

# Regresión Lineal Múltiple

# 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:

X1 : superficie cubierta [m2],

X2 : superficie descubierta [m2],

X3 : antigüedad del inmueble [años]

X4 : 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	115	85,5	50	0

# Regresión Lineal Múltiple

## Objetivo

$$\tilde{y} = f(x | \theta_1, \dots, \theta_p) + \tilde{\varepsilon}$$

$$\tilde{y}_i = f(x_{1i}, \dots, x_{ki} | \theta_1, \dots, \theta_p) + \tilde{\varepsilon}_i$$

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

$x_{1i}, \dots, x_{ki}$  : Conjunto de  $k$  variables explicativas

$\theta_1, \dots, \theta_p$  : Parámetros del modelo ( $p$  parámetros)

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

# Regresión Lineal Múltiple

## Transformaciones

$$f(\vec{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

$$f(\vec{X}) = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k$$

$$f(\vec{X}) = \beta_0 + \beta_1 \sin(X_1) + \beta_2 X_2$$

$$f(\vec{X}) = \alpha X_1^{\beta_1} X_2^{\beta_2}$$

# Regresión Lineal Múltiple

## Planteo Matricial

$$\tilde{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \tilde{\varepsilon}_i \quad k \text{ variables y } p \text{ parámetros}$$

$$\begin{array}{ccccccc} \tilde{\mathbf{Y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} & \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} & \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} & \tilde{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} & & & \tilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \\ n \times 1 & n \times p & p \times 1 & n \times 1 & & & \end{array}$$

En la materia utilizaremos siempre las expresiones para variables sin centrar

# Planteo Matricial

## Modelo Muestral

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki}$$

**k** variables y **p** parámetros

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

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$n \times 1$

$n \times p$

$p \times 1$

$n \times 1$

**En la materia utilizaremos siempre las expresiones para variables sin centrar**

# Método de mínimos cuadrados

## Planteo Matricial

$$Q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left( y_i - b_0 - \sum_{j=1}^k b_j x_{ji} \right)^2$$

$$\frac{\partial Q}{\partial b_j} = 0 \quad \forall j : 0 \dots k \quad \longrightarrow \quad \begin{array}{l} \text{Sistema de } k=p+1 \text{ ecuaciones con } k=p+1 \text{ incógnitas} \\ \text{Obtenemos } b_j \forall j: 0 \dots k \end{array}$$

# Método de mínimos cuadrados

## Planteo Matricial

$$e_i = y_i - \hat{y}_i$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

$$Q = \sum_{i=1}^n e_i^2 = \mathbf{e}^t \mathbf{e} = (\mathbf{Y} - \mathbf{X}\mathbf{b})^t (\mathbf{Y} - \mathbf{X}\mathbf{b})$$

$$\frac{\partial Q}{\partial \mathbf{b}} [(\mathbf{Y} - \mathbf{X}\mathbf{b})^t (\mathbf{Y} - \mathbf{X}\mathbf{b})] = 0$$

$$\frac{\partial Q}{\partial \mathbf{b}} [\mathbf{Y}^t \mathbf{Y} - \mathbf{b}^t \mathbf{X}^t \mathbf{Y} - \mathbf{Y}^t \mathbf{X} \mathbf{b} + \mathbf{b}^t \mathbf{X}^t \mathbf{X} \mathbf{b}] = 0$$

$$\frac{\partial Q}{\partial \mathbf{b}} [\mathbf{Y}^t \mathbf{Y} - 2\mathbf{b}^t \mathbf{X}^t \mathbf{Y} + \mathbf{b}^t \mathbf{X}^t \mathbf{X} \mathbf{b}] = 0$$

$$-2\mathbf{X}^t \mathbf{Y} + 2(\mathbf{X}^t \mathbf{X}) \mathbf{b} = 0$$

$$(\mathbf{X}^t \mathbf{X}) \mathbf{b} = \mathbf{X}^t \mathbf{Y}$$

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$$



# Regresión Lineal Múltiple

## Supuestos del modelo

$$E(\vec{\epsilon}) = \vec{0}$$

Ausencia de vicio

$$\text{Var}(\vec{\epsilon}) = \sigma^2 \mathbf{I}$$

Homocedasticidad

$$\text{Cov}(\tilde{\epsilon}_i; \tilde{\epsilon}_j) = 0 \quad \forall i \neq j$$

Ausencia de autocorrelación

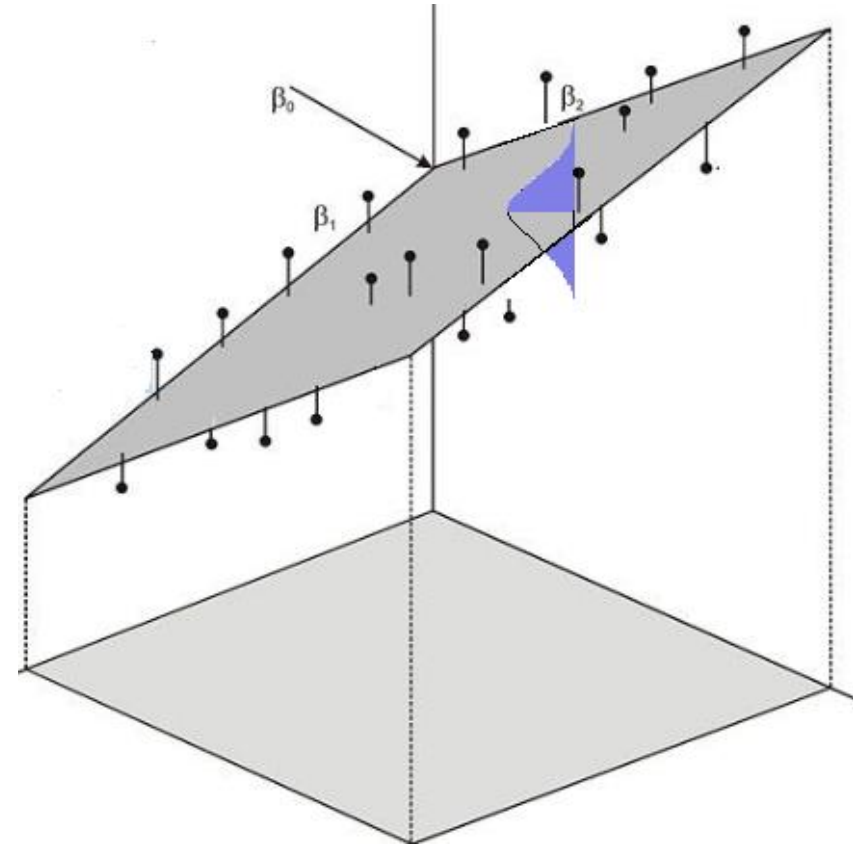
$$\vec{\epsilon} \square \text{Normal}(\vec{0}; \sigma^2 \mathbf{I})$$

Normalidad de las perturbaciones

De lo que se deduce que:

$$E(\tilde{y}_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} \quad \forall i: 1..n$$

$$\text{Var}(\tilde{y}_i) = \sigma^2 \quad \forall i: 1..n$$



# Regresión Lineal Múltiple

## Tipos de Análisis

### **Análisis Confirmatorio**

Se busca establecer la validez de un modelo de Regresión

### **Análisis Exploratorio**

Se exploran los distintos modelos que se pueden generar con un conjunto de variables explicativas y establecer cuál de ellos es el mejor

# Regresión Lineal Múltiple

ANÁLISIS CONFIRMATORIO

# Regresión Lineal Múltiple

## Análisis Confirmatorio

- Análisis de la Bondad de Ajuste
- Análisis de Multicolinealidad
- Diagnóstico de residuos
- Validación de supuestos

# Análisis Confirmatorio

BONDAD DE AJUSTE

# Regresión Lineal Múltiple

## Análisis Confirmatorio

- Análisis de la Bondad de Ajuste
- Análisis de Multicolinealidad
- Diagnóstico de residuos
- Validación de supuestos

# Regresión lineal múltiple

## Bondad de ajuste

### Procedimientos para establecer la bondad de ajuste

- **Coeficiente de correlación  $R$ :** Solo lo utilizaremos en regresión lineal simple
- **Coeficiente de determinación  $R^2$ :** Se interpreta de la misma manera que en regresión lineal simple.
- **Varianza de los residuos  $S^2$ :** Es una medida de bondad de ajuste, pero es difícil establecer los límites para determinar si un  $S^2$  es alto o bajo.
- **Test de significación para los coeficientes de regresión ( $\beta_i$ ):** Se utilizan los mismos principios que en Regresión lineal simple, para cada uno de los coeficientes  $\beta_i$

# Bondad de Ajuste

## Coeficiente de Determinación

***Variación total = Variación explicada + Variación Residual***

***SC<sub>tot</sub> = SC regresión + SC residuos***

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

- Un valor alto de  $R^2$  es condición necesaria pero no suficiente para un buen ajuste.
- La inclusión de variables explicativas siempre eleva el  $R^2$  aunque la variable no aporte información



# Bondad de Ajuste

Estimación de la varianza residual

$$\text{Var}(\tilde{\boldsymbol{\varepsilon}}) = \sigma^2 \mathbf{I}$$

Se estima con

$$S^2 = \frac{Q}{n-p} = \frac{(\mathbf{Y} - \mathbf{X}\mathbf{b})^t (\mathbf{Y} - \mathbf{X}\mathbf{b})}{n-p} = \frac{\mathbf{Y}^t \mathbf{Y} - \mathbf{b}^t \mathbf{X}^t \mathbf{Y}}{n-p}$$

$$\frac{(n-p) S^2}{\sigma^2} \square \chi_{n-p}^2$$

# Regresión Lineal Múltiple

Media y varianza de los coeficientes de regresión

$$\mathbf{E}(\mathbf{b}) = \mathbf{E}\left[\left(\mathbf{X}^t \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{Y}\right] = \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{E}(\mathbf{Y}) = \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}$$

$$\text{Var}(\mathbf{b}) = \begin{bmatrix} \sigma_{b_0}^2 & \text{Cov}(b_0; b_1) & \cdots & \text{Cov}(b_0; b_k) \\ \text{Cov}(b_1; b_0) & \sigma_{b_1}^2 & \cdots & \text{Cov}(b_1; b_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(b_k; b_0) & \text{Cov}(b_k; b_1) & \cdots & \sigma_{b_k}^2 \end{bmatrix} = \sigma^2 \left(\mathbf{X}^t \mathbf{X}\right)^{-1}$$

# Regresión Lineal Múltiple

Inferencia sobre los coeficientes de regresión

$$H_0) \beta_j = \beta_{j0} \quad \text{vs} \quad \begin{array}{l} H_1) \beta_j \neq \beta_{j0} \\ H_1) \beta_j < \beta_{j0} \\ H_1) \beta_j > \beta_{j0} \end{array} \quad \forall j : 1 \dots k$$

$$\frac{b_j - \beta_{j0}}{S_{bj}} \square t_{(n-p)} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad S_{bj} = S \sqrt{C_{jj}} \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} C_{00} & & & \\ & C_{11} & & \\ & & \ddots & \\ & & & C_{kk} \end{bmatrix}$$

# Regresión Lineal Múltiple

## Tabla ANOVA

	Suma Cuadrados	Grados libertad	Cuadrados medios
Regresión	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$p - 1$	$CM_{Reg} = [\mathbf{b}^t \mathbf{X}^t \mathbf{Y} - (\sum y_i)^2 / n] / (p - 1)$
Residual	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - p$	$CM_{Res} = (\mathbf{Y}^t \mathbf{Y} - \mathbf{b}^t \mathbf{X}^t \mathbf{Y}) / (n - p)$
Total	$\sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$	

$$H_o) \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$CR: F = \frac{CM_{Reg}}{CM_{Res}} > F_{p-1; n-p} \quad 1-\alpha$$

# Pronósticos

BONDAD DE AJUSTE

# Modelos de Regresión Múltiple

Estimación puntual

$$\hat{\mathbf{y}}_o = \mathbf{x}_o^t \mathbf{b} = \begin{bmatrix} 1 & x_{1o} & x_{2o} & \cdots & x_{ko} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = b_0 + \sum_{j=1}^k b_j x_{jo}$$

# Modelos de Regresión Múltiple

## Intervalos de Confianza

$$E(\tilde{y}|\vec{\mathbf{x}}_o) = \beta_0 + \sum_{j=1}^k \beta_j x_{0j}$$

$$t_{n-p} = \frac{\mathbf{x}_o^t \mathbf{b} - \mathbf{x}_o^t \boldsymbol{\beta}}{S \sqrt{\mathbf{x}_o^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_o}}$$

$$\begin{aligned}\mathbf{x}_o^t \boldsymbol{\beta} &= \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_k x_{0k} \\ \mathbf{x}_o^t \mathbf{b} &= b_0 + b_1 x_{01} + b_2 x_{02} + \dots + b_k x_{0k}\end{aligned}$$


La expresión del intervalo es :

$$b_0 + \sum_j b_j x_{j0} \pm t_{n-p;1-\alpha/2} \sqrt{S^2 \left( \mathbf{x}_o^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_o \right)}$$

# Modelos de Regresión Múltiple

## Intervalos de predicción

$$\tilde{y}|\vec{\mathbf{x}}_0 = \beta_0 + \sum_{j=1}^k \beta_j x_{0j} + \tilde{\varepsilon}$$

$$t_{n-p} = \frac{\mathbf{x}_0^t \mathbf{b} - \mathbf{x}_0^t \boldsymbol{\beta}}{S \sqrt{1 + \mathbf{x}_0^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_0}}$$


$$\begin{aligned}\mathbf{x}_0^t \boldsymbol{\beta} &= \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_k x_{0k} \\ \mathbf{x}_0^t \mathbf{b} &= b_0 + b_1 x_{01} + b_2 x_{02} + \dots + b_k x_{0k}\end{aligned}$$

