©Journal of Business & Economic Statistics, Vol. 2, No. 1, January 1984

Robust Methods of Building Regression Models—An Application to the Housing Sector

Daniel Peña
Escuela de Ingenieros Industriales, Universidad Politécnica de Madrid, Madrid, Spain

Javier Ruiz-Castillo
Departmento de Teoría Económica, Universidad Complutense de Madrid, Madrid, Spain

This article studies robustification strategies for the linear model in the presence of outliers. The advantages of an internal analysis of the robustness of least squares for a given sample are pointed out. The application of this methodology is illustrated by building an explicit model of the determinants of rental housing values in the Madrid Metropolitan Area.

KEY WORDS: Outliers; Influential observations; Robust regression; Cook distance; Hedonic price function; Housing market.

1. INTRODUCTION

The question of introducing some degree of objectivity in the rejection of outlying observations has been the subject of considerable research in the statistics literature. This is a fundamental problem with cross-section samples where one typically has a large body of data on numerous variables. A few atypical observations may make the data distribution nonnormal, destroying the optimality of the least squares estimation procedure, which could become very inefficient.

In this article we consider the problem from the point of view of building an explanatory model of market rental values in terms of the observed traits of each housing unit in an urban area. Hence, this exercise belongs to the vast literature on hedonic price functions in urban economics, which has been reviewed by Griliches (1971), Ball (1973), and Quigley (1979). The microeconomic underpinnings of the empirical work in this area can be found in Rosen (1974), who provides a model of price determination of a differentiated and indivisible product under competitive conditions.

The rest of the article is organized as follows. Section 2 summarizes the effects of outliers in the context of maximum likelihood estimation of the linear model. Section 3 briefly surveys the possible solutions and shows the advantages of a robustification of the model's construction methodology, consisting of an internal sensitivity analysis of a model estimated by least squares with a particular sample. Its empirical application is illustrated in Section 4, where we present a model of the determinants of housing rental values for the Madrid Metropolitan Area. The final section, Section 5, contains some concluding comments.

2. THE EFFECTS OF OUTLIERS

We begin by briefly reviewing for later reference the maximum likelihood estimation of the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U} \tag{1}$$

where $\mathbf{Y}$ is a $(n \times 1)$ vector of responses, $\mathbf{X}$ is a $(n \times k)$ matrix of predetermined variables with rank k, β is a $(k \times 1)$ vector of parameters, and $\mathbf{U}$ is a $(n \times 1)$ vector of disturbances.

Let f be the density function of $\mathbf{U}$, and assume that $E[\mathbf{U}] = \mathbf{0}$ and $E[\mathbf{U}\mathbf{U}'] = \sigma^2 \mathbf{I}$. The maximum likelihood estimation of (1) leads to

$$\max \sum_{i=1}^{n} \ln f(e_i) = \min \sum_{i=1}^{n} -g(e_i), \tag{2}$$

where $-g = \ln f$ and $e_i = y_i - \mathbf{x}_i' \beta$ are the sample residuals.

If f is differentiable, the maximum likelihood estimator of β is the solution (assumed unique) to the system

$$\sum_{i=1}^{n} \psi(e_i)\mathbf{x}_i' = \mathbf{0}', \tag{3}$$

where ψ is the first derivative of g, and $\mathbf{x}_i$ is the ith row

of **X**. Another way of writing (3) is

$$\sum_{i=1}^{n} e_i \mathbf{x}_i' w_i = \mathbf{0}', \tag{4}$$

where $w_i = \psi(e_i)/e_i$.

Thus, the maximum likelihood estimation of the linear model can be interpreted as (a) the minimization of a certain function g of the sample residuals; (b) the choice of a function ψ of sample residuals, whose components are orthogonal to the linear space generated by the columns of **X**; and (c) as weighted least squares with weights w_i determined iteratively.

If f is symmetric, we may assume that it belongs to the potential exponential family—a general form suggested by Diananda (1949) and Box (1953), and studied by Box and Tiao (1973). In this case

$$f(u) = k_1(\alpha)\sigma^{-1}\exp\{-k_2(\alpha) \mid u/\sigma \mid^{2/(1+\alpha)}\},$$
$$-1 < \alpha \leq 1, \quad \sigma < \infty, \quad -\infty\, u < \infty, \tag{5}$$

where σ is the standard deviation, and α indicates the kurtosis of the distribution.

For $\alpha = 0$, the distribution is the normal; for $\alpha = 1$, it is the Laplace distribution; and as α approaches -1, one obtains in the limit the uniform distribution. Moreover, expression (5) includes leptokurtic distributions with tails wider than the normal when $\alpha > 0$, and platokurtic distributions when $\alpha < 0$.

Taking α as known, the maximization of the likelihood of a linear model with disturbances given by (5) leads to

$$\min \sum_{i=1}^{n} \left| y_i - \mathbf{x}_i' \beta \right|^{2/(1+\alpha)}.$$

This includes as particular cases the minimization of absolute deviations ($\alpha = 1$), least squares ($\alpha = 0$), and the minimization of the maximum deviation (as $\alpha \to -1$). Therefore, the decision on an adequate estimation criterion strongly depends on the specific characteristics of the distribution with which one is working.

