Investigaciones Económicas (Segunda época). Vol. XI, n.º 1 (1987) págs. 3-24

OBSERVACIONES INFLUYENTES EN MODELOS ECONOMETRICOS

Daniel PEÑA*

Universidad Politécnica de Madrid

La importancia de cada observación en la construcción de un modelo econométrico es generalmente muy distinta: es frecuente que un pequeño número de observaciones, que llamaremos observaciones influyentes, sea responsable de las propiedades más relevantes del modelo. La identificación de estas observaciones es vital para juzgar la robustez del modelo construido y para evitar graves errores de especificación tanto en los modelos estáticos (sección transversal) como dinámicos. Este trabajo ilustra este problema en modelos de regresión, presenta algunas medidas de influencia, analiza el efecto de las observaciones influyentes en la selección de modelos mediante el coeficiente de correlación y presenta un criterio de robustez interna para elegir entre modelos.

1. Introducción

Una ley empírica sobre la dimensión es que el «tamaño» de las cosas es muy desigual: existen siempre unos pocos elementos «grandes» que son responsables de una gran parte de la dimensión agregada y muchos elementos «pequeños» que contribuyen individualmente muy poco al conjunto total. Esta ley, que se conoce a veces en Economía como el principio de Pareto, se ha encontrado en campos tan distintos como la Economía (distribución de la renta, tamaño empresarial), Geografía (tamaño de ríos, montañas, ciudades), Ingeniería (tipos de defectos en Ingeniería de Procesos) o Lingüística (uso de las palabras de un idioma), y se resume a veces en la expresión: «pocos relevante y muchos triviales».

Un descubrimiento reciente es que esta ley aparece también al estudiar la información de cada observación en un modelo econométrico: unas pocas observaciones suelen ser responsables de la mayoría de las propiedades relevantes del modelo. Podemos tener un modelo con 1.000 observaciones y que tenga, sin embargo, en los aspectos más importantes, la fiabilidad y precisión de uno de 10.

Estas ideas se han desarrollado y hecho operativas en los últimos años para modelos de regresión estáticos, Belsley *et. al.* (1980), y Cook y Weisberg (1982)

* Este trabajo ha sido parcialmente financiado por una ayuda de la CAICYT. Agradezco los comentarios de Ricardo Sanz y de un evaluador anónimo que han contribuido a mejorarlo.

han sido pioneros en el desarrollo de métodos para medir la influencia de las observaciones. La extensión de estas ideas a modelos dinámicos está actualmente en sus inicios (Peña (1984), (1985), (1986a), 1986b)).

Este trabajo está estructurado como sigue: la sección 2 presenta un ejemplo de un modelo de regresión de sección transversal para relacionar indicadores de tamaño empresarial donde las propiedades del modelo dependen crucialmente de una observación. Este ejemplo pone de manifiesto los riesgos de no tener en cuenta la robustez interna, o estudio de la influencia en la muestra, y la necesidad de un estudio diagnóstico del efecto de cada observación.

Los fundamentos estadísticos de estos instrumentos diagnósticos se desarrollan brevemente en las secciones tres y cuatro y se aplican al ejemplo anterior. La sección cinco analiza los riesgos de utilizar un criterio automático, como el coeficiente de correlación corregido por grados de libertad o el criterio de Mallows para seleccionar entre modelos cuando hay observaciones influyentes. Finalmente, la sección seis define el coeficiente de robustez de un modelo como un criterio útil para la selección de modelos de regresión.

2. Análisis de algunos indicadores de tamaño empresarial por países

Para ilustrar los problemas que introducen las observaciones influyentes en los modelos econométricos de corte transversal o estáticos, utilizaremos unos datos de Berges (1986) sobre los valores medios de cuatro indicadores de tamaño en las empresas de quince países, que se presentan en cuadro 1. Supondremos que se desea construir un modelo para prever las ventas medias de las empresas de un país en función de los otros tres indicadores de dimensión.

CUADRO 1
Datos medios de dimensión empresarial

País	Ventas	Activos	Empleados	R. Propios
España	249	454	3.358	166
EE.UU.	3.334	2.612	15.230	1.209
Alemania	707	542	7.391	119
Inglaterra	511	352	7.307	243
Francia	477	535	6.306	91
Suecia	142	137	2.075	34
Suiza	494	475	6.163	215
Holanda	301	227	3.517	70
Italia	109	100	874	16
Bélgica	167	124	1.267	37
Noruega	100	81	894	14
Dinamarca	84	67	978	20
Finlandia	119	100	1.350	15
Portugal	35	46	1.302	16
Irlanda	237	283	3.668	80

Nota: Las medidas monetarias van en miles de dólares USA. Los datos provienen de una muestra de las 8.500 empresas europeas con mayor cifra de ventas y las 500 mayores americanas. Tomados de Berges (1986).

Existe considerable evidencia de que para describir adecuadamente o relacionar variables que representan dimensión (sea esta de ciudades, empresas, contribuciones científicas o consumo de electricidad) es conveniente utilizar logaritmos (véase Simon, H. (1978)). Los gráficos bivariantes de las ventas respecto a los otros indicadores parecen confirmarlo, aunque no hay que conceder mucho peso a esta información que puede ser engañosa. (Véase Peña (1987, cap. 13)). En consecuencia se han ajustado distintos modelos a los datos sin y con logaritmos (cuadros 2 y 3). El mejor modelo del cuadro 2 es el que incluye los Activos y Recursos propios como variables explicativas: tiene la menor varianza residual, ($\hat{s}_R^2$) y, por tanto, el mayor coeficiente de correlación corregido por grados de libertad ($\bar{R}$). (Para simplificar la exposición, no incluimos el criterio C_p de Mallows ya que su comportamiento es similar al de $\bar{R}^2$, véase Kennard (1971)). El mejor modelo para las variables en logaritmos es, con los mismos criterios, el que incluye únicamente log A como variable explicativa.

CUADRO 2
Resumen de modelos de regresión del tipo $V = \beta_0 + \Sigma\beta_i X_i$

β_0	$\beta(A)$	$\beta(NE)$	$\beta(RP)$	$\hat{s}_R$	R^2	$\bar{R}^2$
−34	0,86	0,006	0,81	104	0,987	0,984
	(0,27)	(0,017)	(0,51)			
−23,3	0,91	—	0,79	100	0,987	0,985*
	(0,23)		(0,49)			
−50	1,27	—	—	106	0,984	0,983
	(0,04)					

Nota: Las variables explicativas son Activos (A); Número de empleados (NE) y Recursos Propios (RP). Entre paréntesis errores estándar de los coeficientes. R^2 es el coeficiente de determinación y $\bar{R}^2$ el corregido por grados de libertad, y $\hat{s}_R^2$ la desviación típica residual. El * indica mejor modelo con los criterios de la tabla.

CUADRO 3
Resumen de modelos de regresión del tipo log $\hat{V} = \beta_0 + \Sigma\beta_i \ln x_i$

β_0	$\beta \log A$	$\beta \log NE$	$\beta \log RP$	$\hat{s}_R$	R^2	$\bar{R}^2$
−0,28	0,95	0,086	−0,015	0,296	0,942	0,927
	(0,26)	(0,27)	(0,22)			
0,1	0,982	—	0,017	0,285	0,942	0,932
	(0,23)		(0,19)			
0,064	1	—	—	0,274	0,942	0,937*
	(0,07)					

Nota: Las variables explicativas son los logaritmos de las de el Cuadro 2. El * indica el mejor modelo con los criterios de la tabla.