Refleja la varianza de la variable Y

La expresión del intervalo es :

$$\underbrace{b_0 + \sum_{j=1}^k b_j x_{0j}}_{\hat{y}(\vec{\mathbf{x}}_0)} \pm t_{n-p; 1-\alpha/2} \sqrt{S^2 \left( 1 + \vec{\mathbf{x}}_0^t (\mathbf{X}^t \mathbf{X})^{-1} \vec{\mathbf{x}}_0 \right)}$$



# Análisis Confirmatorio

MULTICOLINEALIDAD

# Regresión Lineal Múltiple

## Análisis Confirmatorio

- Análisis de la Bondad de Ajuste
- **Análisis de Multicolinealidad**
- Diagnóstico de residuos
- Validación de supuestos

# Multicolinealidad

## Definición

La multicolinealidad es la presencia de asociaciones lineales entre las variables explicativas de un modelo de regresión

- La multicolinealidad no es un problema de presencia o ausencia sino de grado
- Su origen puede ser muestral o poblacional

# Multicolinealidad

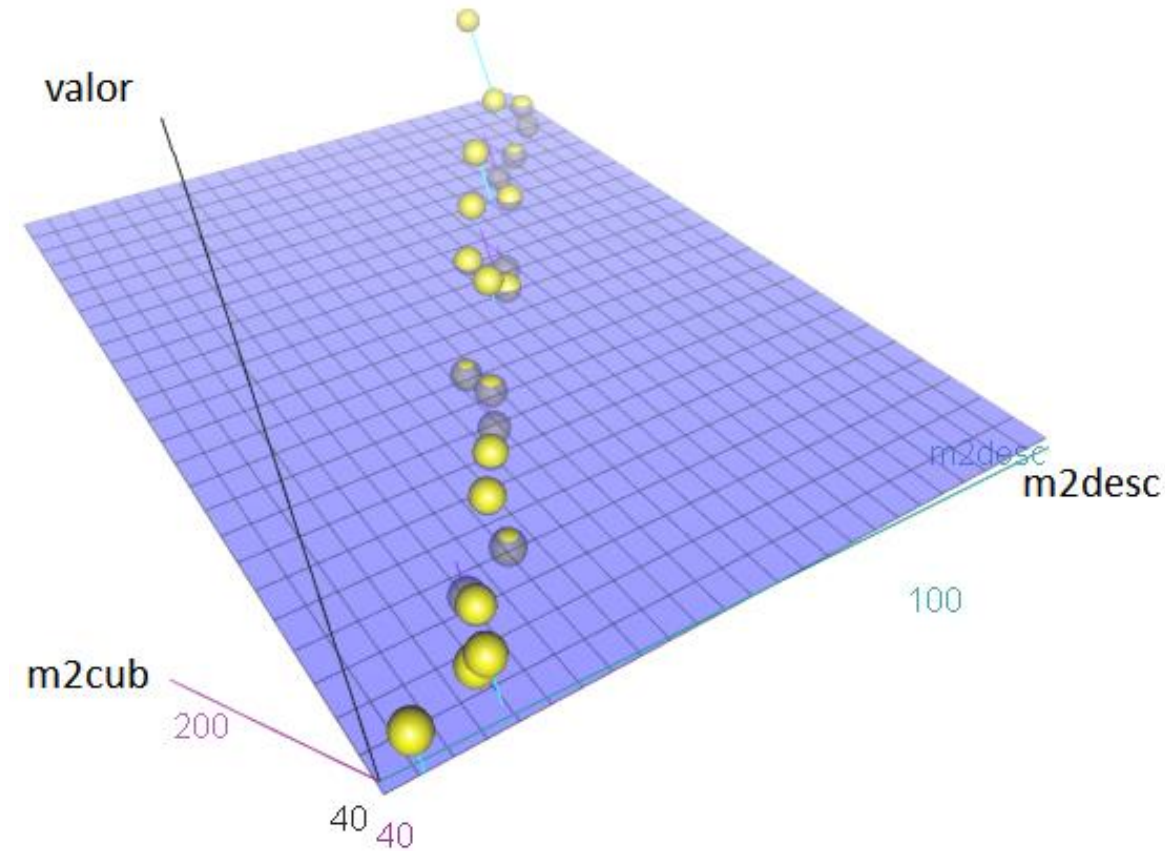
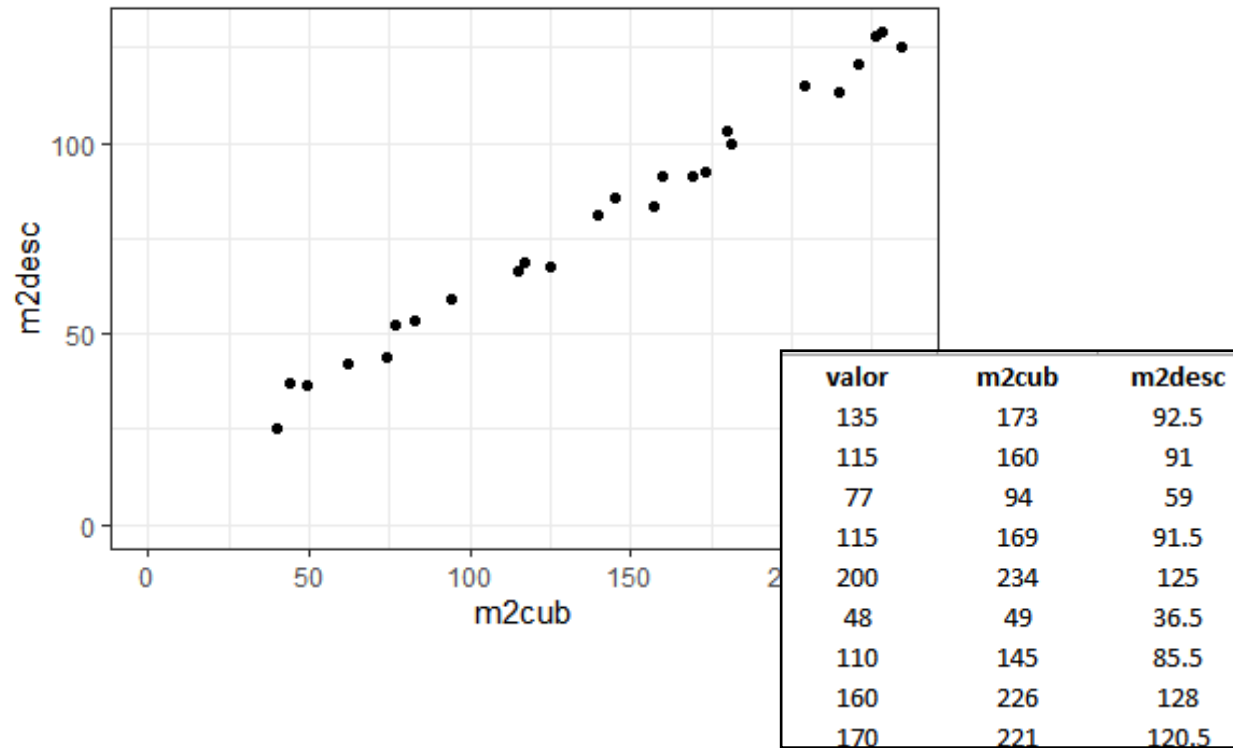
## Definición

### Grados de Multicolinealidad

- La **multicolinealidad perfecta** se produce cuando dos o mas variables explicativas son linealmente dependientes. En estos casos los modelos de regresión no son estimables.
- La **ausencia de multicolinealidad** se produce si todas las variables explicativas son independientes entre sí. Solo se aplica a casos diseñados experimentalmente.
- En la mayoría de los casos nos encontraremos en situaciones intermedias donde lo que intentamos saber es si la multicolinealidad es lo suficientemente alta para afectar los resultados.

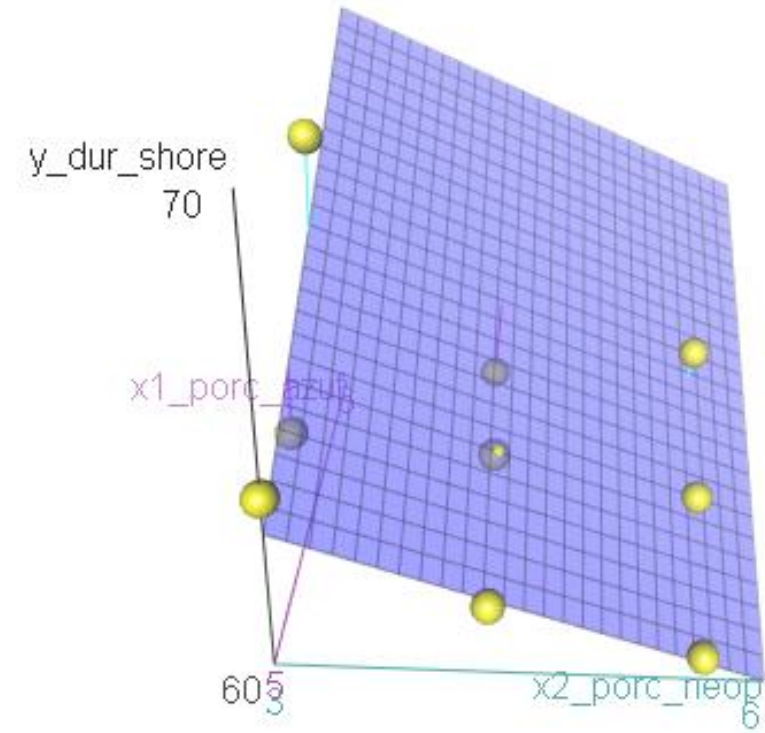
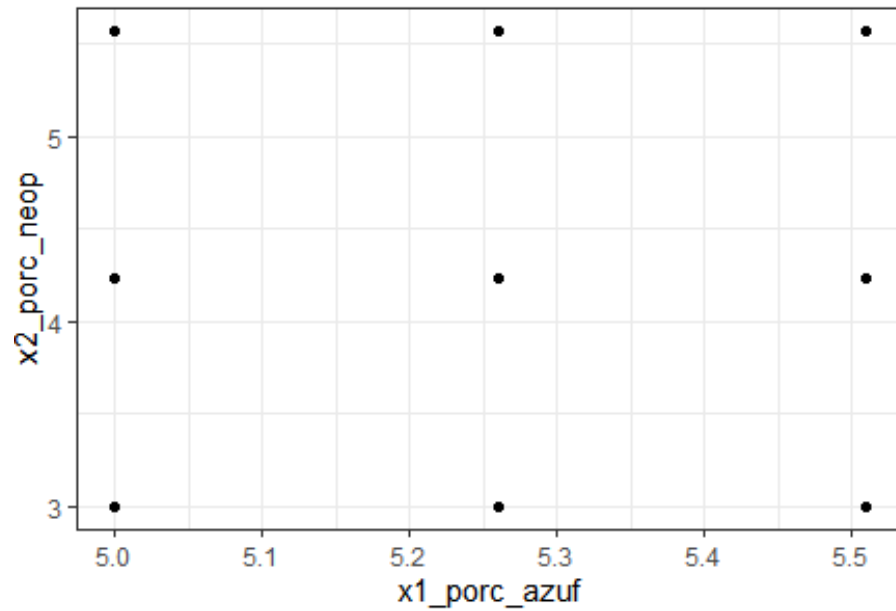
# Multicolinealidad

Ejemplo de multicolinealidad elevada



# Multicolinealidad

## Ejemplo de ausencia de multicolinealidad



Es posible diseñar un experimento de manera que un modelo de regresión múltiple tenga multicolinealidad nula

# Multicolinealidad

## Efectos adversos de la multicolinealidad

El efecto principal de la multicolinealidad es el aumento en la varianza estimada de los coeficientes de regresión. Esto a su vez trae varios problemas:

- Test de significación para los  $\beta_i$  que pueden resultar no significativos, aún cuando la variable tenga es importante y se tienen valores de  $R^2$  altos
- Las estimaciones de los coeficientes pueden presentar signos distintos a los esperados y magnitudes poco razonables. No podemos interpretar estos coeficientes
- Pequeños cambios en los datos o en la especificación provocan grandes cambios en las estimaciones de los coeficientes
- Intervalos de confianza de los coeficientes de regresión son más amplios de lo que deberían ser
- El modelo es sumamente inestable frente a la inclusión de nuevos datos
- Es muy riesgoso extrapolar

# Multicolinealidad

## Diagnóstico

### Matriz de correlaciones de las variables explicativas

- El análisis de la matriz de correlaciones permite identificar dependencias entre variables explicativas
- El análisis de esta matriz no es suficiente ya que no permite identificar relaciones que afecten tres o mas variables explicativas

$$\begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$



# Multicolinealidad

## Diagnóstico

### DET

$$DET = \begin{vmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{vmatrix} \quad 0 \leq DET \leq 1$$

$r_{ij}$  Coeficiente de Correlación entre  $x_i$  y  $x_j$

- El determinante de la matriz de correlaciones es un buen indicador global de multicolinealidad
  - **DET  $\geq 0,20$** : la multicolinealidad no es lo suficientemente elevada para afectar las estimaciones
  - **$0,10 \leq DET \leq 0,20$** : se debe proceder con cautela
  - **DET  $\leq 0,10$** : la multicolinealidad es severa y afecta las estimaciones

DET > 0.2

0,1 < DET < 0.2

DET < 0.1

# Multicolinealidad

## Diagnóstico

### Factores de Inflación de Varianza (VIF)

$$D^2(\beta_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_e^2}{(n-1)S_j^2}$$

$$VIF_j = \frac{1}{1 - R_j^2}$$

**Propiedad** Los VIF son los elementos de la diagonal principal de la inversa de la matriz de correlaciones

$R_j^2$  Coeficiente de determinación donde  $x_j$  es la variable explicada y el resto las variables explicativas

- **VIF < 5:** la multicolinealidad no es lo suficientemente elevada para afectar las estimaciones
- **$5 \leq \text{VIF} \leq 10$ :** se debe proceder con cautela
- **VIF > 10:** la multicolinealidad es severa y afecta las estimaciones

VIF < 5

$5 < \text{VIF} < 10$

DET > 10

# Multicolinealidad

## Diagnóstico

- Si el problema es muestral, cambiar la muestra o agregar observaciones cuyas valores de  $x_i$  no estén muy correlacionados con los anteriores, puede resolver el problema
- Si el problema es poblacional y además las predicciones se van a realizar cerca del baricentro, la multicolinealidad no debería ser un obstáculo importante. En este caso, deben observarse los  $h_{ii}$  detenidamente.
- Si se observan dos variables muy correlacionadas, quitar alguna de ellas puede ser una solución
- Otros procedimientos de estimación (Estadística Aplicada III)
  - Regresión Ridge
  - Componentes Principales

# Análisis Confirmatorio

DIAGNÓSTICO DE RESIDUOS

# Regresión Lineal Múltiple

## Análisis Confirmatorio

- Análisis de la Bondad de Ajuste
- Análisis de Multicolinealidad
- **Diagnóstico de residuos**
- Validación de supuestos

# Regresión Lineal Múltiple

## Diagnóstico

- Outliers
- Alto Leverage
- Influencia

Se utilizan las mismas métricas que en Regresión Lineal Simple y se interpretan de igual forma.

