Chapter 11

ESTIMATION FOR DIRTY DATA AND FLAWED MODELS

WILLIAM S. KRASKER*

Harvard University

EDWIN KUH and ROY E. WELSCH*

Massachusetts Institute of Technology

Contents

*This research was supported, in part, by the National Science Foundation, U.S. Department of Energy, and I.B.M.

Handbook of Econometrics, Volume I, Edited by Z. Griliches and M.D. Intriligator

1. Introduction

We are concerned with the econometric implications of the sensitivity to data of coefficient estimates, policy analyses, and forecasts in the context of a regression model. In contrast to the emphasis in standard treatments of the linear model paradigm described subsequently, we are interested in data, how they are generated, and particular data configurations in the context of a specified regression model. The focus of this chapter is on resistant estimation procedures and methods for evaluating the impact of particular data elements on regression estimates. While terminology is not yet firmly fixed in this rapidly evolving area, resistant estimation here is presumed to include classical robust estimation for location [Andrews et al. (1972)] or regression [Huber (1977)] and bounded-influence regression [Krasker and Welsch (1982a)]. "Classical robust" estimation reduces the effect of outliers in error space. Bounded-influence regression, in addition, limits the permissible impact of outliers in explanatory-variable space.

The time-honored point of departure in econometrics is the ordinary least squares (OLS) estimator $b = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ for the linear regression model $y = X\beta + \varepsilon$, where y is the response variable data vector, X is the explanatory variable data matrix, β are coefficients to be estimated, and ε conditional on X is a random vector with $E(\varepsilon\varepsilon^{T}) = \Sigma = \sigma^2 I$ and $E(\varepsilon) = 0$. The widespread appeal of this model lies in its simplicity, its low computational cost, and the BLUE (Best Linear Unbiased Estimator) property shown by the Gauss–Markov theorem. When ε is normally distributed, there is the added theoretical imprimatur of maximum likelihood and attendant full efficiency. Also, for fixed X, exact small sample tests of significance are possible.

More elaborate estimators are needed when the simple assumptions that motivate OLS are considered invalid. Thus, generalized least squares (GLS) replaces OLS when $\Sigma \neq \sigma^2 I$ leading to the Aitkin estimator, $b = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}y$, when the errors are heteroscedastic or autocorrelated. GLS estimates are BLUE for known Σ and have desirable asymptotic properties when Σ has been consistently estimated.

When the explanatory variables cannot be viewed as fixed, the choice of estimator depends on the sources of random behavior and whatever further assumptions the econometrician considers tenable. Random behavior in the explanatory variables includes observational errors, endogenous variables that are part of a simultaneous-equation system, variance-component models, lagged endogenous variables, stochastic regressors with some joint distribution, and stochastic parameter variation. Failure to recognize these statistical attributes can lead to one or more of the following shortcomings: inefficiency, finite sample bias, inconsistency, and incorrect tests of significance. Generally speaking, correct estimation procedures differ from OLS/GLS when these circumstances prevail

and estimators are tailored to whatever specific stochastic conditions are deemed the most important.

This perspective can be extended to encompass estimators that avoid undue reliance on small segments of the data when there are large but isolated departures from the maintained statistical hypotheses. Thus, reliable estimation sometimes calls for explicit consideration of the X matrix so as to limit the permissible influence of any one of its rows. At the same time, one would also like protection against occasional large ε. A class of resistant estimators that restricts unusually influential components of X and ε, called bounded-influence estimators, offers protection against several types of common specification problems and requires less restrictive assumptions about stochastic properties than those customarily required in the more complex regression structures enumerated above.

Robust regression has appeared in econometrics literature since the mid-1950s, mainly in the guise of Least Absolute Residuals, an estimator that minimizes the sum of the absolute values rather than the square of errors. According to a fine survey by Lester D. Taylor (1974): "LAR has the same illustrious progenitors as least squares (Gauss and Laplace)... but has historically never attracted much attention." Even though coefficient computations became practical through linear programming, as initially pointed out by Charnes, Cooper and Ferguson (1955), Karst (1958), Wagner (1959), and W. Fisher (1961), distribution theory has remained a problem, although a recent paper by Koenker and Basset (1978) provides asymptotic theory.

