Tema 4
Regresión lineal simple

José R. Berrendero

Departamento de Matemáticas
Universidad Autónoma de Madrid

Estructura de este tema

- Planteamiento del problema. Ejemplos
- Recta de regresión de mínimos cuadrados
- El modelo de regresión lineal simple
- IC y contrastes para los parámetros del modelo
- Predicción

Ejemplo: consumo de vino y dolencias cardíacas

Consideramos dos variables (fichero `vino.sav`):

- X: Consumo anual de vino en litros por habitante
- Y: Número de muertes por enfermedad cardíaca, por cada 100.000 habitantes

¿Qué podemos decir sobre la relación entre las dos variables?

¿Podemos afirmar que valores altos en consumo de vino están asociados con valores bajos en número de muertes por enfermedad cardíaca?

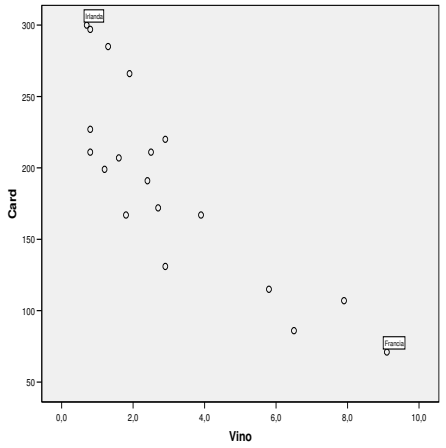
¿Podemos predecir aproximadamente el valor de la variable Y si sabemos el valor de X?

Estadísticos

		Vino	Card
N	Válidos	19	19
	Perdidos	0	0
Media		3,026	191,05
Desv. típ.		2,5097	68,396

Correlaciones

		Vino	Card
Vino	Correlación de Pearson	1	-,843
	Sig. (bilateral)		,000
	N	19	19
Card	Correlación de Pearson	-,843	1
	Sig. (bilateral)	,000	
	N	19	19



Ejemplo: renta y fracaso escolar en la CAM

EL PAÍS, martes 18 de octubre de 2005

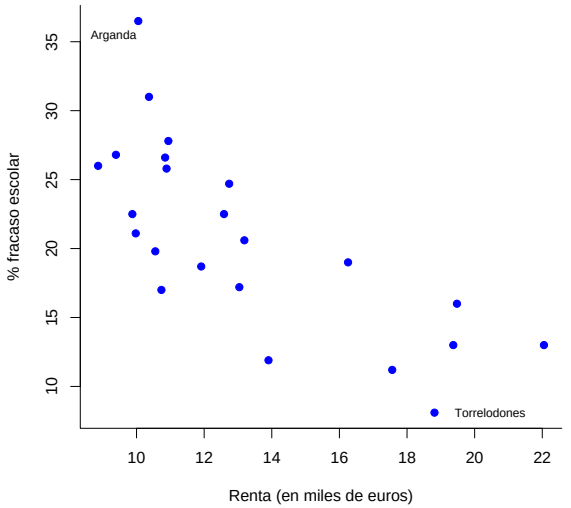
El fracaso escolar es más alto en las zonas con menor renta

Fracaso escolar en la Comunidad de Madrid

Renta per cápita bruta media en 2003: 13.095 euros

	CURSO 2003/2004	
	Renta (euros)	Fracaso escolar (%)
Parla	8.864	26,0
Fuenlabrada	9.391	26,8
Leganés	9.877	22,5
Móstoles	9.977	21,1
Arganda	10.052	36,5
Torrejón	10.369	31,0
Getafe	10.555	19,8
Coslada	10.736	17,0
Pinto	10.846	26,6
Alcorcón	10.888	25,8
Alcalá de Henarés	10.942	27,8
Collado	11.913	18,7
Colmenar Viejo	12.587	22,5
Arroyomolinos	12.740	24,7
S. Sebastián de los Reyes	13.041	17,2
S. Lorenzo del Escorial	13.189	20,6
Rivas	13.903	11,9
Alcobendas	16.256	19,0
Tres Cantos	17.562	11,2
Torrelodones	18.812	8,1
Boadilla	19.368	13,0
Majadahonda	19.477	16,0
Pozuelo	22.050	13,0

EL PAÍS



Problema de regresión

Observamos dos variables, X e Y, el objetivo es analizar la relación existente entre ambas de forma que podamos predecir o aproximar el valor de la variable Y a partir del valor de la variable X.