Una práctica frecuente en la literatura econométrica es seleccionar el modelo final mediante R^2 o $\bar{R}^2$. Esta regla carece de justificación cuando la variable dependiente no es la misma (por ejemplo, en uno es V y otro log V), y conduce

con alta frecuencia a modelos mal especificados cuando hay observaciones influyentes, como veremos en la sección 5. El criterio a seguir para comparar modelos con distintas transformaciones de la respuesta es ver cuál de ellos verifica la hipótesis básica de linealidad, normalidad, homocedasticidad e independencia. Cuando no pueda discriminarse entre ellos en estas bases, un criterio adicional es efectuar un estudio de sensibilidad para escoger el más robusto.

El gráfico 1 muestra los residuos de la regresión con datos sin transformar y el cuadro 4 lista estos datos ordenados, como en el gráfico, por valores previstos de $\hat{y}$. Aunque en el gráfico no se aprecia bien, el cuadro 4 muestra claramente curvatura: los valores positivos se concentran al principio y al final, y los negativos aparecen juntos entre ellos. Esto sugiere que la relación no es lineal. Además, el histograma de los residuos, gráfico 2, muestra que su distribución es muy asimétrica, lo que de nuevo sugiere la necesidad de transformar los datos.

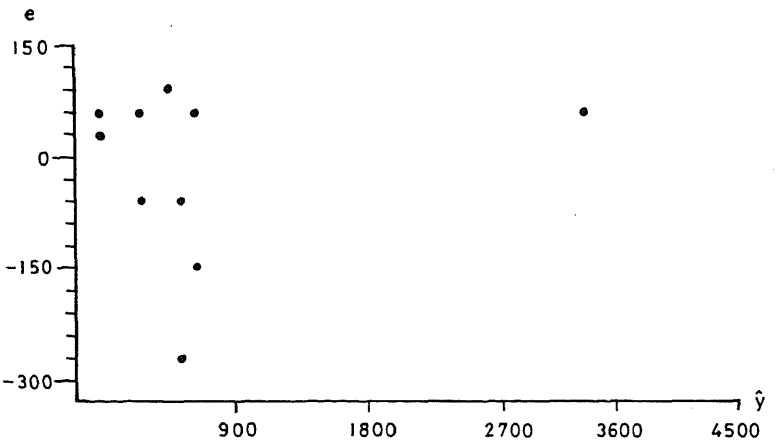


Gráfico 1. Residuos de la mejor regresión de datos sin transformar.

Centro del Intervalo	Número de Observaciones	
−300	1	•
−250	0	
−200	0	
−150	1	•
−100	0	
−50	2	• •
0	1	•
50	9	• • • • • • • • •
100	1	•

Gráfico 2. Histograma de los residuos del cuadro 4.

CUADRO 4
Datos del gráfico 1

$\hat{y}$	e
5,6	29,4
30,7	53,3
47,9	52,1
71,6	37,3
73,4	45,6
103,2	63,8
122,5	19,5
240,7	60,3
311,4	−74,5
411,5	99,5
542,7	−275,7
561,4	−67,4
637,2	−160,2
649,9	57,0
3.274,2	59,8

Al aplicar este mismo análisis al modelo con las variables en logaritmos, se obtienen los resultados que se resumen en los gráficos 3 y 4 y en el cuadro 5. El modelo en logaritmos parece más adecuado a pesar de tener un $\bar{R}^2$ menor.

Una relación bien establecida no debe modificarse radicalmente al eliminar o introducir un nuevo dato muestral. Para confirmar este aspecto, eliminaremos de la muestra el dato de EE.UU. y reestimaremos el modelo con los 14 datos europeos con y sin transformaciones. El cuadro 6 resume el resultado de este estudio de sensibilidad para los datos sin transformar. Se observa, que la relación es muy inestable, ya que los coeficientes cambian sustancialmente. De hecho, la variable más importante ahora para explicar las ventas de las empresas europeas es el número de empleados, en lugar de las variables financieras, y el mejor modelo, señalado con un * en el cuadro 6, incluye además los recursos propios pero con coeficiente negativo. Además la combinación de variables que condujo al mejor modelo con todos los datos (A y RP, véase cuadro 2) es claramente el peor modelo para los datos europeos.

Centro del Intervalo	Número de Observaciones
−0,7	1 •
−0,6	0
−0,5	0
−0,4	0
−0,3	1 •
−0,2	2 • •
−0,1	0
0,0	3 • • •
0,1	2 • •
0,2	5 • • • • •
0,3	1 •

Gráfico 3. Histograma de los residuos del modelo en logaritmos.

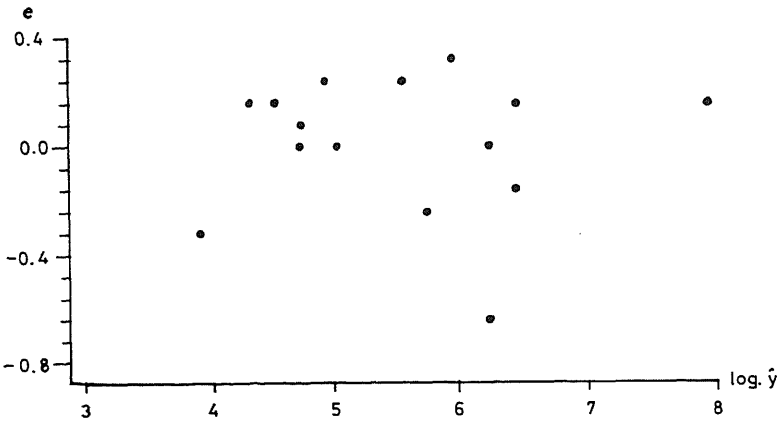


Gráfico 4. Residuos frente a valores previstos para el modelo en logaritmos.

CUADRO 5
Residuos y valores previstos por el modelo en logaritmos

$\log \hat{y}$	e
3,89	−0,34
4,27	0,15
4,46	0,14
4,67	0,10
4,67	0,02
4,89	0,22
4,99	−0,03
5,49	0,21
5,71	−0,25
5,93	0,30
6,19	−0,67
6,23	−0,03
6,35	−0,19
6,37	0,19
7,94	0,17

CUADRO 6
Modelos para datos sin transformar excluyendo EE.UU. Debajo de cada coeficiente entre paréntesis su error estándar de estimación

β_0	$\beta(A)$	$\beta(NE)$	$\beta(RP)$	s_R	R^2	$\bar{R}^2$
−2,7	0,18	0,08	−0,45	53,56	0,947	0,931
	(0,19)	(0,02)	(0,35)			
19,1	0,85	—	0,40	95,66	0,779	0,813
	(0,22)		(0,54)			
2,6	—	0,09	−0,42	53,38	0,942	0,931*
		(0,01)	(0,35)			
4,6	—	0,08	—	54,4	0,934	0,928
		(0,006)				

Por tanto, las conclusiones que podemos extraer de estos datos dependen totalmente de la inclusión o exclusión de la observación de EE.UU. que, en consecuencia, es muy influyente. Además, si analizamos los residuos del mejor modelo de el cuadro 6 (indicado por (*)) (gráficos 5 y 6) se observa que la distribución de los residuos es bastante simétrica y que ha desaparecido la curvatura indicativa de no linealidad. En resumen, una sola observación (el dato de EE.UU.) es responsable de las siguientes propiedades del modelo:

(1) qué variables son significativas;

(2) signo de los coeficientes;

(3) falta de linealidad;

(4) falta de normalidad de los residuos.

