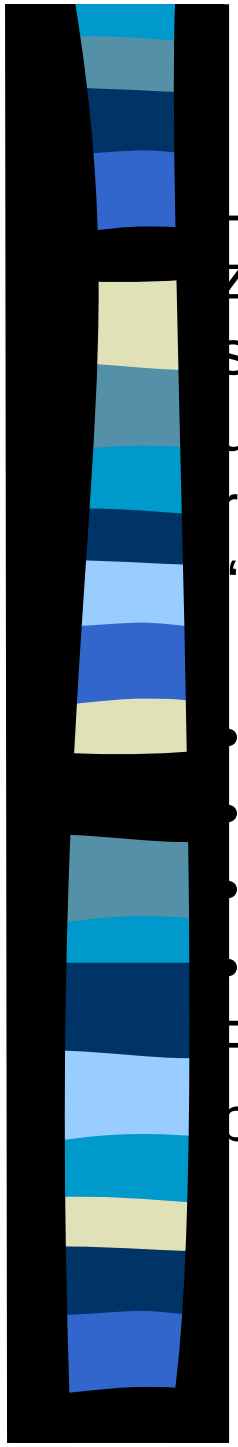
ESTADISTICA TECNICA

MODELOS LINEALES DE REGRESION

CLASE I I: REGRESION LINEAL MULTIPLE

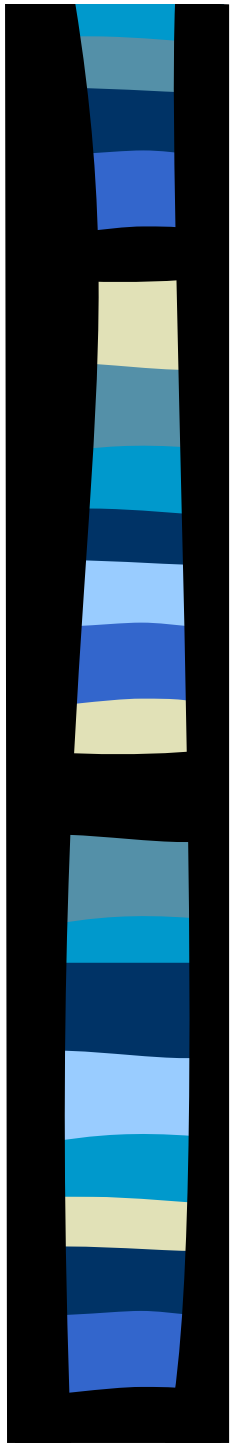


CASO DE DISCUSION I

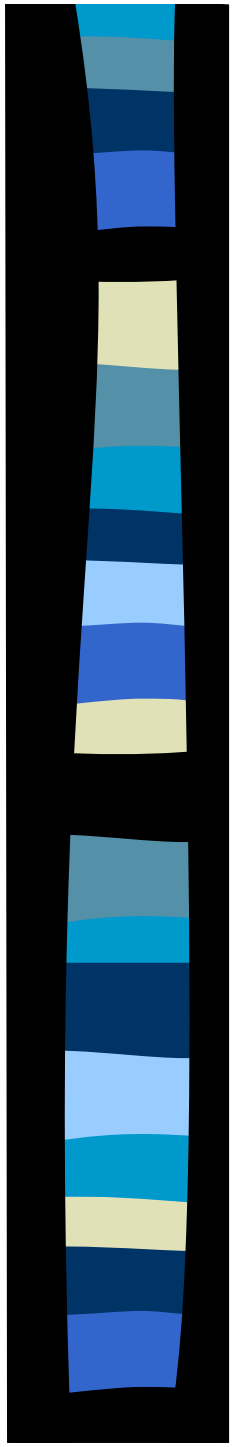
La Inmobiliaria GAUSS & MARKOV que opera en una zona residencial del Gran Buenos Aires desea mejorar su sistema de tasación basado en la experiencia por uno científico. Se propone adoptar un modelo de regresión lineal para explicar el valor de la propiedad "y" [K $], por las siguientes variables:

- X_1 : superficie cubierta [m2],
- X_2 : superficie descubierta [m2],
- X_3 : antigüedad del inmueble [años]
- X_4 : distancia al centro comercial [cuadras].

