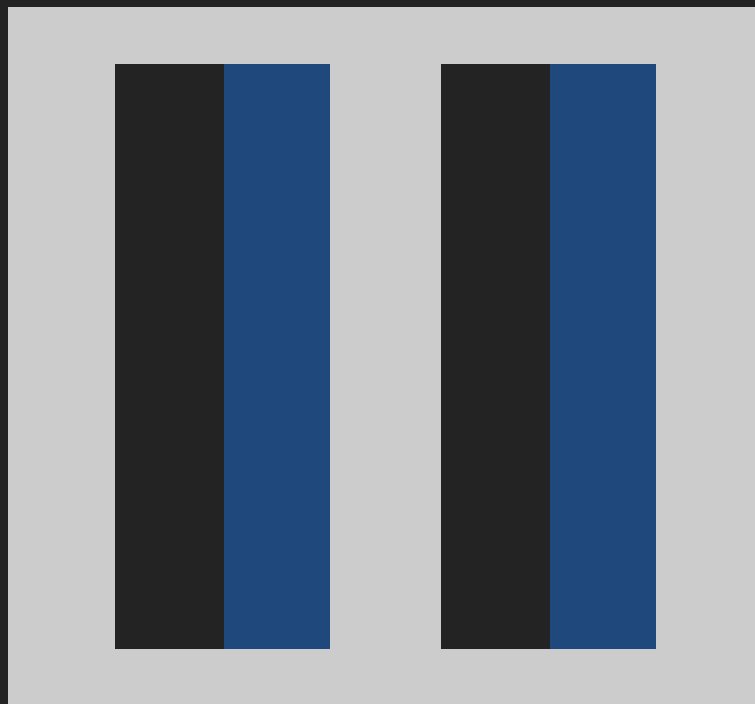


# ECONOMETRÍA

JOSÉ ALBERTO MAURICIO



3

## REGRESIÓN LINEAL MÚLTIPLE II

### 3.2 DIAGNOSIS - RESIDUOS - OBSERVACIONES INFLUYENTES



EctrGr-JAM-3-2.pdf

Copyright © 2022 - 2024 J.A.M.

[ucm.randomshock.com/ectrgr](http://ucm.randomshock.com/ectrgr)

Versión 2.4 - Enero 2024

### REQUISITOS

En esta Sección 3.2 se suponen bien conocidos [1] los aspectos teóricos y prácticos del Análisis de Regresión Lineal cubiertos en los Temas 1 - 2 y en la Sección 3.1 del Tema 3, y [2] todos los procedimientos descritos en las Secciones 1 - 14 de la guía *Introducción al Uso de EViews 4.1*.

### BIBLIOGRAFÍA PARA LA SECCIÓN 3.2



Hill, Griffiths, Lim (2018): Apartados 4.3.4 - 4.3.6.

Wooldridge (2020): Apartado 9-5C.

Sección 15 de la guía *Introducción al Uso de EViews 4.1*.

En adelante IEV41 ...

## PARTE 1 - DIAGNOSIS: RESUMEN

### 1 Parámetros (efectos causales):

1.1 Valores/signos: Variables omitidas (Sección 2.4), observaciones influyentes (PARTE 3).

1.2 Significación estadística: Variables irrelevantes, multicolinealidad (Sección 2.4).

### 2 Residuos (lo que queda "fuera" de un modelo estimado):

2.1 Los residuos no deben contener ningún tipo de información que resulte útil para mejorar el modelo estimado del que proceden. De manera equivalente, los residuos deben tener propiedades análogas a las que se imponen sobre las perturbaciones (HC3-HC5) para garantizar la optimalidad de MCO (los residuos deben ser "puramente aleatorios", o "ruido blanco"). Cualquier discrepancia significativa debe resolverse modificando el modelo estimado en la dirección adecuada [...].

#### 2.2 Gráficos (PARTE 2):

A: Forma funcional.

B: También aportan información general sobre A-D en 2.3.

#### 2.3 Operaciones específicas para detectar:

A: Heteroscedasticidad (HC4) (Sección 3.3).

B: Autocorrelación (HC4) (Tema 4).

C: Ausencia de Normalidad (HC5) (PARTE 2).

D: Observaciones atípicas (PARTE 3).

**Observación importante:** Con frecuencia, el incumplimiento de HC3-HC5 se debe a **errores de especificación** en la **forma funcional** del modelo y/o en el contenido del **conjunto de variables explicativas**, como se irá viendo en esta sección y en las siguientes [...].

