ESTADISTICA ESPAÑOLA
Núm. 97, 1982, págs. 47 a 76

Métodos robustos de construcción de modelos de regresión. Una aplicación al sector de la vivienda (*)

por DANIEL PEÑA
Universidad Politécnica de Madrid

y JAVIER RUIZ-CASTILLO
Universidad Complutense de Madrid

RESUMEN

Este trabajo analiza procedimientos de robustificación del modelo lineal. Se comparan las ventajas de utilizar un procedimiento de estimación robusto basado en M-estimadores con un análisis interno de la robustez de los mínimos cuadrados a la muestra. Se señalan las ventajas teóricas de este último enfoque y se ilustra su aplicación mediante la construcción de un modelo explicativo de los determinantes del alquiler en el Area Metropolitana de Madrid.

Palabras clave: Métodos robustos, observaciones con influencia, regresión, distancias generalizadas.

1. INTRODUCCION

La construcción de un modelo estadístico con datos que no han sido recogidos mediante un diseño experimental cuidadoso requiere, en primer lugar, dedicar una

* Este trabajo forma parte de una investigación más amplia financiada por el Ministerio de Economía y Comercio. Los autores desean agradecer la colaboración de José Antonio Quintero, del Centro de Cálculo de ese Ministerio, que se ocupó de los aspectos informáticos de la parte empírica de este artículo.

atención especial a aquellas observaciones no homogéneas con el resto de la muestra y que pueden introducir sesgos importantes en los resultados del análisis.

Este es un problema fundamental en la construcción de modelos lineales con muestras de corte transversal donde, típicamente, se dispone de un número elevado de datos de muchas variables. El problema es grave, ya que unas pocas observaciones atípicas pueden invalidar la hipótesis de normalidad en el modelo lineal y destruir la optimalidad del procedimiento de estimación mínimo-cuadrático que puede convertirse en muy poco eficiente.

En esta sección vamos a presentar la formulación general del problema de estimación por máxima verosimilitud de un modelo lineal. En la sección 2 comentamos el efecto de las anomalías en el modelo de regresión y una panorámica de las distintas soluciones posibles. En la sección 3 se discuten las ventajas y limitaciones de los métodos robustos de estimación, y en particular de los M-estimadores de Huber. En la sección 4 presentamos la necesidad de robustificar la metodología de construcción del modelo, y resumimos las aportaciones existentes para realizar un estudio interno de sensibilidad a la muestra utilizada de un modelo estimado por mínimos cuadros. Estas ideas se aplican en la sección 5 a la construcción de un modelo explicativo de las variables determinantes del alquiler. Finalmente, la sección 6 incluye algunos comentarios finales. La descripción de los datos utilizados se recoge en un apéndice.

Comencemos revisando brevemente la estimación del modelo lineal:

$$\underline{Y} = \underline{\underline{X}}\underline{\beta} + \underline{U} \qquad [1.1]$$

donde $\underline{Y}$ es un vector de respuestas $n \times 1$, $\underline{\underline{X}}$ es una matriz de rango k de variables predeterminadas, con dimensión $n \times k$, $\underline{\beta}$ es un vector de k parámetros y $\underline{U}$ es un vector $n \times 1$ de perturbaciones. La estimación máximo-verosímil de [1.1], llamando f a la función de densidad de Y y suponiendo $E[\underline{U}] = \underline{0}$ y $E[\underline{U}\underline{U}'] = \sigma^2\underline{\underline{I}}$, conducirá a

$$\text{máx} \sum_{i=1}^{n} \ln f(y_i - \underline{x}_i'\underline{\beta}) = \text{mín} \sum_{i=1}^{n} g(y_i - \underline{x}_i'\underline{\beta}) \qquad [1.2]$$

donde $-g = \ln f$. En la hipótesis de que f es derivable, el estimador máximo-verosímil de $\underline{\beta}$ es la solución (supuesta única), del sistema de ecuaciones

$$\sum_{i=1}^{n} \psi(y_i - \underline{x}_i'\underline{\beta})\underline{x}_i' = \underline{0}' \qquad [1.3]$$

donde $\psi = g' = -f'f^{-1}$ y $\underline{x}_i'$ es el vector $(1 \times k)$ de la fila i de la matriz $\underline{\underline{X}}$. El sistema anterior puede escribirse de forma compacta definiendo $\underline{\psi}'(i, \beta)$ como el vector fila $1 \times n$, cuyos componentes son los valores de la función ψ en los n puntos muestrales. Entonces [1.3] equivale a:

$$\underline{\psi}'(i, \beta) \cdot \underline{\underline{X}} = \underline{0}' \qquad [1.4]$$

Si $e_i = y_i - \underline{x}_i\beta$ y $w_i = \psi(e_i)/e_i$, otra manera ilustrativa de escribir estos resultados es

$$\sum e_i \underline{x}_i' w_i = \underline{0}' \qquad [1.5]$$

que revela cómo la estimación máximo-verosímil del modelo lineal conduce a mínimos-cuadrados ponderados, pero donde los pesos w_i dependen de los parámetros desconocidos β_i. Por tanto, la solución general del sistema de ecuaciones [1.5] tendrá que obtenerse iterativamente.

En resumen, concluimos que la estimación máximo-verosímil del modelo lineal puede interpretarse: 1) como la minimización de una cierta función g de los residuos muestrales; 2) como la determinación de una función ψ de los residuos cuyos componentes sean ortogonales al espacio vectorial generado por las columnas de $\underline{\underline{X}}$, y 3) como mínimos cuadrados ponderados con los pesos determinados iterativamente.

Si admitimos las hipótesis [1.2] respecto a $\underline{U}$ y convenimos en que su distribución es simétrica, una formulación general de la misma es la familia exponencial potencial propuesta por Diananda (1949) y Box (1953), y estudiada por Box y Tiao (1973). Su función de densidad es

$$f(u) = k_1(\alpha)\sigma^{-1} \exp\left\{-k_2(\alpha)\left|\frac{u}{\sigma}\right|^{\frac{2}{1+\alpha}}\right\} \qquad [1.6]$$

$$-1 < \alpha \leq 1 \qquad \sigma > \infty \qquad -\infty < u < \infty$$

donde σ es la desviación típica y el parámetro α es indicativo de la kurtosis o aplastamiento de la distribución. Para $\alpha = 0$ la distribución es la normal, para $\alpha = +1$ la distribución es la doble exponencial o de Laplace y para $\alpha \to -1$ se obtiene, como límite, la distribución uniforme. La función [1.6] incluye además desde distribuciones con colas más amplias que la normal (leptokúrticas para $\alpha > 0$), hasta distribuciones con colas muy poco marcadas (platikúrticas si $\alpha < 0$). La maximización de la verosimilitud de un modelo lineal con perturbaciones dadas por [1.6] conduce a:

$$\sum_{i=1}^{n} |y_i - \underline{x}_i\beta|^{\frac{2}{1+\alpha}} \qquad [1.7]$$

que incluye como casos particulares la minimización de desviaciones absolutas ($\alpha = 1$), mínimos cuadrados ($\alpha = 0$) y, en el límite cuando $\alpha \to -1$, la minimización de la desviación máxima.

Observamos que para una distribución de $\underline{U}$ simétrica el criterio adecuado es muy dependiente de la propia distribución de $\underline{U}$. Un criterio robusto ante anomalías, como la desviación de desviaciones absolutas, es poco eficiente si los datos son normales, mientras que mínimos cuadrados es muy poco eficiente si existen unas pocas observaciones atípicas que contaminan la distribución haciéndola leptokúrtica. Este efecto se comenta en la sección siguiente.

2. EL EFECTO DE LAS ANOMALIAS

En esta sección supondremos que las perturbaciones $\underline{U}$ del modelo lineal son $N(0, \sigma^2)$, pero que existe una pequeña proporción $\in$ desconocida de observaciones atípicas. Este hecho puede modelarse mediante el siguiente enfoque: supongamos que estas observaciones anómalas provienen de una distribución también normal, de media cero y varianza $k\sigma^2$, donde $k > 1$. Entonces la función de densidad de las perturbaciones observadas es

$$f(u) = (1 - \in) f_N (u | 0, \sigma^2) + f_N (u | 0, k\sigma^2) \qquad [2.1]$$

donde $f_N (\cdot | \mu, \sigma^2)$ representa la función de densidad normal de media μ y varianza σ^2. Es inmediato que entonces

$$\text{Var } (u) = \sigma^2 (1 + \in (k - 1) \qquad [2.2]$$

y f será simétrica con kurtosis

$$\gamma = \frac{E[u^4]}{\{E[u^2]\}^2} - 3 = 3\left[\frac{1 + \in (k^2 - 1)}{(1 + \in (k - 1))^2} - 1\right] = 3(\delta - 1) \qquad [2.2]$$

donde $\delta < 1$. Por tanto, la distribución de u será leptokúrtica. Por ejemplo, para $\in = 0{,}1$ y $k = 9$, se obtiene $g = 5.33$, y las colas de la distribución son más abiertas que las de la distribución de Laplace (que tiene $\gamma = 3$). Este modelo para las anomalías es el considerado por Tukey (1960), Box y Tiao (1968, 1973) y Guttman (1973), entre otros.

Por tanto, concluimos que con este modelo tendremos: a) una mayor varianza del error, y b) una distribución con colas más largas que la normal.

La varianza de la estimación de los parámetros β del modelo lineal es

$$\text{Var } (\underline{\beta}) = (\underline{\underline{X}}' \underline{\underline{X}})^{-1} \sigma^2$$

y si σ^2 viene dado por [2.2], por ejemplo, con k grande, los parámetros estimados serán muy pocos fiables, muy inestables de muestra en muestra y con una gran varianza.