Two empirical studies suggest that in some cases LAR (or variants) may outperform OLS. Again, quoting Taylor: "What Meyer and Glauber (1964) did was first to estimate their investment models by LAR as well as least squares and then test the equations on post-sample data by using them to forecast the 9 (and sometimes 11) observations subsequent to the period of fit. They found that, with very few exceptions, the equations estimated by LAR outperformed the ones estimated by least squares even on criteria (such as the sum of squared forecast errors) with respect to which least squares is ordinarily thought to be optimal (p. 171)." Another study by Fair (1974) used approximations to LAR and adaptations of other robust estimators in a fifteen-equation macro model. His comparisons had an outcome similar to that of Meyer and Glauber: LAR outperformed OLS in post-sample forecasts.

While these isolated instances of empirical research are suggestive of potentially attractive results, resistant estimation (in its LAR garb or any other) has remained peripheral to mainstream econometric work because of computational costs as well as the absence of widely available code designed for this purpose, and the lack of convincing theoretical support. These deficiencies, along with more intense concerns about other econometric issues and widespread acceptance of OLS, help to explain the relative neglect of resistant regression.

Resistant estimators offer protection against certain fairly general model failures while preserving high efficiency in well-behaved situations. This approach differs from the more standard econometric approach where an alternative estimator is devised to cope with specific departures from a more standard specification.

There is an inevitable gap between a model and reality; it is one thing to write down a model and another to believe it. Three model/data problems are of immediate concern.[1] First, there may be "local errors", such as round-off errors or groupings of observations. Second, there may be "gross errors" in the data, e.g. incorrectly recorded numbers, keypunch errors, or observations made on the wrong quantity. Finally, the model itself is typically thought to be only an approximation. In regression, for example, the linearity of the model and the normality of the disturbance distribution are both good approximations, at best.

Local errors occur in virtually all data sets, if for no other reason than the fact that we work with only finitely many significant digits. However, local errors do not ordinarily cause serious problems for the classical regression procedures, so we will not be too concerned with them.

Gross errors occur more often in some types of data sets than in others. A time-series model using National Income Accounts data and a moderate number of observations is unlikely to contain data with gross errors (provided the numbers which are actually read into the computer are carefully checked). However, consider a large cross section for which the data were obtained by sending questionnaires to individuals. Some respondants will misinterpret certain questions, while others will deliberately give incorrect information. Further errors may result from the process of transcribing the information from the questionnaires to other forms; and then there are the inevitable keypunch errors. Even if the data collectors are careful, some fraction of the numbers which are ultimately fed into the computer will be erroneous.

The third category – the approximate nature of the model itself – is also a serious problem. Least squares can be very inefficient when the disturbance distribution is heavy tailed. Moreover, although the linear specification is often adequate over most of the range of the explanatory variables, it can readily fail for extreme values of the explanatory variables; unfortunately, the extreme values are typically the points which have the most influence on the least squares coefficient estimates.

Gross errors – even if they are a very small fraction of the data – can have an arbitrarily large effect on the distribution of least squares coefficient estimates. Similarly, a failure of the linear specification – even if it affects only the few observations which lie in extreme regions of the X space – can cause OLS to give a misleading picture of the pattern set by the bulk of the data.

[1] The discussion which follows goes back to Hampel (1968).

While general considerations about data appear in Chapter 27 by Griliches we need to examine in more detail those circumstances in which statistical properties of the data – in isolation or in relation to the model – counsel the use of resistant estimators. These will often be used as a check on the sensitivity of OLS or GLS estimates simply by noting if the estimates or predictions are sharply different. Sometimes they will be chosen as the preferred alternative to OLS or GLS.

The common practice in applied econometrics of putting a dummy variable into a regression equation to account for large residuals that are associated with unusual events, requires a closer look. The inclusion of a single dummy variable with zeros everywhere except in one period forces that period's residual to zero and is equivalent to deleting that particular row of data. The resulting distribution of residuals will then appear to be much better behaved. [See Belsley, Kuh and Welsch (1980, pp. 68–69).] Should dummy variables be used to downweight observations in this manner? Dummy variables are often an appealing way to increase estimation precision when there are strong *prior* reasons for their inclusion, such as strikes, natural disasters, or regular seasonal variation. Even then, dummy variables are often inadequate. When a strike occurs in a particular quarter, anticipations will influence earlier periods and unwinding the effects of the strike will influence subsequent periods. As an interesting alternative to OLS, one might wish to consider an algorithm that downweights observations smoothly according to reasonable resistant statistical criteria instead of introducing discrete dummy variables after the fact, which has the harsher effect of setting the row weight to zero.