## PARTE 2 - DIAGNOSIS: GRÁFICOS DE RESIDUOS

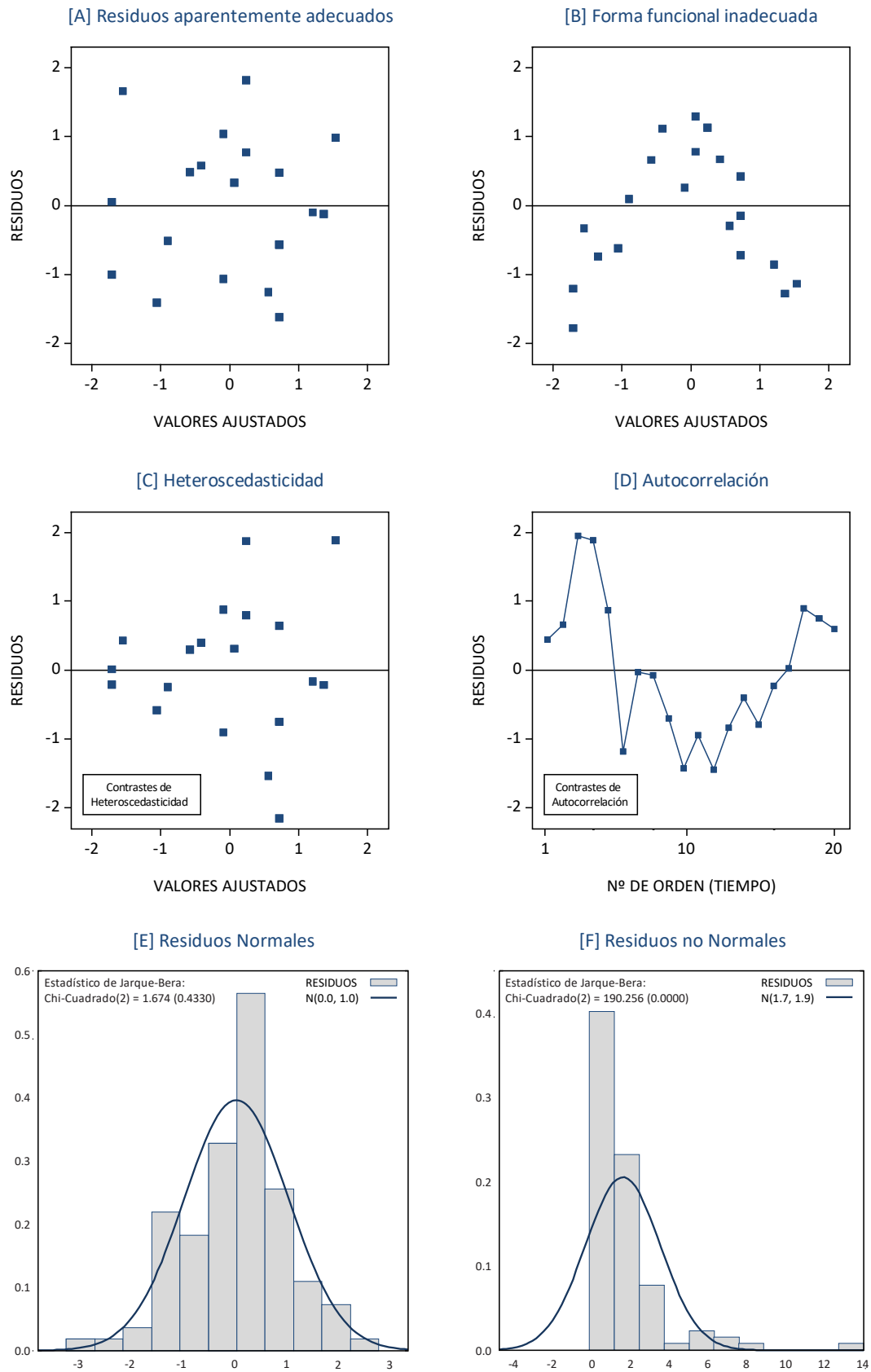
⇒ Las hipótesis  $E[U_i | \mathbf{X}] = 0$  (HC3, exogeneidad estricta) y  $E[U_i^2 | \mathbf{X}] = \sigma^2$  (HC4, homoscedasticidad) sugieren que el **nivel medio** y la **dispersión** de los residuos deben ser **constantes** (en particular, deben ser independientes de los datos sobre todas las variable explicativas). Ambas hipótesis pueden contrastarse (informalmente) examinando las nubes de puntos de los residuos sobre cada variable explicativa y/o sobre los valores ajustados del modelo estimado (que son una combinación lineal de los datos sobre todas las variables explicativas). Ver Figura 1.