In this context, the problem with least squares is that it may become very inefficient in the presence of a few atypical data that make the distribution leptokurtic. To see this, assume that the disturbances in the linear model are $N(0, \sigma^2)$ but there exists an unknown small proportion ϵ of atypical observations. This fact can be modeled, following, among others, Tukey (1960), Box and Tiao (1968, 1973), and Guttman (1973), by assuming that these anomalous observations come from a normal distribution with zero mean and variance $h\,\sigma^2$ with $h > 1$. Then the density function will be

$$f(u) = (1 - \epsilon) f_N(u \mid 0, \sigma^2) + \epsilon\, f_N(u \mid 0, h\,\sigma^2). \tag{6}$$

It is immediate that

$$\text{var}(u) = \sigma^2(1 + \epsilon(h - 1))$$

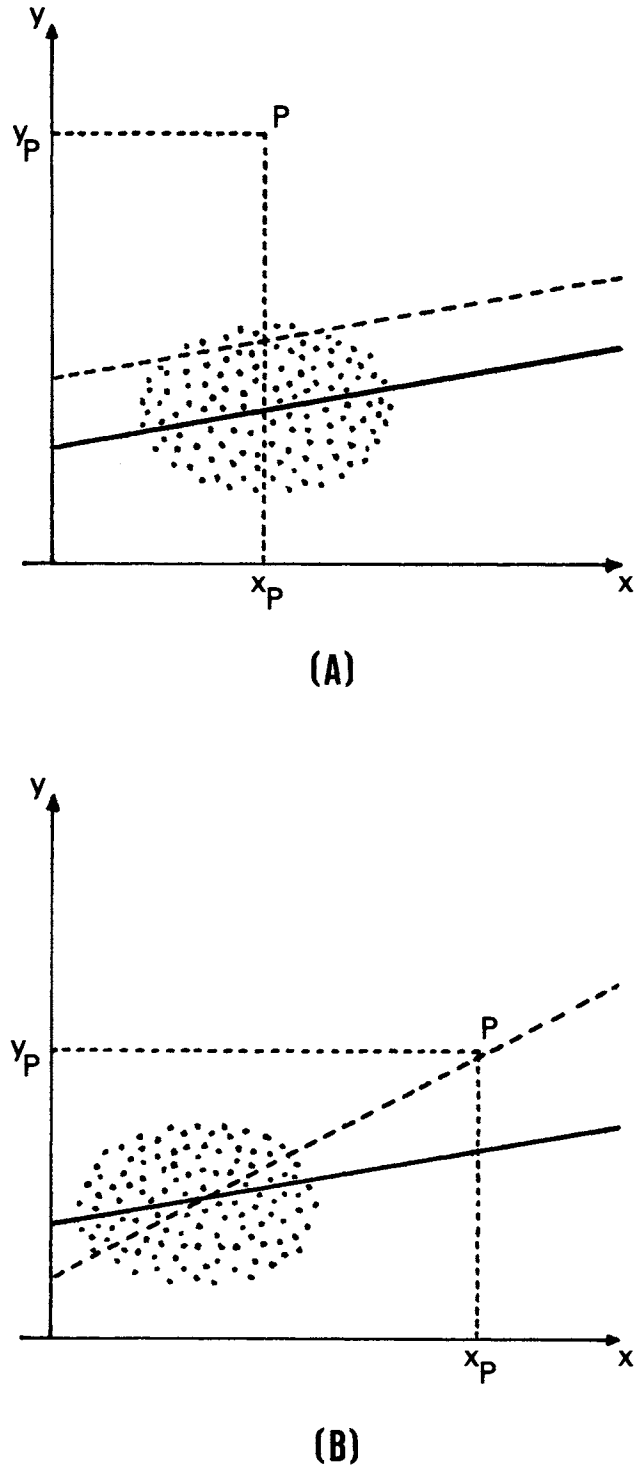

Figure 1. TWO TYPES OF OUTLIERS. In case (A), the anomalous value of the response leads to a vertical displacement of the regression line and a large residual. In case (B), the atypical value of the explicative variables determines the slope of the regression line but leads to a small residual.

and f will be symmetric with kurtosis

$$\gamma = 3\left(\frac{1 + \epsilon(h^2 - 1)}{(1 + \epsilon(h - 1))^2} - 1\right) = 3\,(\delta - 1)$$

with $\delta > 1$. Therefore, the distribution will be leptokur-

tic. In this case, least squares is no longer optimal. Also, since the variances of the parameter estimators depend directly on the error variance, which is greater than σ^2, such estimates will be unreliable and very unstable in different samples.

Finally, it is worthwhile to note that there are two types of possible outliers. If we consider sample points $(y_i, \mathbf{x}_i')$, one may find an anomalous value of y_i for the corresponding $\mathbf{x}_i$, as in Figure 1 (A). The residual for point P will be large, and its effect will be a vertical displacement of the regression line. Alternatively, we may have an atypical value of the vector of explanatory variables that may not be associated with an atypical response, as in Figure 1 (B). Here, point P alone essentially determines the slope of the regression line, so that in spite of the anomalous nature of the situation the residual may be very small or even zero.

3. DIFFERENT APPROACHES TO SOLVE THE PROBLEM

The practical approaches to deal with the problems posed by outliers can be summarized as follows:

1. Appeal to the central limit theorem to justify the normality hypothesis in order to use least squares. Once the model has been estimated, use residual plots against the estimated values or the explanatory variables to detect possible outliers.
2. Use a Bayesian approach that involves building a formal model which incorporates the a priori expected deviation with respect to the standard linear model by means of parameters in an extended model.
3. Reject least squares in favor of a robust estimation procedure by selecting a function g that yields reasonably efficient estimates under the normality assumption without suffering the instability of least squares in the presence of outliers.
4. Robustify, rather than the estimation criterion, the methodology followed in the construction of the linear model.

This requires checking at each stage that decisions are not determined by a small group of anomalous observations. Hence, least squares is not abandoned, but instead the estimation process is supplemented by a battery of diagnostic checks that permit detection of potentially influential observations, measurements of their effects on estimated coefficients, and tests of whether they are significantly atypical.

In the next section we briefly review these approaches.

3.1 The Use of Residual Plots

This is the alternative suggested by the vast majority of statistics and econometric textbooks. Its main limitation is that, at best, residual plots can only serve to detect outliers of type A in Figure 1. However, in the context of a large sample of data on numerous variables, residual plots by themselves are not very helpful for detecting atypical multivariant values with several coordinates far from the mean values of the explanatory variables. Unfortunately, these outliers of type B in Figure 1 may have a great influence on the regression results and are, therefore, particularly damaging.

3.2 The Bayesian Approach

This has been used by Jeffreys (1961), Box and Tiao (1968), Chen and Box (1979 a, b, c), Box (1979, 1980) and others. It is possibly the most general and thorough approach to the problem, but we have been unable to implement it because of its computational complications and the requirements of adequate software for its efficient application. Thus, we abstain here from further comments on it.