Los enfoques prácticos para resolver este problema pueden resumirse en:

1. Acudir al teorema central del límite para justificar la hipótesis de normalidad y utilizar, por tanto, mínimos cuadrados. Una vez estimado el modelo, utilizar gráficos de residuos frente a valores estimados para detectar anomalías y realizar finalmente un test de normalidad de los residuos.

2. Desechar mínimos cuadrados y utilizar un procedimiento robusto de estimación eligiendo una función g que proporcione estimadores razonablemente eficientes en la hipótesis de normalidad, sin padecer la inestabilidad de los mínimos cuadrados ante anomalías.

3. Utilizar un enfoque bayesiano construyendo un modelo formal que incorpore las desviaciones «a priori» previsibles respecto al modelo lineal estándar mediante parámetros en un modelo ampliado.

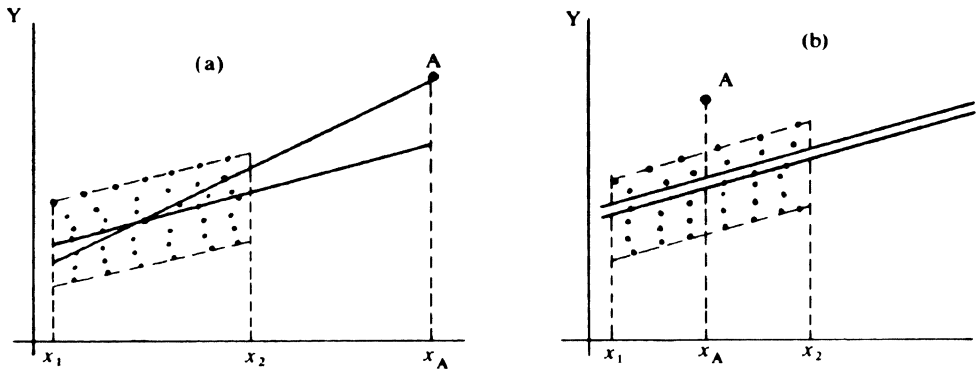
4. Robustificar no tanto el criterio de estimación como la metodología de construcción del modelo lineal. Esto requiere comprobar en cada etapa de construcción del modelo que la decisión adoptada no está determinada por un pequeño grupo de observaciones atípicas. Por tanto, no se abandona «a priori» mínimos cuadrados, como en 2, pero se complementa la estimación con un conjunto de contrastes diagnósticos que permitan: a) detectar las observaciones potencialmente influyentes; b) medir sus efectos sobre las coeficientes, y c) contrastar si son significativamente atípicas.

De estas cuatro alternativas, la 1 es la generalmente propuesta por la inmensa mayoría de los textos de Estadísticas y Econometría. Sus limitaciones son de dos tipos. La primera es que los gráficos bidimensionales clásicos son incapaces de revelar cierto tipo de observaciones claramente atípicas, que son además las que pueden producir una mayor distorsión en la estimación mínimo-cuadrática del modelo lineal. Si consideramos puntos muestrales (y_i, x'_1), el carácter atípico puede mostrarse por: 1) un valor anómalo de y_i para el x_i correspondiente. Este es el tipo de anomalía más fácil de detectar y corresponde al caso b) de la figura 1; 2) un valor atípico x'_i del vector de variables explicativas. Este segundo caso puede o no ir acompañado de una respuesta y_i atípica. La figura la representa a esta situación.

Se observa que el punto A es de hecho el que determina la pendiente de la recta y es destacadamente atípico, ya que sin él la recta obtenida sería la marcada por la línea de puntos. Sin embargo, el residuo en el punto A será muy pequeño, e incluso cero.

Figura 1

Tipos de anomalías



El enfoque clásico de análisis de residuos va encaminado principalmente a encontrar anómalos asociados, por tanto, a la situación 1b, lo que constituye una de sus limitaciones principales.

En efecto, los gráficos de residuos frente a la variable prevista o frente a las variables explicativas, aunque muy útiles para detectar errores de especificación (véase, por ejemplo, Draper y Smith, 1980); no pueden detectar aquellos valores atípicos multivariantes que se caracterizan por tener varias coordenadas x alejadas de los valores medios para estas variables. Como veremos en la sección 3, estos puntos son especialmente influyentes en la regresión y requieren un análisis particularmente cuidadoso.

El enfoque 3 ha sido utilizado por Box y Tiao (1968, 1973); Abraham y Box (1978); Chen y Box (1978 a, b y c) y Box (1979, 1980). Desde nuestro punto de vista, éste es el enfoque más general y completo para el tratamiento del problema, aunque presenta mayores inconvenientes computacionales y requiere disponer de un software para su aplicación eficiente.

En las secciones siguientes comentamos las soluciones 2 y 4.

3. ESTIMADORES ROBUSTOS DE REGRESION

Los inconvenientes apuntados de los mínimos cuadrados han conducido en los últimos veinte años a una extensa investigación de métodos que superen estas dificultades. Los libros de Mosteller y Tukey (1977), Huber (1981), Barnett y Lewis (1978) introducen el problema y contienen abundantes referencias.

La inestabilidad de los mínimos cuadrados es debida a la forma particular de las funciones g y ψ de [1.3] y [1.4]. La función g es en este caso cuadrática, por lo que aquellas observaciones con residuo más grande, en valor absoluto, entran en la suma a minimizar al cuadrado, lo que «arrastrará» la ecuación mínimo-cuadrática hacia esas observaciones, efecto obviamente indeseable. Por considerarse intuitivas, parece claro que una función g que crezca más lentamente que $g(u) = u^2$ cuando u sea grande, dará un peso menor a las observaciones atípicas y, por tanto, conducirá a estimadores más robustos.

Huber (1964) planteó este problema formalmente. Dada una familia de normales contaminadas según [2.1] puede investigarse la forma de g de manera que la máxima varianza posible de la estimación máximo-verosímil de β sea mínima. El resultado de este ejercicio consiste en obtener la siguiente función a minimizar

$$g(u) = \begin{cases} u^2/2, & |u| \leq a \\ a|u| - a^2/2, & |u| > a \end{cases} \qquad \psi(u) = \begin{cases} -a, & u < -a \\ u, & |u| < a \\ a, & u > a \end{cases} \qquad [3.1]$$

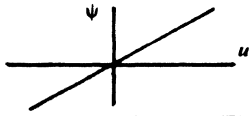
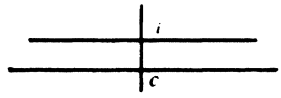
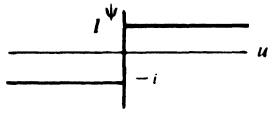
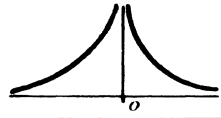
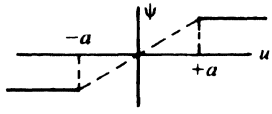
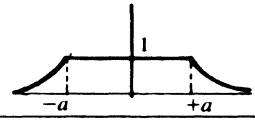
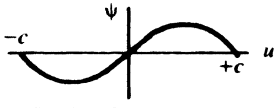
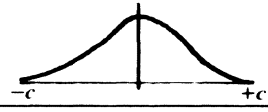
donde el valor de a se toma en la práctica entre 1 y 2. Observemos que si a es grande, este método será similar a mínimos cuadrados, mientras que si $a \to 0$ se obtiene la maximización de las desviaciones absolutas.

Para comparar esta función con criterios conocidos, la tabla 1 presenta la forma de la función ψ y los pesos w_i del sistema de mínimos cuadrados ponderados equivalente [1.5]. Mínimos cuadrados corresponde a $\psi(u) = u$ y, por tanto, $w_i = \psi(u) \cdot u^{-1} = 1$, que da peso idéntico a todas las observaciones. La minimización de las desviaciones absolutas da un peso muy grande a los residuos muy pequeños y un peso que decrece geométricamente a los más grandes. Es, por tanto, un criterio muy drástico. La función de Huber introduce ponderaciones que decrecen geométricamente a partir de a. Por último, hemos incluido la función bicuadrada propuesta por Beaton y Tukey (1974) (véase Mosteller y Tukey, 1977, para aplicaciones)

$$w_i(x) = \begin{cases} \left[1 - \left(\frac{x}{c}\right)^2\right]^2 & |x| \leq c \\ 0 & |x| > c \end{cases} \qquad \psi(u) = \begin{cases} x\left[1 - \left(\frac{x}{c}\right)^2\right]^2 & |x| \leq c \\ 0 & |x| > c \end{cases} \qquad [3.2]$$

Las tres últimas funciones se consideran robustas, ya que ponderan menos a las observaciones más alejadas. Otras funciones análogas han sido propuestas por Hampel (1974) y Andrews (1972). (Véase Hogg, (1979) para una revisión de estas funciones.)

TABLA 1

Distribución	$\psi(u)$	$w_i(u)$
Normal		
Laplace		
Huber		
Beaton y Tukey (1974)		

Para conseguir que la estimación sea independiente de la escala de medida, las ecuaciones [1.3] y [1.5] se modifican

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \underline{x}_i'\beta}{s}\right)\underline{x}_i' = 0 \qquad [3.3]$$

$$\sum\left(\frac{y_i - \underline{x}'_i\,\beta}{s}\right)\underline{x}'_i\, w_i = 0 \qquad [3.4]$$

con pesos $w_i = \psi\left(\frac{e_i}{s}\right)\cdot\left(\frac{e_i}{s}\right)^{-1}$, donde s es un estimador robusto de la dispersión. La obtención de los estimadores se realiza iterando sucesivamente con [3.4] mediante un programa de mínimos cuadrados ponderados. Huber (1981) ha demostrado la convergencia del método en condiciones generales, así como la consistencia de los estimadores obtenidos (ver también Yohai y Maronna (1979).