Model-builders using macroeconomic time series are often plagued by occasional unusual events, leading them to decrease the weights to be attached to these data much in the spirit of resistant estimation. Even when there are good data and theory that correspond reasonably well to the process being modeled, there are episodic model failures. Since it is impractical to model reality in its full complexity, steps should be taken to prevent such model failures from contaminating the estimates obtainable from the "good" data. Some of these breakdowns are obvious, while others are not. At least some protection can be obtained through diagnostic tests. Where the aberrant behavior is random and transitory, estimators that restrict the influence of these episodes should be seriously considered. We do not view resistant estimation as a panacea: some types of model failure require different diagnostic tests and different estimators.

Other types of model difficulties are sometimes associated with cross sections, quite apart from the sample survey problems mentioned earlier. Cross-sectional data are often generated by different processes than those which generate time series. This hardly startling proposition is a belief widely shared by other econometricians, as evidenced by the proliferation of variance-components models which structure panel data error processes precisely with this distinction in mind. (See Chapter 22 by Chamberlain on panel data.)

To some extent these differences reflect the aggregation properties of the observational unit rather than different (i.e. intertemporal versus cross-sectional) behavior. Time series often are aggregates, while cross sections or panel data often are not. There is a tendency for aggregation to smooth out large random variations which are so apparent in disaggregated data. However, time series of speculative price changes for stock market shares, grains, and non-ferrous metals are often modeled as heavy-tailed Pareto–Levy distributions which are poorly behaved by our earlier definition. These constitute a significant exception, and there are doubtless other exceptions to what we believe, nevertheless, is a useful generalization.

Cross-sectional individual observations reflect numerous socio-demographic, spatial, and economic effects, some of which can reasonably be viewed as random additive errors and others as outlying observations among the explanatory variables; many of these are intertemporally constant, or nearly so. Such particularly large cross-sectional effects have four principal consequences in econometrics. One already mentioned is the burst of interest during the last twenty years in variance-component models. A second effect is the natural proclivity in empirical research to include a great many (relative to time series) subsidiary explanatory variables, i.e. socio-demographic and spatial variables of only minor economic interest. Their inclusion is designed to explain diverse behavior as much as possible, in the hope of improving estimation accuracy and precision. Third, the relative amount of explained variation measured by R^2 is characteristically lower in cross sections than in time series despite the many explanatory variables included. Fourth, anomalous observations are likely to appear in cross sections more often than in time series.

Thus, with individual or slightly aggregated observations, resistant estimation appears especially promising as an alternative estimator and diagnostic tool since ideosyncratic individual behavior – i.e. behavior explained poorly by the regression model or a normal error process – pervades cross-section data.

A strong trend exists for exploiting the information in large data sets based on sample surveys of individuals, firms, establishments, or small geographic units such as census tracts or countries. Often these are pooled time series and cross sections. A volume of more than 700 pages, containing 25 articles, was devoted to this subject alone [*Annales de l'Insee* (1978)].[2] Research based on social security records by Peter Diamond, Richard Anderson and Yves Balcer (1976) has 689,377 observations. This major evolution in the type of data used in econometrics is a consequence of several factors, not the least of which has been enormous reductions in computational costs.

[2]It includes empirical studies on investment by M. Atkinson and J. Mairesse with about 2300 observations and R. Eisner with 4800 observations; economic returns to schooling by G. Chamberlain with 2700 observations as well as an article on a similar topic by Z. Griliches, B. Hall and J. Hausman with 5000 observations.

Since survey data are notoriously prone to various kinds of mistakes, such as response or keypunch errors, it is essential to limit their effects on estimation. Some gross errors can be spotted by examining outliers in each particular data series, but it is often impossible to spot multivariate outliers. The isolated point in Figure 1.1 would not be observed by searches of this type. Thus, observational errors compound the effects of sporadic model failures in ways that are not overcome by large sample sizes (law of large numbers). Resistant estimation is a major innovation with the potential for reducing the impact of observational error on regression estimates.

To drive home the point that the likelihood of a slightly incorrect model and/or some bad data force us to change the way we look at those extremely large cross-section data sets, consider this example: via questionnaires, we obtain a sample from a certain population of individuals to estimate the mean value of some characteristic of that population, which is distributed with mean μ and standard deviation σ. However, there are "bad" observations occurring with probability ε in the sample due, for example, to keypunch errors, or forms sent to inappropriate people. The bad points are distributed with mean $\mu + \theta$ and standard deviation $k\sigma$. The mean squared error for the sample mean $\bar{X}_n$ is $\{(1-\varepsilon+\varepsilon k^2)+\theta^2\varepsilon(1-\varepsilon)\}\sigma^2/n$. Without loss of generality, suppose $\sigma = 1$. Then if $\theta = 1$, $k = 2$, and $\varepsilon = 0.05$ (which are not at all unreasonable), the mean squared error is $0.0025 + 1.20/n$. Obviously there is very little payoff to taking a sample larger than 1000 observations. Effort would be better spent improving the data.