- La variable Y se llama **variable respuesta**
- La variable X se llama **variable regresora o explicativa**

En un problema de regresión (a diferencia de cuando calculamos el coeficiente de correlación) el papel de las dos variables no es simétrico.

Recta de regresión

Frecuentemente, existe entre las variables una relación aproximadamente lineal:

$$Y_i \approx \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

- La recta $y = \beta_0 + \beta_1 x$ es una **recta de regresión**
- El parámetro β_1 es la **pendiente** de la recta. Indica cómo cambia la variable respuesta cuando $\Delta X = 1$
- El parámetro β_0 es el **término independiente** de la recta. Indica el valor de Y cuando $X = 0$

Problema estadístico: estimar los parámetros β_0 y β_1 a partir de los datos (x_i, Y_i), $i = 1, \ldots, n$.

La recta de mínimos cuadrados

Si estimamos β_0 y β_1 mediante $\hat{\beta}_0$ y $\hat{\beta}_1$, la predicción de la variable respuesta Y_i en función de la regresora x_i es:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

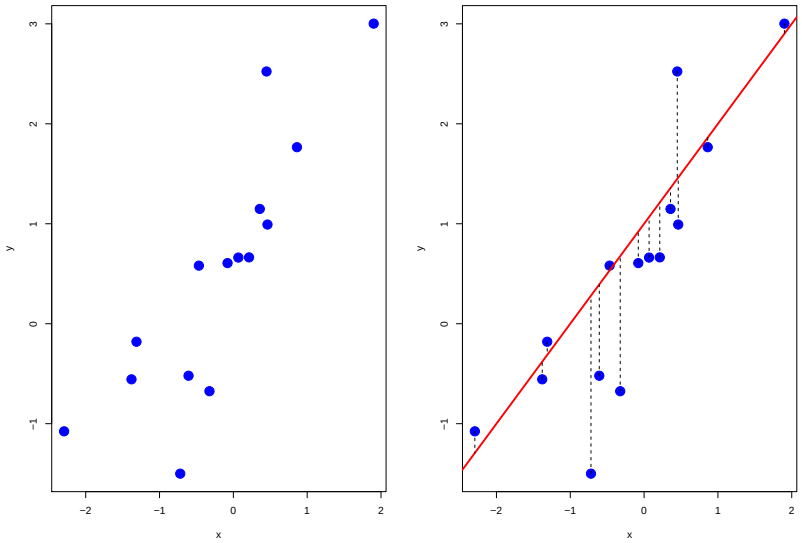
Unos buenos estimadores deben ser tales que los errores de predicción

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

sean pequeños

La **recta de regresión de mínimos cuadrados** viene dada por los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ para los que se minimiza:

$$\sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2$$



Estimadores de mínimos cuadrados

Pendiente:

$$\hat{\beta}_1 = r \frac{S_y}{S_x}$$

donde r es el coeficiente de correlación, S_y es la desviación típica de la variable respuesta y S_x es la desviación típica de la variable regresora.

Término independiente:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

A los errores $e_i = Y_i - \hat{Y}_i$ se les llama **residuos**

A las predicciones $\hat{Y}_i$ se les llama **valores ajustados**

Ejemplo: consumo de vino

Estimadores de los parámetros:

$$\hat{\beta}_1 = r \frac{S_y}{S_x} = -0.843 \frac{68.396}{2.5097} = -22.974$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = 191.05 - (-22.974) \times 3.026 = 260.57$$