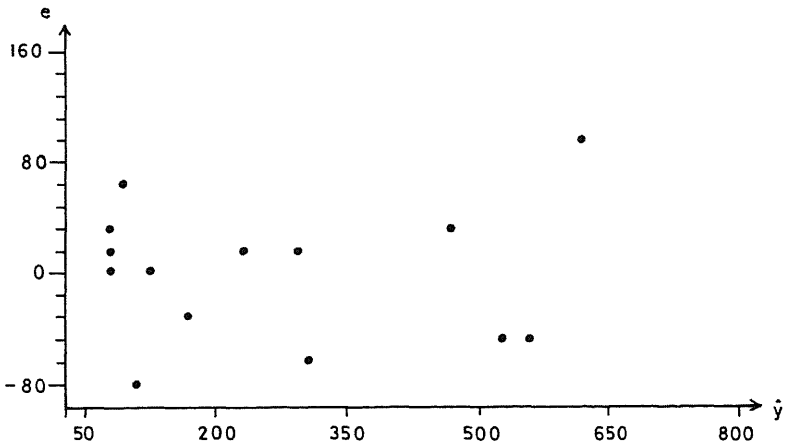


Gráfico 5. Residuos frente a valores estimados para el modelo (*) de el cuadro 6.

Centro del Intervalo	Número de Observaciones
−80	1 •
−60	2 • •
−40	2 • •
−20	0
0	2 • •
20	4 • • • •
40	1 •
60	1 •
80	0
100	1 •

Gráfico 6. Histograma de los residuos del modelo (*) del cuatro 6.

Si analizamos los datos europeos en logaritmos, se obtiene un modelo similar al construido para toda la muestra, por lo que concluimos que el modelo en logaritmos no cambia sustancialmente.

Este ejercicio muestra la importancia de un análisis de la influencia de las observaciones para obtener conclusiones razonables de los datos. Tenderemos a desconfiar de una teoría basada en una evidencia empírica tan débil como es una única observación. El lector puede pensar que los problemas del ejemplo son debidos al pequeño tamaño muestral y que se resolverían con una muestra grande que nos pondrá a salvo de esta inestabilidad. Como veremos en las secciones siguientes esta intuición es errónea y el lector puede encontrar ejemplos de observaciones influyentes en muestras a partir de 450 datos en Peña y Ruiz-Castillo (1982, 1984).

3. La influencia potencial de cada observación

3.1. *Fundamentos*

Vamos a demostrar que la influencia potencial de cada observación en un modelo de regresión depende de la similitud entre los valores de las variables explicativas en dicho punto y los valores medios en la muestra. Cuanto más «heterogéneo», en este sentido, es un punto mayor será su capacidad de influencia. Los resultados que presentamos a continuación, son debidos a Box y Draper (1975), Mosteller y Tukey (1977) y Huber (1980) entre otros.

Consideremos el modelo de regresión lineal estándar:

$$Y = X\beta + U$$

donde Y es un vector de n observaciones de una variable respuesta o endógena, X es una matriz $nx(k + 1)$ con los valores en columnas de k variables explicativas y una columna de unos, β es un vector de $(k + 1)$ parámetros y U un vector de variables aleatorias independientes normales, de media cero y varianza σ^2 constante. En estas condiciones, es bien conocido que la estimación mínimo-cuadrática del vector Y es:

$$\hat{Y} = X(X'X)^{-1}X'Y = VY \qquad [1]$$

llamando V a la matriz cuadrada de dimensión n, simétrica e idempotente que proyecta Y sobre el espacio definido por las columnas de X. Los residuos de la relación se definen por:

$$e = Y - \hat{Y} = (I - V)Y \qquad [2]$$

donde I es la matriz unidad. Por tanto, la matriz de varianzas y covarianzas para las predicciones será, según [1]:

$$\text{Var}\,(\hat{Y}) = V\sigma^2 \qquad [3]$$

y para los residuos

$$\text{Var } (e) = (I - V)\sigma^2 \qquad [4]$$

que implica que la varianza de la predicción en la observación iésima definida por la fila x_i' de la matriz X será:

$$\text{Var } (\hat{y}_i) = \sigma^2 v_{ii} \qquad [5]$$

siendo v_{ii} el elemento iésimo de la diagonal de la matriz V que, según [1], se calcula como:

$$v_{ii} = x_i'(X'X)^{-1}x_i \qquad [6]$$

Además, la varianza del residuo iésimo será:

$$\text{Var } (e_i) = \sigma^2(1 - v_{ii}) \qquad [7]$$

como se deduce de [4] y de la descomposición ortogonal $y_i = \hat{y}_i + e_i$, donde se verifica el teorema de Pitágoras.

$$\sigma^2 = \text{Var } (y_i) = \text{Var } (\hat{y}_i) + \text{Var } (e_i)$$

Como la varianza es siempre positiva, las ecuaciones [5] y [7] indican que los términos v_{ii} serán siempre positivos y menores que la unidad. También muestran que estos términos v_{ii} determinan la precisión de la estimación de cada punto. Según [5] cuanto mayor sea v_{ii} mayor será la varianza, o equivalentemente, menor la precisión, en la estimación y viceversa. La importancia de estos términos radica en que en la estimación de los parámetros del modelo cada observación interviene con un peso que es función de ellos. Por ejemplo, en regresión lineal simple la pendiente de la recta se estima mediante:

$$\hat{\beta} = \Sigma\left(v_{ii} - \frac{1}{n}\right)b_i \qquad [8]$$

donde $b_i = (y_i - \bar{y})/(x_i - \bar{x})$ es la pendiente de la recta que pasa por el punto $(y_i x_i)$ (véase Peña (1987) Cap. 12). Por tanto, la pendiente estimada es una media ponderada de las rectas que unen cada punto con el centro de gravedad de las observaciones. La expresión [8] corresponde a un promedio ya que los coeficientes que multiplican a cada pendiente parcial suman la unidad. En efecto, en regresión lineal:

$$v_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \qquad [9]$$

se demuestra que, en el caso general de k regresores, los términos v_{ii} verifican las propiedades siguientes:

a) $v_{ii} = \frac{1}{n}(1 + d_i)$, siendo $d_i = (x_i - \bar{x})'S^{-1}(x_i - \bar{x})$ la distancia entre cada punto x_i y la media $\bar{x}$ estandarizada por su matriz de varianzas y covarianzas S. Observemos que si las observaciones x estuviesen incorreladas (S fuese diagonal) d_i se reduce a:

$$d_i = \sum_{j=1}^{K} \frac{(x_{ji} - \bar{x}_j)^2}{s_j^2}$$

y es por tanto una suma de las distancias estandarizadas para las k variables. Si las variables exógenas o explicativas están incorrelacionadas d_i tiene en cuenta la dependencia ya que las líneas de nivel de la distancia d_i son elipsoides con centro $\bar{x}$. En efecto, entonces la matriz de varianzas y covarianzas de las observaciones, S, no será diagonal y el conjunto de puntos definido por d_i = cte., será un elipsoide ya que S se supone definida positiva.