3.3 Robust Regression Estimates

The shortcomings of the least squares approach already mentioned have led in the last 20 years to an extensive literature that aims to overcome these difficulties. Books by Mosteller and Tukey (1977), Huber (1981), and Barnett and Lewis (1978) present the problem and contain numerous references.

The instability of least squares in the presence of outliers is due to the form of the functions g and ψ in expressions (2) and (3). In this case, $g(u) = u^2$, $\psi(u) = u$, and $w_i(u) = \psi(u)/u = 1$. Therefore, since all observations are given equal weight, those data with a large residual in absolute value carry the least squares equation towards them—an obviously undesirable effect. It is intuitively clear that a function g that grows more slowly when u is large will give a smaller weight to such atypical observations, leading, consequently, to more robust estimates. This solution has been advocated by Huber (1964) and others (see Stigler 1973 for historical comments). Hogg (1979) and Huber (1981) present a good summary of this approach. See also Jeffreys (1961, p. 214 ff.).

These robust procedures are subject to three types of criticisms. First, the heuristic nature of the functions g or ψ introduce a certain arbitrariness in the formulation. Second, the small-sample properties of the estimates are unknown. Third, these methods are useful in dealing with outliers of type A in Figure 1, but they do not solve the problem posed by atypical values with small residuals.

With respect to the first criticism, Chen and Box (1979a) have established that the functions g and ψ suggested in the literature are optimal for particular types of contamination. For instance, Huber's function g is optimal for a normal distribution with Laplace tails, which can be closely approximated by the contaminated normal model presented in (6). Therefore, it can be argued that the methodology we use should depend

on the specific structure of each particular sample. The third criticism leads to generalized M-estimates in which the weights w_i in (4) depend not only on the residual but also on the observation's influence measured by its distance to the center of the scatter of points as in Krasker and Welsch (1982). Although this approach partially solves the problem, the solution remains heuristic and obtaining sampling properties of estimates is difficult.

3.4 Robustification of the Methodology

The main reason for constructing robust estimation methods is to guarantee that our results will not be fundamentally dependent on a few anomalous observations. However, the fact that an estimate *might* be very sensitive to a small set of outliers does not mean that it is inefficient in every conceivable case. Before rejecting an estimation procedure, it is reasonable to investigate whether its good properties are preserved in each particular sample.

Therefore, given a data set susceptible to being treated by means of a linear model, it is pertinent to ask the following questions: (a) Does this sample contain observations whose a priori influence is much greater than the rest in the construction of the model? (b) Is it possible to measure the actual influence that each individual observation has a posteriori on the parameter estimates? (c) Does there exist a test to determine whether an observation constitutes an outlier?

We now review the answers that have been given to these questions. The first issue has been approached with the help of the "hat" matrix, whose properties have been discussed by Huber (1975), Hoaglin and Welsch (1978), Cook (1977, 1979), Belsley, Kuh, and Welsch (1980), and Weisberg (1980).

The "hat" matrix $\mathbf{V}$ projects the vector $\mathbf{Y}$ on the linear space generated by the columns of $\mathbf{X}$:

$$\hat{\mathbf{Y}} = \mathbf{V}\,\mathbf{Y}, \quad \mathbf{V} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{8}$$

The matrix $\mathbf{V}$ is symmetric and idempotent. Its importance for our purpose lies in the fact that $\mathbf{e} = (\mathbf{I} - \mathbf{V})\mathbf{U} = (\mathbf{I} - \mathbf{V})\mathbf{Y}$, from which one obtains

$$\operatorname{var}(e_i) = \sigma^2(1 - v_{ii}) \tag{9}$$

with $v_{ii} = (\mathbf{x}_i - \bar{\mathbf{x}})'(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$, where $\tilde{\mathbf{X}}$ is the centered matrix of the observation, and $1/n\ (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})$ is the variance and covariance matrix for the explanatory variables.

Therefore, except for a constant term, v_{ii} represents the Mahalanobis distance of an observation $\mathbf{x}_i$ to the center of gravity of the scatter of points, $\bar{X}$. If a point $\mathbf{x}_i$ is very far from $\bar{X}$, its v_{ii} will be large and the variance of the corresponding residual will be small, as (9) indicates. In the limit, if $v_{ii} = 1$, the variance will be zero, which means that the point's position relative to the rest forces the regression equation to go through it, irrespective of the observed value for y_i.

It can be concluded that sample points with high v_{ii} are, potentially, influential. Since $\mathbf{V}$ is a projection matrix, $0 < v_{ii} \leq 1$. Moreover, since the trace of an idempotent matrix is equal to its rank, $\sum_{i=1}^{n} v_{ii} = k$, where k is the rank of $\mathbf{X}$. Consequently, the average value of the v_{ii}'s is k/n. Following Belsley, Kuh, and Welsch (1980), in practice an observation is considered potentially influential if $v_{ii} > 2k/n$.

Huber (1981) has suggested another interesting interpretation for the v_{ii} terms. Since $\mathbf{V}$ is idempotent, $v_{ii} = \sum_{j=1}^{n} v_{ij}^2$. Thus, taking (8) into account

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

Therefore, recalling that the sample mean of h independent observations with common variance σ^2 has variance σ^2/h, it is clear that $1/v_{ii}$ can be interpreted as the number of equivalent observations used to compute y_i. If $v_{ii} = 1$, then y_i is computed with a single observation, its residual is zero (see Equation 9).

In an alternative approach to determine a priori influential observations, Andrews and Pregibon (1978) use the change of "volume" of the scatter of points when one eliminates a subset of observations. However, Draper and John (1981) have established that a measure of a single point's influence in this approach is precisely $1 - v_{ii}$.

The second issue is how to determine the actual influence on the model of each observation in a given sample. There are several ways of doing this based on the empirical influence function $IE_A = \hat{\beta}_A - \tilde{\beta}$, where $\hat{\beta}_A$ is the estimate obtained after eliminating the subset A of observations, and $\tilde{\beta}$ is the estimate with the full sample (see Cook and Weisberg 1980).