El procedimiento expuesto es el más utilizado y corresponde a la utilización de M-estimadores, llamados así por conducir a una estimación máximo-verosímil modificada. Existen otras dos grandes familias de métodos robustos, los L-estimadores basados en estadísticos ordinales y los R-estimadores a partir de los rangos de las observaciones, pero su aplicación a problemas de regresión es mucho menor. El lector interesado puede acudir a Hogg (1979).

Los procedimientos robustos que hemos comentado están sujetos a tres tipos de críticas. La primera es el carácter heurístico de las funciones g o ψ que conducen a una cierta arbitrariedad en la formulación. La segunda es que las propiedades muestrales de los estimadores para muestras pequeñas son desconocidas, con lo que se pierde la flexibilidad de los contrastes en el modelo lineal. La tercera es que los métodos así formulados, aunque útiles para tratar observaciones atípicas del tipo de la figura 1b, no resuelven el problema de los valores atípicos con residuos pequeños (figura 1a) que, en la práctica, son las más perjudiciales en muchos casos.

Respecto a la primera crítica. Chen y Box (1979 a) han mostrado que las funciones g o ψ propuestas en la literatura para estos problemas son óptimas para un tipo concreto esperado de contaminación. Por ejemplo, la función g de Huber [3.1] es óptima para una distribución normal en el centro, pero con colas de Laplace, que puede aproximarse bien por el modelo normal contaminado [2.1]. Puede, por tanto, argumentarse que la metodología a utilizar debe depender del tipo de estructura previsible o presente en la muestra concreta. La tercera crítica conduce a los M-estimadores generalizando en los que los pesos w_i de [3.4] se hacen depender no sólo del residuo, sino también de su capacidad de influencia, medida por su distancia al centro de la nube de puntos. Aunque este enfoque resuelve parcialmente el problema, acrecienta el carácter *ad hoc* del método y hace más problemático la consideración de las propiedades muestrales de las estimaciones.

4. ROBUSTIFICACION DE LA METODOLOGIA

La razón principal de acudir a métodos robustos de estimación es garantizar que los resultados obtenidos no van a depender fundamentalmente de unas pocas observaciones anómalas. Así, una de las ideas centrales en la teoría de la estimación robusta clásica es evaluar la sensibilidad de los estimadores. La curva de sensibilidad, CS, de un estimador T, se define, para un tamaño muestral fijo, n, por

$$CS_n(x) = n[T_{n+1}(x_1, \ldots, x_n, x) - T_n(x_1, \ldots, x_n)]$$

y es una función de los valores de una observación adicional x. Se calcula a partir de la diferencia entre los valores que toma el estimador T en cuestión cuando se tiene o no en cuenta el dato x. El límite de esta expresión cuando $n \to \infty$ es la curva de influencia introducida por Hampel (1974) para comparar estimadores. Lambert (1981) ha utilizado estas nociones para comparar test robustos.

Sin embargo, el hecho de que un estimador *pueda* ser muy sensible a una observación anómala no indica en absoluto que tenga un comportamiento poco eficiente ante una muestra dada. Antes de abandonar un procedimiento de estimación que puede ser

óptimo, parece razonable preguntarse si ante una muestra concreta es susceptible o no de serlo.

Por tanto, dado un conjunto de observaciones susceptibles de modelarse mediante un modelo lineal es pertinente preguntarse: 1) ¿Existen en esta muestra observaciones que, «a priori», tienen una capacidad de influencia mucho mayor que las demás en la construcción del modelo? 2) ¿Cómo podríamos medir la influencia concreta que, «a posterior», ha tenido un dato o conjunto de datos en la estimación de los parámetros? 3) ¿Cómo construir un test para determinar si una observación es atípica?

La primera pregunta puede responderse desde dos puntos de vista complementarios. El primero es mediante la matriz «sombrero», cuyas propiedades han sido discutidas por Huber (1975), Hoaglin y Welsh (1978), Cook (1977, 1979), Belsley, Kuh y Welsh (1980) y Weisberg (1980). El segundo, mediante el cociente de dos determinantes, fue introducido por Andrews y Pregibon (1978) y ha sido discutido por Draper y John (1981), y Belsley y otros (1980).

La matriz «sombrero» es la matriz $\underline{\underline{V}}$, que proyecta el vector Y sobre el espacio vectorial generado por las columnas de $\underline{\underline{X}}$

$$\underline{\hat{Y}} = \underline{\underline{V}} \cdot \underline{Y} \qquad [4.1]$$

$$\underline{\underline{V}} = \underline{\underline{X}} (\underline{\underline{X}}'\underline{\underline{X}})^{-1} \underline{\underline{X}}' \qquad [4.2]$$

Por tanto $(\underline{\underline{I}} - \underline{\underline{V}})$ proyecta los vectores del espacio n-dimensional sobre la variedad lineal ortogonal a la generada por $\underline{\underline{X}}$. La matriz $\underline{\underline{V}}$ es simétrica e idempotente con elementos

$$v_{ij} = \underline{x}'_i(\underline{\underline{X}}'\underline{\underline{X}})^{-1}\underline{x}_j \qquad [4.3]$$

donde $\underline{x}_i$ y $\underline{x}_j$ son los vectores de observación de las variables $\underline{x}$ en los individuos i, j, respectivamente. La importancia de esta matriz en el modelo lineral es que

$$\underline{e} = (\underline{\underline{I}} - \underline{\underline{V}}) \cdot \underline{U} = (\underline{\underline{I}} - \underline{\underline{V}}) \underline{Y} \qquad [4.4]$$

de donde se deduce en particular que

$$\text{Var } (e_i) = \sigma^2(1 - v_{ii}) \qquad [4.5]$$

donde v_{ii} tiene una interpretación inmediata. Si las $\underline{x}$ están expresadas en desviaciones a su media

$$v_{ii} = (\underline{x}_i - \underline{\bar{x}})'(\underline{\underline{\tilde{X}}}'\underline{\underline{\tilde{X}}})^{-1}(\underline{x}_i - \underline{\bar{x}}) \qquad [4.6]$$

siendo ahora $(\tilde{X}'\tilde{X})n^{-1}$ la matriz de varianzas y covarianzas entre las observaciones x. Por tanto, v_{ii} representa la distancia de Mahalanobis de una observación $\underline{x}_i$ respecto al centro de la nube de puntos $\bar{\underline{x}}$. Como los v_{ii} son términos de una matriz proyección, $0 \leq v_{ii} \leq 1$, y como la traza de una matriz idempotente es igual a su rango, $\Sigma v_{ii} = k$, siendo k el rango de $\underline{\underline{X}}$. Por tanto, el valor medio de los términos v_{ii} será k/n.

Si uno de los puntos $\underline{x}_i$ está muy alejado del centro de la nube, $\bar{\underline{x}}$, su v_{ii} será grande y la varianza del residuo correspondiente a ese punto será muy pequeña, según [4.5]. En el límite, si $v_{ii} = 1$, la varianza del residuo será cero, lo que indica que la posición de ese punto respecto al centro de los datos fuerza a la ecuación de regresión a pasar por él, sea cual sea el valor observado de y. Por tanto, aquellos puntos muestrales que tengan v_{ii} altos son, potencialmente, influyentes. Es posible calcular la distribución muestral de ciertas funciones de los v_{ii} (Belsley y otros, 1980), lo que conduce a considerar una observación potencialmente influyente si $v_{ii} > 2k/n$.

Otra interesante interpretación de los v_{ii} es debida a Huber (1981). Según [4.1]

$$\text{Var}\ (\hat{y}_i) = \sum_{j=1}^{n} v_{ij}^2\ \text{Var}\ (y_j) = \sigma^2 v_{ii}$$

dado que al ser indempotente $\underline{\underline{V}}$, $v_{ii} = \sum_{j=1}^{n} v_{ij}^2$. Por tanto, si recordamos que la media muestral de h observaciones, cada una de ellas con varianza σ^2 e independientes, es σ^2/h, es claro que v_{ii}^{-1}puede interpretarse como el número equivalente de observaciones utilizadas para calcular la estimación $\hat{y}_i$. Si $v_{ii} = 1$, la estimación $\hat{y}_i$ está, pues, calculada con una observación, por lo que lógicamente su residuo es cero (compárese con [4.5]).

El segundo enfoque para determinar observaciones influyentes, Andrews y Pregibon (1978), utiliza el cambio de «volumen» que experimenta la nube de puntos al eliminar un conjunto de observaciones. Si tomamos $|\underline{\underline{X}}'\underline{\underline{X}}|$, (determinante de $\underline{\underline{X}}'\underline{\underline{X}}$) como medida del volumen inicial y $|\underline{\underline{X}}'_{(i)}\underline{\underline{X}}_{(i)}|$ el resultante de eliminar la observación i, el cociente

$$\Delta = \frac{|\underline{\underline{X}}'_{(i)}\ \underline{\underline{X}}_{(i)}|}{|\underline{\underline{X}}'\underline{\underline{X}}|} \qquad [4.7]$$

recuerda el estadístico Δ de Wilks y puede usarse como una medida de la influencia del punto i. Una ventaja de este procedimiento es que se extiende obviamente a la influencia simultánea de calcular conjunto de puntos $i, j, k...$ Para un único punto, puede demostrarse (Draper y John, 1981) que $\Delta = 1 - v_{ii}$, con lo que ambos enfoques son coincidentes.