Profundizaremos sobre el concepto de Leverage en Regresión Lineal Múltiple.

# Regresión Lineal Múltiple

Error estándar de la predicción y la matriz H

$$\hat{\mathbf{y}}_o = \mathbf{x}_o^t \mathbf{b} = \begin{bmatrix} 1 & x_{1o} & x_{2o} & \cdots & x_{ko} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = b_0 + \sum_{j=1}^k b_j x_{jo}$$

$$\hat{\mathbf{Y}} = \mathbf{X} \mathbf{b} = \underbrace{\mathbf{X} (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t}_{\mathbf{H}} \mathbf{Y} = \mathbf{H} \mathbf{Y}$$

$$\mathbf{H} = \mathbf{X} (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H} \mathbf{Y} = (\mathbf{I} - \mathbf{H}) \mathbf{Y}$$

$$h_{ii} = \mathbf{h} = \mathbf{x}_i^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_i$$

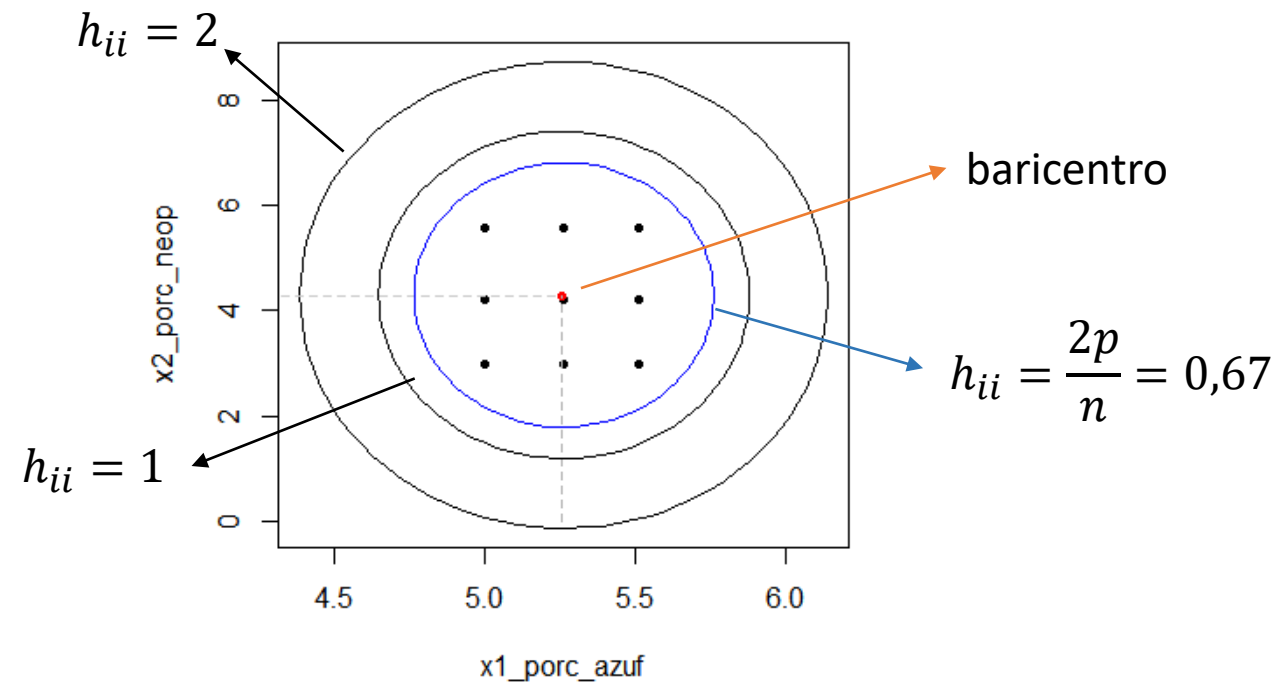
## Propiedades de H

- Tiene dimensión  $n \times n$
- Simétrica
- Idempotente
- $\text{Traza}(\mathbf{H}) = p$
- Es una matriz de proyección de los valores observados en valores pronosticados

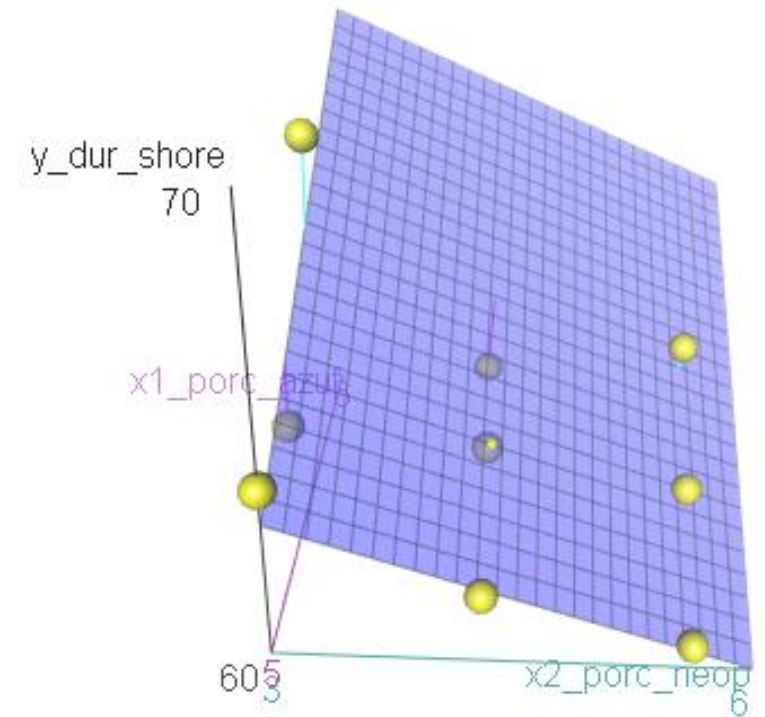
$$\text{En RLS } h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

# Leverage y Multicolinealidad

Ejemplo de ausencia de multicolinealidad



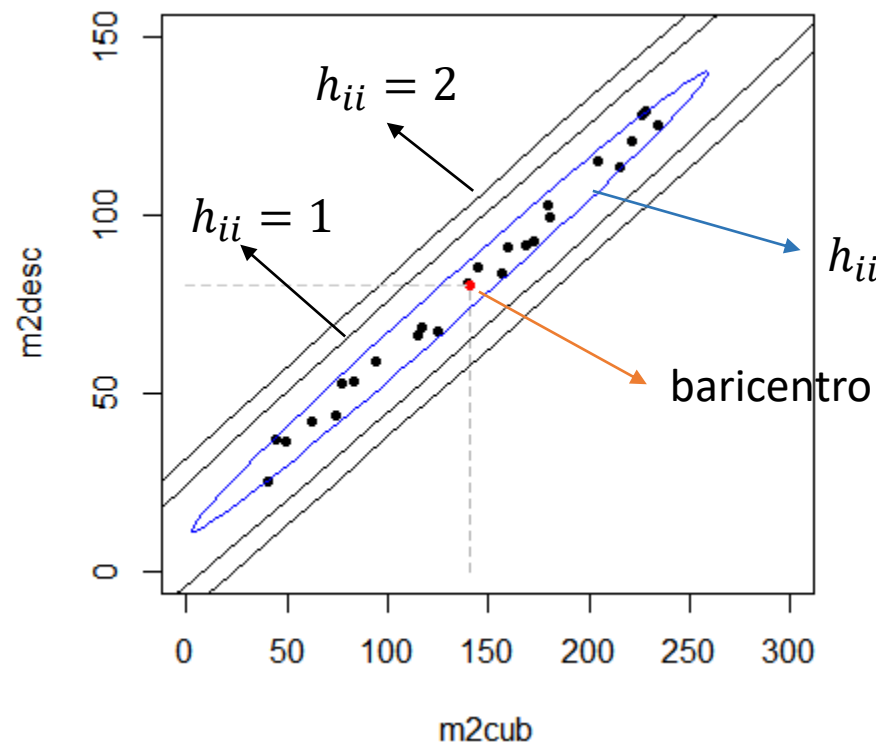
$$\begin{pmatrix} 1 & x & y \end{pmatrix} \begin{pmatrix} 72.76 & -13.47 & -0.43 \\ -13.47 & 2.56 & 0 \\ -0.43 & 0 & 0.1 \end{pmatrix} \begin{pmatrix} 1 \\ x \\ y \end{pmatrix} = \begin{pmatrix} 2.56x^2 - 26.94x + 0.1y^2 - 0.86y + 72.76 \end{pmatrix}$$



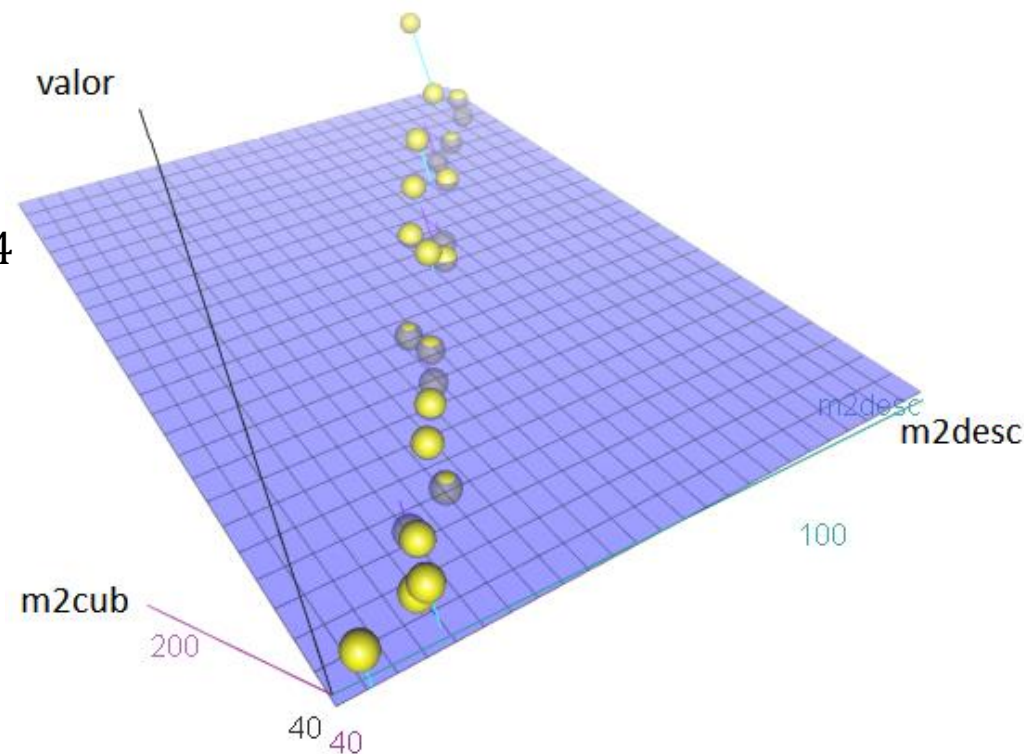


# Leverage y Multicolinealidad

Ejemplo de ausencia de multicolinealidad

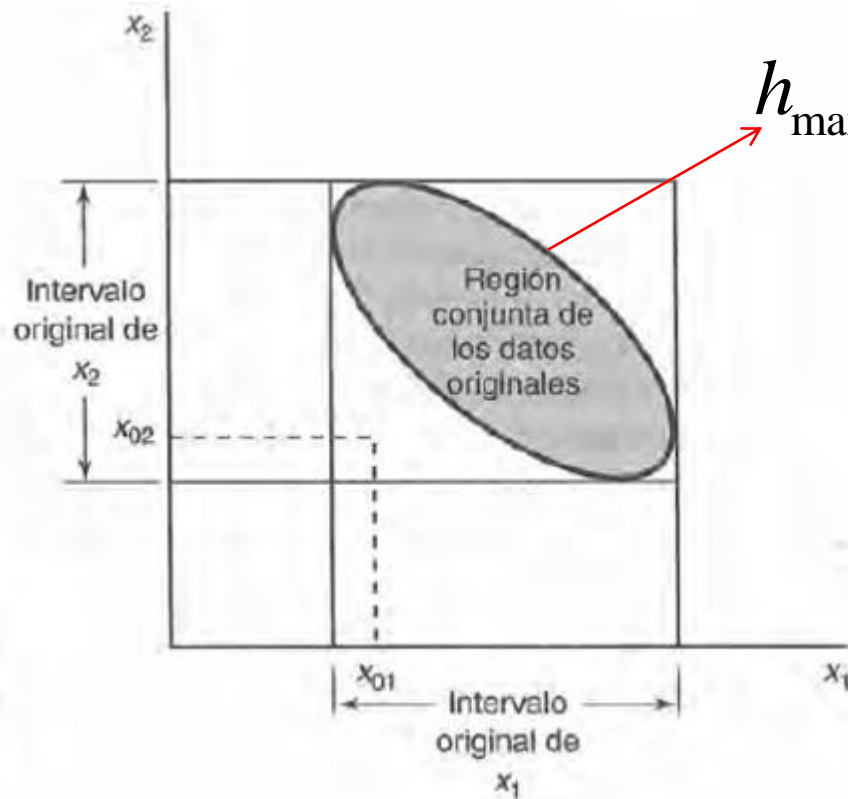


$$h_{ii} = \frac{2p}{n} = 0,24$$



$$\begin{pmatrix} 1 & x & y \end{pmatrix} \begin{pmatrix} 0.6422 & 0.01785 & -0.03868 \\ 0.01785 & 0.000959 & -0.001897 \\ -0.03868 & -0.001897 & 0.003797 \end{pmatrix} \begin{pmatrix} 1 \\ x \\ y \end{pmatrix} = 0.000959x^2 - 0.003794xy + 0.0357x + 0.003797y^2 - 0.07736y + 0.6422$$