A simple way of obtaining a scalar measure of A's influence is to consider the distance between $\hat{\beta}_A$ and $\tilde{\beta}$ in a metric with statistical meaning. Cook (1977) introduced such a measure by

$$D_A = (\hat{\beta}_A - \tilde{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\beta}_A - \tilde{\beta})/ks^2,$$

where s^2 is the regression residual variance and $(X'X)^{-1}s^2$ is an estimate of the variance covariance matrix for $\tilde{\beta}$.

Using the subindex (i) to indicate that a referred-to characteristic has been calculated without the ith observation, the Cook distance can easily be obtained from

$$\begin{aligned} D_i &= (\hat{\beta}_{(i)} - \tilde{\beta})'\ (X'X)(\hat{\beta}_{(i)} - \tilde{\beta})/ks^2 \\ &= e_i^2 v_{ii}/ks^2(1 - v_{ii})^2. \end{aligned}$$

It is interesting to note that D_i can also be written as

$$D_i = (\hat{\mathbf{Y}}_{(i)} - \mathbf{Y})'\ (\hat{\mathbf{Y}}_{(i)} - \mathbf{Y})/ks^2.$$

indicating that D_i measures the Euclidean distance in which the prediction vector $\mathbf{Y}$ is translated after eliminating the ith observation from the regression.

Finally, the construction of tests to determine the

atypical data in regression models has used numerous approaches (see Barnett and Lewis 1978 for a survey of this topic). When the problem is considered within a likelihood ratio testing approach, the resulting test statistic is a monotonic function of the Studentized residuals

$$r_i = e_i/s\sqrt{1 - v_{ii}}, \tag{10}$$

where the least squares residual has been divided by its estimated standard deviation.

A shortcoming of this approach is that the distribution of r_i under the normality assumption is not Student t, because numerator and denominator are not independent. However, the substitution of $s_{(i)}$ for s in (10) yields a Student t distribution with $n - k - 1$ degrees of freedom. For computational reasons (see Weisberg 1980), it is convenient to express it as

$$t_i = r_i\sqrt{(n - k - 1)/(n - k - r_i^2)},$$

where r_i is given by (10). The relevant distribution for obtaining the test's significance level is that of the maximum value of a sample of t statistics with $n - k - 1$ degrees of freedom, which value is unknown. However, approximate critical values have been tabulated using the Bonferroni inequality (see Miller 1977, and Cook and Prescott 1981).

In conclusion, the statistics v_{ii}, D_i, and t_i constitute the basis for the methodological robustification of the linear model. The v_{ii} terms depend only on the predetermined variables and measure the potential influence of each observation taking into account its position relative to the rest of the sample. We would have a robust design if all points had analogous v_{ii} values. The Cook D_i statistic captures the actual influence of each observation on the estimated parameters of the prediction vector $\hat{\mathbf{Y}}$. The statistic is interesting because it indicates the practical irrelevance of worrying about sample observations that, although anomalous, have little influence on the model. Finally, the t statistic summarizes both features and is used as a formal test of whether a single observation is an outlier. The next section contains an application of this way of attacking the problems posed by outliers.

4. A MODEL FOR THE DETERMINANTS OF RENTAL HOUSING VALUES IN THE MADRID METROPOLITAN AREA

4.1 The Problem and the Data

In Spain, government intervention in the rental housing sector takes two forms. First, several public institutions promote—directly or indirectly—the construction of public housing at rents below the market level. Second, since 1920 the government has enforced compulsory lease renewal and rent controls in the private sector. In 1964 rents were liberalized on new contracts. Therefore, the rental market sector includes only private housing units occupied after 1964.

Our problem in this section is to build an explanatory model of market rental values in terms of the observed traits of dwelling units in the Madrid Metropolitan Area (MMA hereafter). This is not a behavioral relationship but a function that gives the rent resulting from the interaction of supply and demand for each variety of the differentiated product. The partial derivatives of that function are interpreted as the implicit or hedonic prices of the corresponding characteristics.

The final aim is to use the estimated model to assess the economic advantages and the distributional conse-

Table 1. Structural Housing Traits

Name	Description	Mean	Standard Deviation
	a. Continuous Variables		
AGE	Building age in years since its construction	21.9	23.0
OCUP	Years of occupancy of the housing unit	3.6	2.3
M2	Space in squared meters	68.0	42.8
ROOM[a]	Space in number of rooms	3.6	1.2
NFL	Number of floors	5.1	2.8
DET	Deterioration state of the building	4.6	17.5
	b. Dummy Variables		*Percent*
Age of building			
AXIX	Built in XIX century		7
—	Built in 1900–1940		19
A4164	Built in 1941–1964		21
A6574	Built in 1965–1974		53
			100
Type of building			
MAGL	"Marginal" housing (in bad condition)		3
CHTW	Chalet or townhouse		4
APT	Detached apartment building		38
—	Other apartment building		55
			100
Type of promotion			
PRI[a]	Private firm		27
—	User's cooperative, particular individual, selfconstruction		45
UNK[a]	Unknown		28
			100
Hygienic services			
LESS	Less than a full bathroom		19
—	One full bathroom		70
TWOM	Two or more bathrooms		11
			100
Payments of utilities			
HOTW[a]	Hot water bill included in other concept		47
HEAT[a]	Heating bill included in other concept		38
BEXP[a]	Building expenditures included in other concept		19
Other variables			
TELPH	With telephone		36
CHEAT	With central heating		20
GAR	With garage		7
JAN[a]	With a janitor		34
FURN	With furniture		15
FTEN[a]	First tenancy		18

[a] Nonsignificant variables in the exploratory analysis.

quences of public housing and rent control policies in Spain. The results of such an assessment will be reported elsewhere in Peña and Ruiz-Castillo (1983).

Our data come from a 1974 survey of 4,067 housing units in the MMA (or .4% of the total number for that area). The sample used here consists of 460 private rental dwellings occupied between 1964 and 1974. Such data will be made available by the authors upon request.

Hedonic price functions for urban areas usually involve two types of explanatory variables: a set of traits characterizing dwelling units of the buildings to which they belong, and a set of locational characteristics. Tables 1 and 2 describe the variables of either type that we could measure.