El segundo aspecto es determinar, para una muestra concreta, la influencia de cada observación sobre el modelo obtenido. Para ello existen varios enfoques basados en la función de influencia empírica (Cook y Weisberg, 1980)

$$IE_A = \hat{\underline{\beta}}_A - \hat{\underline{\beta}} \qquad [4.8]$$

donde $\hat{\underline{\beta}}_A$ es el estimador obtenido eliminando las observaciones A y $\underline{\beta}$ el correspondiente a la muestra completa. Una forma simple de obtener una medida escalar de la influencia de A es midiendo la distancia, en una métrica con sentido estadístico como la de Mahalonobis, entre $\hat{\beta}_A$ y β. Esta medida fue introducida por Cook (1977)

$$D_A = \frac{1}{k}(\hat{\underline{\beta}}_A - \underline{\beta})' \frac{(\underline{\underline{X}}'\underline{\underline{X}})}{s^2}(\hat{\underline{\beta}}_A - \underline{\beta}) \qquad [4.9]$$

donde $s^2 = \dfrac{1}{n-k}\underline{Y}'(\underline{\underline{I}} - \underline{\underline{V}})\underline{Y}$ es la varianza residual de la regresión y $(\underline{\underline{X}}'\underline{\underline{X}})^{-1}s^2$ la estimación de la matriz de varianzas, convarianzas de β. Welsh y Kuh (1977) y Belsley y otros (1980) definen medidas de distancia similares, pero tomando como matriz definidora de la métrica $(\underline{\underline{X}}'\underline{\underline{X}})s^2_{(A)}$, donde $s^2_{(A)}$ corresponde a la regresión sin las observaciones A.

Utilizando el subíndice (i) para indicar que la característica afectada ha sido calculada sin la observación i, la distancia de Cook se obtiene fácilmente de

$$D_i = \frac{(\hat{\underline{\beta}}_{(i)} - \underline{\beta})(\underline{\underline{X}}'\underline{\underline{X}})(\hat{\underline{\beta}}_{(i)} - \underline{\beta})}{ks^2} = \frac{1}{k}\frac{e_i^2}{s^2}\frac{v_{ii}}{(1-v_{ii})^2} \qquad [4.10]$$

donde $e_i = y_i - \underline{x}'_i\hat{\underline{\beta}}$. Es interesante que D_i puede también escribirse

$$D_i = \frac{(\hat{\underline{Y}}_{(i)} - \hat{\underline{Y}})'(\hat{\underline{Y}}_{(i)} - \hat{\underline{Y}})}{ks^2} \qquad [4.11]$$

que indica que D_i mide la distancia euclidea (hecha adimensional por ks^2) en que se traslada el vector de predicción $\hat{Y}$ al eliminar la observación i de la regresión.

De una manera análoga puede estudiarse el efecto de cada observación sobre los coeficientes individuales β_j.

La construcción de tests para determinar observaciones atípicas en modelos de regresión ha sido objeto de numerosos enfoques. (Véase Barnett y Lewis, 1978, para una panorámica del tema.) Planteado como un contraste de la razón de verosimilitudes, el test resultante es función monótona de los residuos estudentizados, definidos por

$$r_i = \frac{e_i}{s\sqrt{1-v_{ii}}} \qquad [4.12]$$

donde cada residuo mínimocuádrico ha sido dividido en su varianza (véase [4.5]). Un inconveniente de esta construcción es que la distribución de r_i en la hipótesis de normalidad de u no es la t de Student, ya que el numerador y el denominador no son independientes. Sin embargo, el estadístico

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - v_{ii}}} \qquad [4.13]$$

tiene, en la hipótesis de normalidad, una distribución t de Student con $n - k - 1$ grados de libertad. Computacionalmente, es conveniente expresarlo (Weisberg, 1980)

$$t_i = r_i \sqrt{\frac{n - k - 1}{n - k - r_i^2}} \qquad [4.14]$$

donde r_i está definida por [4.12].

Un problema clásico para realizar el contraste mediante [4.14] es determinar el nivel de significación adecuado, ya que la distribución relevante para obtenerlo es la del máximo valor de un estadístico t con $n - k - 1$, en una muestra de tamaño n. Esta distribución exacta se desconoce, pero se han tabulado valores críticos aproximados utilizando la desigualdad de Bonferroni. (Véase Miller, 1977; Cook y Prescott, 1981.)

En resumen, los estadísticos v_{ii}, D_i y t_i constituyen la base de la robustificación metodológica del modelo lineal. Los términos v_{ii} dependen únicamente de las variables predeterminadas y miden la influencia potencial de cada observación por su posición dentro del espacio de las variables. Tendremos un diseño robusto si todos los puntos tienen coeficientes v_{ii} análogos. Esta misma idea ha sido expresada de formas distintas aunque con el mismo sentido final por Anscombe y Tukey (1963), Huber (1973), Box y Draper (1975) y Belsley y otros (1980).

El estadístico D_i de Cook recoge la influencia *práctica* de cada observación sobre los parámetros estimados o el vector de predicción $\hat{\underline{Y}}$, y mide, en consecuencia, la sensibilidad de la estimación a la muestra concreta utilizada. El interés de D_i es que puede indicar la poca relevancia práctica de preocuparse por observaciones muestrales que, aunque quizá anómalas, no influyen en el modelo. Un alto valor de D_i resalta la importancia de investigar en profundidad aquellos datos atípicos, según el estadístico t que tengan una apreciable influencia en los resultados del modelo.

Una de las ventajas del estudio de la influencia empírica de las observaciones que hemos descrito en los párrafos precedentes, es que puede desvelar no linealidades que de otro modo quedarían ocultas. La anomalía A de la figura 1a sobresaldrá en el análisis por un alto coeficiente v_{ii}, asociado a un residuo muy pequeño. Una vez aceptado este punto, caben dos interpretaciones: 1) se trata de un error de datos y lo óptimo es

desecharlo y repetir el ajuste sin ese punto; 2) el punto A no es erróneo y lo que revela es una relación *no lineal* sobre un rango más amplio de variables. Esta posibilidad debe tenerse muy en cuenta ante observaciones con v_{ii} y D_i altos.

5. UN MODELO DE LOS DETERMINANTES DEL ALQUILER

5.1. El problema y los datos

En nuestro país, el Estado interviene en el sector de la vivienda en arrendamiento a través de dos políticas distintas. En primer lugar, varios organismos públicos promueven —directa o indirectamente— la construcción de viviendas que denominaremos de protección oficial, cuyos alquileres vienen regulados con mayor o menor eficacia de acuerdo con complejos procedimientos que no es necesario revisar aquí. El objetivo central de esta forma de intervención es proporcionar viviendas a un alquiler inferior al que se daría en el mercado libre. En segundo lugar, desde 1920 a 1956 el Estado intervino en el sector privado de las viviendas en alquiler, estableciendo la prórroga forzosa de los contratos y congelando el precio del arrendamiento. La Ley de Bases de 1955 inicia la liberalización de los alquileres, que se completa en la Ley de Arrendamientos Urbanos de 1964. Así pues, de resultas de la intervención del Estado, es útil distinguir entre tres tipos de viviendas en arrendamiento: las viviendas de protección oficial y las viviendas ocupadas antes o después de la promulgación de la LAUR de 1964.

El problema que vamos a estudiar a continuación es cómo construir un modelo que permita explicar el alquiler de mercado de las viviendas post-64 en términos de sus características físicas. Esta tarea se enmarca dentro de un estudio, para analizar tanto el alcance de las ventajas económicas como las consecuencias distributivas de las políticas de protección oficial y de control de alquileres. Los beneficios que perciben los ocupantes de los dos tipos de viviendas, al amparo de la legislación vigente, se estiman mediante la diferencia entre el alquiler de mercado que les correspondería en el año de la muestra y el alquiler efectivamente pagado en esa fecha.

Para una discusión de los fundamentos microeconómicos del intento de explicar las diferencias observadas en el alquiler de las viviendas liberalizadas en términos de las cantidades que entrañan de determinadas características, puede consultarse el trabajo de Ruiz-Castillo (1982a), que incluye también una justificación de la noción de beneficio mencionada. El lector interesado en los resultados empíricos del cálculo de los beneficios y el análisis de su distribución desde el punto de vista de la equidad, puede consultar los trabajos de Peña y Ruiz-Castillo (1982a, 1982b).

TABLA 2

CARACTERISTICAS ESTRUCTURALES

Denominación	Descripción	Viviendas post-64	
		Media	Desviación típica
	VARIABLES CUANTITATIVAS		
1. EDAD	Edad del edificio en años desde su construcción	21,9	23,0
2. OCUP	Años que lleva ocupada la vivienda	3,6	2,3
3. M2	Superficie habitable en m^2	68,0	42,8
4. NHAB (*)	Número de habitaciones	3,6	1,2
5. NPL	Número de plantas del edificio	5,1	2,8
6. ECON	Estado de conservación del edificio	4,6	17,5
	VARIABLES CUALITATIVAS		
	Tipo de edificio	Porcentaje en %	
7. MAGL	Vivienda «marginal»	3	
8. CHHI	Chalet o hilera unifamiliar	4	
9. BLOQ	Bloque de pisos	38	
—	Medianería	55	
	Modalidad de promoción		
10. PPRI (*)	Sociedad promotora privada	27	
—	Cooperativa de usuarios, comunidad de propietarios, particular(es), autoconstrucción, otros	45	
11. NS (*)	No sabe	28	
	Servicios higiénicos		
12. POCOS	Sin servicios higiénicos, con solo WC, sólo lavabo, sólo WC y lavabo, o sólo ducha o baño y WC	19	
—	Con cuarto de aseo (lavabo, ducha y WC) o con un cuarto de baño completo	70	
13. DOMAS	Con dos, tres, cuatro o más piezas de aseo o baño	11	
14. TELF	Con teléfono	36	
15. CALC	Con calefacción central	20	
16. GAR	Con garage	7	
17. PORT (*)	Con portería	34	
18. MUEB	Con muebles	15	
19. ESTR (*)	Vivienda a estrenar en el momento de su ocupación	18	
20. GCOM (*)	Con gastos comunes incluidos en otro concepto	19	
21. GCAL (*)	Con el recibo de la calefacción incluido en otro concepto	38	
22. GAGC (*)	Con los pagos por agua caliente incluidos en otro concepto	47	
23. EDS19	Vivienda construida en el siglo XIX	7	
24. ED4164	Vivienda construida entre 1941 y 1964	21	
25. ED6574	Vivienda construida entre 1965 y 1974	53	

(*) Variables que no resultaron significativas inicialmente (véase el texto).