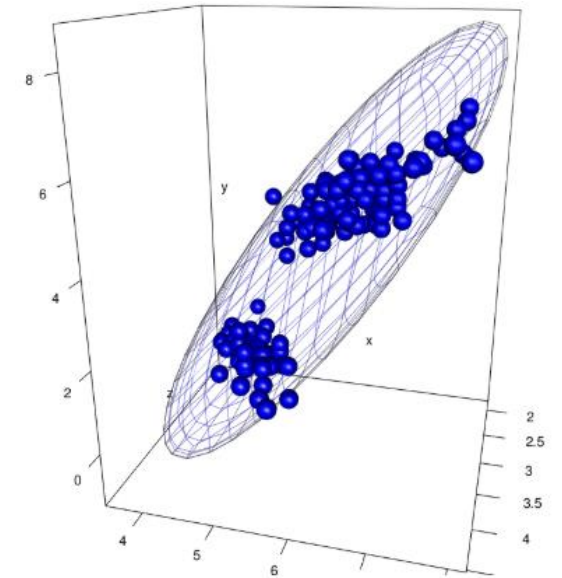
# Leverage y Extrapolación



$$h_{\max} = \mathbf{x}_j^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_j$$

Estamos extrapolando si:

$$\mathbf{x}_o^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_o \geq h_{\max}$$



Si  $k = 3$

# Análisis Exploratorio

# Análisis Exploratorio

## Definición

A partir de una variable respuesta  $y$  y un conjunto de  $k$  variables explicativas  $x_j$  (con  $j: 1..k$ ), se trata de encontrar el modelo que mejor se ajuste a los datos observados

- Si se tienen  $k$  variables explicativas, la cantidad de modelos posibles es  $2^k - 1$

¿Cuál modelo será el mejor?

Si se tienen  $k$  variables

explicativas:  $x_1, x_2, x_3, \dots, x_k$

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \tilde{\varepsilon}_i$$

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \tilde{\varepsilon}_i$$

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \tilde{\varepsilon}_i$$

$$\tilde{y} = \dots$$

# Análisis Exploratorio

## Metodología

1. Estimar todos los posibles modelos
2. Calcular distintos indicadores de calidad de los modelos
  - Descartar modelos incompletos o con elevada multicolinealidad
  - Ordenar los modelos y seleccionar los mejores

A partir de este análisis conseguimos uno o mas modelos candidatos.  
Para terminar de confirmar a estos modelos candidatos:

- Análisis de la Bondad de Ajuste
- Diagnóstico de residuos
- Validación de supuestos

# Análisis Exploratorio

## Ordenamiento y Descarte

### Ordenadores de modelos

Nos permiten comparar modelos y determinar cual es mejor de acuerdo con algún criterio:

- $R^2$  ajustado
- $S^2$
- PRESS

### Indicadores de descarte

Miden características indeseables y nos indican si el modelo tiene problemas que pueden invalidar el mismo

- $DET$
- $C_p$

# Comparación de Modelos de Regresión

## $R^2$ Ajustado

El  $R^2$  siempre aumenta al agregar variables al modelo. Esto se debe a que  $T$  no varía (solo depende de la variable respuesta) y  $Q$  siempre se reduce, aunque estas variables no aporten información. Por lo tanto, si aumenta el número de variables en el modelo va a aumentar artificialmente el valor de  $R^2$ .

$$R^2 = 1 - \frac{Q}{T} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

*Como consecuencia*

- Un  $R^2$  no es un buen comparador de modelos ya que nos indicará que el modelo con mas variables es mejor
- $R^2$  no sirve como “ordenador” de modelos

# Comparación de Modelos de Regresión

## $R^2$ Ajustado

Para poder comprar modelos se corrige la expresión del  $R^2$  teniendo en cuenta los grados de libertad de  $T$  y  $Q$ , dando lugar al  $R^2$  *ajustado*:

$$R^2_{Ajustado} = 1 - \frac{S_e^2}{S_y^2} = 1 - \frac{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n-p}}{\sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n-1}} = 1 - (1 - R^2) \left( \frac{n-1}{n-p} \right)$$

- $R^2$  *ajustado* permite comparar modelos de regresión que contengan distinta cantidad de parámetros
- Se dice que  $R^2$  *ajustado* es un indicador parsimonioso
- El orden de modelos obtenido con  $R^2$  *ajustado* es el mismo que el obtenido con  $S^2$



# Comparación de Modelos de Regresión

## PRESS

El *PRESS* es la suma de los residuos PRESS al cuadrado:

$$PRESS = \sum_{i=1}^n e_{-i}^2 = \sum_{i=1}^n (y_i - \hat{y}_{-i})^2$$

$\hat{y}_{-i}$ : pronóstico para  $y_t$  obtenido a partir del modelo estimado sin el dato  $(\mathbf{x}_i; y_i)$

- A menor valor de *PRESS*, mejor es el modelo
- El *PRESS* es un indicador que mide la **capacidad de predicción** de un modelo. Cuando se va a utilizar el modelo para realizar predicciones, es un ordenador a tener presente

# Comparación de Modelos de Regresión

## CP

$CP$  es un indicador que nos mide la ausencia de variables importantes en el modelo. Se lo usa para eliminar modelos demasiado simplificados, que dejan de lado variables con poder explicativo:

$$CP = \frac{Q_P}{S_k^2} - n + 2p$$

$Q_P$ : Suma de cuadrados residual del modelo

$S_k^2$ : Varianza residual del modelo completo, con todas las variables del sistema

$n$ : Tamaño de muestra

$p$ : Número de parámetros del modelo

- $CP$  debería tomar valores cercanos a  $p$  para modelos completos, es decir, sin ausencia de variables importantes (entre las  $x_j$  disponibles inicialmente)
- En la medida que al modelo le falten variables importantes  $CP$  aumenta
- Si  $CP > 5P$  se puede considerar que el modelo es incompleto, en el sentido que le falta incluir variables importantes

**$CP/P < 5$**

**$CP/P > 5$**

# Análisis Exploratorio

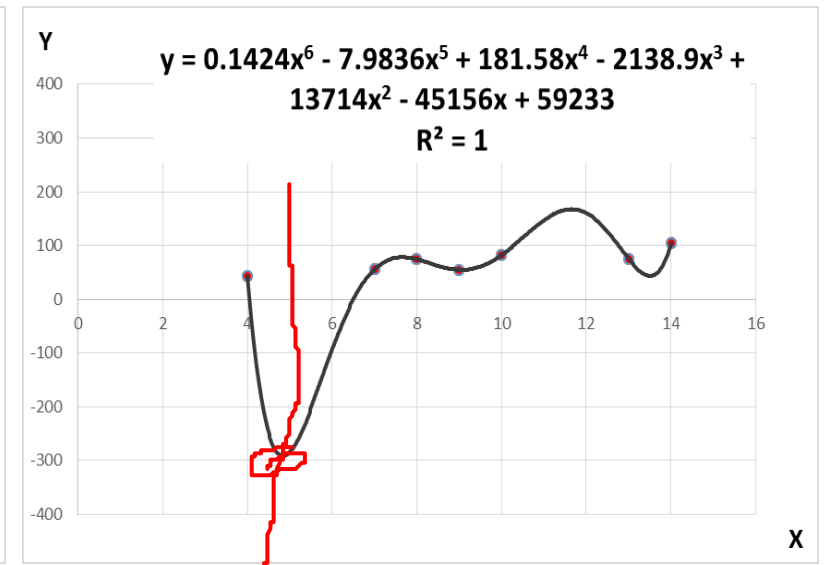
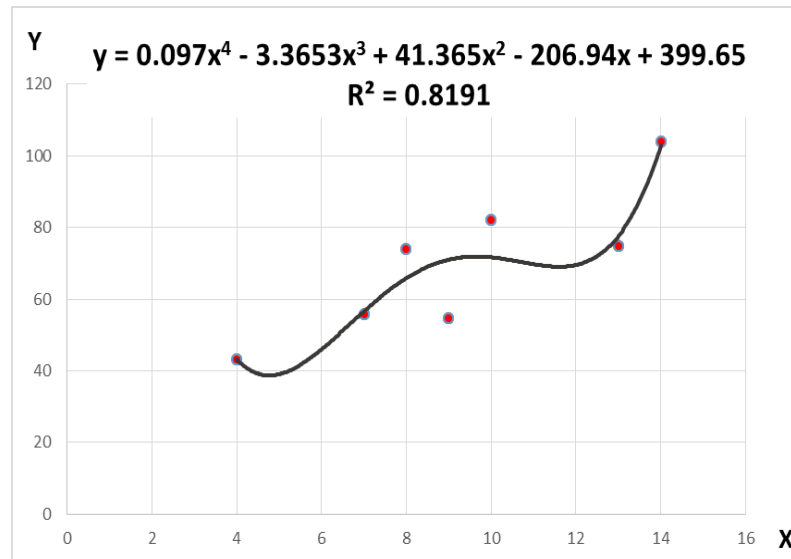
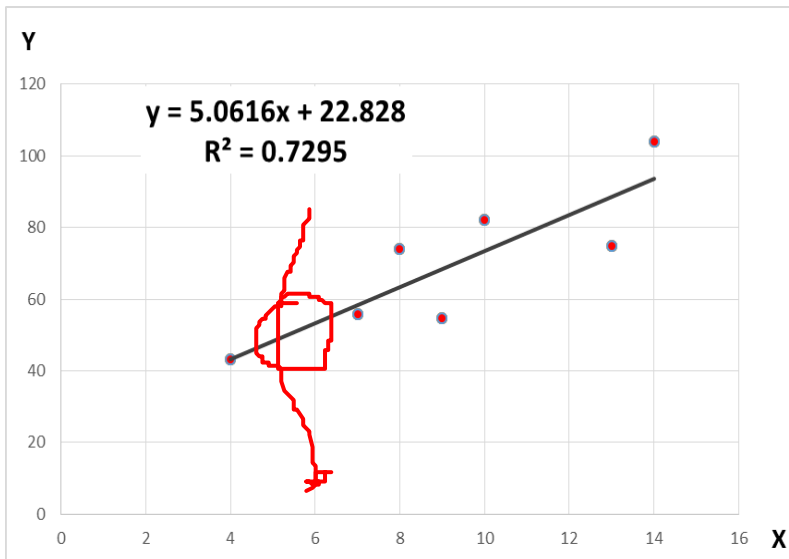
## Ejemplo

Modelo	$R^2$	$S^2$	DET	PRESS	p	$C_p$	$C_p/p$
X1 X3	0,990015	0,723514	0,465897	12,56803	3	4,832549	1,61085
X1 X2 X3	0,991921	0,634125	0,049393	14,83343	4	4	1
X1 X2	0,990014	0,723583	0,814085	15,16263	3	4,833969	1,611323
X1	0,935309	4,352465	1	73,26319	2	84,09233	42,04616
X2 X3	0,84814	11,00326	0,208837	200,5016	3	215,5747	71,85822
X3	0,750711	16,77245	1	308,6936	2	358,2968	179,1484
X2	0,394421	40,74409	1	763,4074	2	887,535	443,7675

# Regresión Lineal Múltiple

## Overfitting

*¿Cuál de los modelos es mejor?*



# Regresión Lineal Múltiple

## Overfitting

Si la cantidad de parámetros es elevada respecto de la cantidad de observaciones que tenemos corremos peligro de “sobreajustar” el modelo.

Un modelo sobreajustado se ajusta muy bien a los datos de la muestra pero no es capaz de realizar buenas predicciones para nuevos valores desconocidos para el modelo.

En la práctica, el problema se manifiesta en una sub-estimación de la varianza residual y los pronósticos realizados carecen de precisión.

# Regresión Lineal Múltiple

## Overfitting

Para evitar el sobreajuste:

- La relación  $n/k$  no debería ser inferior a 10
- Si la cantidad de información disponible es escasa, podemos relajar este requerimiento a  $n/k \geq 5$ . De todas formas, no es recomendable y en la medida de lo posible debe evitarse. Los modelos estimados pueden estar sujetos a problemas de sobreajuste.

# Análisis Exploratorio

## Principio de Parsimonia

Entre dos modelos de características similares, será preferido aquel que presente mayor sencillez, es decir aquel que tenga menor cantidad de parámetros.

A la inversa: solo elegimos un modelo más complejo, si tenemos fuertes indicios de que realmente es mejor que el modelo con menos parámetros.