Some comments on measurement problems are in order:

1. In many cases, to get a value for a building's age, we had to consider the midpoint of the known construction interval. This led to certain discontinuities for this variable. When we only knew that the building dated from the 19th century, the *AGE* variable was assigned the value 85, which implies that the construction date was assumed to be 1880.

2. For each building, interviewers recorded a number of points for each of eight types of observed defects. Thus, the greater the value of the state of deterioration variable (*DET*), the worse the condition of the corresponding building.

3. The accessibility variable is a weighted index of transportation times from each of the 85 zones that make up the MMA to six centrally located neighborhoods where 25% of all employment is concentrated. The weights reflect the relative importance of employment in each of those neighborhoods relative to total employment in the six. Accessibility is measured in minutes of private and public transportation, weighted by the utilization rate of both modes for all transportation purposes in the metropolitan area.

4. The index *HIGH* is the first principal component explaining 34% of the variance of a set of 12 variables representing different socioeconomic aspects of the 85 zones in the MMA.

5. We also applied principal components analysis to 5 variables describing several buildings' characteristics in each of the 85 zones. The first principal component (*OLD*), explaining 38% of the variance, was dominated by the average age of buildings in each zone. The second component (*INFRA*), explaining an additional 21% of the variance, was interpreted as an index of the importance of housing in bad condition in each zone.

6. Of all the variables that might conceivably represent local public sector activities, we could only measure primary and secondary school enrollment (*SCHOOL*).

7. It is always difficult to ascertain whether the rental amount in monthly expenditures in surveys of this type reflects gross or net rent. In our case, we only knew whether the payments for certain utilities were included or not in other measures, but did not know whether such a measure was the rent bill itself. We tried to account for these effects by constructing three dummy variables that take the value 1 if the person interviewed declared, respectively, that payments for hot water, heating, or building expenditures were included in the other measure.

Table 2. Locational Variables[a]

Name	Description	Mean	Standard Deviation
ACC	Accessibility index in minutes of transportation time	41.2	17.6
POPD	Population density in inhabitants per squared kilometer	19,950	20,683
RENT[b]	Average monthly family income in pesetas	18,826	4,764
HIGH	Socioeconomic index	.06	.88
OLD	Buildings age index	.13	1.14
INFRA[b]	Index of housing in bad conditions	−.07	.78
SCHOOL[b]	Primary and secondary school enrollments	7,351	4,345

[a] All variables take values in the 85 zones that make up the Madrid Metropolitan Area.
[b] Nonsignificant variables in the explanatory analysis.

4.2 The Selection of the Functional Form

To decide on the best functional form, we followed an iterative process that began with an exploratory analysis of the data to obtain a reasonable first representation. Next, we concentrated on the identification of possible outliers. Finally, we carried out the maximum likelihood estimation of the dependent variable's best transformation and performed various checks to find an adequate metric for the predetermined variables.

For the initial exploratory analysis, we used three types of tools: bivariate plots of the response variable with respect to each explanatory variable, the empirical distribution of each variable, and residual plots from preliminary regressions with several sets of explanatory variables. The results were the following:

1. The rent variable required transformation, possibly a logarithmic transformation. Plots of $e_i = F(y_i)$ for the untransformed y variable showed curvature and heteroskedasticity. Moreover, the distribution of both the rent variable and the regression residuals had a strong positive asymmetry. Finally, the logarithm has a clear economic interpretation, indicating that each trait's effect depends on the level reached by the other housing attributes.

2. To obtain linearity for the response, once rents were expressed in logs, the logarithmic transformation was also applied to the continuous variables *OCUP*,

M2, *NFL*, *DET*, *ACC*, and *POPD*. Since the case for *OCUP* and *NFL* was not clear, the decision to transform them was maintained only provisionally.

3. Variables denoted by *a* in Tables 1 and 2 were initially rejected because they did not supply additional information.

4. The building age variable showed a complex and highly nonlinear influence, probably because it captures very different effects and acts as a proxy for other variables. Moreover, as already indicated, its construction was not free of difficulties. To identify nonlinear effects, its range was broken down into several intervals represented by a set of dummy variables. The results were that both 19th-century and very modern housing showed rents significantly higher, while housing from 1940 was the least expensive. In a first approximation, we represented this effect by a second degree polynomial. To avoid the expected multicollinearity, the following variables were defined: $AGDM = AGE - \tilde{AGE}$, and $AGDM2 = AGDM^2$, where $\tilde{AGE}$ is the mean age for all housing.

With these decisions made, the resulting model appears in column (1) of Table 3. The residuals' distribution is asymmetric, with asymmetry and kurtosis coefficients equal to −1.95 and 7.5. The Kolmogorov-Smirnov test leads to the rejection of the residuals'