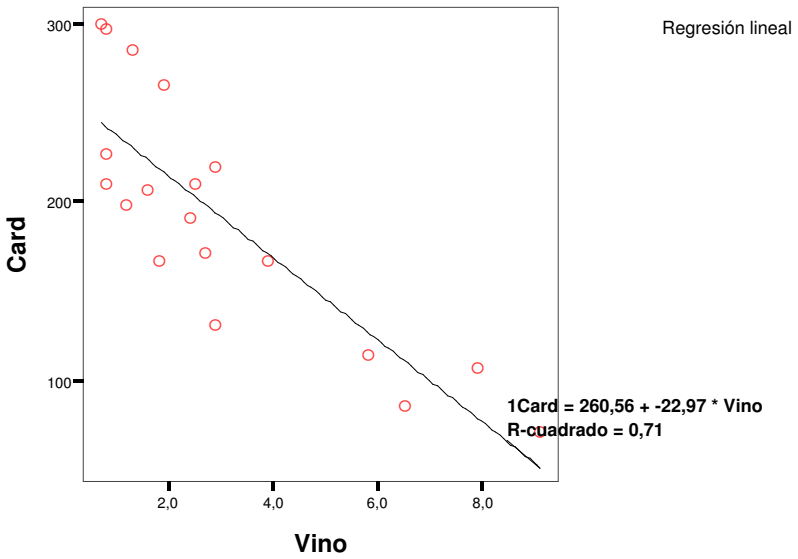
Recta de regresión:

$$y = 260.57 - 22.974x$$

Predicción de Y_0 para $x_0 = 4$:

$$\hat{Y}_0 = 260.57 - 22.974 \times 4 = 168.674$$

Diagrama de dispersión y recta estimada



Observaciones

- La recta de mínimos cuadrados pasa por el punto cuyas coordenadas son las medias: $(\bar{x}, \bar{Y})$.
- Si la variable regresora se incrementa en una desviación típica $\Delta x = S_x$, entonces la predicción de la variable respuesta se incrementa en r desviaciones típicas: $\Delta \hat{Y} = rS_y$
- Puede demostrarse que la suma de los residuos siempre vale cero.
- La recta para predecir Y en función de X no es la misma que la recta para predecir X en función de Y.
- Como medida de lo bien que se ajusta la recta a los datos, se utiliza el **coeficiente de determinación (o R-cuadrado)**: el cuadrado del coeficiente de correlación.

El modelo de regresión lineal simple

Para poder hacer inferencia (IC y contrastes) sobre los parámetros, suponemos que se verifica el siguiente modelo:

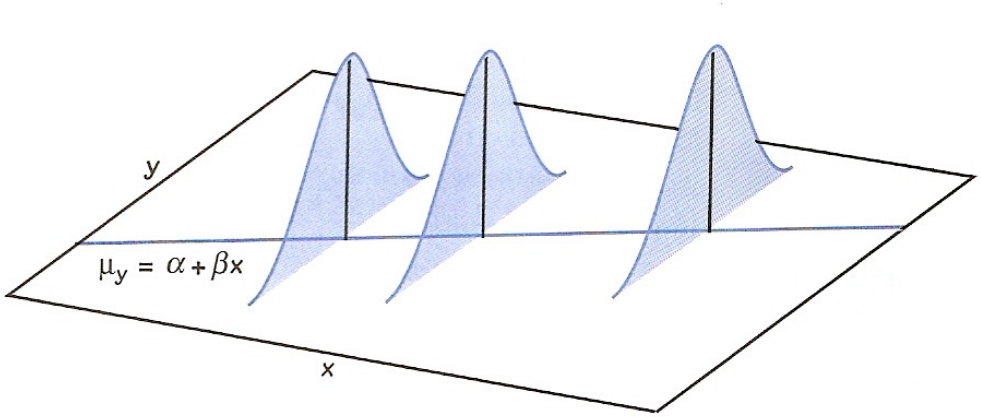
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

donde:

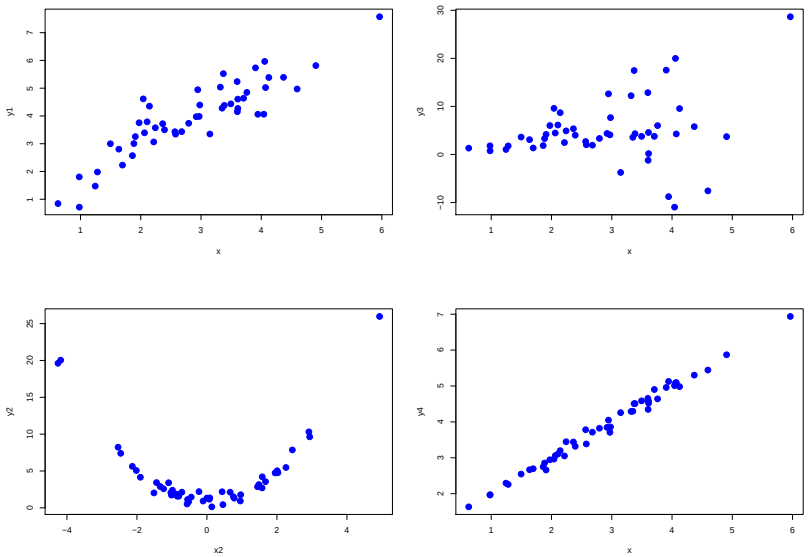
- El valor esperado de los errores ϵ_i es cero.
- Todos los errores tienen la misma varianza σ^2.
- Los errores tienen distribución normal y son independientes.

En resumen:

$$\epsilon_1, \ldots, \epsilon_n \equiv \mathrm{N}(0, \sigma) \text{ independientes}$$



¿En cuáles de las 4 situaciones se verifica el modelo?