Since bounded-influence estimators are designed to limit the influence that any small segment of the data can have on the estimated coefficients, it is not surprising that these estimators also contain diagnostic information (much as a first-order autoregressive coefficient is both part of the standard GLS transformation and also contains diagnostic/test information). Thus, the GLS compensation for heteroscedasticity, when computed by weighted least squares (WLS), has a parallel to an algorithm used in bounded-influence estimation (hereafter often

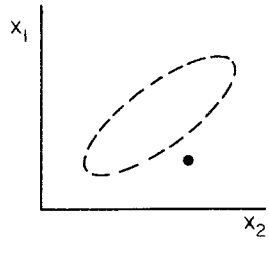

Figure 1.1

referred to as BIF) that gives weights to the rows of the data matrix: large error variances are downweighted in WLS while highly influential observations are downweighted in bounded-influence estimation. Hence, small weights in BIF point to influential data. Although computational complexity and costs are higher for BIF, they are decidely manageable.

Section 2 considers in more detail some model failures that can arise in practice. Section 3 describes recent developments in methods for the detection of influential data in regression. Section 4 is a sketch of the Krasker–Welsch BIF estimator. Section 5 raises several issues about inference in the resistant case. Section 6 considers some of the main theoretical foundations of robust and BIF estimation. Section 7 presents an example of BIF applied to the Harrison–Rubinfeld large cross-section hedonic price index. Section 8 gives some recent results on instrumental-variables bounded-influence estimation, and Section 9 discusses resistant estimation for time-series models.

2. Sources of model failure

In this section we discuss the ways in which the classical assumptions of the linear regression model are often violated. Our goal is to determine what types of data points must be downweighted in order to provide protection against model failures. Specifically, under what conditions should we downweight observations which have large residuals, "extreme" X rows, or both.

As we mentioned above, there are two categories of model failures that are potentially serious. The first consists of "gross errors", e.g. keypunch errors, incorrectly recorded numbers, or inherently low precision numbers. The second derives from the fact that the model itself is only an approximation. Typically an econometrician begins with a response (dependent) variable together with a list of explanatory (independent) variables with the full realization that there are in truth many more explanatory variables that might have been listed. Moreover, the true functional form is unknown, as is the true joint distribution of the disturbances.

A reasonable, conventional approach is to hypothesize a relatively simple model, which uses only a few of the enormous number of potential explanatory variables. The functional form is also chosen for simplicity; typically it is linear in the explanatory variables (or in simple functions of the explanatory variables). Finally, one assumes that the disturbances are i.i.d., or else that their joint distribution is described by some easily parameterized form of autocorrelation. All of these assumptions are subject to errors, sometimes very large ones.

We have described this procedure in detail in order to establish the proposition that *there is no such thing as a perfectly specified econometric model.* Proponents of robust estimation often recommend their robust estimators for "cases in which

gross errors are possible", or "cases in which the model is not known exactly". With regard to gross errors the qualification is meaningful, since one can find data sets which are error-free. However, to point out that robust procedures are not needed when the model is known exactly is misleading because it suggests that an exactly known model is actually a possibility.

If the model is not really correct, what are we trying to estimate? It seems that this question has a sensible answer only if the model is a fairly good approximation, i.e. if the substantial majority of the observations are well described (in a stochastic sense) by the model. In this case, one can at least find coefficients such that the implied model describes the bulk of the data fairly well. The observations which do not fit that general pattern should then show up with large residuals. If the model does not provide a good description of the bulk of the data for any choice of coefficients, then it is not clear that the coefficient estimates can have any meaningful interpretation at all; and there is no reason to believe that bounded-influence estimators will be more useful than any other estimator, including ordinary least squares.

The hard questions always arise after one has found a fit for the bulk of the data, and located the outliers. To gain some insight, consider a data configuration which arises often enough in practice (Figure 2.1). Most of the data in Figure 2.1 lie in a rectangular area to the left; however, some of the observations lie in the circled region to the right. Line A represents the least-squares regression line, whereas line B would be obtained from a bounded-influence estimator, which restricts the influence of the circled points.