Table 3. Regression Results

Variables	Coefficients (and standard errors)					Other Alternative Variables
	(1)	(2)	(3)	(4)	(5)	
CONSTANT	5.77	7.14	7.57	7.81	8.13	
	(.77)	(.60)	(.35)	(.31)	(.33)	
AGDM	−.012	−.010	−.012	−.008	−.09	AGE[a]
	(.002)	(.002)	(.002)	(.002)	(.03)	
AGDM2	.0002	.0002	.0002	.0001	.12	AXIX
	(.00005)	(.00004)	(.00004)	(.0003)	(.08)	
OCUP[a]	−.25	−.25	−.25	−.25	−.08	OCUP
	(.04)	(.03)	(.03)	(.02)	(.007)	
M2[a]	.46	.42	.42	.40	.39	
	(.07)	(.05)	(.05)	(.05)	(.05)	
NFL[a]	.26	.20	.20	.18	.19	
	(.06)	(.05)	(.04)	(.04)	(.04)	
DET[a]	−.06	−.06	−.06	−.07	−.08	
	(.03)	(.02)	(.02)	(.02)	(.02)	
MAGL	.11	−.0008	—	—	—	
	(.15)	(.11)				
CHTW	.66	.53	.54	.48	.49	
	(.15)	(.12)	(.11)	(.10)	(.10)	
APT	−.08	.02	—	—	—	
	(.06)	(.05)				
LESS	−.19	−.23	−.23	−.23	−.25	
	(.08)	(.07)	(.07)	(.06)	(.06)	
TWOM	.06	.07	.08	.18	.19	
	(.09)	(.07)	(.07)	(.06)	(.06)	
TELPH	.05	.14	.14	.14	.15	
	(.06)	(.05)	(.05)	(.04)	(.04)	
CHEAT	.11	.09	.09	.11	.13	
	(.07)	(.06)	(.06)	(.05)	(.05)	
GAR	.28	.29	.30	.26	.25	
	(.10)	(.08)	(.08)	(.07)	(.07)	
FURN	.33	.29	.29	.24	.26	
	(.07)	(.05)	(.05)	(.05)	(.05)	
BEXP	—	—	.06	.11	.12	
			(.05)	(.05)	(.04)	
ACC[a]	−.15	−.33	−.38	−.41	.41	
	(.13)	(.10)	(.08)	(.07)	(.07)	
POPD[a]	.04	.0009	—	—	—	
	(.02)	(.017)				
HIGH	.16	.10	.10	.09	.08	
	(.04)	(.03)	(.03)	(.03)	(.03)	
OLD	−.06	−.05	−0.7	−.06	−.07	
	(.03)	(.02)	(.03)	(.02)	(.02)	
R^2	.71	.79	.79	.82	.82	
Standard error	.48	.37	.37	.33	.32	
Number of observations	460	451	451	443	443	

[a] In logarithms.

Table 4. Possible Outliers

Observation number	v_{ii}	D_i	t_i
1	.03	.06	−6.6
2	.04	.05	−5.3
3	.06	.08	−5.4
4	.03	.02	−4.1
5	.04	.03	−3.7
6	.05	.03	−3.7
7	.06	.04	−3.7
8	.03	.02	−3.6
9	.05	.03	−3.3
10	.05	.02	−3.0
11	.04	.02	−3.0
12	.11	.04	−2.8
13	.03	.01	−2.8
14	.03	.01	−2.5
15	.03	.00	2.4
16	.04	.01	−2.4
17	.12	.04	2.4
18	.15	.00	−0.7
19	.15	.00	.07

normality with $\alpha = .01$. The distribution appears to be a normal contaminated by a small number of negative values, since it is symmetric around the median (whose value is .07) and reasonably normal in the range $.07 \pm 1.5\ \hat{\sigma}$, where $\hat{\sigma}$ is the residuals standard deviation.

The internal analysis of the model's robustness yielded 19 observations worthy of attention, which are included in Table 4. The last two (numbers 18 and 19) are the potentially more influential with the largest v_{ii} values, although their actual influence is negligible according to the D_i statistic. The first 17 observations with the largest t_i values were carefully reviewed, with the result that the first 9 appeared to suffer from data transcription errors (omission of a zero in the rent figure). Since several of the next 8 observations were open to doubt, we decided to maintain them provisionally. Consequently, we estimated a new regression with 451 data points with results summarized in column (2) of Table 3. As can be seen, the elimination of the first 9 observations improves the results without changing them substantially. The coefficients of most variables remain essentially constant with the following exceptions: (a) the coefficients of *MAGL* and *APT* become practically zero, suggesting that they should be eliminated from the model; (b) the influence of *POPD* appears to be captured now by the accessibility index; and (c) the coefficients of *TELPH* and *ACC* increase, making them significant.

In view of this information, we estimated a new model without the variables *MAGL*, *APT*, and *POPD*, but introducing the variables previously rejected for the model with 460 data points. As a result, *BEXP* was provisionally included in the model because, although not significant (it has a t value of 1.2), it appears to be potentially important. Column (3) of Table 3 summarizes the final model fitted with 451 data points.

Next, we repeated the robustness analysis for this model in order to detect new anomalous data. We confirmed the atypical nature of the 8 observations previously commented upon, although their actual influence appeared to be generally small judging by their D_i values. At any rate, we reestimated the model without these 8 observations, obtaining the results presented in column (4) of Table 3. We should remark that (a) the variables *TWOM*, *CHEAT*, and *BEXP*, which were not formally significant with $\alpha = .05$, become significant without any doubt; (b) the rest of the coefficients are not significantly altered; and (c) the Kolmogorov-Smirnov test, as well as tests on the asymmetry and kurtosis, lead to the acceptance of the hypothesis of the residuals' normality with $\alpha = .10$.

In conclusion, if we compare the latter with the initial model, it can be observed that after eliminating the 18 observations that we considered as outliers (3.7% of the total), the residual variance has diminished by 55%, the proportion of explained variability has increased by 17%, and we can reasonably accept the hypothesis that the residuals are normally distributed. Most coefficients have changed very slightly, and when this is not the case and the model becomes more compatible with a priori economic information: the distance to the center of Madrid measured by the logarithm of the accessibility index, and the fact that a housing unit has two or more bathrooms, telephone, central heating, and building expenditures included in another measure, become significant variables in the model presumably free from atypical values.

To test the former specification we have performed the maximum likelihood estimation of the Box-Cox transformation parameter λ for the rent variable. This has been done for the models with 460, 451, and 443 data. The results, which are practically insensitive to different specifications of the continuous explanatory variables, are presented in Table 5.

As we keep eliminating atypical observations, the maximum of the likelihood function for λ gradually

Table 5. Maximum Values of the Likelihood Function for Different Specifications of the Box-Cox Parameter and the Sample Size

	λ							
n	−.2	−.1	.0	.1	.2	.3	.4	.5
460	−4866	−4816	−4775	−4745	−4727	−4721	−4729	−4750
451	−4635	−4605	−4585	−4574	−4574	−4584	−4605	−4637
443	−4493	−4469	−4454	−4447	−4450	−4464	−4489	−4524

approaches zero. The maximum is reached for $\lambda = .3$ with the full sample, .2 with 451 data, and .1 with 443 data. In the latter case, a 95% confidence interval does not include the logarithmic transformation ($\lambda = 0$). Although this suggests that the model might still contain further outliers, we did accept the logarithmic transformation as adequate because it is reasonable from an economic point of view and is not dramatically rejected by the empirical evidence (see Atkinson 1982 for a recent analysis of the transformation parameters' sensitivity to outliers).