b) $1/n \leqslant v_{ii} \leqslant 1$. Esta propiedad es inmediata a partir de la anterior. Como el valor mínimo de d_i es obviamente cero, v_{ii} no puede ser menor de $1/n$, además, por [7] debe ser menor que la unidad.

c) Σv_{ii} = rango $(X) = k + 1$. Este resultado se deduce teniendo en cuenta que si los productos existen, traza (AB) = traza (BA), por tanto:

$$\Sigma v_{ii} = \text{traza } (V) = \text{traza } \left((X'X)^{-1}X'X\right) = \text{traza } (I) = k + 1$$

3.2. *El número equivalente de observaciones*

La varianza al estimar una media con n datos es σ^2/n. Como la varianza al estimar la media en el punto x_i es σ^2/v_{ii}^{-1}, v_{ii}^{-1} puede interpretarse como el número equivalente de observaciones existente para esta estimación.

La estimación para $\bar{x}$ es $\bar{y}$. En este punto, al ser $d_i = 0$, $v_{ii} = 1/n$ y las n observaciones contribuyen igualmente a la estimación. Para puntos alejados de la media con $1/n < v_{ii} < 1$, el número equivalente de observaciones estará entre 1 y n ($n_e = v_{ii}^{-1}$).

Si $v_{ii} = 1$, $n_e = 1$, y la precisión en la estimación es equivalente a la que tendríamos con una observación. Para interpretar este resultado, observemos que según [7], la varianza del residuo en dicho punto será cero, es decir, el residuo tomará siempre su valor esperado que es cero y por lo tanto, la ecuación de regresión pasará *siempre* por dicho punto, *sea cual sea el valor observado para y*. En consecuencia, puntos con valor de v_{ii} próximo a uno, o con un número equivalente de observaciones sensiblemente menor que las demás, son potencialmente influyentes.

La ecuación [8] muestra que los v_{ii} equivalen también en la estimación de la pendiente al número de observaciones disponibles en cada punto. En efecto, la media de k observaciones $x_1, \ldots, x_K$ que aparecen con frecuencias relativas $fr(i)$ es:

$$\bar{x} = \Sigma fr(i)x_i \qquad [10]$$

comparando esta fórmula con la [8] concluimos que los términos $v_{ii} - 1/n$ se comportan como frecuencias relativas en la estimación de la pendiente a partir de las pendientes parciales.

Señalaremos por último que los valores v_{ii} de influencia potencial pueden interpretarse, en términos intuitivos aunque poco precisos, como la probabilidad que tienen a priori los puntos de ser influyentes: si v_{ii} es muy próximo a uno, el punto es con seguridad influyente, mientras que si v_{ii} es pequeño (próximo a $1/n$) es muy difícil que lo sea. Esta interpretación proviene de que si llamamos $\hat{y}_{(i)}$ a la previsión que haría un modelo sin incluirle para el punto (y_i, x_i) cuando se construye se verifica según [1]:

$$\hat{y}_i = v_{ii}y_i + \sum_{j \neq i} v_{ij}y_j \qquad [11]$$

por tanto, según [10] v_{ii} representa la «frecuencia relativa» asignada a la observación y_i en la estimación de la media $\hat{y}_i$. Esta expresión puede también escribirse (Peña (1987, Cap. 13)):

$$\hat{y}_i = v_{ii}y_i + (1 - v_{ii})\hat{y}_{(i)} \qquad [12]$$

que presenta la media como una combinación lineal del valor «a priori» y_i con peso v_{ii} y el valor estimado con el resto de la muestra, $\hat{y}_{(i)}$, con peso $(1 - v_{ii})$.

Por tanto, el valor v_{ii} puede interpretarse como el peso relativo que se asigna a la observación y_i frente al resto de la muestra para determinar las propiedades del modelo para $x = x_i$.

3.3. *Aplicación al estudio de la dimensión empresarial*

El gráfico 7 muestra las curvas de nivel de los coeficientes v_{ii} de influencia potencial en cualquier modelo de regresión lineal que contenga los 15 datos de activos y recursos propios de el cuadro 1 como variables explicativas. Se observa que la influencia potencial de EE.UU. es enormemente mayor que la de cualquier otra observación y, por tanto, cualquier modelo basado en estos datos tendrá que analizarse con gran precaución. Se observa también que un punto no necesita tener coordenadas «grandes» para ser muy influyente. Un país con las coordenadas del punto x tendría tanta influencia como EE.UU., aunque sus activos y recursos propios medios, al mirarlos aisladamente, no destacarían del resto de los países. A gran distancia del efecto de EE.UU.

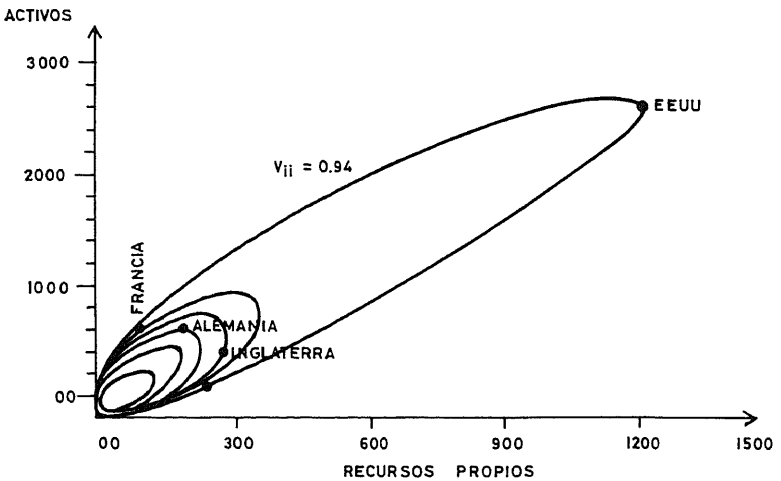


Gráfico 7. Influencia potencial de los 15 países.

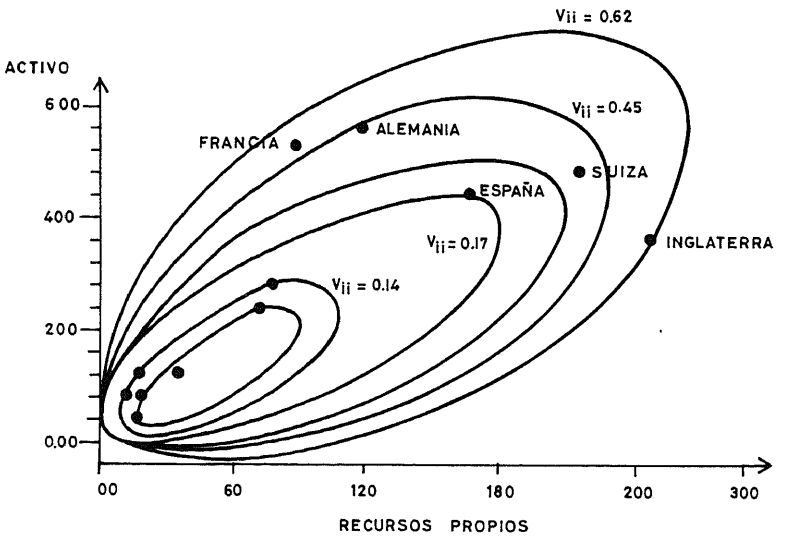


Gráfico 8. Influencia potencial de los países europeos.

aparecen Francia, Alemania e Inglaterra. Los dos primeros países tiene relativamente empresas con activos muy altos con relación a sus propios recursos, mientras que a Inglaterra le ocurre lo contrario. La influencia de estos tres países es, sin embargo, mucho menor que la de EE.UU., teniendo: v(Francia) = 0,43; v(Inglaterra) = 0,37; v(Alemania) = 0,30.

