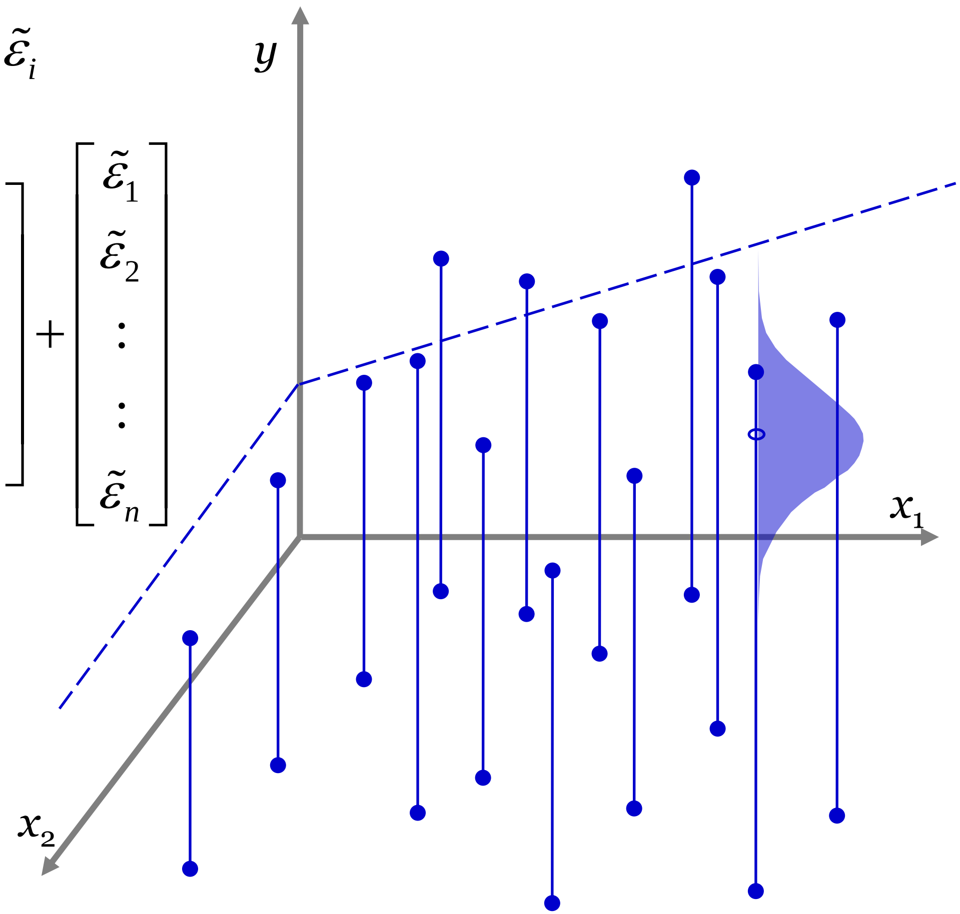
El Modelo

$$\tilde{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \tilde{\varepsilon}_i$$

$$\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \\ \\ \tilde{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \\ \vdots \\ \vdots \\ \tilde{\varepsilon}_n \end{bmatrix}$$