As regards the explanatory variables, we have already pointed out that the first exploratory models did not indicate whether to express the *OCUP* and *NFL* variables with or without a logarithmic transformation. To decide this issue, we performed the following 2×2 factorial experiment in which the sums of squared residuals are presented for each possible specification:

	OCUP	*ln OCUP*
NFL	45.34	46.25
ln NFL	44.22	45.14

The results suggest that the number of floors should be in logs, while the years of occupancy should not be transformed.

The last variable whose specification was open to doubt was the building age. We searched for the best nonlinear specification using the procedure suggested by Box and Tidwell (1962), but unfortunately the computation algorithm did not converge. Finally, we applied the following criterion. First, among the plausible transformations, choose the one generating the smallest sum of square residuals. Next, study the possibility of supplementing that specification with one or more of the dummy variables *AXIX*, *A4164*, or *A6574*.

This procedure leads to the logarithmic transformation, corrected by the dummy *AXIX*. Since the coefficient of the log of *AGE* was negative and that for *AXIX* was positive, this formulation is consistent with our information on the relationship's pattern: ceteris paribus, the greater the building age, the smaller the housing rent except for the 19th-century buildings, whose solid construction (or other unobserved characteristics) require an upward correction.

4.3 The Selection of the Final Model

Since we want to predict market rents for housing units whose rents are government controlled, a relevant criterion to choose the number of regressors is the mean squared prediction error. An estimate of this error serving to compare different models is the Mallows statistic: $C_p = (SSR_p/\hat{\sigma}^2) + 2p - n$, where SSR_p is the sum of squared residuals with p regressors, $\hat{\sigma}^2$ is an unbiased estimator of the residual variance in the model with the largest number of variables, and n is the sample size. The C_p statistic permits the selection of the subset that maximizes the model's predictive capacity (or minimizes the mean squared error).

This criterion did not lead to the inclusion of new variables to our previous list. The best model is presented in column (5) of Table 3, and has a C_p of 12.6 with 17 explanatory variables. Once this model was selected, we repeated the internal analysis of each observation and searched for other sources of specification errors. The results were as follows:

1. The maximum value for the t statistic for the Studentized residuals was 3.5. The two next values were 3.1, while the rest of the data presented no problems. These three observations are close to the explanatory variables' center of gravity, so that their influence on parameter estimates is small. At any rate, there was no observation with a high D_i value. Therefore, we concluded that the final model is robust to outliers.
2. Residual plots did not show any evidence of specification errors. The residual distribution is normal according to the Kolmogorov-Smirnov test with $\alpha = .05$.
3. Finally, the estimation situation is adequate without multicollinearity problems: the condition index of the $\mathbf{X'X}$ matrix was only 8.8 (see Belsley, Kuh, and Welsch 1980).

4.4 The Economic Interpretation

In the first place, the above analysis indicates that 82% of rent differences for market housing in the MMA can be explained by the 17 characteristics that were empirically relevant. While the information on structural traits is rather rich, data on attributes referring to housing location in the 85 zones of the MMA were very poor. Thus, it is not surprising that the latter, the accessibility index *ACC*, the socioeconomic index *ALTA*, and the buildings age index *OLD* explained only 4% of the observed variability, while the 14 structural characteristics explained the remaining 78%. If we had data on local public goods levels, pollution of different types, and the distribution of nonresidential land uses, we should expect that the location variables' relative importance would have been greater.

In the second place, all variables appear with the expected algebraic sign. As for the coefficients' interpretation, the following comments are in order:

1. For the variables in logarithms *AGE*, *M2*, *NFL*, *DET*, and *ACC* the coefficients measure the elasticity directly. Thus, a 10% increase in housing size measured in squared meters leads to a 4% increase in rents, which indicates that there are decreasing returns to scale in this variable. The $-.4$ elasticity for the accessibility index is somewhat low; two dwellings identical in every respect, except for a difference of 50% on transportation time to the Central Business District, would have a 20% difference in rent. The .19 elasticity for number of

floors does not have an immediate interpretation; perhaps taller buildings are more desirable on average because they possess some characteristic not reflected in our survey. Finally, the −.08 elasticity for housing deterioration state seems reasonable.

2. For the continuous untransformed variables *OCUP*, *ALTA*, and *ANTIG* the coefficients, multiplied by 100, represent the percentage in rent increase attributable to a unit increase in the corresponding characteristic. The 8% for occupancy years can be interpreted as the annual rate of rent inflation during the 1965–1974 period. The market premiums for location in more modern zones or for better socioeconomic conditions are, respectively, 7% to 9%.

3. When the dependent variable appears in logs, the expression $(\exp \beta_j - 1)\cdot 100$, where β_j is the coefficient of a dummy variable, is interpreted as the percentage change in rents due to the presence of the attribute in question. For the 9 significant dummy variables, such effects, expressed in percent, are as follows:

FURN	*AXIX*	*CHTW*	*LESS*	*TWOM*
29.0	12.7	62.9	−21.8	20.5

TELF	*CHEAT*	*GAR*	*BEXP*
16.0	14.0	28.3	13.3

In conclusion, the goodness of fit is very satisfactory, and the economic explanation of rent differences in terms of the final model's 17 significant variables is, on balance, quite reasonable.

5. CONCLUSIONS

To prevent least squares' great sensitivity to outliers, an internal study of the model's robustness was recommended in Section 3 highlighting the potentially influential observations, as well as those that, in fact, clearly affect the estimation results. The diagonal terms of the projection matrix V is a good indication of the former, while the Cook D_i statistic is adequate to measure the latter. Moreover, a t statistic serves to determine which observations can be considered atypical.