Al eliminar el dato de EE.UU. se obtienen las curvas del gráfico 8. La observación más influyente es Inglaterra (v_{ii} = 0,62) por las razones anteriores, pero su efecto a priori es moderado.

La razón de que al expresar las variables en logaritmos disminuya la influencia potencial de EE.UU. es clara a la vista del gráfico 7. Los logaritmos disminuyen las distancias relativas entre los puntos aproximando su influencia potencial. Observemos que si la muestra contuviese un punto de alta influencia del tipo presentado por el país imaginario X, su efecto no se eliminaría mediante ninguna transformación de las variables.

4. Influencia real de cada observación

4.1. *El Estadístico D. de Cook*

Hemos visto en la sección anterior que un punto será influyente a priori si las coordenadas de sus variables explicativas (o sus coordenadas de diseño) son atípicas. Sin embargo, su influencia real en la estimación es también función de la respuesta realmente observada según la ecuación [12], si ésta es análoga a la prevista para ese punto con el resto de los datos, su influencia será baja, mientras que en caso contrario, ésta será alta. Es intuitivo que, en consecuencia, la influencia real puede medirse:

a) Por el cambio del vector de predicción $\hat{Y}$ cuando se elimina dicho punto. Una posible medida sería

$$D_1(i) = \frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})}{(k+1)s_R^2} \qquad [13]$$

donde $\hat{Y} = VY$ es la estimación estándar y $\hat{Y}_{(i)} = X\hat{\beta}_{(i)}$ con $\hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)}$, es la estimación sin incluir la observación i. Hemos dividido por s_R^2 para hacer la medida adimensional y por el número estimado de parámetros para normalizar la expresión.

b) Por el cambio en la predicción del punto i. Si $\hat{y}_{(i)} = x_i'\hat{\beta}_{(i)}$, entonces, dividiendo por la varianza de predicción en dicho punto

$$D_2(i) = \frac{1}{(k+1)}\left(\frac{\hat{y}_i - \hat{y}_{(i)}}{s_R\sqrt{v_{ii}}}\right)^2 \qquad [14]$$

c) Por el cambio en el vector de parámetros $\hat{\beta}$. Estandarizando por su matriz de varianzas como en casos anteriores, se obtiene:

$$D_3(i) = \frac{1}{(k+1)}\frac{(\hat{\beta} - \hat{\beta}_{(i)})'X'X(\hat{\beta} - \hat{\beta}_{(i)})}{s_R^2} \qquad [15]$$

Es fácil comprobar que en un modelo de regresión estas tres medidas son idénticas, es decir, siempre:

$$D = D_1 = D_2 = D_3$$

Para jugar con el estadístico D —debido a Cook (1977)— cuando una observación es influyente utilizaremos la expresión [15]. Observemos que el intervalo de confianza conjunto a nivel $1 - \alpha$ para el vector de parámetros β viene dado por

$$\frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{(k + 1)s_R^2} < F(k + 1, n - k + 1; 1 - \alpha) \qquad [16]$$

donde $F(a, b; 1 - \alpha)$ es el percentil $1 - \alpha$ de la distribución F con a y b grados de libertad. Diremos que un punto (y_i, x_i) es influyente a nivel $1 - \alpha$, cuando $\hat{\beta}_{(i)}$ no está incluido en el intervalo de nivel $1 - \alpha$ construido a partir de $\hat{\beta}$, es decir, cuando

$$D(i) > F(k + 1, n - k + 1; 1 - \alpha)$$

Esta influencia se refiere al *conjunto* del vector de parámetros. Cuando existan muchas variables es conveniente construir medidas de influencia parcial sobre conjuntos de coeficientes. Estas medidas serán del tipo

$$D_p(i) = \frac{(\hat{\beta}_p - \hat{\beta}_{p(i)})'(X'X)_p(\hat{\beta}_p - \hat{\beta}_{p(i)})}{ps_R^2} \qquad [17]$$

donde $\hat{\beta}_p$ contiene p de los $(k + 1)$ componentes de $\hat{\beta}$ y $(X'X)_p$ es su matriz de varianzas y covarianzas que corresponde a una submatriz de $(X'X)^{-1}$. El estadístico $D_p(i)$ es especialmente útil cuando tenemos una muestra grande ya que entonces es raro que una observación tenga gran influencia en *todos* los parámetros del modelo, y, por tanto, con el criterio [16] una observación será raramente influyente aunque su efecto en un grupo de coeficientes sea muy destacado. La expresión [17] elimina este inconveniente. En particular, para un coeficiente individual su cambio estandarizado medido por $D_p(i)$ deberá compararse con un t^2 de Student y, por tanto, podremos considerar influyentes a aquellas observaciones que produzcan un cambio en un coeficiente individual mayor de 2 desviaciones típicas.

4.2. *Aplicación a los datos de dimensión*

Las medidas de influencia para los datos de indicadores del tamaño de las empresas ajustados al modelo que contiene A y RP como variables explicativas sin transformar se resume en el cuadro 7. Se observa que una sola observación, que corresponde a EE.UU., determina muchos parámetros del modelo ($D_2 = 11{,}24$). Esta observación no es atípica, sin embargo, de acuerdo con un test de

valores atípicos ($r_2 = 1{,}42$). La observación era potencialmente muy influyente ($\hat{n}_i = 1{,}05$) y su influencia se confirma por el valor de D_i. Al eliminar este dato y estimar el modelo anterior para los 14 datos europeos se obtuvo (cuadro 6):

$$\text{M1:}\quad \hat{V} = \underset{}{19{,}1} + \underset{(0{,}22)}{0{,}85\,A} + \underset{(0{,}54)}{0{,}40\,RP}$$