# Análisis Exploratorio

PROCEDIMIENTOS DE SELECCIÓN DE VARIABLES



# Análisis Exploratorio

## Procedimientos de Selección de Variables

- Origen Algorítmico
  - Stepwise Forward
  - Stepwise Backward
  - Stepwise Both (Forward+Backward)
  - Best Subsets Regression
- Origen Estadístico:
  - Regression Lasso
  - Regresión Ridge

# Análisis Exploratorio

## Algoritmo Stepwise Forward

1. Comenzamos con un modelo sin variables
2. Se determina la mejor variable para incluir en el modelo, por ejemplo:
  - Aquella que tenga el menor valor P al incluirla en el modelo o
  - Aquella que al agregarse produzca la mayor reducción en  $R^2$
3. La variable seleccionada se agrega al modelo
4. Se repiten los pasos 2 y 3 hasta agregando mas variables al modelo hasta que se cumpla una condición de corte, por ejemplo:
  - Que los valores P de las variables a ingresar sean todos mayores a un determinado valor
  - Que el incremento en el  $R^2$  al agregar variables sea menor a un determinado valor

# Análisis Exploratorio

## Algoritmo Stepwise Backward

1. Comenzamos con un modelo con todas las variables
2. Se determina la mejor variable para quitar del modelo, por ejemplo:
  - Aquella que tenga el mayor valor P
  - Aquella que al quitarse produzca la menor reducción en  $R^2$
3. La variable seleccionada se quita del modelo
4. Se repiten los pasos 2 y 3 hasta quitando variables del modelo hasta que se cumpla una condición de corte, por ejemplo:
  - Que los valores P de todas las variables del modelo sean todos menores a un determinado valor
  - Que la reducción en el  $R^2$  al quitar nuevas sea mayor a un determinado valor

# Análisis Exploratorio

## Algoritmo Stepwise Both (Forward+Backward)

Se trata de una combinación de los métodos Forward y Backward.

Se procede los procedimientos Forward y Backward alternativamente hasta llegar a un modelo donde no se deben agregar ni quitar variables.

# Análisis Exploratorio

## Algoritmo Best Subsets

Consiste en realizar los  $2^k - 1$  modelos posibles y mediante uno o mas criterio seleccionar el modelo mas conveniente.

# Análisis Confirmatorio

DIAGNÓSTICO DE RESIDUOS

# Regresión Lineal Múltiple

## Análisis Confirmatorio

- Análisis de la Bondad de Ajuste
- Análisis de Multicolinealidad
- Diagnóstico de residuos
- **Validación de supuestos**

# Regresión Lineal Múltiple

## Supuestos del modelo

$$E(\vec{\varepsilon}) = \vec{0}$$

$$\text{Var}(\vec{\varepsilon}) = \sigma^2 \mathbf{I}$$

$$\text{Cov}(\tilde{\varepsilon}_i; \tilde{\varepsilon}_j) = 0 \quad \forall i \neq j$$

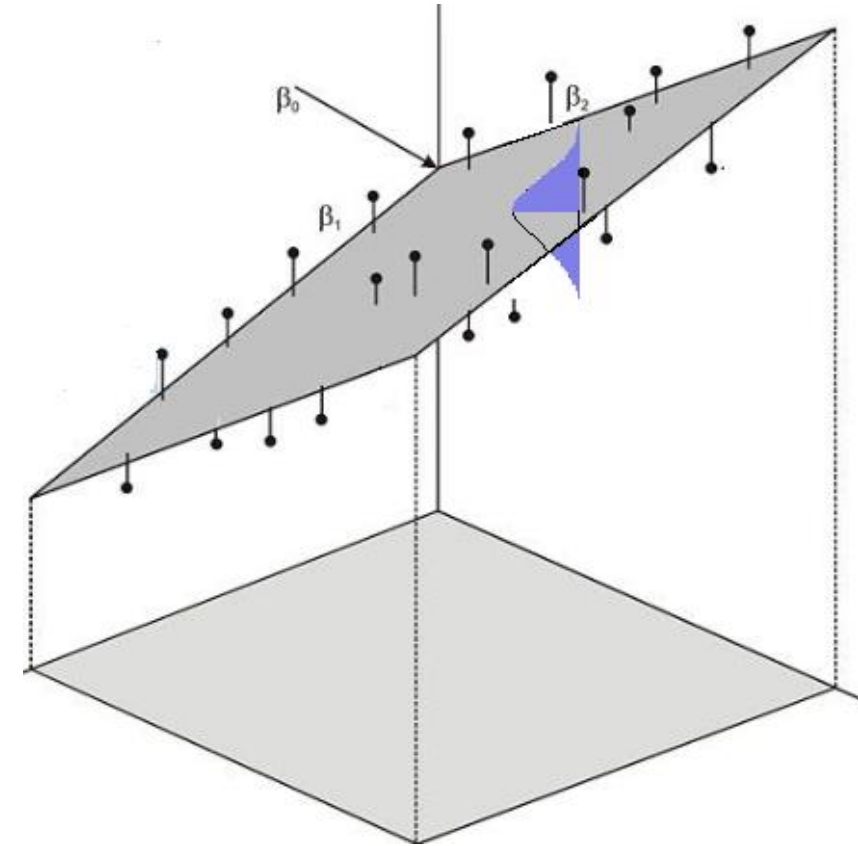
$$\vec{\varepsilon} \sim \text{Normal}(\vec{0}; \sigma^2 \mathbf{I})$$

Ausencia de vicio

Homocedasticidad

Ausencia de autocorrelación

Normalidad de las perturbaciones



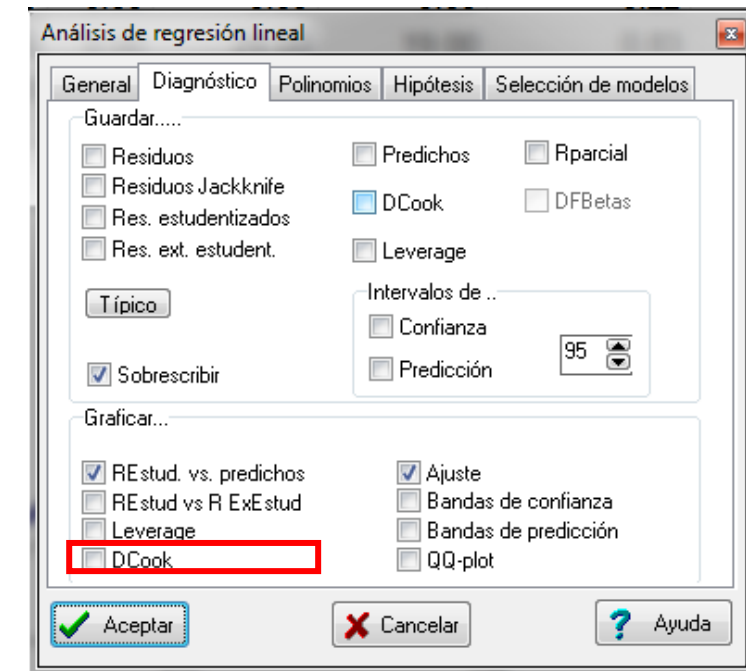
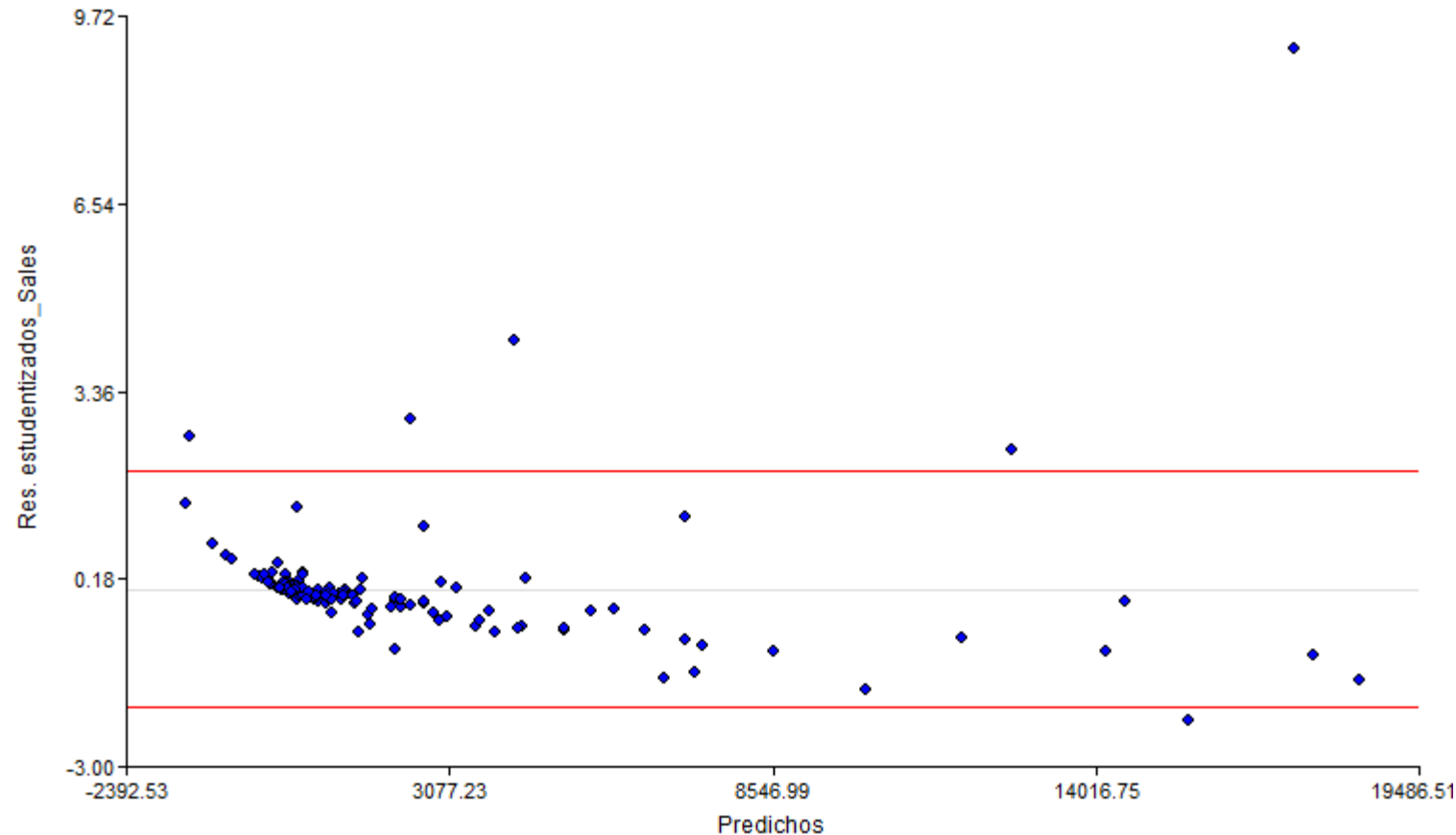


# Verificación de Supuestos

No linealidad - Heterocedasticidad

$$E(\varepsilon_i) \neq 0$$

$$D^2(\varepsilon_i) \neq \sigma^2$$



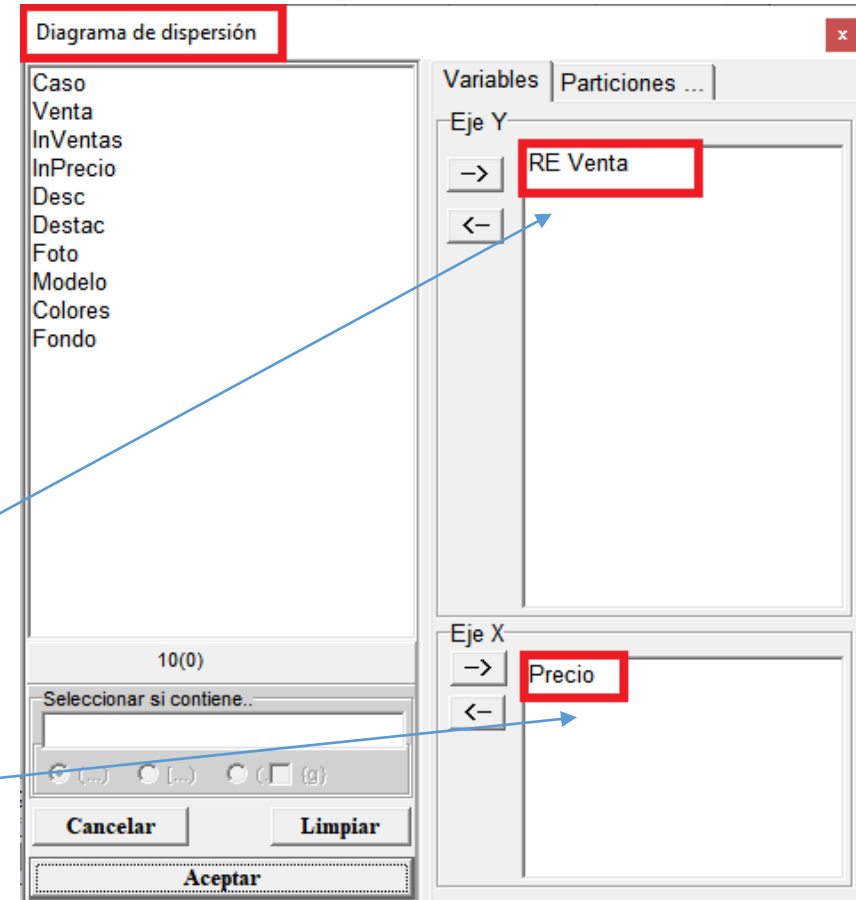
# Verificación de Supuestos

## No linealidad - Heterocedasticidad

Si al realizar el gráfico de *residuos vs predichos* se detecta heterocedasticidad o no linealidad, es posible obtener información sobre que variables lo generan realizando los gráficos de residuos vs  $x_j$ :