Our empirical application of this methodology could be placed in the context of Chapter 4 of Belsley, Kuh, and Welsch (1980). These authors analyze the Harrison and Rubinfeld (1978) data on market values and characteristics of 506 owner-occupied dwelling units in the Boston Metropolitan Area. In the first place, they detect the nonnormality of OLS residuals in such a regression. Then they compute M-estimators, observe that some coefficients are considerably modified, and verify that the weighted Studentized residuals follow a normal distribution. On the other hand, using some statistics different from ours, they detect that 10% of the sample consists of influential observations, and informally analyze the consequences of applying OLS after deleting 5 data points. In neither case do they investigate the effect of outliers on the functional form.

In Section 4, we also found nonnormality of OLS residuals in an exploratory model. However, when we apply the robustification strategy we recommend, we find that (a) decisions regarding the model's functional form and the variables to include in it can be considerably affected by a few outliers; (b) the residuals' lack of normality can be attributed to some identified data coding errors and other anomalous observations; and (c) the elimination of outliers improves the statistical model of rent housing values in the MMA, enhancing its economic meaning.

ACKNOWLEDGMENTS

This work is part of a larger project financed by the Spanish Ministerio de Economía y Comercio. The authors wish to acknowledge the cooperation of José Antonio Quintero, who was responsible for the computational aspects of this work, as well as helpful comments by this journal's referees.

[*Received September 1982. Revised July 1983.*]

REFERENCES

ANDREWS, P. F., and PREGIBON, D. (1978), "Finding the Outliers that Matter," *Journal of the Royal Statistical Society*, Ser. B, 40, 85–93.

ATKINSON, A. C. (1982), "Regression Diagnostics, Transformations and Constructed Variables," *Journal of the Royal Statistical Society*, Ser. B, 44, 1–36.

BALL, M. J. (1973), "Recent Empirical Work on the Determinants of Relative House Prices," *Urban Studies*, 10, 213–231.

BARNETT, V., and LEWIS, T. (1978), *Outliers in Statistical Data*, John Wiley.

BELSLEY, D. A., KUH, E., and WELSCH, R. E. (1980), *Regression Diagnostics*, John Wiley.

BOX, G. E. P. (1953), "Non-normality and Tests on Variances," *Biometrika*, 40, 318–335.

——— (1979), "Robustness and Modeling," in *Robustness in Statistics*, eds. R. L. Launer and G. N. Wilkinson, Academic Press.

——— (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 383–430.

BOX, G. E. P., and TIAO, C. G. (1968), "A Bayesian Approach to Some Outlier Problems," *Biometrika*, 55, 119–129.

——— (1973), *Bayesian Inference in Statistical Analysis*, Reading, Mass.: Addison-Wesley.

BOX, G. E. P., and TIDWELL, P. W. (1962), "Transformation of the Independent Variables," *Technometrics*, 4, 531–550.

CHEN, G. G., and BOX, G. E. P. (1979a), "Implied Assumptions for Some Proposed Robust Estimators," Technical Report No. 568, University of Wisconsin, Madison, Dept. of Statistics.

——— (1979b), "A Study of Real Data," Technical Report No. 569, University of Wisconsin, Madison, Dept. of Statistics.

——— (1979c), "Further Study of Robustification Via a Bayesian Approach," Technical Report No. 570, University of Wisconsin, Madison, Dept. of Statistics.

COOK, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

——— (1979), "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74, 169–17.

COOK, R. D., and PRESCOTT, P. (1981), "On the Accuracy of

Bonferroni Significance Levels for Detecting Outliers in Linear Models," *Technometrics*, 23, 59–63.

COOK, R. D., and WEISBERG, S. (1980), "Characterization of an Empirical Influence Function for Detecting Influential Cases in Regression," *Technometrics*, 22, 495–508.

DIANANDA, P. H. (1949), "Note on Some Properties of Maximum Likelihood Estimates," *Proceedings of the Cambridge Philosophical Society*, 45, 536–544.

DRAPER, N. R., and JOHN, J. A. (1981), "Influential Observations and Outliers in Regression," *Technometrics*, 23, 21–26.

GRILLICHES, Z. (ed.) (1971), *Price Indexes and Quality Change*, Harvard University Press.

GUTTMAN, I. (1973), "Premium and Protection of Several Procedures for Dealing With Outliers When Sample Sizes are Moderate to Large," *Technometrics*, 15, 385–404.

HARRISON, D., and RUBINFELD, D. L. (1978), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81–102.

HOAGLIN, D. C., and WELSCH, R. E. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17–22.

HOGG, R. V. (1979), "An Introduction to Robust Estimation," in *Robustness in Statistics*, eds. R. L. Launer and G. N. Wilkinson, Academic Press.

HUBER, P. J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 135, 73–101.

——— (1975), "Robustness and Designs," in *A Survey of Statistical Design and Linear Models*, ed. J. N. Srirastarra, North-Holland.

——— (1981), *Robust Statistics*, John Wiley.

JEFFREYS, H. (1961), *Theory of Probability*, Oxford Clarendon Press.

KRASKER, W. S., and WELSCH, R. E. (1982), "Efficient Bounded-influence Regression Estimation," *Journal of the American Statistical Association*, 77, 595–604.

MILLER, R. G. (1977), "Developments in Multiple Comparisons 1966–1976," *Journal of the American Statistical Association*, 72, 779–788.

MOSTELLER, F., and TUKEY, J. W. (1977), *Data Analysis and Regression*, Addison-Weslsy.

PEÑA, D., and RUIZ-CASTILLO, J. (1983), "Distributional Aspects of Public Rental Housing and Rent Control Policies in Spain," *Journal of Urban Economics*, forthcoming.

QUIGLEY, J. M. (1979), "What Have We Learned about Urban Housing Markets?," in *Current Issues in Urban Economics*, eds. P. Mieszkowski and M. Straszheim, The Johns Hopkins University Press.

ROSEN, S. (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economics*, 82, 34–55.

STIGLER, S. M. (1973), "Simon Newcomb, Perey Daniel, and the History of Robust Estimation 1855–1920," *Journal of the American Statistical Association*, 63, 872–879.

TUKEY, J. W. (1960), "A Survey of Sampling from Contaminated Distributions, in *Contributions to Probability and Statistics*, ed. J. Olkin, Stanford University Press.

WEISBERG, S. (1980), *Applied Linear Regression*, John Wiley.