TABLA 3

VARIABLES CUANTITATIVAS REPRESENTATIVAS DE LOS ATRIBUTOS DE LAS ZONAS DE ANALISIS

Denominación	Descripción	Viviendas post-64	
		Media	Desviación típica
1. ACC	Indice de accesibilidad en minutos de tiempo de transporte	41,2	17,6
2. DENPO	Densidad de población en habitantes por km^2	19.950	20.683
3. RENT(*)	Renta familiar media en pesetas mensuales	18.826	4.734
4. ALTA	Indice del nivel socioeconómico	0,06	0,88
5. ANTIG	Indice de la antigüedad media de los edificios	0,13	1,14
6. MARCH(*)	Indice de marginalidad y chabolismo	−0,07	0,78
7. ESCOL(*)	Puestos de preescolar y EGB por 1.000 habitantes	7.351	4.345

(*) Variables que no resultaron significativas inicialmente (véase el texto).

Los datos disponibles provienen de la encuesta que la empresa CETA realizó para COPLACO en 1974 sobre las necesidades de vivienda en el Area Metropolitana de Madrid. Esos datos fueron completados con cierta información que facilitó COPLACO. La tabla 2 describe las variables estructurales que fue posible medir, mientras que la tabla 3 recoge las variables representativas de las características relacionadas con la localización de las viviendas dentro del Area Metropolitana de Madrid. Los detalles de la construcción de determinadas variables se resumen en el apéndice. Contamos con 460 observaciones de viviendas arrendadas con posterioridad a 1964, libres, por tanto, del control de alquileres.

5.2. La selección de la forma funcional.

Para decidir la forma funcional adecuada se siguió un proceso iterativo que comenzó con un análisis exploratorio de los datos para obtener una primera representación razonable e identificar posibles observaciones anómalas. A continuación se estimó por máxima verosimilitud la mejor transformación de la variable dependiente y se realizaron diversas pruebas para encontrar la métrica adecuada para las variables dependientes.

En el análisis exploratorio inicial se utilizaron tres tipos de herramientas. La primera fueron gráficas bivariantes de la variable respuesta respecto a cada una de las variables explicativas. La segunda fue la distribución empírica de cada variable. La tercera, el análisis de los residuos de regresiones preliminares que incluían diversos conjuntos de variables explicativas. Se estudió la distribución de los residuos y los gráficos de éstos respecto a cada una de las variables explicativas y el alquiler estimado. Los resultados de este análisis inicial fueron:

a) La variable alquiler requiere transformación, probablemente mediante el logaritmo. Esta conclusión es clara y se basa en que los gráficos $e_i = f(y)$ para la variable y sin transformar mostraban curvatura y heterocedasticidad. Además la distribución del

alquiler —así como la de los residuos e_i de las regresiones— tiene fuerte asimetría positiva. Finalmente, el logaritmo tiene una clara interpretación económica en este caso, indicando que el efecto de cada factor depende del nivel que alcancen los demás.

b) Las variables OCUP, M2, ECON, ACC6 y DENPO se expresaron en logaritmos. Esta decisión se tomó para obtener linealidad en la respuesta una vez expresado el alquiler en logaritmos. Como la decisión no fue clara respecto a OCUP y NPL, se optó por mantenerla provisionalmente a expensas de una revisión posterior.

c) Las variables señaladas con un arterisco se desecharon en una primera etapa por no aportar información adicional.

d) La edad del edificio mostró una influencia compleja y marcadamente no lineal, probablemente porque esta variable recoge efectos muy distintos y es «proxy» de otras variables. Además, como se indica en el apéndice, su construcción no estuvo exenta de dificultades. Para identificar posibles efectos no lineales, se discretizó su recorrido mediante variables cualitativas, con el resultado de que las viviendas construidas en el siglo XIX y las muy modernas (posteriores a 1965) mostraron alquileres significativamente más altos, mientras que las viviendas del período 1900-1940 resultaron ser las más baratas. En una primera aproximación intentamos una representación simple de este efecto mediante un polinomio de segundo grado. Para prevenir la previsible multicolinealidad, se definieron las variables

$$\text{EDM} = \text{EDAD} - \bar{\text{E}}, \qquad \text{EDM2} = (\text{EDAD} - \bar{\text{E}})^2$$

donde E es la media de la edad de las viviendas post-64.

Con estas decisiones, el modelo estimado resultante se presenta en la primera columna de la tabla 5. La distribución de los residuos es asimétrica con coeficiente de asimetría −1,95 y 7,5 de kurtosis. El test de Kolmogorov Smirnov condujo a rechazar la hipótesis de normalidad de los residuos con $\alpha = 0{,}01$. La distribución parece la de una normal contaminada por un pequeño número de valores negativos ya que, tomando la mediana como centro (que es 0,07), la distribución es simétrica y razonablemente normal en el rango $0{,}07 \pm 1{,}5\ \hat{\sigma}$, siendo $\hat{\sigma}$ la desviación típica de los residuos.

El análisis interno de robustez del modelo mostró 19 observaciones destacables que se incluyen en la tabla siguiente.

TABLA 4
VALORES ATIPICOS

Obs. n.º	1	2	3	4	5	6	7	8	9	10
v_{ii}	0,03	0,04	0,06	0,03	0,04	0,05	0,06	0,03	0,05	0,05
D_i	0,06	0,05	0,08	0,02	0,03	0,03	0,04	0,02	0,03	0,02
t	−6,6	−5,3	−5,4	−4,1	−3,7	−3,7	−3,7	−3,6	−3,3	−3,0

TABLA 4 (continuación)

Obs. n.º	11	12	13	14	15	16	17	18	19
v_{ii}	0.04	0.11	0.03	0.03	0.03	0.04	0.12	0.15	0.15
D_i	0.02	0.04	0.01	0.01	0.00	0.01	0.04	0.00	0.00
t_i	−3.0	−2.8	−2.8	−2.5	2.4	−2.4	2.4	−0.7	0.7

De ellas, nueve tienen estadísticos t mayores que 3,3 y otras ocho tienen valores entre 3 y 2,4. La tabla recoge también las observaciones potencialmente más influyentes por tener mayor v_{ii} [18 y 19], aunque esta influencia no se da de hecho en la muestra. Se revisaron cuidadosamente las 17 primeras observaciones con el resultado de que las nueve mayores parecían ser errores en la perforación de los datos (omisión de un cero en el alquiler). Sobre las siguientes ocho, existían dudas en algunas por lo que se decidió conservarlas y estimar una nueva regresión con 451 datos, cuyos resultados se presentan en la segunda columna de la tabla 5. Como puede verse, la eliminación de estas nueve observaciones mejora los resultados sin alterarlos sustancialmente. Los coeficientes de las variables no sufren prácticamente variación, con las excepciones siguientes:

— Los coeficientes de MAGL, BLOQ y LDENPO se hacen prácticamente cero, lo que sugiere fuertemente su eliminación del modelo.

— Aumenta el coeficiente de TELF, que pasa de no ser significativa a serlo, y el del índice de accesibilidad LACC, que deja la zona de dudas para convertirse en una variable significativa.

A la vista de esta información, estimamos un nuevo modelo sin las variables MAGL, BLOQ y LDENPO, con los resultados siguientes: los coeficientes de todas las variables incluidas son prácticamente idénticos a los anteriores, y la varianza residual disminuye ligeramente debido a que la suma de cuadrados es casi idéntica, pero tiene ahora más grados de libertad. Por otra parte, introducimos las variables que habíamos desechado anteriormente para el modelo de 460 observaciones, con el resultado de que GCOM, aunque no significativa (tiene un $t = 1{,}22$), parece prometedora, por lo que se incorporó provisionalmente al modelo. La columna (3) de la tabla 5 resume este modelo final con 541 observaciones.

A continuación, sometimos los residuos de este modelo a un análisis detallado encaminado a investigar la presencia de nuevos valores anómalos. La repetición del análisis de robustez que hicimos para la regresión inicial, condujo a confirmar el carácter atípico de las ocho observaciones que señalamos anteriormente. Reestimamos

el modelo eliminando estas ocho observaciones con los resultados que se presentan en la columna (4) de la tabla 5. Se observará que:

— las variables DOMAS, CALC y GCOM, que no eran formalmente significativas con $\alpha = 0{,}05$, pasan a serlo, sin lugar a dudas;

— el resto de los coeficientes no se modifican sustancialmente;