Para estimar al modelo se recopila información de 25 propiedades de la zona de influencia.



valor	m2 cub	m2 desc	antigüedad	distancia
y	x1	x2	x3	x4
135	173	92,5	32	2
115	160	91	66	7
77	94	59	33	3
115	169	91,5	69	7
200	234	125	14	7
48	49	36,5	48	18
110	145	85,5	59	9
160	226	128	66	7
170	221	120,5	20	14
80	115	66,5	65	17
135	180	103	51	8
55	62	42	40	6
70	83	53,5	25	2
150	181	99,5	33	3
45	40	25	16	16
93	125	67,5	27	16
113	140	81	16	10
88	117	68,5	53	8
50	44	37	37	11
149	204	115	66	6
60	77	52,5	44	10
167	228	129	32	20
55	74	44	57	11
145	215	113,5	70	17
120	157	83,5	57	5



ENFOQUES: **CONFIRMATORIO**

EXPLORATORIO

$$\tilde{y} = f(X_1 \quad X_2 \quad X_3 \quad \ldots \quad X_n) + \tilde{e}$$

$\tilde{y}$: Variable explicada o de respuesta. Se considera aleatoria.

X_i : Variables explicativas (o independientes). Se consideran como no aleatoria.

$\tilde{e}$: Perturbación o error. Es una variable aleatoria

MODELOS LINEALES

$$f(\vec{X}) = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

$$f(\vec{X}) = b_0 + b_1 X + b_2 X^2 + ... + b_k X^k$$

$$f(\vec{X}) = b_0 + b_1 \operatorname{sen}(X_1) + b_2 X_2$$

$$f(\vec{X}) = a X_1^{b_1} X_2^{b_2}$$

La variable x puede estar bajo formas no lineales

El modelo debe ser lineal en los parámetros

MODELOS LINEALES

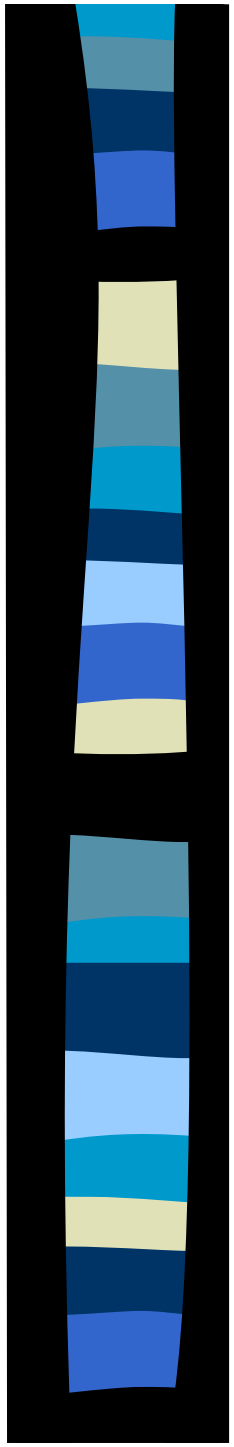
$$\tilde{y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki} + \tilde{e}_i$$

k varibles y ***p*** parámetros

$$\tilde{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{k1} \\ 1 & x_{12} & x_{22} & x_{k2} \\ & & & \\ 1 & x_{1n} & x_{2n} & x_{kn} \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ \\ b_k \end{bmatrix} \qquad \tilde{e} = \begin{bmatrix} e_1 \\ e_2 \\ \\ e_n \end{bmatrix}$$

n x 1 n x p p x 1 n x 1

$$\tilde{Y} = X.b + \tilde{e}$$



Supuestos del modelo:

$$E(\tilde{e}) = \vec{0}$$ Ausencia de vicio

$$Var(\tilde{e}) = s^2 \, I$$ Homocedasticidad

$$Cov(\tilde{e}_i ; \tilde{e}_j) = 0 \quad \forall \quad i \neq j$$ Ausencia de autocorrelación