Residuos estudentizados del modelo

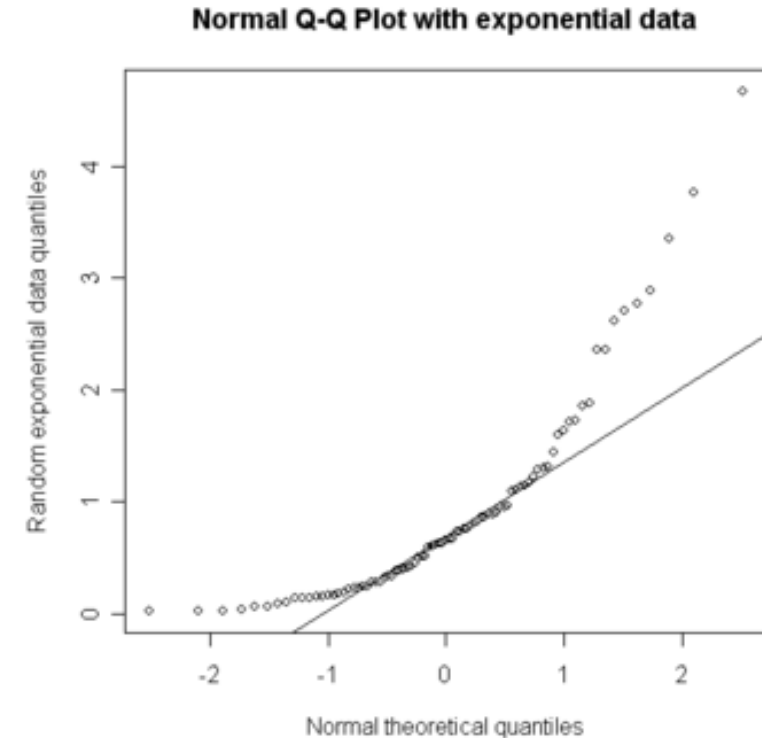
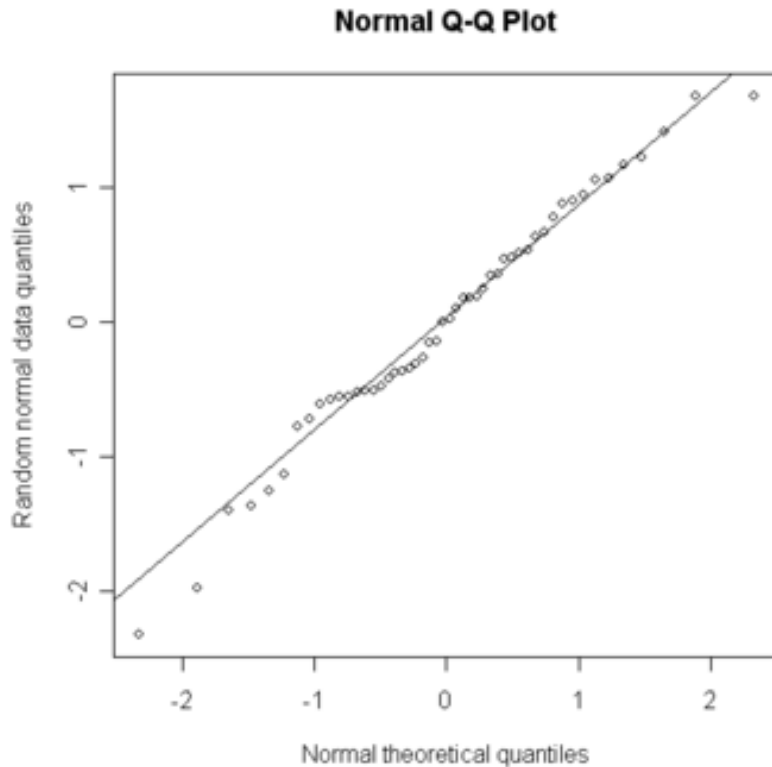
Variables X



# Verificación de Supuestos

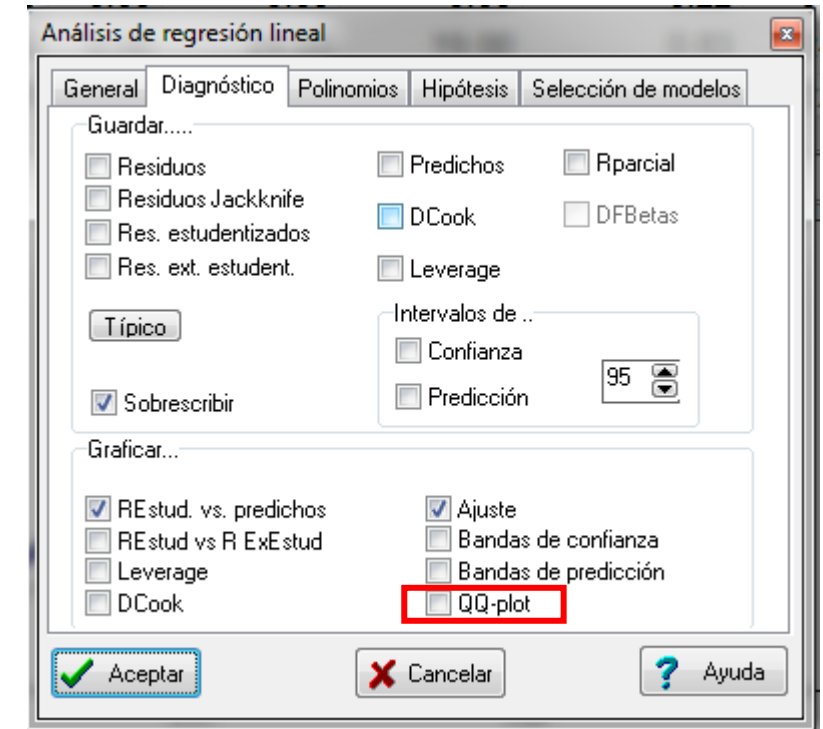
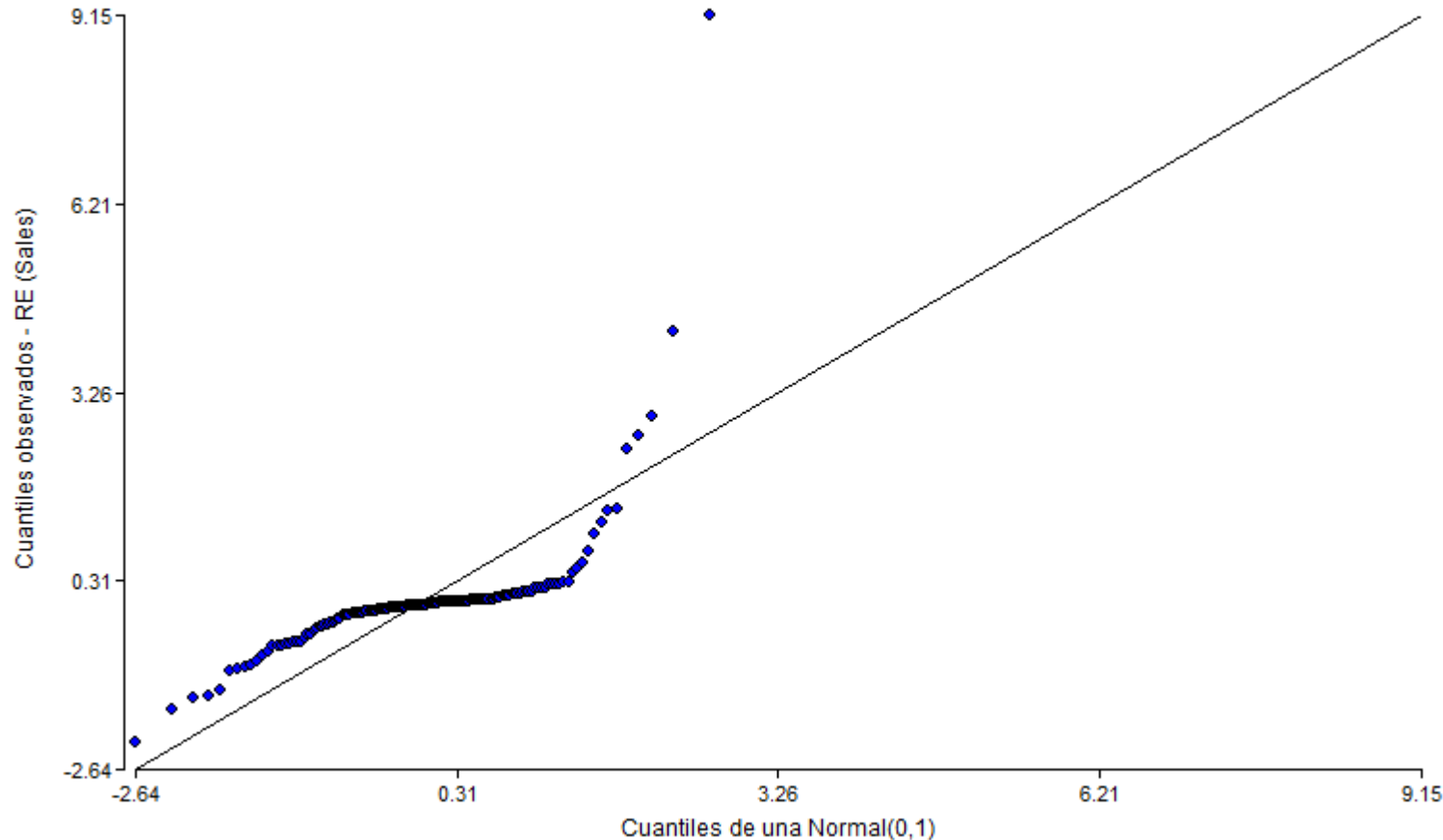
## Normalidad de las perturbaciones

$$\varepsilon_i \sim \text{Normal}(0; \sigma_\varepsilon^2)$$



# Verificación de Supuestos

## Normalidad de las perturbaciones



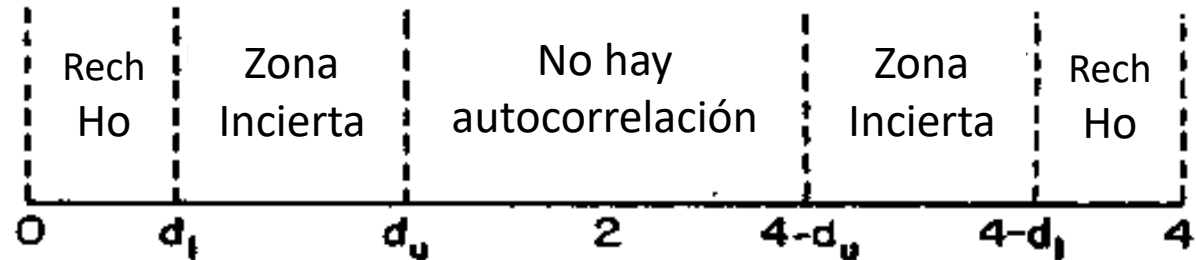
# Verificación de Supuestos

## Autocorrelación

Se suelen producir si los datos están ordenados en el tiempo o espacio

$$Cov(\varepsilon_i; \varepsilon_j) = 0$$

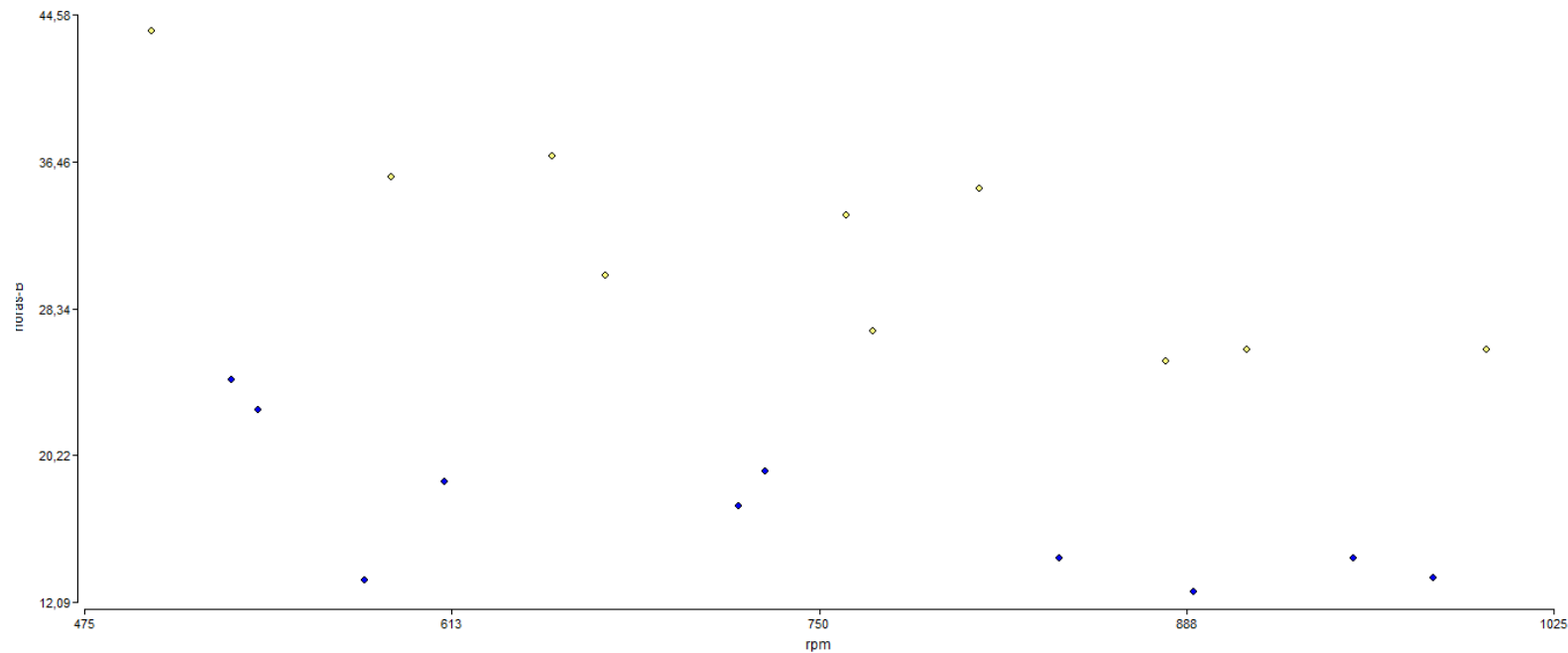
$$H_0) \rho_1 = 0$$
$$d_k = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$



# Modelado con variables independientes cualitativas

# CASO DE DISCUSION I

Un ingeniero estudia la duración de la herramienta de un torno en función de la velocidad de corte y del tipo de material (A o B)



i	$y_i$ (horas)	$x_{i1}$ (rpm)	Material
1	18,73	610	A
2	14,52	950	A
3	17,43	720	A
4	14,54	840	A
5	13,44	980	A
6	24,39	530	A
7	13,34	580	A
8	22,71	540	A
9	12,68	890	A
10	19,32	730	A
11	30,16	670	B
12	27,09	770	B
13	25,4	880	B
14	26,05	1000	B
15	33,49	760	B
16	35,62	590	B
17	26,07	910	B
18	36,78	650	B
19	34,95	810	B
20	43,67	500	B

material	Variable	N	R <sup>2</sup>	R <sup>2</sup> Aj	ECMP	AIC	BIC
A	horas	10	0,54	0,48	14,11	53,91	54,82

#### Coefficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	30,18	4,38	20,07	40,28	6,89	0,0001		
rpm	-0,02	0,01	-0,03	-4,3E-03	-3,05	0,0158	9,39	1,00

#### Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	82,10	1	82,10	9,32	0,0158
rpm	82,10	1	82,10	9,32	0,0158
Error	70,49	8	8,81		
Total	152,59	9			

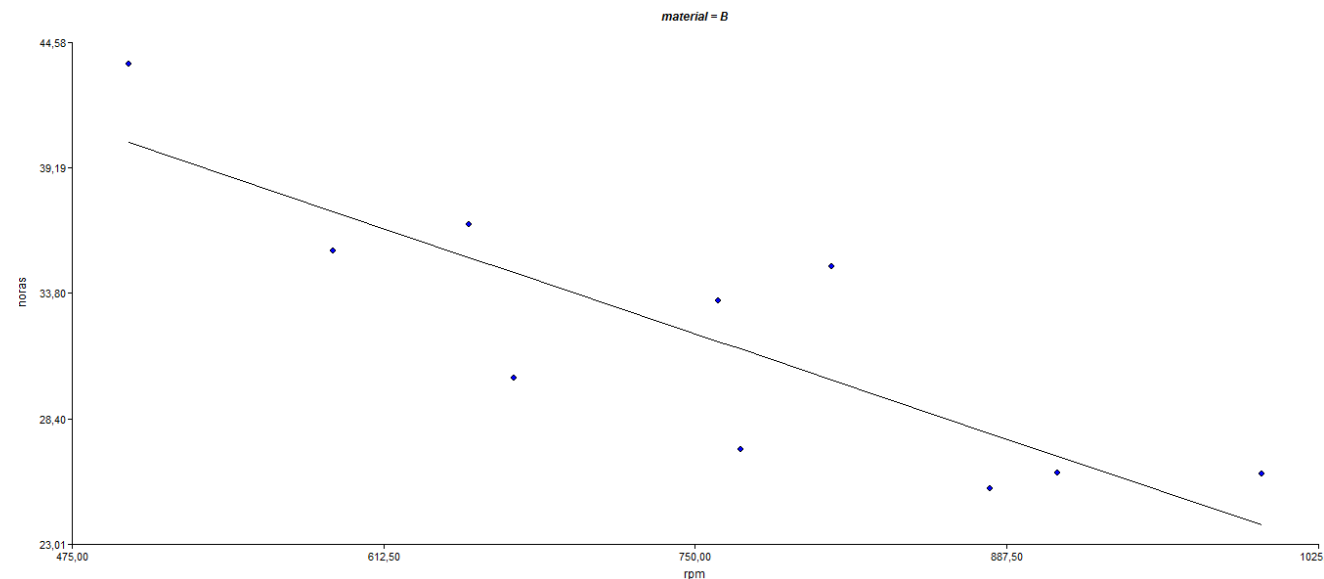
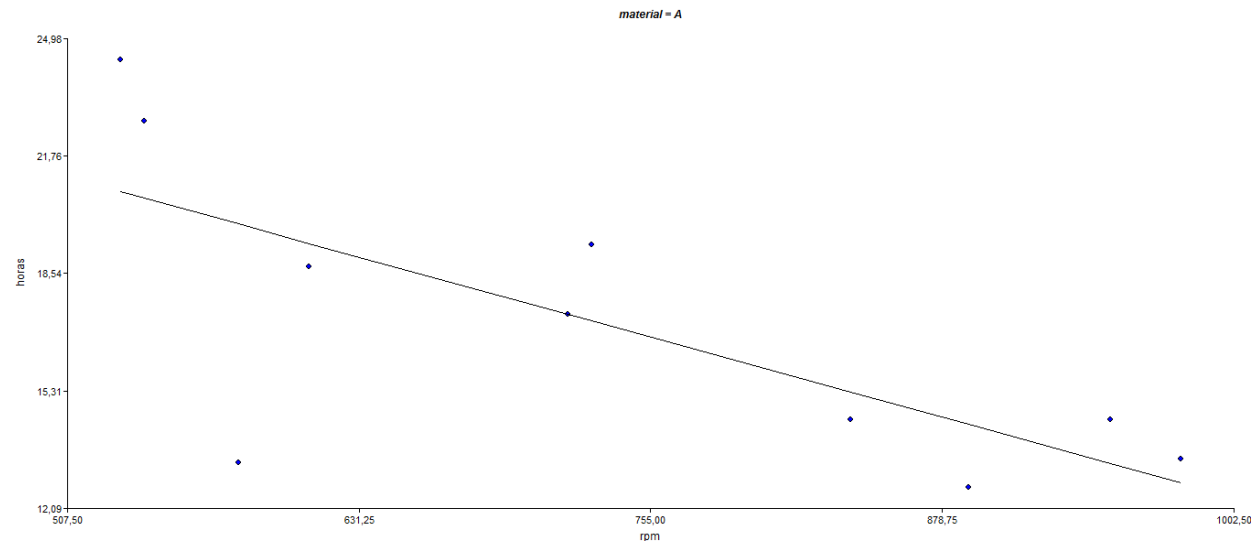
material	Variable	N	R <sup>2</sup>	R <sup>2</sup> Aj	ECMP	AIC	BIC
B	horas	10	0,71	0,68	18,27	56,72	57,63

#### Coefficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	56,75	5,68	43,65	69,84	9,99	<0,0001		
rpm	-0,03	0,01	-0,05	-0,02	-4,45	0,0021	18,72	1,00

#### Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	231,23	1	231,23	19,80	0,0021
rpm	231,23	1	231,23	19,80	0,0021
Error	93,40	8	11,68		
Total	324,63	9			



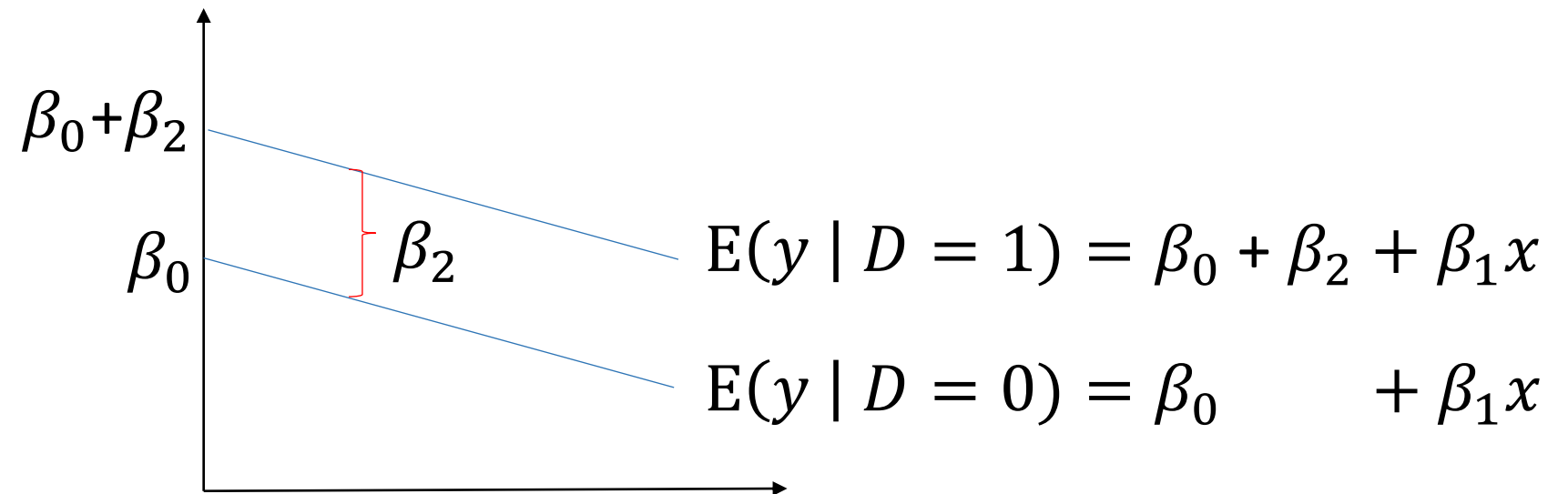


# Modelado con variables indicadoras

Efecto sobre la ordenada al origen

1	610	0
1	950	0
1	720	0
1	840	0
1	980	0
1	530	0
1	580	0
1	540	0
1	890	0
1	730	0
1	670	1
1	770	1
1	880	1
1	1000	1
1	760	1
1	590	1
1	910	1
1	650	1
1	810	1
1	500	1

$$\tilde{y} = \beta_0 + \beta_1 x + \beta_2 D + \tilde{\varepsilon}$$



# Modelado con variables indicadoras

## Efecto sobre la ordenada al origen

### Análisis de regresión lineal

Variable	N	R <sup>2</sup>	R <sup>2</sup> Aj	ECMP	AIC	BIC
horas	20	0,88	0,86	15,81	109,89	113,87

### Coefficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	35,21	3,74	27,32	43,10	9,42	<0,0001		
rpm	-0,02	4,9E-03	-0,03	-0,01	-5,05	0,0001	26,12	1,00
material B	15,24	1,50	12,07	18,40	10,15	<0,0001	99,33	1,00

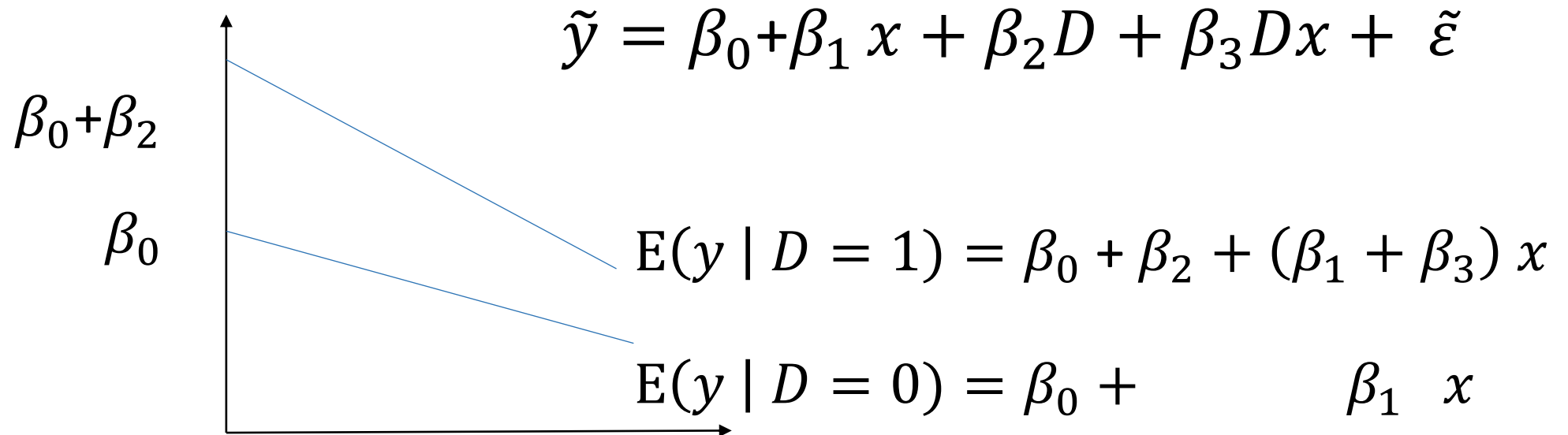
Al emplear el material B la duración de la herramienta se incrementa en promedio entre 12 y 18,4 horas (NC=95%)

### Cuadro de Análisis de la Varianza (SC tipo I)

F.V.	SC	gl	CM	F	p-valor
Modelo.	1384,11	2	692,05	61,60	<0,0001
rpm	227,03	1	227,03	20,21	0,0003
material_B	1157,08	1	1157,08	103,00	<0,0001
Error	190,98	17	11,23		
Total	1575,09	19			

# Modelado con variables indicadoras

Efecto sobre la pendiente (Interacción)



# Modelado con variables indicadoras

## Efecto sobre la pendiente (Interacción)

1	610	0	0
1	950	0	0
1	720	0	0
1	840	0	0
1	980	0	0
1	530	0	0
1	580	0	0
1	540	0	0
1	890	0	0
1	730	0	0
1	670	1	670
1	770	1	770
1	880	1	880
1	1000	1	1000
1	760	1	760
1	590	1	590
1	910	1	910
1	650	1	650
1	810	1	810
1	500	1	500

### Análisis de regresión lineal

Variable	N	R <sup>2</sup>	R <sup>2</sup> Aj	ECMP	AIC	BIC
Columna1	20	0,83	0,80	23,53	118,52	123,50

### Coefficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI (95%)	LS (95%)	T	p-valor	CpMallows	VIF
const	33,25	5,92	20,70	45,79	5,62	<0,0001		
Columna2	-0,02	0,01	-0,04	-3,7E-03	-2,58	0,0200	9,34	1,79
Columna3_B	25,57	9,17	6,15	45,00	2,79	0,0131	10,39	25,00
Columna4	-0,01	0,01	-0,04	0,01	-1,21	0,2432	4,44	26,68

No es posible concluir que el material influye en la pendiente

### Cuadro de Análisis de la Varianza (SC tipo I)

F.V.	SC	gl	CM	F	p-valor
Modelo.	1308,94	3	436,31	26,23	<0,0001
Columna2	227,03	1	227,03	13,65	0,0020
Columna3_B	1057,49	1	1057,49	63,57	<0,0001
Columna4	24,42	1	24,42	1,47	0,2432
Error	266,14	16	16,63		
Total	1575,09	19			

# Modelado con variables indicadoras

Efecto sobre la ordenada al origen mas de 2 categorías

Se estudia el rendimiento de un proceso en función de la temperatura de la reacción y con tres catalizadores diferentes. Se supone que el tipo de catalizador y la temperatura son independientes.