Una simulación

Supongamos que $\sigma = 1$, $\beta_0 = 0$ y $\beta_1 = 1$

Entonces el modelo es

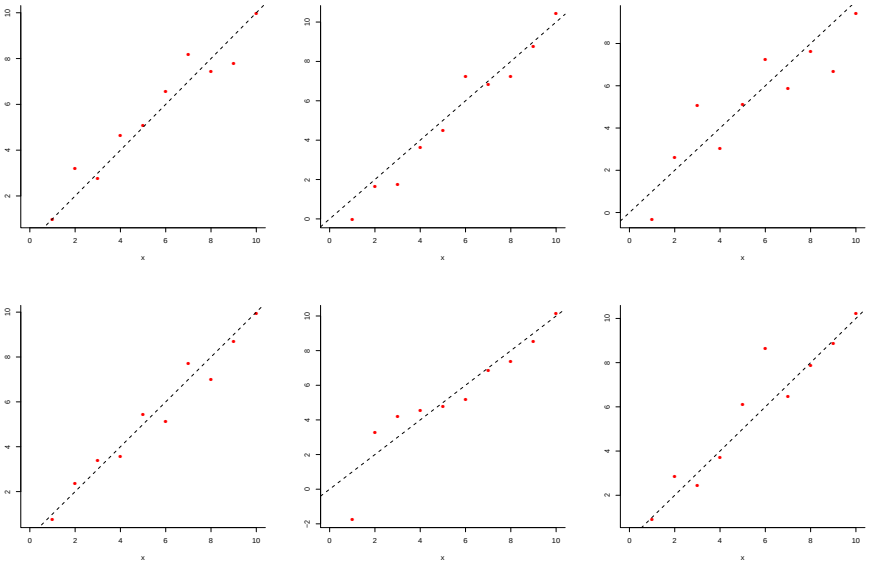
$$Y_i = x_i + \epsilon_i,$$

donde los errores ϵ_i tienen distribución normal estándar y son independientes

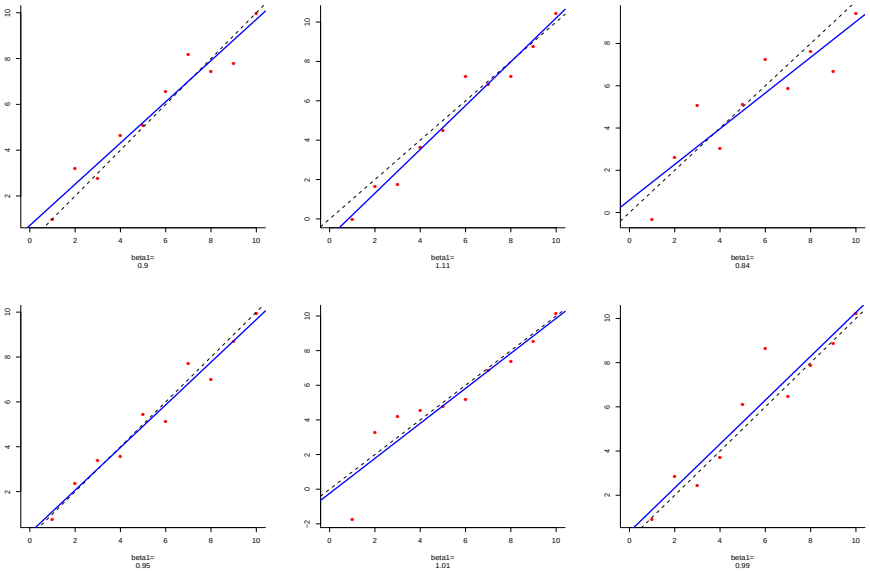
Fijamos $x_i = 1, 2, \ldots, 10$ ($n = 10$) y generamos las respuestas correspondientes de acuerdo con este modelo

Posteriormente calculamos la recta de mínimos cuadrados y la representamos junto con la *verdadera recta* $y = x$

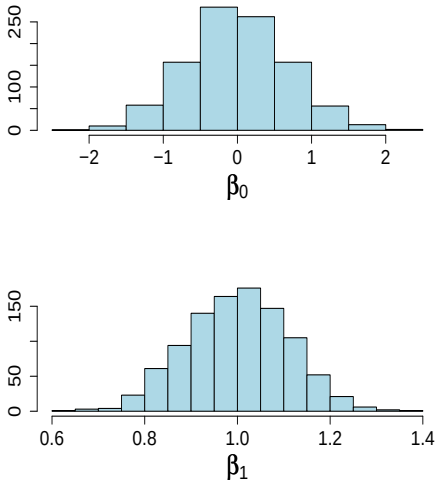
Repetimos 6 veces el experimento



Repetimos 6 veces el experimento



Repetimos 1000 veces el experimento



- Los estimadores van cambiando para las diferentes muestras
- Existen fórmulas del error típico de $\hat{\beta}_0$ y $\hat{\beta}_1$ que recogen esta variabilidad
- Estas fórmulas son las que se utilizan para calcular IC y llevar a cabo contrastes en lo que sigue.

Intervalos de confianza

Los intervalos de confianza de nivel $1-\alpha$ para los parámetros $\hat{\beta}_i$ $(i = 0, 1)$ tienen la estructura habitual:

$$\left[\hat{\beta}_i \mp t_{n-2,\alpha/2} \times \text{E.T.}(\hat{\beta}_i)\right]$$

En comparación con los intervalos de confianza para la media:

- Los grados de libertad son $n-2$ en lugar de $n-1$.
- La fórmula del error típico es más complicada (siempre lo calcularemos con el ordenador).

Contrastes para los parámetros

Contrastes unilaterales:

- Hipótesis: $H_0 : \beta_i \leq \beta_i^*$ frente a $H_1 : \beta_i > \beta_i^*$
- Región crítica:

$$R = \left\{ \frac{\hat{\beta}_i - \beta_i^*}{\text{E.T.}(\hat{\beta}_i)} > t_{n-2,\alpha} \right\}.$$

- Hipótesis: $H_0 : \beta_i \geq \beta_i^*$ frente a $H_1 : \beta_i < \beta_i^*$
- Región crítica:

$$R = \left\{ \frac{\hat{\beta}_i - \beta_i^*}{\text{E.T.}(\hat{\beta}_i)} < -t_{n-2,\alpha} \right\}.$$

Contraste bilateral:

- Hipótesis: $H_0 : \beta_i = \beta_i^*$ frente a $H_1 : \beta_i \neq \beta_i^*$
- Región crítica:

$$R = \left\{ \frac{|\hat{\beta}_i - \beta_i^*|}{\text{E.T.}(\hat{\beta}_i)} > t_{n-2,\alpha/2} \right\}.$$

Ejemplo: consumo de vino

Sabiendo que el error típico del estimador de mínimos cuadrados de la pendiente es 3.557, calcula un IC para β_1 de nivel 95%:

$$\left[\hat{\beta}_1 \mp t_{n-2,\alpha/2} \times \text{E.T.}(\hat{\beta}_1)\right] \equiv [-22.974 \mp 2.11 \times 3.557]$$

ya que $t_{17,0.025} = 2.11$.