De lo que se sigue:

$$E(\tilde{y}_i) = b_0 + b_1 \, X_{1i} + b_2 X_{2i} + ... + b_k X_{ki}$$

$$Var(\tilde{y}_i) = s^2 \quad \forall \quad i$$

ESTIMACION DE LOS PARAMETROS

$$e_i^{\,2} = (y_i - \hat{y}_i)^2$$

Aplicamos los mínimos cuadrados de Gauss:

$$Q = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

MINIMIZAR

$$Q = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \left(b_0 + b_j\, x_{j_i} \right) \right)^2$$

$$\frac{\partial Q}{\partial b_j} = 0 \qquad j : 0\ a\ k$$

ESTIMACION DE LOS PARAMETROS

$$B = \left(X^{t} X \right)^{-1} X^{t} Y$$

$$\hat{Y}_0 = B\vec{x}_0$$

VALIDACION DEL MODELO

Coeficiente de determinación

$$r^2 = \frac{\textit{Variación explicada}}{\textit{Variación total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{Q}{S_{yy}}$$

Variación total = Variación explicada + Variación Residual

SCtot = SC regresión + SC residuos

$$R^2 = \frac{SC\ regresión}{SC\ total}$$

ESTIMACION DE LA VARIANZA RESIDUAL

$$Var(\tilde{e}) = s^2 I$$

Se estima por:

$$S^2 = \frac{Q}{n-p}$$

$$\frac{n\ S^2}{s^2} \approx \quad c^2_{n-p}$$

VALIDACION DEL MODELO

Prueba de significación de los coeficientes de regresión

$$H_o : b_j = 0 \qquad j : 1 \; a \; k$$

Requerimos un supuesto teórico: $\tilde{e} \approx N(0; s^2)$

Se demuestra que :

$$b_j \approx N\left(b_j ; s^2{}_b\right)$$

Estadístico de prueba:

$$\frac{b_j - b_{j1}}{S_{bj}} \approx t_{(n-p)}$$

$$con \quad S_{bj} = s\, C_{jj}$$

$$\left(X^t X\right)^{-1} = \begin{bmatrix} C_{oo} & & & \\ & . & & \\ & & C_{jj} & \\ & & & . \\ & & & & C_{kk} \end{bmatrix}$$

INFERENCIA

INFERENCIA SOBRE LA MEDIA

$$E(\tilde{y} / \vec{x}_o) = b_0 + \sum_j b_j x_{j0}$$

Estadístico

$$t_n = \frac{\vec{x}_o.b - .\vec{x}_0.b}{s..\sqrt{\vec{x}^t{}_0.\left(X^t.X\right)^{-1}.\vec{x}_0}} \qquad n = n - p$$

Entonces como se desconoce la varianza, la expresión del intervalo es :

$$b_0 + \sum_j b_j x_{j0} \pm t_{n\,1-\frac{a}{2}} \sqrt{s^2\left(\frac{1}{n} + \vec{x}^t{}_o(X^t X)^{-1}\vec{x}_o\right)}$$

PREDICCION

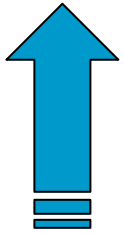
PREDICCION SOBRE UN VALOR EN PARTICULAR

$$\tilde{y} / x_o = b_0 + \sum_j b_j x_{j_o} + \tilde{e}$$

La expresión del intervalo es :

$$b_0 + \sum_j b_j x_{j0} \pm t_{n\,1-\frac{a}{2}} \sqrt{s^2 \left(1 + \frac{1}{n} + \vec{x}^t{}_o (X^t X)^{-1} \vec{x}_o \right)}$$

$$n = n - p$$



Refleja la varianza de la variable Y, además de la muestra.

ANALISIS EXPLORATORIO

$$\tilde{y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + \tilde{e}_i$$

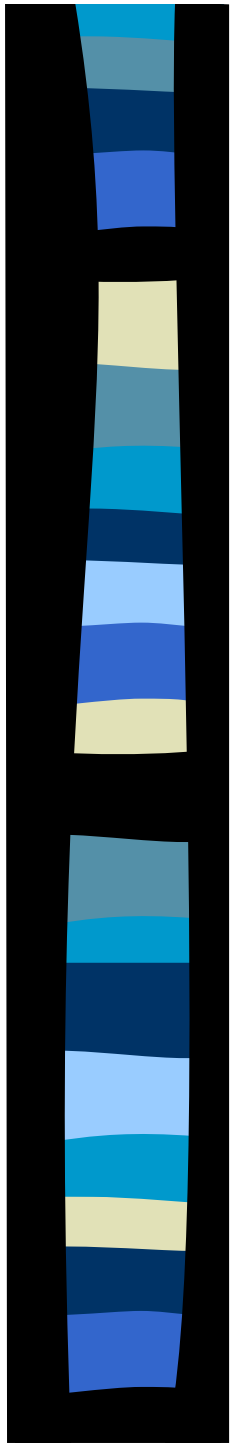
Para K variables hay 2^k-1 posibles:

¿Cuál es el mejor?

$$\tilde{y} = b_0 + b_1 X_1 + b_2 X_2 + \tilde{e}_i$$

$$\tilde{y} = b_0 + b_1 X_1 + \tilde{e}_i$$

$$\tilde{y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \tilde{e}_i$$



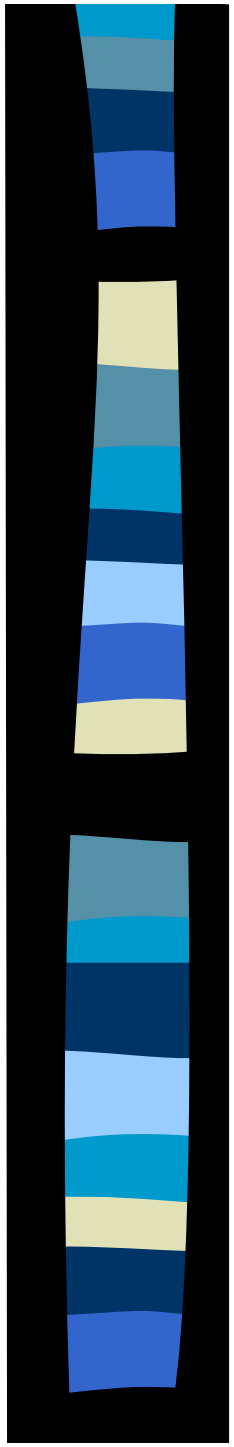
LAS TRES REGLAS DEL ANALISIS EXPLORATORIO

- **Un valor alto de R^2 es condición necesaria pero no suficiente para un buen ajuste.**

- **De los modelos con un R^2 aceptable, deberán considerarse como candidatos los de menor varianza (s^2) .**

- **Principio de Parsimonia:**

  En la selección del mejor modelo deberá tener prioridad la sencillez del mismo, dada por el menor número de variables explicativas.



MULTICOLINEALIDAD

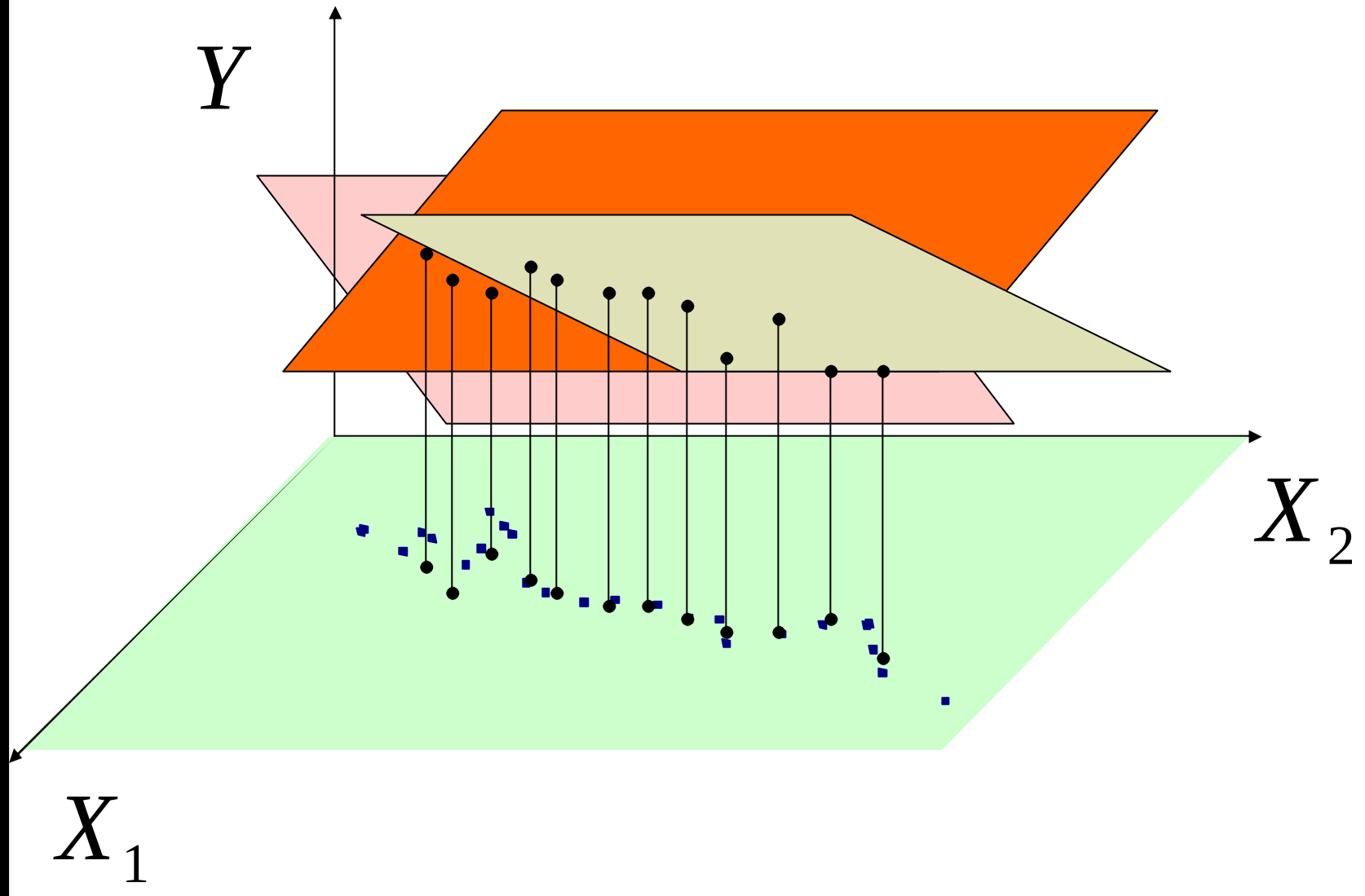
Es la existencia de asociaciones lineales ***aproximadas*** entre los datos de las variables ***explicativas***.