CUADRO 7
Estadísticos de Influencia

	V	$\hat{V}$	e	r_i	v_{ii}	$\hat{n}_i$	D_i
1	249	519,51	−270,506	−2,79401	0,070080	14,2695	0,1961
2	3.334	3.300,19	33,813	1,41915	0,943679	1,0597	11,2485
3	707	562,17	144,829	1,73186	0,306208	3,2658	0,4413
4	511	487,85	23,154	0,29167	0,374811	2,6680	0,0170
5	477	533,71	−56,705	−0,75371	0,438452	2,2808	0,1478
6	142	127,79	14,206	0,14754	0,080206	12,4678	0,0006
7	494	577,26	−83,256	−0,86748	0,086180	11,6036	0,0237
8	301	237,84	63,160	0,65326	0,072594	13,7753	0,0111
9	109	80,03	28,975	0,30152	0,083867	11,9236	0,0028
10	167	118,38	48,624	0,50639	0,085316	11,7217	0,0080
11	100	61,22	38,783	0,40456	0,088287	11,3267	0,0053
12	84	53,26	30,737	0,32270	0,099928	10,0072	0,0039
13	119	79,24	39,765	0,41377	0,083723	11,9441	0,0052
14	35	31,06	3,939	0,04158	0,109690	9,1166	0,0001
15	237	296,52	−59,517	−0,61703	0,076980	12,9903	0,0106

y al estimar el modelo con las tres variables:

$$\text{M2:}\quad \hat{V} = -2{,}7 + \underset{(0{,}19)}{0{,}18\,A} + \underset{(0{,}02)}{0{,}08\,NE} - \underset{(0{,}35)}{0{,}45\,RP};$$

Por tanto para los datos Europeos la variable más significativa para explicar las ventas es el número de empleados ($t = 5{,}01$) mientras que, como vimos previamente, al incluir EE.UU. esta variable deja de ser significativa. Observemos que la desviación típica de predicción se reduce a la mitad al eliminar EE.UU.

El cuadro 8 indica las medidas relevantes para el modelo M2 (que no utiliza el dato de EE.UU.). Se incluye su valor previsto por el modelo, (1.115,69), que está muy por debajo del valor observado y la distancia de EE.UU. al resto de los datos utilizando la expresión:

$$v_{\text{EE.UU.}}(\text{sin EE.UU.}) = x'_{\text{EE.UU.}}(X'X)^{-1}{}_{(\text{sin EE.UU.})}x'_{\text{EE.UU.}}$$

Se observa que el número equivalente de observaciones para prever las ventas en EE.UU. (observación 2 del cuadro 7) cuando no se incluye en la muestra es casi cero (0,0188). Por tanto, los datos europeos no permiten extrapolaciones para EE.UU.

CUADRO 8
Estadísticos de influencia al eliminar el dato de EE.UU.
Modelo con las tres variables como explicativas y sin transformar (M2)

	V	$\hat{V}$	e_i	r_i	v_{ii}	$\hat{n}_i$	D_i
1	249	267,49	−18,4868	−0,71032	0,7639	1,3091	0,40814
2	*	1.115,69	*	*	53,1295	0,0188	*
3	707	621,62	85,3842	2,19929	0,4746	2,1068	1,09250
4	511	525,49	−14,4940	−0,58646	0,7871	1,2705	0,31790
5	477	547,59	−70,5877	−1,82234	0,4771	2,0962	0,75737
6	142	169,62	−27,6208	−0,54645	0,1095	9,1322	0,00918
7	494	469,90	24,0962	0,54828	0,3268	3,0602	0,03648
8	301	282,88	18,1206	0,35609	0,0974	10,2669	0,00342
9	109	76,64	32,3601	0,65432	0,1475	6,7801	0,01852
10	167	102,38	64,6195	1,29163	0,1276	7,8364	0,06101
11	100	75,74	24,2599	0,49001	0,1457	6,8645	0,01024
12	84	77,17	6,8344	0,13826	0,1483	6,7417	0,00083
13	119	114,55	4,4530	0,08925	0,1324	7,5545	0,00030
14	35	100,73	−65,7325	−1,35833	0,1838	5,4415	0,10385
15	237	300,21	−63,2061	−1,22916	0,0784	12,7624	0,03211

Analicemos ahora los modelos en logaritmos. El cuadro 9 indica las medidas de influencia para un modelo de todos los datos que contiene el activo en logaritmos como una única variable explicativa del logaritmo de las ventas. Se observa que, al transformar, ningún punto es muy influyente, lo que sugiere que este modelo puede representar bien los datos en todo el rango de variación. Para comparar con el caso anterior, vamos a estimar el modelo en logaritmos para las tres variables eliminando el dato de USA. El resultado es:

$$\log \hat{V} = \underset{}{-0{,}50} + \underset{(0{,}28)}{0{,}86} \ln A + \underset{(0{,}29)}{0{,}20} \ln NE - \underset{(0{,}24)}{0{,}08} \ln RP;$$

$$\hat{s}_R = 0{,}297$$

$$\bar{R}^2 = 0{,}88$$

CUADRO 9
Medidas de influencia para el modelo en logaritmos

	$\ln V$	$\ln \hat{V}$	e	r	v	$\hat{n}$	D_i
1	5,51745	6,19031	−0,672861	−2,58539	0,099100	10,0908	0,367638
2	8,11193	7,94251	0,169520	0,83445	0,451723	2,2137	0,286844
3	6,56103	6,36773	0,193303	0,75030	0,117152	8,5359	0,037351
4	6,23637	5,93550	0,300874	1,14409	0,080132	12,4793	0,057013
5	6,16752	6,35471	−0,187193	−0,72598	0,115692	8,6436	0,034477
6	4,95583	4,99054	−0,034713	−0,13209	0,081425	12,2812	0,000773
7	6,20254	6,23559	−0,033058	−0,12732	0,103330	9,6778	0,000934
8	5,70711	5,49621	0,210903	0,79618	0,066698	14,9929	0,022651
9	4,69135	4,67529	0,016054	0,06196	0,106957	9,3495	0,000230
10	5,11799	4,89070	0,227291	0,86808	0,088152	11,3440	0,036425

	ln V	ln $\hat{V}$	e	r	v	$\hat{n}$	D_i
11	4,60517	4,46428	0,140889	0,55122	0,131063	7,6299	0,022914
12	4,43082	4,27426	0,156554	0,62207	0,157586	6,3457	0,036195
13	4,77912	4,67529	0,103829	0,40070	0,106957	9,3495	0,009615
14	3,55535	3,89769	−0,342343	−1,41698	0,223629	4,4717	0,289173
15	5,46806	5,71701	−0,248950	−0,94168	0,070403	14,2040	0,033579

y las medidas de influencia se indican en el cuadro 10. Comparando este modelo con el que incluye EE.UU. se observa que no cambia sustancialmente.

La conclusión de este ejercicio es que la mejor variable para prever las ventas medias de las empresas de un país en todo el rango de variación de los datos, parece ser el Activo medio de dichas empresas. Este resultado parece consistente con la muestra y confirma una elasticidad unitaria entre ambas medidas del tamaño.

CUADRO 10
Medidas de influencia, datos en logaritmos, sin incluir EE.UU.

	ln V	ln $\hat{V}$	e	r	v	$\hat{n}$	D
1	5,51745	6,02196	−2,58539	−0,504508	0,48423	2,0651	1,31088
2	*	7,68089	0,83445	*	1,20031	0,8331	*
3	6,56103	6,36119	0,75030	0,199845	0,33537	2,9818	0,08579
4	6,23637	5,93300	1,14409	0,303369	0,46542	2,1486	0,42408
5	6,16752	6,33825	−0,72598	−0,170736	0,41747	2,3954	0,10147
6	4,95583	5,01678	−0,13209	−0,060949	0,09619	10,3960	0,00124
7	6,20254	6,16503	−0,12732	0,037503	0,23618	4,2341	0,00161
8	5,70711	5,50270	0,79618	0,204412	0,09868	10,1336	0,01436
9	4,69135	4,62771	0,06196	0,063643	0,29808	3,3548	0,00693
10	5,11799	4,82372	0,86808	0,294276	0,23553	4,2457	0,09875
11	4,60517	4,46164	0,55122	0,143527	0,21728	4,6023	0,02067
12	4,43082	4,28949	0,62207	0,141322	0,22148	4,5151	0,02065
13	4,77912	4,72161	0,40070	0,057511	0,25034	3,9946	0,00417
14	3,55535	4,04224	−1,41698	−0,486891	0,55246	1,8101	1,85010
15	5,46806	5,69038	−0,94168	−0,222323	0,09130	10,9531	0,01546

Hemos incluido junto a cada regresión el R^2, para mostrar como este coeficiente *disminuye* cuando pasamos de una relación con una observación muy influyente, que indica no linealidad, a otra que elimina la observación influyente. Analizaremos las razones teóricas de este hecho en la sección siguiente.