⇒ Si el orden de las observaciones es relevante (series temporales), las dos hipótesis anteriores y la hipótesis  $E[U_{i_1} U_{i_2} | \mathbf{X}] = 0$  (HC4, ausencia de autocorrelación) también pueden contrastarse examinando un gráfico ordenado (temporal) de los residuos.

⇒ La hipótesis de Normalidad (HC5) puede contrastarse con el histograma de los residuos y el estadístico de Jarque-Bera.

⇒ El examen de los gráficos mencionados también ayuda a detectar **observaciones atípicas** que pueden influir notablemente en un modelo estimado (PARTE 3).

**FIGURA 1**  
Gráficos de Residuos



### PARTE 3 - OBSERVACIONES INFLUYENTES

Una **observación** (punto muestral) es **influyente** si los resultados de la estimación de un modelo cambian notablemente al eliminar de la muestra dicha observación [Figuras 2-5]. En la práctica, la presencia de una observación influyente en una muestra puede deberse a:

- ⇒ Un error en los datos que conforman dicha observación.
- ⇒ La existencia de un punto muestral (una entidad o un momento) que es muy diferente del resto en algún aspecto relevante.

Una observación influyente del primer tipo debe corregirse (cuando es posible), o bien eliminarse del análisis (cuando no es posible corregirla). Una observación influyente del segundo tipo **no** debe eliminarse de manera rutinaria:

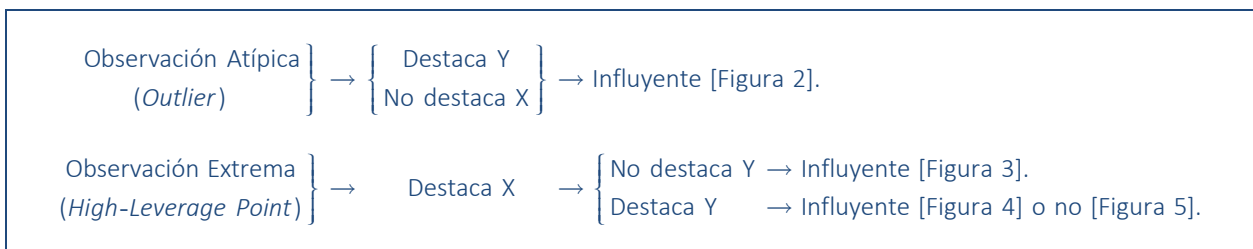
- ⇒ Siempre debe ser examinada para intentar explicar su carácter especial.
- ⇒ Su presencia puede indicar algún error de especificación que debe ser corregido.
- ⇒ Sólo debe eliminarse en casos suficientemente justificados (cuando hace referencia a una entidad o a un momento que es especial por determinados motivos que carecen de interés en el análisis, o cuando distorsiona las conclusiones generales del mismo).

### TIPOS DE OBSERVACIONES INFLUYENTES

⇒ Una **observación** es **atípica** o **anómala** (*outlier*) cuando su dato de la **variable dependiente** destaca sobre los datos de dicha variable en otras observaciones que son, por el contrario, similares en cuanto a los datos de las variables explicativas. En general, una observación atípica es al mismo tiempo una observación influyente, que se manifiesta a través de un valor grande (atípico o anómalo) en el **residuo** correspondiente (Figura 2).

⇒ Una **observación** es **extrema** o **potencialmente influyente** (*high-leverage point*) cuando sus datos de las **variables explicativas** destacan sobre los datos de dichas variables en el resto de la muestra. Una observación extrema es influyente cuando su dato de la variable dependiente **no** destaca del resto de la muestra (Figura 3). A diferencia del caso anterior, una observación extrema influyente **no** suele tener asociado un residuo atípico.

⇒ Una observación extrema cuyo dato de la variable dependiente **sí** destaca del resto de la muestra, puede ser (Figura 4) o no (Figura 5) una observación influyente; cuando sí lo es, tampoco (como en el caso anterior) suele tener asociado un residuo atípico.