$$\boxed{\tilde{y} = X\beta + \tilde{\varepsilon}}$$

n datos
p=k+1 parámetros



Supuestos

$S1)\ E\tilde{\varepsilon} = 0$

$S2/S3)\ V\tilde{\varepsilon} = \sigma^2 I$

$S4)\ \tilde{\varepsilon} : Normal$

$S5)\ Rango(X) = p$

Emilio Picasso

Estimación

$$\tilde{y} = Xb + \tilde{e}$$

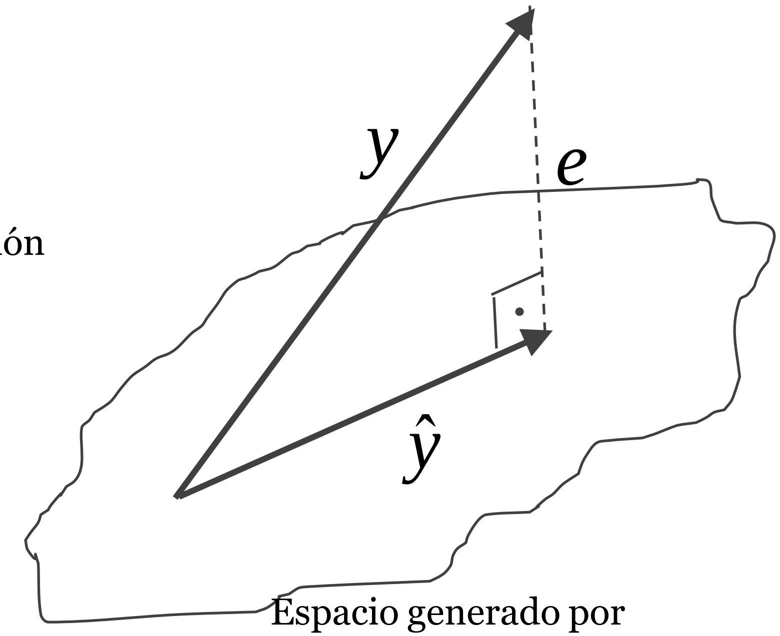
$$b = ({}^{t}XX)^{-1}\,{}^{t}Xy = \mathfrak{I}^{-1}\,{}^{t}Xy$$

$$\hat{\sigma}^{2} = \tfrac{1}{n}\,{}^{t}ee$$

Matriz de Proyección

$$\hat{y} = Xb = X({}^tXX)^{-1}\,{}^tXy = Hy$$

con $\quad H = X({}^tXX)^{-1}\,{}^tX \quad$ Matriz de proyección



Espacio generado por las columnas de **X**

<u>Propiedades de la matriz de proyección</u>

Es simétrica

Es idempotente, luego es semidefinida positiva con valores propios: 0 o 1.

$$e = y - \hat{y} = y - Hy = (I - H)y$$

$$^tXe = {}^tX(I-H)y = ({}^tX - {}^tXX({}^tXX)^{-1}\,{}^tX)y = 0$$

entonces los residuos suman 0.

Además: $\quad \frac{1}{n} \le h_{ii} \le 1$

Emilio Picasso

Análisis de la Varianza

$$\hat{y} = Hy$$
$$e = (I - H)y$$
$$y = \hat{y} + e$$

$$\begin{aligned} {}^{t}yy &= {}^{t}(\hat{y} + e)(\hat{y} + e) \\ &= {}^{t}\hat{y}\hat{y} + 2\,{}^{t}\hat{y}e + {}^{t}ee \\ &= {}^{t}\hat{y}\hat{y} + 2\,{}^{t}yH(I - H)y + {}^{t}ee \\ &= {}^{t}\hat{y}\hat{y} + 2\,{}^{t}y(H - H^{2})y + {}^{t}ee = {}^{t}\hat{y}\hat{y} + {}^{t}ee \end{aligned}$$

Haciendo este desarrollo sobre la suma de cuadrados centrada se obtiene:

$$\boxed{Q_T = Q_E + Q}$$

Propiedades de los Estimadores

<u>Ausencia de Sesgo</u>

$$E\tilde{b} = E(\Im^{-1}\,{}^{t}X\tilde{y}) = \Im^{-1}\,{}^{t}XE\tilde{y} = \Im^{-1}\,{}^{t}XX\,\beta = \beta$$

<u>Varianza</u>

$$V\tilde{b} = V(\Im^{-1}\,{}^{t}X\tilde{y}) = \Im^{-1}\,{}^{t}XV\tilde{y}X\,\Im^{-1} = \Im^{-1}\,{}^{t}X\,\sigma^{2}IX\,\Im^{-1} = \Im^{-1}\,{}^{t}XX\,\Im^{-1}\sigma^{2} = \Im^{-1}\sigma^{2}$$

<u>Teorema de Gauss-Markov</u>

$\tilde{b}$ es BLUE

Propiedades de los Estimadores

<u>Otras</u>

$$Es^2 = \sigma^2 \qquad \text{donde: } s^2 = \tfrac{1}{\nu}\,{}^t e e, \quad \nu = n - p$$

$$\tilde{b} \text{ y } \tilde{s}^2 \quad \text{independientes}$$

$$\tilde{b} \text{ y } \tilde{s}^2 \quad \text{suficientes}$$

$$\tilde{b} : Normal$$

$$\frac{\nu \tilde{s}^2}{\sigma^2} : \chi^2_\nu$$

Inferencia

Sobre β

$$\tilde{b}_j : Normal(\beta_j, \mathfrak{I}^{jj}\sigma^2) \Rightarrow \frac{(\tilde{b}_j - \beta_j)^2}{\mathfrak{I}^{jj}\sigma^2} : \chi_1^2$$

$$\Rightarrow \frac{(\tilde{b}_j - \beta_j)^2}{\mathfrak{I}^{jj}s^2} : F_{1,\nu}$$