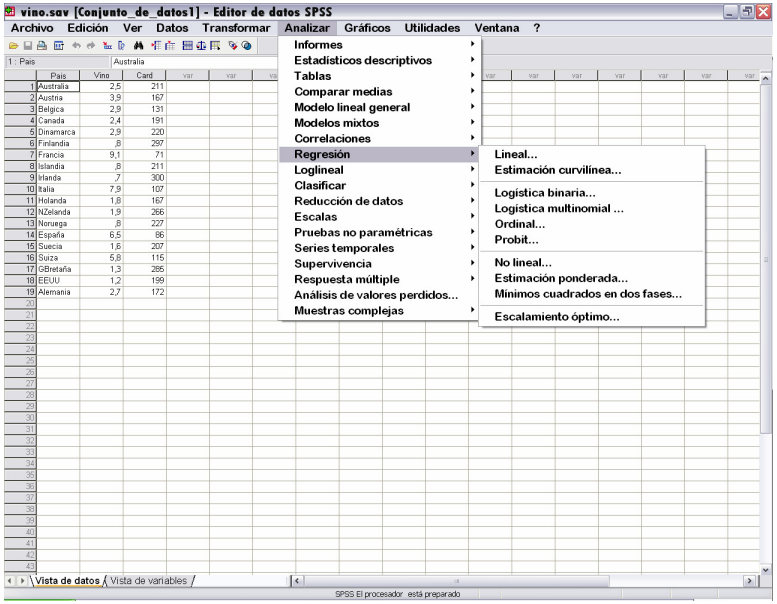
¿Aportan los datos evidencia para afirmar ($\alpha = 0.01$) que un incremento en el consumo de vino está asociado a un menor número de muertes por enfermedad cardíaca?

Queremos contrastar $H_0 : \beta_1 \geq 0$ frente a $H_1 : \beta_1 < 0$. Calculamos:

$$t = \frac{-22.974}{3.557} = -6.457 \quad \text{y} \quad -t_{17,0.01} = -2.567$$

Como $-6.457 < -2.567$ estamos en la región crítica y podemos rechazar H_0.

Con SPSS



Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,843[a]	,710	,693	37,879

a. Variables predictoras: (Constante), Vino

b. Variable dependiente: Card

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	260,563	13,835		18,833	,000
	Vino	-22,969	3,557	-,843	-6,457	,000

a. Variable dependiente: Card

Ejemplo: renta y fracaso escolar

Para los datos de nivel de renta (en miles de euros) y fracaso escolar (%) en la CAM se tiene la siguiente salida de SPSS:

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,742[a]	,550	,528	4,7566

a. Variables predictoras: (Constante), Renta

b. Variable dependiente: Fracaso

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	38,494	****		10,562	,000
	Renta	-1,347	,266	-,742	-5,065	,000

a. Variable dependiente: Fracaso

Cuestiones

- Escribe la ecuación de la recta de mínimos cuadrados que describe el nivel de fracaso escolar como función de la renta.
- Calcula intervalos de confianza de nivel 95% para la pendiente y el término independiente de la recta de regresión.
- ¿Podemos afirmar, a nivel $\alpha = 0.05$ que niveles más altos de renta están asociados a niveles más bajos de fracaso escolar?
- ¿Cuánto vale el coeficiente de correlación entre el nivel de renta y el porcentaje de fracaso escolar?
- ¿Qué porcentaje de fracaso escolar se predice en una población cuya renta es $x_0 = 13000$ euros?
- ¿Cuál es el residuo correspondiente a Colmenar Viejo?