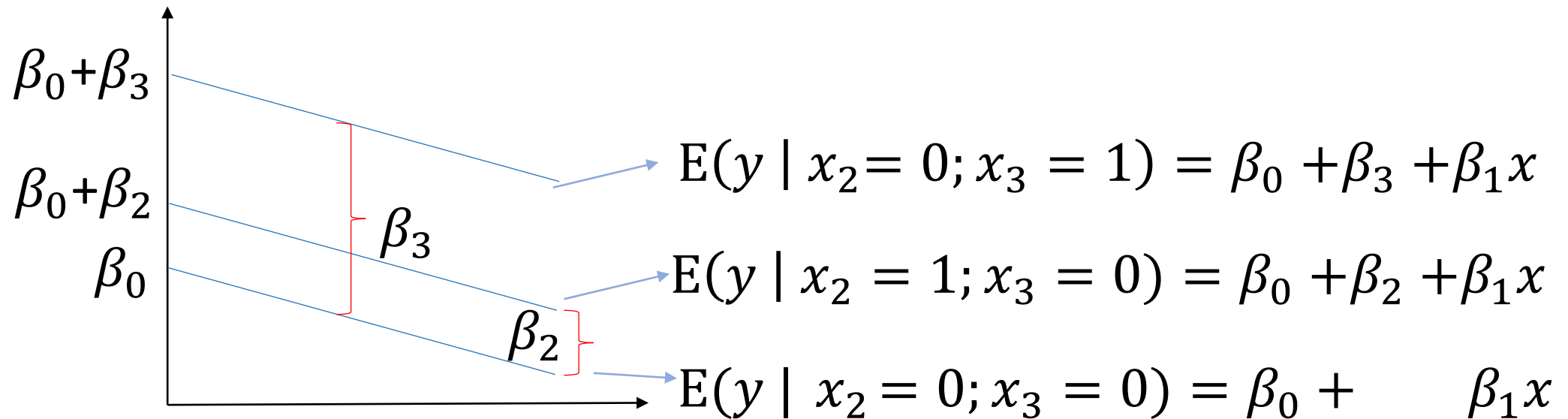
$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \tilde{\varepsilon}$$

Catalizador	$x_2$	$x_3$
A	0	0
B	1	0
C	0	1

# Modelado con variables indicadoras

Efecto sobre la ordenada al origen mas de 2 categorías

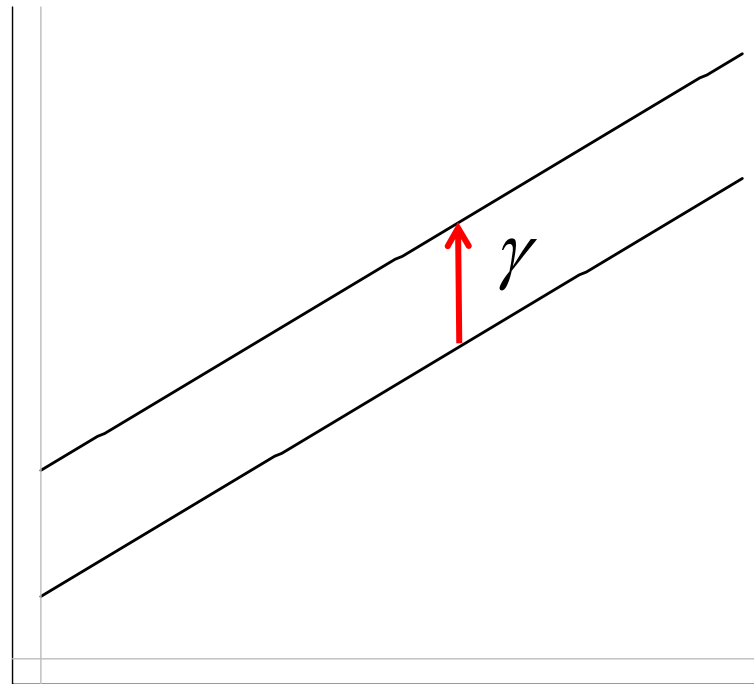
$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \tilde{\varepsilon}$$



# Modelado con variables indicadoras

## Resumen de casos

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \gamma D_i$$



$$\hat{y}_{i \ D_i=1} = (\beta_0 + \gamma) + \beta_1 x_i$$

$$\hat{y}_{i \ D_i=0} = \beta_0 + \beta_1 x_i$$

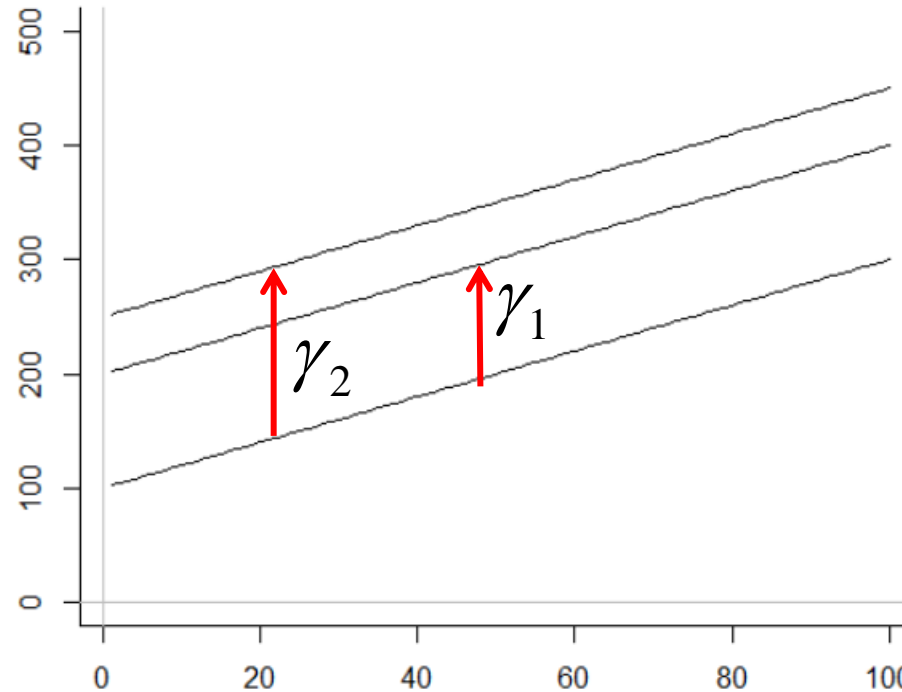
# Modelado con variables indicadoras

## Resumen de casos

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \gamma_1 D_{1i} + \gamma_2 D_{2i}$$

Material	Variables Indicadoras		
	D1	D2	D3
Bronce	1	0	0
Acero	0	1	0
Hierro	0	0	1

Referencia



$$\hat{y}_i(x|D_1=0;D_2=1) = (\beta_0 + \gamma_2) + \beta_1 x_i$$

$$\hat{y}_i(x|D_1=1;D_2=0) = (\beta_0 + \gamma_1) + \beta_1 x_i$$

$$\hat{y}_i(x|D_1=0;D_2=0) = \beta_0 + \beta_1 x_i$$



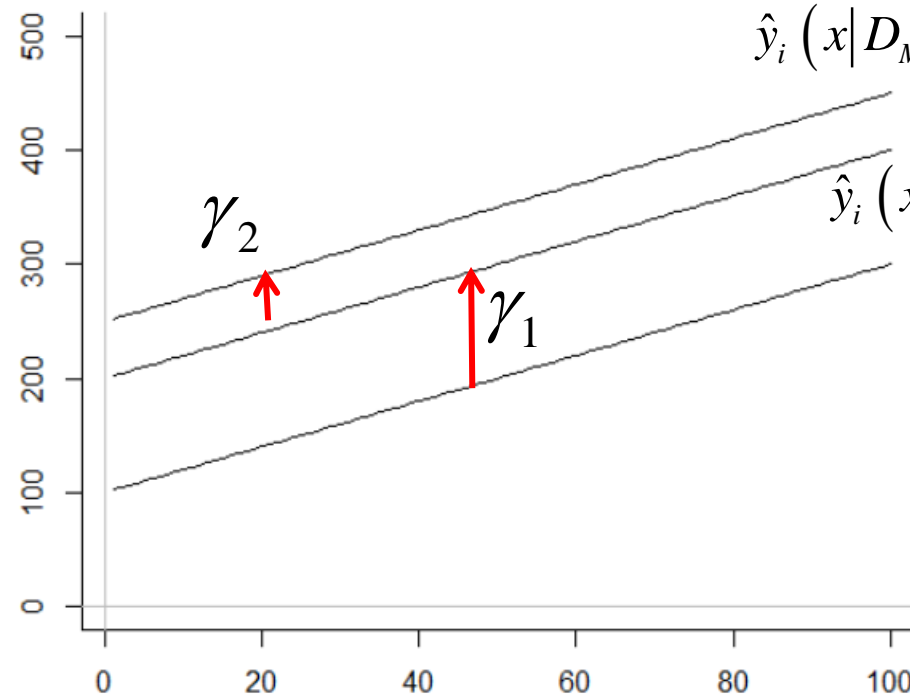
# Modelado con variables indicadoras

## Resumen de casos

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \gamma_1 D_{Mi} + \gamma_2 D_{Ai}$$

Dureza	Variables Indicadoras		
	D_baja	D_Media	D_alta
Baja	0	0	0
Media	0	1	0
Alta	0	1	1

Referencia



$$\hat{y}_i (x | D_M = 1; D_A = 1) = (\beta_0 + \gamma_1 + \gamma_2) + \beta_1 x_i$$

$$\hat{y}_i (x | D_M = 1; D_A = 0) = (\beta_0 + \gamma_1) + \beta_1 x_i$$

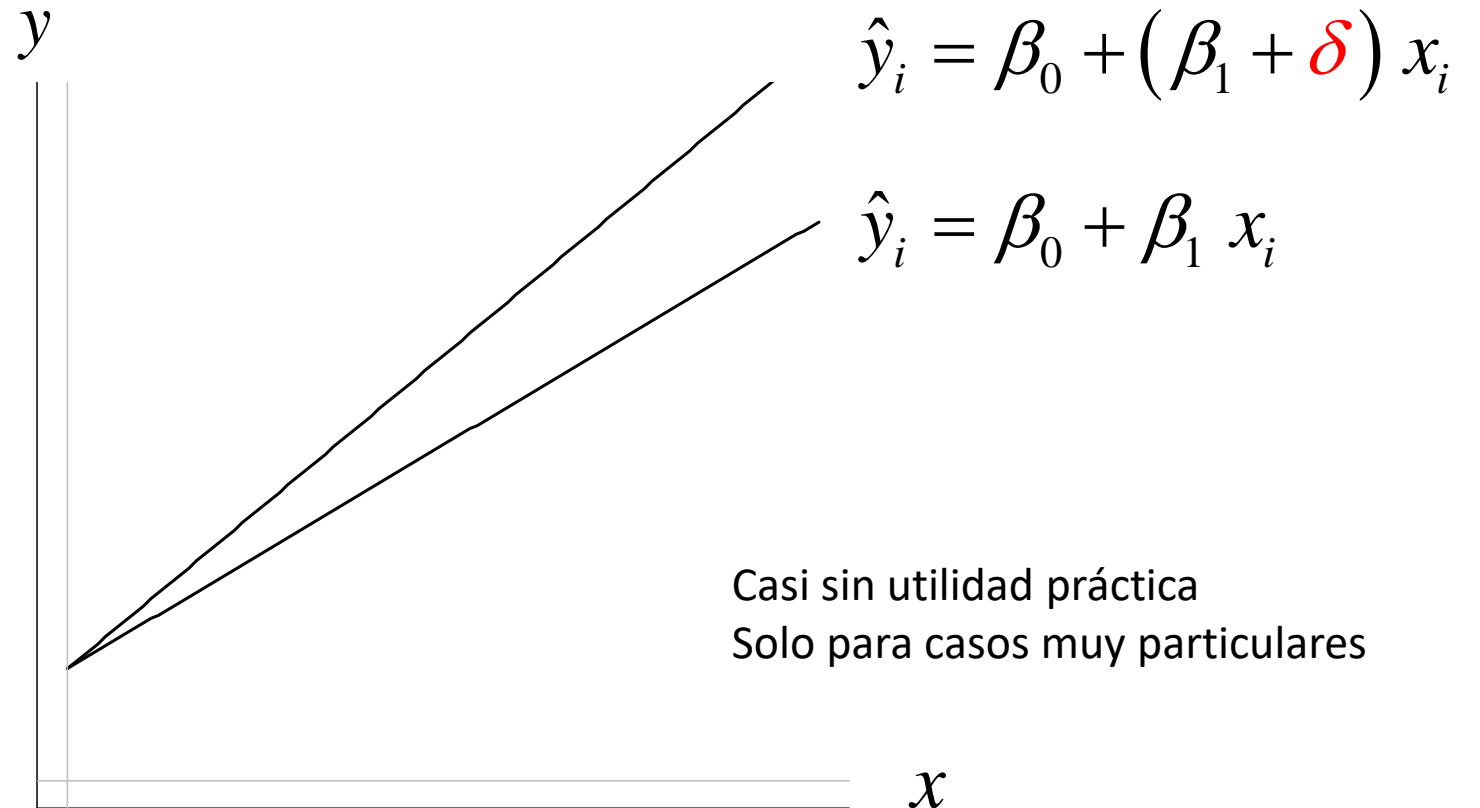
$$\hat{y}_i (x | D_M = 0; D_A = 0) = \beta_0 + \beta_1 x_i$$

Esta codificación permite medir la significancia de la diferencia entre durezas Media y Alta

# Modelado con variables indicadoras

## Resumen de casos

$$y_i = \beta_0 + \beta_1 x_i + \delta x_i D_i$$



# Modelado con variables indicadoras

## Resumen de casos

$$y_i = \beta_0 + \beta_1 x_i + \gamma D_i + \delta x_i D_i$$