Inferencia

Sobre $E(\tilde{y}/x_0) = {}^t x_0 \beta$

$$\hat{y}/x_0 = {}^t x_0 b : Normal({}^t x_0 \beta, \sigma^2\, {}^t x_0 \mathfrak{I}^{-1} x_0)$$

$$\Rightarrow \quad \frac{\hat{y}/x_0 - E(\tilde{y}/x_0)}{\sqrt{{}^t x_0 \mathfrak{I}^{-1} x_0}\;\, s} : t_\nu$$

Predicción de $\tilde{y}/x_0 = {}^t x_0 \beta + \tilde{\varepsilon}$

$$V(\hat{y}/x_0 - \tilde{y}/x_0) = V(\hat{y}/x_0) + V(\tilde{y}/x_0) = {}^t x_0 \mathfrak{I}^{-1} x_0 \sigma^2 + \sigma^2$$

$$\Rightarrow \quad \frac{\hat{y}/x_0 - \tilde{y}/x_0}{\sqrt{1 + {}^t x_0 \mathfrak{I}^{-1} x_0}\;\, s} : t_\nu$$

Incumplimiento de Supuestos
Análisis de Residuos

- Es la técnica mas potente para diagnosticar el incumplimiento de supuestos.

$$e = (I - H)y \quad \Rightarrow \begin{cases} Ve_i = (1 - h_{ii})\sigma^2 \\ V(e_i ; e_l) = -h_{il}\sigma^2 \end{cases}$$

- Residuos Studentizados:

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

- Residuos PRESS:

$$e_{-i,i} = y_i - \hat{y}_{-i,i} = \frac{e_i}{1 - h_{ii}}$$

- Residuos Externamente Studentizados:

$$t_i = \frac{e_i}{s_{-i}\sqrt{1 - h_{ii}}} \qquad s_{-i}^2 = \tfrac{1}{\nu - 1}\left(\nu s^2 - \frac{e_i^2}{1 - h_{ii}} \right)$$

Incumplimiento de Supuestos
Anormalidad

- Diagnóstico:
  - Analizar si los residuos studentizados siguen una distribución Normal
  - Cuando la muestra es chica esto no se puede hacer pues: $e_i = \varepsilon_i - \sum h_{ij}\varepsilon_j$ tiende a normal por el TLC. Cuando es grande las h_{ij} tienden a 0, y $e_i \cong \varepsilon_i$

- Tratamiento:
  - Regresión Robusta: Consiste en mochar los residuos en la suma de cuadrados, de manera de mitigar la influencia de los outliers. Se resuelve por regresión ponderada.

Incumplimiento de Supuestos
Heterocedasticidad

- Diagnóstico:
  - Se observa la dispersión de los residuos en función de y, y de cada x_j.

- Tratamiento:
  - Regresión ponderada:

$$V\tilde{\varepsilon} = \sigma^2 V$$

$$b = \mathfrak{I}^{-1}\,{}^tXV^{-1}y \quad \text{con} \quad \mathfrak{I} = {}^tXV^{-1}X$$

  - Transformaciones:
    - Box & Cox:

$$y' = \begin{cases} \frac{1}{\lambda}(y^{\lambda} - 1) & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}$$