5. Observaciones influyentes y R^2

Cuando la muestra contiene una observación potencialmente influyente y comparemos distintas transformaciones de la variable dependiente, vamos a comprobar que cualquier transformación que «estire» la nube de puntos y haga,

por tanto, a dicha observación todavía más influyente, tenderá a aumentar el R^2 y viceversa. Por tanto, si partimos de una relación lineal con una observación potencialmente algo influyente y aplicamos una transformación que estira la nube de puntos y convierte la relación en no lineal, la influencia de esta observación y el coeficiente de correlación, tenderán, en general, a aumentar.

Para justificar este resultado que parece ir contra la intuición, consideremos un modelo de regresión lineal simple. Supongamos que la relación es lineal en los datos originales y sea r el coeficiente de correlación en una muestra concreta. Sea (x_i, y_i) el punto más influyente en dicha muestra, que será aquel con $|x_i - \bar{x}|$ máxima. Supongamos, sin pérdida de generalidad, que $x_i - \bar{x} > 0, y_i - \bar{y} > 0$, y desplazamos su coordenada y_i verticalmente, pasando del punto (x_i, y_i) al $(x_i, y_i + \Delta s_y)$ con $\Delta > 0$, sin alternar los $n - 1$ puntos muestrales restantes. Se demuestra en el Apéndice que entonces:

1) Un desplazamiento pequeño de y aumenta siempre el coeficiente de correlación (la derivada de r respecto a Δ es positiva).

2) Si llamamos $r(\Delta)$ al coeficiente de correlación al desplazar y, $r(\Delta)$ aumenta con Δ y luego disminuye tendiendo hacia un valor límite cuando $\Delta \to \infty$ que es $\sqrt{v_{ii} - 1/n}$, siendo v_{ii} la influencia potencial del punto modificado.

3) Si v_{ii} es alto, podemos desplazar mucho y, por ejemplo 15 desviaciones típicas, y aumentar el coeficiente de correlación (véase el gráfico 9).

La primera conclusión de este resultado es la siguiente: si los datos son, por ejemplo, lineales en logaritmos como en el ejemplo, expresarlos en unidades originales equivale a aplicarle una transformación exponencial que «expande» especialmente los puntos extremos y aumenta, en consecuencia, el coeficiente de correlación.

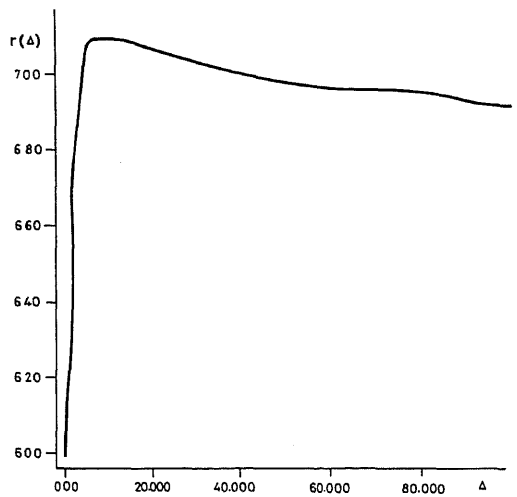


Gráfico 9. Evolución del coeficiente de correlación al desplazar Δ desviaciones típicas un punto con coordenadas estandarizadas $\left((x - \bar{x})/s_x = 3, (y - \bar{y})/s_y = 3\right)$ y 20 datos.

La segunda conclusión es que cuando una muestra contenga un dato erróneo que esté situado en una posición muy influyente —por ejemplo, tomamos mal el valor de y para un x alejado de la media— este error tenderá a inflar el coeficiente de correlación y producirá una sensación de «ajuste» mayor de lo que corresponde a la muestra.

6. Coeficiente de robustez

Un procedimiento atractivo de juzgar la adecuación del modelo es el siguiente: dejar una observación fuera, construir un modelo con las $n - 1$ restantes y prever la observación eliminada. Haciendo esto para las n observaciones se obtiene una medida «externa» de la precisión del modelo dada por el error cuadrático de validación:

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

que utilizando (4,2) puede escribirse:

$$EC_v = \Sigma\left(\frac{1}{1 - v_{ii}}\right)^2 e_i^2$$

Algunos autores —Stone (1974)— han propuesto elegir los estimadores $\hat{\beta}$ de manera que minimicen EC_v. La lógica de este procedimiento es tomar como estimadores los que mejor predigan observaciones externas, y conduce a una estimación que requiere mínimos cuadrados ponderados. El inconveniente de este criterio es que proporciona estimadores poco robustos: las observaciones con mayor influencia potencial por estar más alejadas del resto —por tanto con v_{ii} próximos a la unidad— tendrán un peso enorme en la determinación de los parámetros. Por tanto no recomendamos utilizarlo. Sin embargo, el cociente:

$$B^2 = \frac{\Sigma(y - \hat{y}_i)^2}{\Sigma(y - \hat{y}_{(i)})^2} = \left[\frac{1}{n - (k + 1)}\left(\frac{1}{1 - v_{ii}}\right)^2\left(\frac{e_i}{\hat{s}_R}\right)^2\right]^{-1}$$

representa una medida de la robustez del modelo. Es obvio que:

$$0 < B^2 < 1$$

cuando las predicciones $\hat{y}_{(i)}$ sean próximas a $\hat{y}_i$, el valor de B^2, coeficiente de robustez, será próximo a uno, mientras que si hay gran diferencia entre ambas B^2 será próximo a cero.

Como ejemplo de la utilización de este coeficiente la tabla presenta su aplicación a un conjunto de datos construido por Anscombe (1973).

Los cuatro conjuntos de datos conducen a la misma ecuación de regresión y al mismo valor para los estadísticos R^2, $\hat{s}_R$. Sin embargo, su coeficiente de robus-

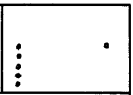
tez, B^2, indica claramente las diferencias entre ellos. Los datos se indican en el cuadro 11. El cuadro 12 los presenta gráficamente conjuntamente con el valor de B^2, que discrimina perfectamente entre el modelo bien especificado (a), los que presentan algún problema (b) y (c), y el que está basado en sólo una observación (d).