TABLA 5

RESULTADOS DEL ANALISIS DE REGRESION CON LA DESVIACION TIPICA DE LOS COEFICIENTES ENTRE PARENTESIS

Variables	(1)	(2)	(3)	(4)	(5)	Otras variables alternativas
CONSTANTE	5,77	7,14	7,57	7,81	8,13	
	(0,77)	(0,60)	(0,35)	(0,31)	(0,33)	
EDM	−0,012	−0,010	−0,012	−0,008	−0,09	LEDAD
	(0,002)	(0,002)	(0,002)	(0,002)	(0,03)	
EDM2	0,00022	0,0002	0,0002	0,0001	0,12	EDS19
	(0,00005)	(0,00004)	(0,00004)	(0,0003)	(0,08)	
LOCUP	−0,25	−0,25	0,35	−0,24	−0,08	OCUP
	(0,04)	(0,03)	(0,04)	(0,02)	(0,007)	
LM2	0,46	0,42	0,42	0,40	0,39	
	(0,07)	(0,05)	(0,05)	(0,05)	(0,05)	
LNPL	0,26	0,20	0,20	0,18	0,19	
	(0,06)	(0,05)	(0,04)	(0,04)	(0,04)	
LECON	−0,06	−0,06	−0,06	−0,07	−0,08	
	(0,03)	(0,02)	(0,02)	(0,02)	(0,02)	
MAGL	0,11	−0,0008	—	—	—	
	(0,15)	(0,11)	—	—	—	
CHHI	0,66	0,53	0,54	0,48	0,49	
	(0,15)	(0,12)	(0,11)	(0,10)	(0,10)	
BLOQ	−0,08	0,02	—	—	—	
	(0,06)	(0,05)	—	—	—	
POCOS	−0,19	−0,23	−0,23	−0,23	−0,25	
	(0,08)	(0,07)	(0,07)	(0,06)	(0,06)	
DOMAS	0,06	0,07	0,08	0,18	0,19	
	(0,09)	(0,07)	(0,07)	(0,06)	(0,06)	
TELF	0,05	0,14	0,14	0,14	0,15	
	(0,06)	(0,05)	(0,05)	(0,04)	(0,04)	
CALC	0,11	0,09	0,09	0,11	0,13	
	(0,07)	(0,06)	(0,06)	(0,05)	(0,05)	
GAR	0,28	0,29	0,30	0,26	0,25	
	(0,10)	(0,08)	(0,08)	(0,07)	(0,07)	
MUEB	0,33	0,29	0,29	0,24	0,26	
	(0,07)	(0,05)	(0,05)	(0,05)	(0,05)	
GCOM	—	—	(0,06)	0,11	0,12	
	—	—	(0,05)	(0,05)	(0,04)	
LACC	−0,15	−0,53	−0,38	−0,41	0,41	
	(0,13)	(0,10)	(0,08)	(0,07)	(0,07)	
LDENPO	0,04	0,0009	—	—	—	
	(0,02)	(0,017)	—	—	—	
ALTA	0,16	0,10	0,10	0,09	0,08	
	(0,04)	(0,03)	(0,03)	(0,03)	(0,03)	
ANTIG	−0,06	0,05	0,07	−0,06	−0,07	
	(0,03)	(0,02)	(0,03)	(0,02)	(0,02)	
R^2	0,71	0,79	0,79	0,82	0,82	
Error estándar	0,48	0,37	0,37	0,33	0,32	
Número de observaciones	460	451	451	443	4.443	

— se acepta con $\alpha = 0{,}10$ la hipótesis de normalidad de los residuos mediante un contraste de Kolmogorov Smirnov y mediante contrastes de su asimetría y kurtosis.

En resumen, comparando este modelo con el inicial, vemos que al eliminar las 17 observaciones que hemos considerado como errores (4 por 100 del total), la varianza de los residuos ha disminuido en un 55 por 100, la proporción de la variabilidad explicada ha aumentado en un 17 por 100 y podemos admitir razonablemente la hipótesis de normalidad en los residuos. Los coeficientes de la mayoría de las variables se han modificado muy ligeramente y, en los casos en que no es así, los cambios hacen el modelo más compatible con la información *a priori:* la distancia al centro medida por LACC, y el hecho de que la vivienda tenga dos o más cuartos de baño, teléfono, calefacción central y los gastos comunes incluidos en otro concepto, aparecen como variables significativas en el modelo presumiblemente limpio de valores atípicos.

Para contrastar esta especificación hemos estimado por máxima-verosimilitud el parámetro λ de la transformación Box-Cox de la variable alquiler (véase Weisberg (1980)). Esta estimación se ha hecho para los modelos con 460, 451 y 443 datos. Los resultados se presentan en la tabla 6.

TABLA 6

FUNCION DE VEROSIMILITUD CON LAS VARIABLES EXPLICATIVAS CONTINUAS EN LOGARITMOS

n \ λ	−0.2	−0.1	0.0	0.1	0.2	0.3	0.4	0.5
460	+4866	−4816	−4775	−4745	−4727	−4721	−4729	−4750
451	−4635	−4605	−4585	−4574	−4574	−4584	−4605	−4637
433	−4493	−4469	−4454	−4447	−4450	−4464	−4489	−4524

De este análisis obtenemos las conclusiones siguientes:

1. Al eliminar valores atípicos, el máximo de la función de verosimilitud para λ se aproxima gradualmente a cero. El máximo es 0,3 con todos los datos, 0,2 con 451 y 0,1 con 443. En este último caso un intervalo de confianza del 95 por 100 no incluye el cero (transformación logarítmica) como valor posible. Esto puede sugerir que todavía existen valores atípicos en el modelo. A pesar de esto, aceptaremos el logaritmo como transformación adecuada, ya que es razonable desde un punto de vista teórico y no se contradice radicalmente por la evidencia empírica.

2. Hemos comprobado que la estimación λ es practicamente insensible a distintas especificaciones de las variables explicativas

Respecto a las variables explicativas, ya hemos comentado que desde los primeros modelos exploratorios existían dudas sobre si expresar las variables OCUP y NPL con o sin logaritmos. Para decidir respecto a esta cuestión, hemos realizado un experimento factorial 2 × 2, probando las combinaciones posibles con y sin logaritmos de estas variables. La tabla siguiente muestra las sumas cuadráticas de los residuos para cada combinación de «factores»:

	OCUP	LOCUP
NPL	45.34	46.25
LNPL	44.22	45.14

Resulta indudable que el número de plantas debe ir en logaritmos, mientras que el año de ocupación produce mejores resultados si no la transformamos. La conclusión es razonable, ya que indica que es el tiempo de ocupación directamente quien influye proporcionalmente sobre el alquiler.

La otra variable cuya especificación no era satisfactoria era la edad del edificio. Se intentaron distintas especificaciones no lineales de acuerdo con el procedimiento sugerido por Box y Tidwell (1962); desgraciadamente, el algoritmo de cálculo correspondiente no resultó convergente. Finalmente, optamos por seguir el criterio siguiente: 1) entre las transformaciones plausibles, elegir aquella que generase la menor suma cuadrática de errores; 2) estudiar a continuación la posibilidad de completar esa especificación con una o varias de las variables cualitativas EDS19, ED4164 ó ED6574. Este camino condujo a elegir la transformación logarítmica corregida por la variable EDS19. Como el coeficiente de LEDAD resultó negativo y el de EDS19 positivo, esta formulación es consistente con la información que teníamos sobre el perfil de la relación: *ceteris paribus*, cuanto mayor es la edad del edificio menor es el alquiler de la vivienda, excepto para las edificaciones del siglo XIX, cuya solidez (u otras características inobservadas) exigen una corrección de alza.

Para prevenir posibles interaciones entre la especificación de la variable EDAD y las variables OCUP y NPL, repetimos el diseño factorial 2 × 2 en presencia de LEDAD y EDS19, confirmando los resultados anteriores.

5.3. La selección del modelo final

El objetivo final del modelo es prever los alquileres de mercado de las viviendas que lo tienen controlado. Por tanto, un criterio relevante para seleccionar el número de

regresores a incluir es el error cuadrático medio de predicción. Una estimación de este error, que sirve para comparar distintos modelos, es el estadístico C_p de Mallows, calculado por:

$$C_p = \frac{SCR_p}{\hat{\sigma}^2} + 2p - n$$

donde SCR_p es la suma cuadrática de los residuos del modelo con p, regresores, $\hat{\sigma}^2$ es una estimación insesgada de la varianza de los residuos en el modelo con el máximo número de variables y n es el número de datos. Si se dispone de K variables explicativas posibles, el estadístico C_p permite seleccionar qué conjunto p de variables maximiza la capacidad predictiva del modelo (minimiza el error cuadrático medio).

Hemos utilizado este procedimiento para seleccionar el conjunto de variables con mayor capacidad predictiva. También hemos contrastado si, con este criterio es adecuado añadir variables booleanas para representar las zonas geográficas, junto con las variables continuas referidas a la localización de las viviendas. El resultado fue negativo. Finalmente, se ha contrastado qué lista de variables geográficas es preferible, las continuas o las cualitativas, resolviéndose la cuestión en favor de las primeras. El mejor modelo obtenido se presenta en la última columna de la tabla 5 y tiene un C_p de 12,6 con 17 variables explicativas.

Una vez seleccionado el modelo, hemos realizado el estudio interno de cada una de las observaciones buscando la presencia de valores anómalos, coeficientes muy dependientes de algún dato y otras fuentes de error de especificación. Los resultados de este estudio de la metodología que comentamos en las secciones anteriores, pueden resumirse como sigue:

a) El máximo valor del estadístico t para los residuos studentizados es 3,5. Hay también dos observaciones con este estadístico igual a 3,1. El resto no presenta problemas. Estas tres observaciones están situadas, sin embargo, cerca del centro de gravedad de los valores de las variables explicativas, por lo que el estadístico D_i de Cook, que mide el efecto de cada observación sobre los parámetros estimados, es bajo. En todo caso, no existe ninguna observación con D_i alto. Así pues, concluimos que el modelo obtenido es robusto ante anomalías.

b) En los gráficos de residuos no aparece ninguna evidencia de que existan errores en la especificación de las variables. Su distribución es normal de acuerdo con un contraste Kolmogorov Smirnov con $\alpha = 0{,}05$.

c) La situación de estimación es adecuada sin que existan problemas de multicolinealidad, ya que el índice de condicionamiento de la matriz $\underline{\underline{X'X}}$ es sólo 8,8 (véase Belsley, Kuh y Welsh, 1981).

5.4. INTERPRETACIÓN ECONÓMICA

Terminado el análisis estadístico del modelo final, sólo resta referirse a la interpretación económica de los resultados de su estimación.