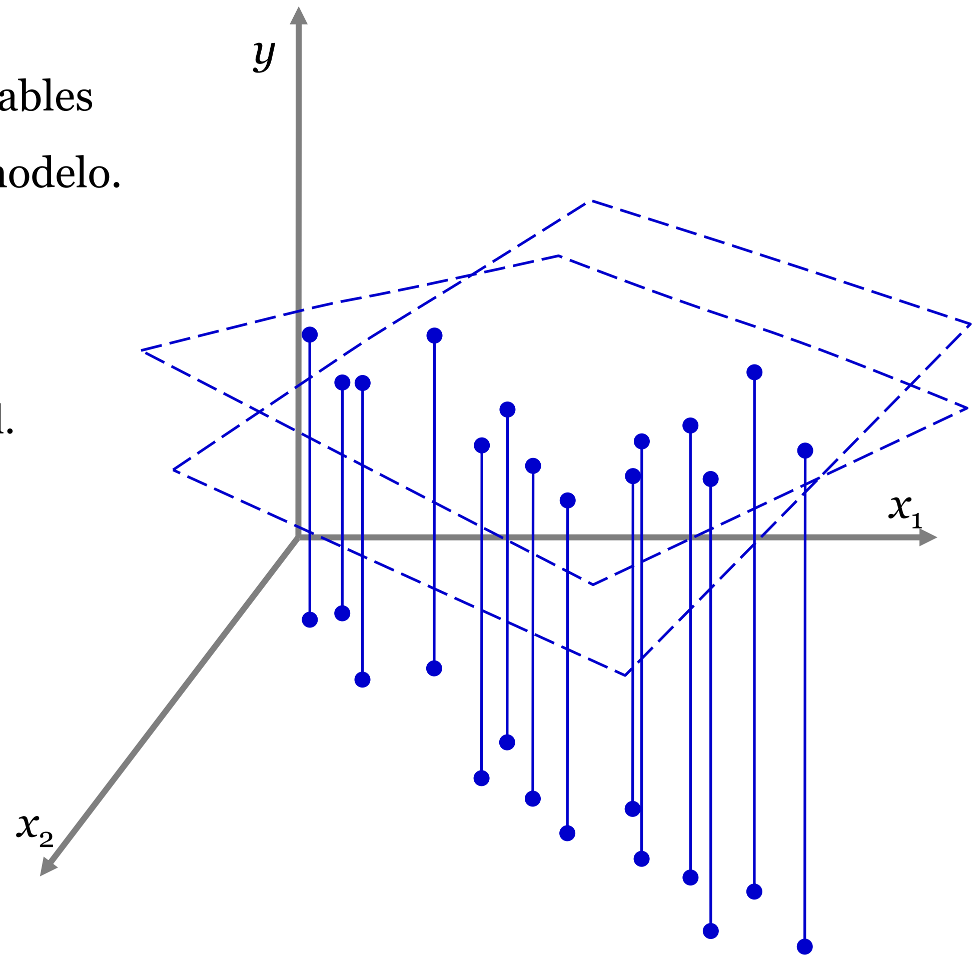
Incumplimiento de Supuestos
Colinealidad

- Las relaciones lineales entre las variables explicativas son deletéreas para el modelo.
  - Desestabilizan los coeficientes.
  - Zonas de predicción inestable.
  - Puede ser poblacional o muestral.
- Diagnóstico:

$$\det(\Im_E) < 0.1$$

$$VIF_j = \Im_E^{jj} > 10$$

- Tratamiento:
  - Sacar variables o agregar datos.
  - Ridge estimation.
  - Componente principales.

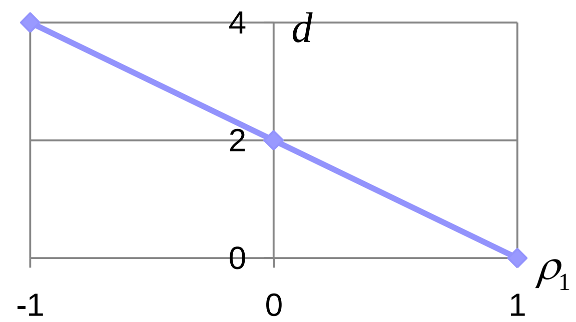


Incumplimiento de Supuestos
Autocorrelación

- Se puede producir cuando los datos están ordenados: *i* es el tiempo.
- Diagnóstico:
  - » Se observan tendencias cíclicas en los residuos en función de *i*.
  - » Ensayo de Durbin-Watson:

$$H_0)\rho_1 = 0$$

$$d_k = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$



    - No es válido cuando hay variables endógenas rezagadas.
- Tratamiento:
  - » Series de tiempo.

Modelización

- Objetivo: Encontrar un modelo parsimonioso que ajuste bien a los datos.
- Indicadores de calidad del modelo:
  - R^2: mide el ajuste pero no la colinealidad.
  - R^2 ajustado (varianzas en vez de Q) o s^2: mide el ajuste y penaliza p, pero no la colinealidad.
  - PRESS: mide la capacidad de predicción en el dominio de los datos. Detecta puntos influyentes pero no otros casos de colinealidad.
  - *det*: Cuando es menor que 0.2 detecta colinealidad.
  - C_p: Cuando supera p detecta sesgo por sub-especificación o bien colinealidad por sobre-especificación. Es un indicador volátil.