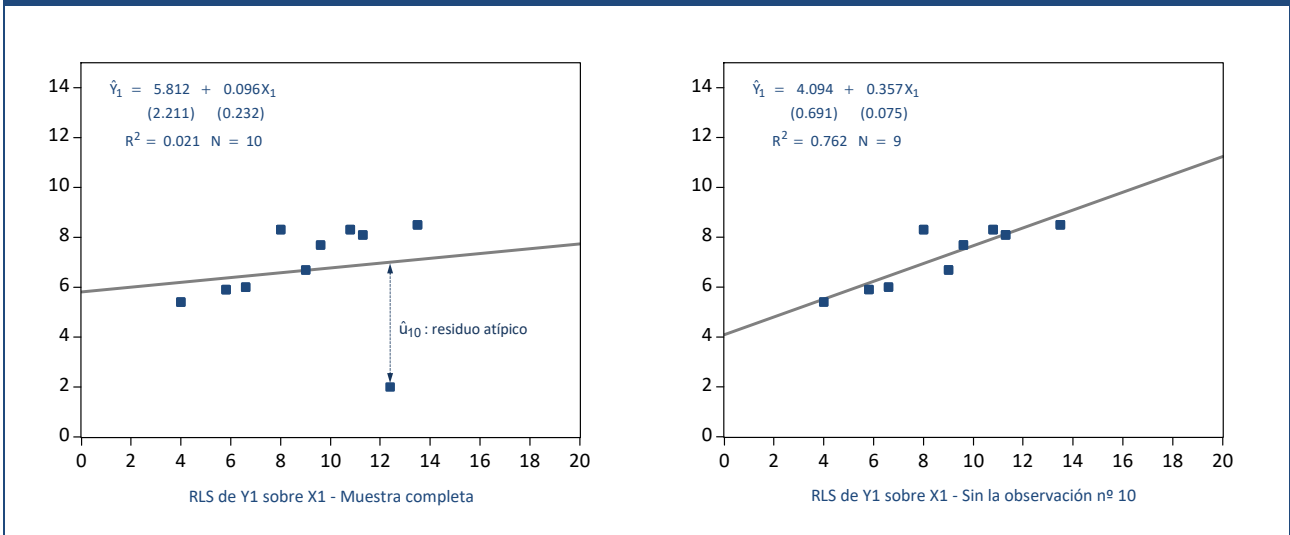


Los datos empleados para estimar los modelos RLS de las Figuras 2-5 son los de la Tabla 1. Las representaciones gráficas se han obtenido con EViews utilizando el programa PRG00-RLS.PRG, ajustando convenientemente la muestra (SAMPLE) utilizada en cada caso. [Ejemplo en clase Figura 2. Tareas Figuras 3-5.]

**TABLA 1**  
 Datos utilizados en las Figuras 2-5 - NUM03-OBSINF.WF1

OBS	Y1	Y2	Y3	Y4	X1	X2
1	8.3	6.6	6.6	6.6	10.8	6.0
2	6.0	5.7	5.7	5.7	6.6	6.6
3	6.7	7.3	7.3	7.3	9.0	7.6
4	7.7	8.5	8.5	8.5	9.6	6.5
5	8.1	8.8	8.8	8.8	11.3	7.8
6	8.5	7.0	7.0	7.0	13.5	7.0
7	5.9	8.0	8.0	8.0	5.8	7.5
8	5.4	7.5	7.5	7.5	4.0	6.5
9	8.3	6.4	6.4	6.4	8.0	5.4
10	2.0 *	8.5	1.0 *	12.6 *	12.4	14.0 *