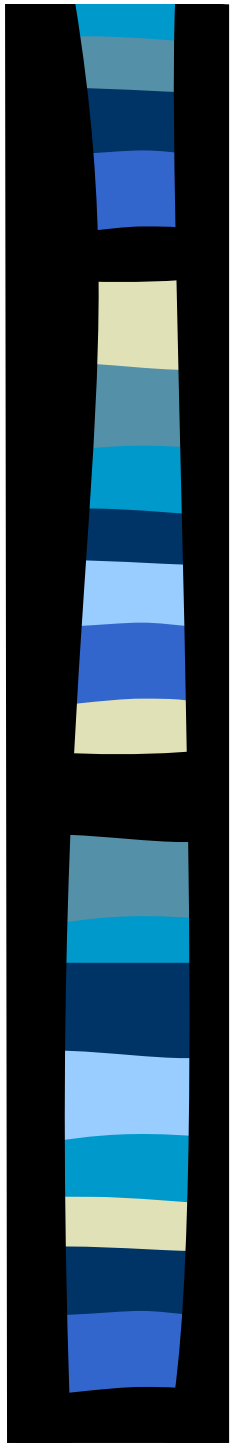
$$X_r = q\ X_s + \tilde{d}_i$$

Cuando hay multicolinealidad en un grado severo las variables aportan la misma información.
Alguna tendrá que ser eliminada......

$$X_1 = q\ X_2 + l\ X_4 + \tilde{d}_i$$

MULTICOLINEALIDAD





MULTICOLINEALIDAD

Consecuencias:

• Las estimaciones de los coeficientes de regresión tienen desvíos elevados y en consecuencia dan pruebas no significativas.

• Las estimaciones de los coeficientes de las variables explicativas pueden tomar signos contrarios a la naturaleza de las variables.

• El modelo es sumamente inestable frente a la inclusión de nuevos datos..

• No deberá extrapolarse

COMO DETECTARLA

Criterio “DET”

$$DET = \left\| \begin{bmatrix} 1 & r_{12} & & & r_{1k} \\ r_{21} & 1 & & r_{2i} & r_{2k} \\ & & \cdots & r_{ji} & \\ & & r_{ij} & 1 & .. \\ & & & .. & \\ r_{k1} & r_{k2} & & & 1 \end{bmatrix} \right\|$$

r_{ij} Coeficente de correlación entre $\boldsymbol{x_i}$ y $\boldsymbol{x_j}$

DET < 0.1

DET > 0.1