CUADRO 11
Datos de Anscombe

Observación	$X(a)\,(b)$ y (c)	$Y(a)$	$Y(b)$	$Y(c)$	$x(d)$	$Y(d)$
1	10	8,04	9,14	7,46	8	6,58
2	8	6,95	8,14	6,77	8	5,76
3	13	7,58	8,74	12,74	8	7,71
4	9	8,81	8,77	7,11	8	8,84
5	11	8,33	9,26	7,81	8	8,47
6	14	9,96	8,10	8,84	8	7,04
7	6	7,24	6,13	6,08	8	5,25
8	4	4,26	3,10	5,39	19	12,50
9	12	10,84	9,13	8,15	8	5,56
10	7	4,82	7,26	6,42	8	7,91
11	5	5,68	4,74	5,73	8	6,89

Nota: El cuadro 11 presenta datos para cuatro modelos distintos. Los datos bajo el encabezamiento $X(a)\,(b)$ y (c) representan los valores numéricos de la variable explicativa que son comunes para las regresiones con $Y(a)$, $Y(b)$ e $Y(c)$. Tenemos pues tres modelos distintos, por serlo la Y, aunque la X sea común a los tres. Por último $Y(d)$ y $X(d)$ definen conjuntamente otra regresión.

CUADRO 12
Aplicación de B^2 a datos de Anscombe (1973)

Gráfico				
Regresión	$\hat{y} = 3 + 0{,}5 \quad ; \quad s_R^2 = 1{,}52 \quad ; \quad r = 0{,}82$			
B^2	0,67	0,57	0,58	0,27

Apéndice

Supongamos que convertimos el punto (x_i, y_i) en $(x_i, y_i + \Delta s_y)$. El nuevo coeficiente de correlación será:

$$r(\Delta) = \frac{\frac{1}{n}\sum_{j \neq i} x_j y_j + \frac{1}{n} x_i (y_i + \Delta s_y) - \bar{x}\left(\bar{y} + \frac{\Delta s_y}{n}\right)}{s_x \sqrt{n \sum_{j \neq i} y_j^2 + \frac{1}{n}(y_i + \Delta s_y)^2 - \left(\bar{y} + \frac{\Delta s_y}{n}\right)^2}}$$

que puede escribirse:

$$r(\Delta) = \frac{\text{Cov}\ (x, y) + \frac{1}{n} s_y \Delta (x_i - \bar{x})}{s_x \sqrt{s_y^2 + \frac{\Delta^2 s_y^2}{n}\left(1 - \frac{1}{n}\right) + \frac{2\Delta s_y}{n}(y_i - \bar{y})}}$$

dividiendo cada término por $s_x s_y$ y llamando $\mathcal{Z}_x = (x_i - \bar{x})/s_x$; $\mathcal{Z}_y = (y_i \bar{y})/s_y$:

$$r(\Delta) = \frac{r + \frac{1}{n}\Delta \mathcal{Z}_x}{\sqrt{1 + \frac{\Delta^2}{n}\left(1 - \frac{1}{n}\right) + \frac{2\Delta}{n}\mathcal{Z}_y}} \qquad [A1]$$

Vamos a obtener la derivada de r. Llamando D al denominador de la fracción anterior:

$$\frac{r(\Delta) - r}{\Delta} = \frac{r(1 - D) + \frac{1}{n}\Delta \mathcal{Z}_x}{D\Delta}$$

y tomando límites cuando $\Delta \to 0$, como $D \to 1$, tendremos:

$$\frac{dr}{d\Delta} = \frac{1}{n}\Delta \mathcal{Z}_x$$

Por tanto, si Δ y $\mathcal{Z}_x$ tienen el mismo signo, el coeficiente de correlación aumentará siempre. Observemos que como el coeficiente de correlación es simétrico en ambas variables, el resultado anterior es válido permutando x por y.

Cuando $\Delta \to \infty$, la expresión [A.1] muestra que:

$$\lim_{\Delta \to \infty} r(\Delta) = \frac{\mathcal{Z}_x}{\sqrt{n - 1}}$$

En regresión lineal, como

$$v_{ii} = \frac{1}{n}(1 + \mathcal{Z}_x^2)$$

sustituyendo en la expresión anterior y suponiendo que $n_1 - 1 \simeq n$

$$\lim_{\Delta \to \infty} r(\Delta) = \sqrt{v_{ii} - \frac{1}{n}}$$

es interesante que los cuadrados de estos valores límites son precisamente los términos de ponderación en la ecuación [8].

Referencias

Anscombe, T. W. (1973): «The analysis of Residuals», *The American Statistician 27*, págs. 17-21.

Barnett, V. y Lewis, T. (1978): *Outliers in Statistical Data*, Wiley.

Belsley, D. A., Kuh, E. y Welsch, R. E. (1980): *Regression Diagnostics*, Wiley.

Berges, A. (1986): «La medición de la dimensión empresarial: una comparación internacional», *Investigaciones Económicas*, Suplemento (1986), págs. 7-18.

Box, G. E. P. y Draper, N. (1975): «Robust Design», *Biometrika 62*, 347-52.

Cook, R. D. (1977): «Detection of influential observations in lineal regression», *Technometrics*, 19, 15-18.

Cook, R. D. y Weisberg, S. (1982): *Residuals and Influence in Regressión*, Chapman y Hall.

Huber, P. (1981): *Robust Statistics*, Wiley.

Kennard, R. W. (1971): «A note on the C_p Statistics», *Technometrics, 13*, 899-900.

Mosteller, F. y Tukey, J. (1977): *Regression: A second course in Statistics*, Addison-Wesley.

Peña, D. (1984): «Influential observations in time series», *Mathematics Research Center Technical Report núm. 2.718.* University of Wisconsin. Madison.

Peña, D. (1985): «A measure of Influence in autorregresive models». *Procedings of the 100th Session of the International Statistical Institute*, 2, 481-482.

Peña, D. (1986a): «Measuring the importance of outliers in ARIMA models», en *New perspectives of Statistics*, Puri, M. *et. al.* editores. Wiley (en prensa).

Peña, D. (1986b): «Comments on local influence by D. R. Cook», *Journal of Roy. Statist. Society, B*, 48, 2, págs. 164-165.

Peña, D. (1987): *Estadística: Modelos y Métodos. Tomo II: Modelos lineales y Series Temporales*, «Alianza Universidad Textos». (Próxima aparición.)

Peña, D. y Ruiz-Castillo, J. (1982): «Métodos Robustos de construcción de modelos de regresión. Una aplicación al sector de la vivienda», *Estadística Española*, 97, 47-76.

Peña, D. y Ruiz Castillo, J. (1984): «Robust methods of Building Regression Models», *Journal of Business and Economic Statistics*, 2, 1, 10-20.

Simon, H. A. (1978): «The sizes of Things», en *Statistics: A guide to the unknown*, Tuner *et. al.* (editores). Holden-Day.

Stone, M. (1974): «Cross-validatory choice and assessment of statistical predictions», *J. Roy, Statist, Soc., B. 36*, 111-47.

Abstract

In building an econometric model the relative importance of each piece of data varies widely. Often, a small number of data points, that are called influential observations, are responsible for the main features of the fitted model. Identification of these points is needed to assess model robustness and to avoid gross misspecification errors. This work shows the problem in regression models, discusses the effect of influential points on the correlation coefficient and presents a robust criteria to choose among models.

Recepción del original, octubre de 1986.
Versión final, noviembre de 1986.