**FIGURA 2**



**FIGURA 3**

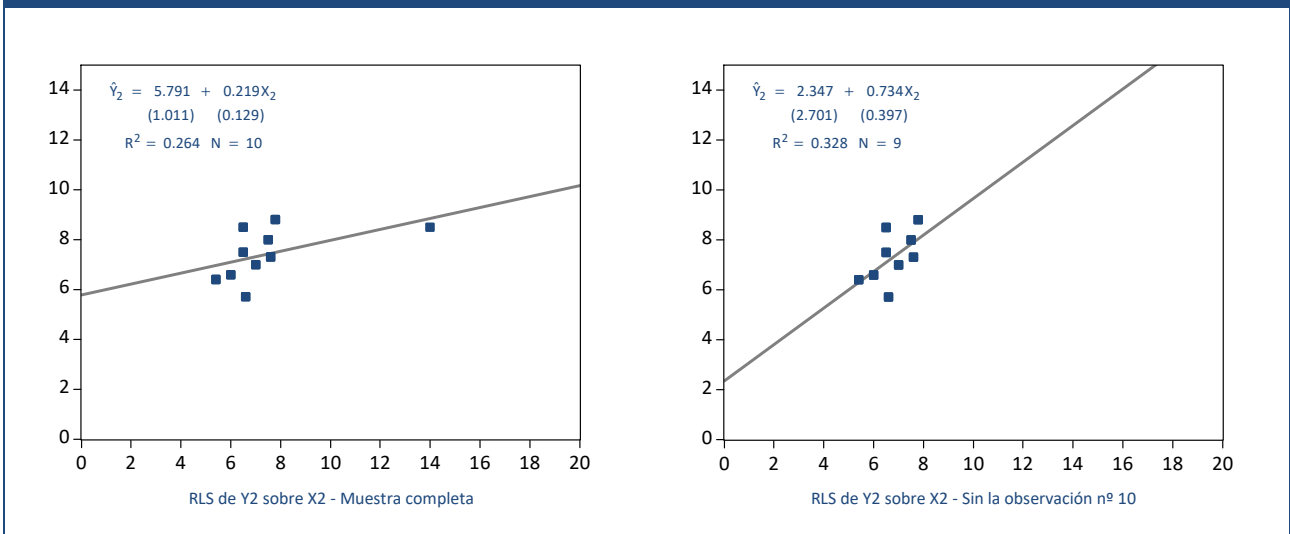


FIGURA 4

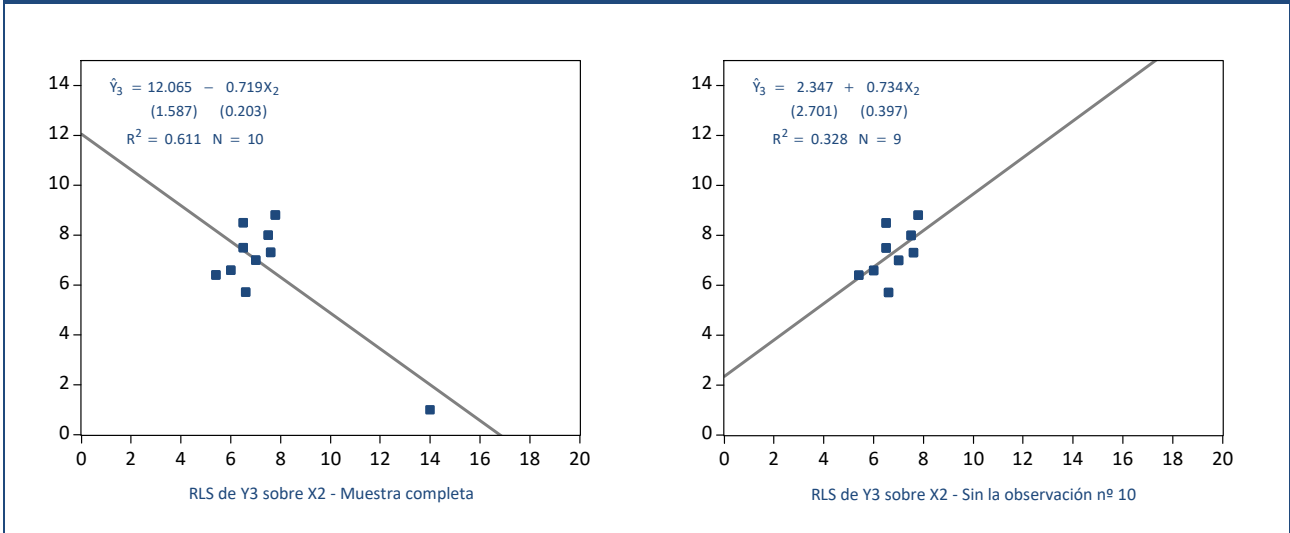
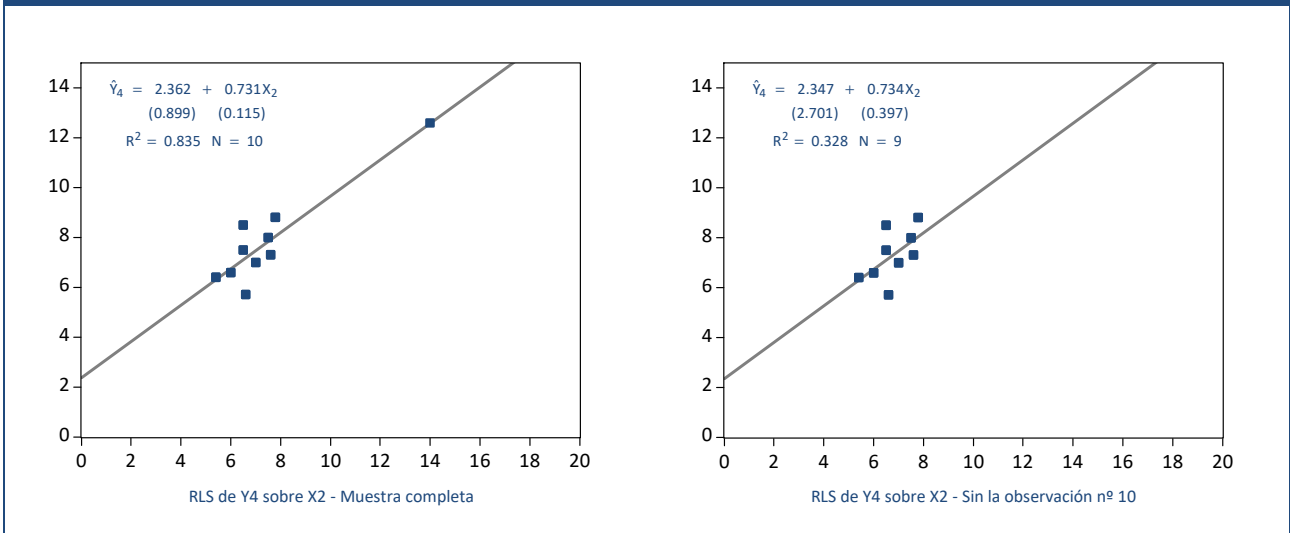