En primer lugar, el análisis realizado permite concluir que el 82 por 100 de las diferencias entre los alquileres de las viviendas post-64 del Area Metropolitana de Madrid puede explicarse por las 17 características que resultaron empíricamente relevantes. Como se indica en el apéndice, mientras que la información sobre características estructurales era bastante completa, la información sobre las características relacionadas con la localización geográfica de las viviendas era muy deficiente. Así, no es de extrañar que estas últimas —el índice de accesibilidad ACC, el índice socioeconómico ALTA y el índice de antigüedad de las edificaciones ANTIG— sólo expliquen el 4 por 100 de la variabilidad observada en los alquileres, frente al 78 por 100 explicado por las 14 características estructurales. (De ese 4 por 100, el 75 por 100 es atribuible a la variable ACC). De haber contado con datos sobre el nivel de los bienes públicos locales, la contaminación de distintos tipos o la distribución de usos del suelo no residenciales, es de esperar que la importancia relativa de las variables geográficas hubiera sido mayor.

En segundo lugar, hay que destacar que todas las variables aparecen con el signo esperado. En cuando a la interpretación de los coeficientes, cabe apuntar los comentarios siguientes:

a) Los coeficientes de las variables explicativas que figuran en logaritmos —EDAD, M2, ECON y ACC— miden directamente la elasticidad. Así, por ejemplo, un aumento del 10 por 100 de la superficie habitable en m^2 conduce a un aumento de casi el 4 por 100 del alquiler —lo que indica que existen rendimientos a escala decreciente respecto a esta variable. La elasticidad de −0,4 del índice de accesibilidad puede que sea algo baja: dos viviendas idénticas en cuanto a sus características estructurales y a las características de los barrios en que estén situadas, pero que difieran en un 50 por 100 del tiempo de transporte a los puntos céntricos de empleo, tendrán una diferencia en el alquiler de un 20 por 100. La elasticidad de 0,19 del número de plantas no tiene una interpretación inmediata; tal vez los edificios más altos sean en promedio más deseables por poseer alguna característica adicional que no quedó recogida en nuestra encuesta. Finalmente, la elasticidad de −0,08 del estado de conservación de las viviendas parece razonable.

b) Los coeficientes de las variables continuas que aparecen sin transformar —OCUP, ALTA y ANTIG— representan el porcentaje en que subiría el alquiler ante un incremento unitario de la característica correspondiente. El 8 por 100 de la variable años de ocupación es interpretable como el índice de inflación anual de los alquileres en el período 1965-1974. El premio que el mercado establece por situarse en zonas más modernas o de mejores condiciones socioeconómicas es del 7 y del 9 por 100, respectivamente.

c) Otra manera de facilitar la interpretación económica de los resultados del análisis de regresión es por medio de los precios implícitos de las características. Dada una relación funcional entre el alquiler y el conjunto de las variables explicativas, el precio implícito de una variable continua no es más que la derivada parcial de esa función respecto de la variable en cuestión. En nuestro caso, como el alquiler aparece en logaritmos, los precios implícitos no son constantes para todas las viviendas. Para una variable explicativa que aparezca también en logaritmos, como los m^2, la función del precio implícito toma la forma

$$\frac{\partial \text{ALQ}}{\partial \text{M2}} = \frac{0{,}39\ \text{ALQ}_i}{\text{M2}_i},\ i = 1, \ldots, 443$$

Cuando la variable está sin transformar, como en el caso del índice ALTA

$$\frac{\partial \text{ALQ}}{\partial \text{ALTA}} = 0{,}09\ \text{ALQ}_i,\ i = 1, \ldots, 443.$$

En la parte superior de la tabla 7 se presentan los precios implícitos de todas las variables continuas evaluados en la media.

d) De acuerdo con Halvorsen y Palmsquit (1980), cuando las variables dependientes aparecen en logaritmos, la expresión $(e^{\beta_j} - 1)\ 100$, donde β_j es el coeficiente de una variable explicativa de tipo cualitativo, se interpreta como el efecto porcentual en porcentaje de la presencia del atributo de que se trate. En la parte inferior de la tabla 7 se presentan estos efectos para el total de las viviendas post-64.

En suma: a) la bondad del ajuste es muy satisfactoria si se tiene en cuenta que se trata de datos de corte transversal; y b) la explicación económica de las diferencias en el alquiler en términos de las 17 variables explicativas del modelo final resulta, en conjunto, razonable.

TABLA 7

	EDAD	OCUP	M2	NPL	ECON	ACC6	ALTA	ANTIG
Precios implícitos de las variables continuas en ptas/mes	−19	−356	26	168	−78	−45	394	−326

	EDS19	CHHI	POCOS	DOMAS	TELF	CALC	GAR	MUEB	GCOM
Efecto porcentual sobre el alquiler en % de la presencia de los atributos cualitativos	12,7	62,9	−21,8	20,5	16,0	14,0	28,3	29,0	13,3

6. CONCLUSIONES

Para prevenir la gran sensibilidad de mínimos cuadrados ante observaciones atípicas, se recomienda realizar un estudio interno de robustez del modelo que ponga de manifiesto, tanto las observaciones potencialmente influyentes como las que, de hecho, ejercen una influencia clara en la estimación. La utilización de los términos diagonales de la matriz de proyección V es un buen indicador del primer aspecto, mientras que el estadístico D_i de Cook es una adecuada medida de lo segundo.

La utilización de estos conceptos es importante porque, como mostramos con un ejemplo concreto, las decisiones sobre la forma funcional del modelo y sobre las variables a incluir en el mismo pueden estar afectados por unas pocas anomalías. El estudio interno de la muestra, permite, además, decidir si la falta de normalidad en los residuos puede ser debida a una contaminación de la distribución normal, o se debe a una falta de normalidad general de la distribución.

Por último, la depuración de anomalías en nuestro ejemplo concreto conduce a un modelo con mayor sentido económico y capacidad predictiva.

APENDICE SOBRE LAS VARIABLES UTILIZADAS

Durante los últimos diez años, en otros países se ha acumulado una gran experiencia en cuanto a los determinantes del valor de mercado de las viviendas (véase, por ejemplo, la revisión de la literatura en Ruiz-Castillo, 1982 b). En general, las características empíricas relevantes se dividen en dos grupos. En primer lugar existe toda una serie de características estructurales de la vivienda o del edificio a la que ésta pertenece, como la superficie útil habitable, las instalaciones de distinto tipo, el año y materiales de construcción, el número de plantas, etc. En segundo lugar, cuando se cuenta con datos para ello, se incluye un conjunto de características relacionadas con la localización espacial de la vivienda. A continuación expondremos la información de que hemos dispuesto sobre los atributos de ambos tipos.

Para las características estructurales se ha utilizado la información suministrada por la empresa CETA. Esta encuesta no estaba depurada. Por consiguiente, de las 473 observaciones correspondientes a viviendas post-64 hubo que prescindir de 13 que carecían de información básica sobre el alquiler, los m^2 de superficie habitable o el tipo de edificios. Se rechazó, por tanto, el 2,7 por 100 de las observaciones. De las 460 restantes, una carecía de datos sobre servicios higiénicos y dos más sobre el número de plantas de edificio. Con objeto de utilizar una muestra del mayor tamaño posible, optamos por asignar a esas tres observaciones los valores medios de las variables que les faltaban.

En muchas ocasiones, las viviendas poseen o no poseen determinados atributos como, por ejemplo, calefacción central o una plaza de garaje. Esa restricción se ha tenido en cuenta a través de variables cualitativas que toman el valor 1 ó 0, según que la vivienda posea o carezca de la característica de que se trate. En general, se ha seguido la convención de tomar la situación que se presenta en mayor número de veces como situación de referencia. Así, por ejemplo, dada la distribución porcentual del tipo de edificios de las viviendas post-64: viviendas «marginales» (0,055 por 100), se ha tomado la medianería como situación de referencia.

En lo que se refiere a las variables continuas, conviene hacer las siguientes precisiones:

a) En caso de subrogación, para el cómputo de la variable años de ocupación (OCUP) se ha tomado la fecha del contrato inicial.

b) Para la variable EDAD disponíamos de dos tipos de datos: los suministrados por el encuestador y los declarados por el inquilino. En ambos casos se daba la fecha de construcción del edificio o el intervalo dentro del cual se llevó a cabo la construcción. Por indicación de la empresa CETA, se concedió mayor credibilidad a la información del encuestador. En muchas ocasiones hubo que tomar el punto medio del intervalo de construcción, lo cual originó ciertas discontinuidades en esta variable. Cuando todo lo que se sabía es que la fecha de construcción pertenecía al siglo XIX, se asignó a la variable EDAD el valor 85, suponiéndose, por tanto, que el edificio se construyó en 1890.

c) La información sobre el estado de conservación de los edificios venía dada por el encuestador, que añadía una serie de puntos por cada uno de ocho tipos de desperfectos que se observaran. Así, cuanto mayor es el valor de la variable ECON, peor es el estado de conservación de la vivienda correspondiente.

En cuanto a las características relacionadas con la localización de la vivienda, se ha utilizado sobre todo información suministrada por COPLACO. Para intentar detectar la influencia sobre el alquiler mensual de la naturaleza del área en que las viviendas post-64 están situadas, seguimos dos rutas alternativas: la construcción de variables cualitativas y la construcción de variables continuas.

Para muchos propósitos, COPLACO agrega las zonas de transporte en 98 zonas de análisis de actividades. Por nuestra parte, combinando criterios socioeconómicos y de proximidad geográfica con las limitaciones que nos imponía el número de observaciones con que contábamos, agregamos simplemente las 98 zonas de análisis en nueve macrozonas y construimos, por tanto, ocho variables cualitativas, cada una de las cuales tomaba el valor 1 si la vivienda pertenecía a la macrozona correspondiente.

Por otra parte, se construyeron siete variables continuas. La primera es un índice ponderado de accesibilidad a los barrios de Cortes y Sol, Justicia y Universidad, Delicias y Legazpi, Recoletos y Castellana, Cuatro Caminos y Argüelles, donde se concentra casi el 25 por 100 de los puestos de trabajo del Area Metropolitana de Madrid. Los coeficientes de ponderación reflejan la importancia relativa del empleo en cada una de esas zonas respecto del total en el conjunto de las mismas. La accesibilidad viene medida en minutos de transporte privado y transporte público, ponderados por la tasa de utilización relativa de ambos modos para los desplazamientos por todos los motivos dentro de nuestra zona geográfica.

Las tres variables siguientes representan características socioeconómicas de las distintas zonas de análisis: la densidad de población medida en habitantes por km^2, la renta familiar media según COPLACO y un índice del nivel socioeconómico, consistente en el primer componente principal que explicaba el 34 por 100 de la varianza de un conjunto de 12 variables que recogían diferentes aspectos de las zonas de análisis.

La quinta y la sexta variables se construyeron también por medio de la técnica de los componentes principales aplicada a cinco variables que describían distintos rasgos de las edificaciones de cada una de las zonas de análisis. Los factores de carga permitieron interpretar el primer componente principal —que explicaba el 38 por 100 de la varianza— como índice de la antigüedad de los edificios de cada zona, y el segundo componente —que explicaba un 21 por 100 adicional de la varianza— como un indicador del grado de chabolismo y marginalidad.

Finalmente, de toda la gama de variables que podría recoger la actuación del sector público local, sólo hemos podido contar con los puestos de preescolar y EGB por 1.000 habitantes, que mide de alguna forma la oferta de servicios educativos de esta naturaleza en las distintas zonas de análisis.

En resumen, esta lista de variables constituye una primera aproximación bastante limitada a la medición de los atributos determinantes de la calidad de las zonas de análisis.

En este apéndice resta tan sólo referirnos a las dificultades de medición del alquiler de las viviendas. En todo estudio de este tipo es siempre difícil decidir si la cantidad que figura en el recibo mensual corresponde solamente al precio del arrendamiento propiamente dicho, o incluye también pagos por otros conceptos como la calefacción central o los gastos comunes del edificio.

En nuestro caso, contábamos con información sobre si los pagos por agua fría o caliente, gas/carbón, calefacción o gastos comunes, estaban o no incluidos en otro concepto. Pero desconocíamos si ese concepto era o no el propio recibo del alquiler.

Además, ni siquiera fue posible estimar con suficiente fiabilidad la media de los pagos por los distintos servicios indicados.

En consecuencia, optamos por no depurar en absoluto la variable alquiler para llegar a una cifra neta de pagos por otros servicios. En su lugar, construimos tres variables cualitativas que toman el valor 1 si el inquilino declara, respectivamente, que los gastos comunes, la calefacción o el agua caliente están incluidos en otro concepto. De esta forma confiamos en controlar los posibles efectos sobre el alquiler de que se den cualquiera de estas situaciones.

BIBLIOGRAFIA

ABRAHAM, B., y BOX, G. E. P.: «Linear Models and Spurious Observations». *Applied Statistics*, 27, 131-138, 1978.

ANDREWS, P. F., y PREGIBON, D.: «Finding the outliers that matter». *JRSS, B*, 40, 85-93, 1978.

ANSCOMBE, F. J., y TUKEY, J. W.: «The estimation and analysis of Residuals». *Technometrics*, 5, 141-160, 1963.

BARNETT, V., y LEWIS, T.: *Outliers in Statistical Data*, Wiley, 1978.

BEATON, A. F., y TUKEY, J. W.: «The fitting of power series, meaning polynomials, illustrated on band-spectrocopic data». *Technometrics*, 16, 147-185, 1974.

BELSLEY, D. A.; KUH, E., y WELSCH, R. E.: *Regression Diagnostics.*, Wiley, 1980.

BOX, G. E. P.: «Non-normality and tests on Variances». *Biometrika*, 40, 318, 1953.

BOX, G. E. P.: «Robutness and Modelling», en *Robutsness in statistics*. R. L. Launer y G. N. Wilkinson, editores. Academic Press, 1979.

BOX, G. E. P.: «Sampling and Bayes'Inference in Scientific Modelling and Robustness». *JRSS, A*, 143, 1980.

BOX, G. E. P., y DRAPER, W. R.: «Robust Designs». *Biometrika*, 62, 347-352, 1975.

BOX, G. E. P., y TIAO, C. G.: «A Bayesian approach to some outlier problems». *Biometrika*, 55, 119-129, 1968.

BOX, G. E. P., y TIAO, C. G.: *Bayesian Inference in Statistical Analysis*, Addison-Wesley, 1973.

COOK, R. D.: «Detection of Influential Observation in Linear Regresion». *Technometrics*, 19, 15-18, 1977.

COOK, R. D.: «Influential Observations in Linear Regressión». *JASA*, 74, 169-17, 1979.

COOK, R. D., y PRESCOTT, P.: «On the accuracy of Bonferroni significance levels for detecting outliers in linear models». *Technometrics*, 23, 59-63, 1981.

COOK, R. D., y WEISBERG, S.: «Characterizations of an empirical influence for detecting influential cases in Regression». *Technometrics*, 22, 495-508, 1980.

CHEN, G. G., y BOX, G. E. P.: «Implied assumptions for some proposed robust estimators». *Technical Report No. 568*, University of Wisconsin, Madison, 1979 a.

CHEN, G. G., y BOX, G. E. P.: «A study of real data». *Technical Report No. 569*, Dept. Statistics, University of Wisconsin, Madison, 1979 b.

CHEN, G. G., y BOX, G. E. O.: «Further study of Robustification via a Bayesian approach». *Technical Report No. 570*, Dep. Statistics, University of Wisconsin, 1979 c.

DIANANDA, P. H.: «Note on some properties of Maximum Likelihood estimates». *Proc. Camb. Phil. Societ.*, 45, 536.

DRAPER, N. R., y SMITH, H.: *Applied Regression Analysis*, 2.ª ed., Wiley, 1980.

DRAPER, N. R., y JOHN, J. A.: «Influential observations and outliers in Regression». *Technometrics*, 23, 21-26, 1981.

GUTTMAN, I.: «Premium and protection of several procedures for dealing with outliers when sample size are moderate to large». *Technometrics*, 15, 385-404, 1973.

HAMPEL, F. R.: «The Influence curve and its role in robust estimation.». *JASA*, 62, 1179-1186, 1974.

HOAGLIN, D. C., y WELSH, R. E.: «The hat matrix in regression and ANOVA». *Amer. Statist*, 32, 17-22, 1978.

HOGG, R. V.: «An Introduction to Robust Estimation», in *Robustness in Statistics*, R. L. Launer, y G. N. Wilkinson, editores. Academic Press, 1979.

HUBER, P. J.: «Robust estimation of a location parameter». *Ann. Math Statist.*, 135, 73-101, 1964.

HUBER, P. J.: «Robust Regression: Asymptotics, conjetures and Monte Carlo». *Ann. Statist.*, 1, 799-821, 1973.

HUBER, P. J.: *Robust Statistics*, Wiley, 1981.

HUBER, P. J.: «Robustness and designs», en *A Survey of Statistical Design and Linear Models*, J. N. Srirastarra, Ed; North-Holland, 1975.

LAMBERT, D.: «Influence Functions for Testing». *JASA*, 76, 649-657, 1981.

MILLER, R. G.: «Developments in multiple comparisons 1966-76». *JASA*, 72, 779-88.

MOSTELLER, F., y TUKEY, J. W.: *Data Analysis and Regression*. Addison-Wesley, 1977.

PEÑA, D., y RUIZ-CASTILLO, J.: «Un análisis econométrico de la legislación sobre el control de alquileres». *Información Comercial Española*, 585, 31-41, 1982 a.

PEÑA, D., y RUIZ-CASTILLO, J.: «Un análisis econométrico de las viviendas en arrendamiento de protección oficial». *Información Comercial Española*, 585, 42-48, 1982 b.

RUIZ-CASTILLO, J.: «Los determinantes del alquiler y los beneficios de la intervención del Estado en el sector de la vivienda en arrendamiento: Una aplicación del enfoque hedónico». *Investigaciones Económicas*, 18, 121-136, mayo-agosto, 1982 a.

RUIZ-CASTILLO, J.: «El enfoque hedónico: Fundamentos microeconómicos y aplicaciones en el sector de la vivienda, próximo a aparecer en un volumen del Instituto de Estudios de Administración Local», 1982 a.

TUKEY, J. W.: «A survey of sampling from contaminated distributions», in *Contributions to Probability and Statistics*, Olkin, edit. University Press, Standford, Calif., 1960.

TUKEY, J. W.: *Exploratory Data Analysis*, Addison-Wesley, 1977.

VALLEMAN, P. F., y WELSCH, R. E.: «Efficient Computing of Regression Diagnostics». *The American Statistician*, 35, 234-242, 1981.

WEISBERG, S.: *Applied linear Regression*, Wiley, 1980.

YOHAI, V. J., y MARONNA, R.: «Asymptotic behavior of M-estimators for the linear model». *Annals of Statistics*, 7, 1980.

SUMMARY

This work analyses procedures for robustification of the lineal model. We compare the advantages of using a robust estimating procedure based on M-estimators with an internal analysis of the stregth of the minimum tables and the sample. Theoretical advantage of this latter point are shown and application is illustrated by means of the construction of an explicative model of the determinating factors of rents in the metropolitan area of Madrid.

Key words: Robustness methods, outliers, regression, generaliced distance.

AMS. 1970, subject classification: 62J05 y 62P20.