FIGURA 5



### DETECCIÓN DE OBSERVACIONES INFLUYENTES

⇒ **Observaciones atípicas** - Gráficos de residuos (⇒ residuos atípicos o anómalos).

⇒ **Observaciones extremas** - Grados de influencia potencial (*leverage*):

$$h_{ii} = \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \quad (i = 1, \dots, N) \Rightarrow 0 \leq h_{ii} \leq 1 \quad (i = 1, \dots, N), \quad \sum_{i=1}^N h_{ii} = K. \quad [1]$$

**Observación I:**  $h_{ii}$  es el elemento en la  $i$ -ésima posición de la diagonal principal de la matriz  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$  [llamada a veces "matriz sombrero" (*hat matrix*) porque convierte a  $\mathbf{y}$  en  $\hat{\mathbf{y}}$ :  $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ ]. En un modelo RLS,

$$h_{ii} = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad (i = 1, \dots, N).$$

En general, un valor destacado de  $h_{ii}$  (por ejemplo,  $h_{ii} > \frac{2K}{N}$ ) implica que los datos de las variables explicativas en la  $i$ -ésima observación destacan al compararlos con la media de

los datos de dichas variables en la muestra completa.

⇒ **Observaciones influyentes** en general - Estadísticos o distancias de Cook:

$$D_i = \frac{1}{K} \times \left[ \frac{\hat{u}_i}{\hat{\sigma} \sqrt{1 - h_{ii}}} \right]^2 \times \left[ \frac{h_{ii}}{1 - h_{ii}} \right] \quad (i = 1, \dots, N). \quad [2]$$

**Observación II:** En  $D_i$  se combina información sobre el posible carácter atípico (a través del residuo  $\hat{u}_i$ ) y/o extremo (a través del grado de influencia potencial  $h_{ii}$ ) de cada observación muestral. La distancia  $D_i$  es una medida de la diferencia entre las estimaciones de los parámetros de un modelo obtenidas con la muestra completa y las obtenidas sin la  $i$ -ésima observación (o, equivalentemente, entre los valores ajustados asociados con ambas estimaciones). Por lo tanto, un valor destacado de  $D_i$  [por ejemplo, mayor que el valor crítico al 5% en una  $F(K, N - K)$ ] suele implicar que la  $i$ -ésima observación muestral es una observación influyente.

### EJEMPLO

Ver IEV41: Sección 15 pp. 58-60. Las medidas de influencia de Hadi (que pueden ayudar a detectar observaciones influyentes **enmascaradas**) son

$$H_i = \left[ \frac{h_{ii}}{1 - h_{ii}} \right] + \left[ \frac{K}{1 - h_{ii}} \right] \times \left[ \frac{\hat{u}_i^2}{\text{SCR} - \hat{u}_i^2} \right] \quad (i = 1, \dots, N). \quad [3]$$

Para otras medidas de influencia, ver Peña, D.; Yohai, V.J. (1995), "The Detection of Influential Subsets in Linear Regression by Using an Influence Matrix", *Journal of The Royal Statistical Society, Series B*, 57: 145-156.