The fit given by line B at least allows us to see that the bulk of the data are well described by an upward-sloping regression line, although a small fraction of the observations, associated with large values of x, deviate substantially from this pattern. Line A, on the other hand, is totally misleading. The behavior of the bulk of the data is misrepresented and, worse yet, the circled outliers do not have large residuals and so might go unnoticed.

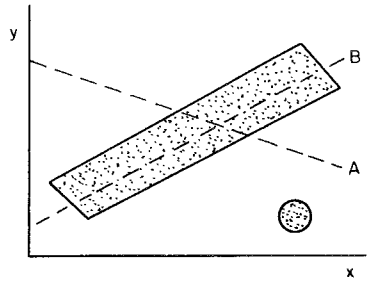

Figure 2.1

What happens as the number of observations in the circle grows large? Eventually, even the bounded-influence fit will pass near the circle. Indeed, an estimator which fits the "bulk" of the sample can hardly ignore the circled observations if they are a majority of the data. In this case there is no linear model which reasonably describes the vast majority of the observations, so that bounded-influence estimation would not help.

With the information provided by fit B, what should we do? There is no unique answer, for it depends on the purpose of the estimation. If our goal is merely to describe the bulk of the data, we might simply use the coefficients from the bounded-influence regression. If we were trying to forecast y conditional upon an x near the center of the rectangle, we would again probably want to use the bounded-influence fit.

If we want to forecast y conditional on an x near the circled points, the situation is entirely different. The circled points really provide all the data-based information we have in this case, and we would have to rely on them heavily. In practice, one would try to supplement these data with other sources of information.

Related to the last point is a well recognized circumstance among applied econometricians, namely that sometimes a small, influential subset of data contain most of the crucial information in a given sample. Thus, only since 1974 have relative energy prices shown large variability. If the post-1974 data have a different pattern from the pre-1974 data (most of the available observations) we might still prefer to rely on the post-1974 information. While this is a dramatic, identifiable (potential) change in regression regime where covariance analysis is appropriate, many less readily identifiable situations can arise in which a minority of the data contain the most useful information. Bounded-influence regression is one potentially effective way to identify these circumstances.

In short, one never simply throws away outliers. Often they are the most important observations in the sample. The reason for bounded-influence estimation is partly that we want to be sure of *detecting* outliers, to determine how they deviate from the general pattern. By trying to fit all the data well, under the assumption that the model is exactly correct, least squares frequently hides the true nature of the data.

3. Regression diagnostics

While realistic combinations of data, models, and estimators counsel that estimators restricting the permissible influence of any small segment of data be given serious consideration, it is also useful to describe a complementary approach designed to detect influential observations through regression diagnostics. While weights obtained from bounded-influence estimation have very important di-

Table 3.1
Notation

Population regression $y = X\beta + \varepsilon$	Estimated regression $y = Xb + e$ and $\hat{y} = Xb$
y: $n \times 1$ column vector for response variable	same
X: $n \times p$ matrix of explanatory variables	same
β: $p \times 1$ column vector of regression parameters	b: estimate of β
ε: $n \times 1$ column vector of errors	e: residual vector: $y - \hat{y}$
σ^2: error variance	s^2: estimated error variance
Additional notation	
x_i: ith *row* of X matrix	$b(i)$: estimate of β when ith row of X and y have been deleted
X_j: jth column of X matrix	$s^2(i)$: estimated error variance when ith row of X and y have been deleted
$X(i)$: X matrix with ith row deleted	

agnostic content, alternative diagnostics that are more closely related to traditional least-squares estimation provide valuable information and are easier to understand.

An influential observation is one that has an unusually large impact on regression outputs, such as the estimated coefficients, their standard errors, forecasts, etc. More generally, influential data are outside the pattern set by the majority of the data in the context of a model such as linear regression and an estimator (ordinary least squares, for instance). Influential points originate from various causes and appropriate remedies vary accordingly (including, but not restricted to, bounded-influence estimation). Diagnostics can assist in locating errors, allowing the user to report legitimate extreme data that greatly influence the estimated model, assessing model failures, and possibly direct research toward more reliable specifications.[3]

Two basic statistical measures, individually and in combination, characterize influential data: first, points in explanatory-variable (X)-space far removed from the majority of the X-data, and scaled residuals which are, of course, more familiar diagnostic fare. We now turn to influential X-data, or *leverage* points. As described above, an influential observation may originate from leverage, large regression residuals, or a combination of the two. A notational summary is given by Table 3.1. We note that $b = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ for OLS and call $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ the *hat matrix* with elements $h_{ik} = x_i(X^{\mathrm{T}}X)^{-1}x_k^{\mathrm{T}}$. Then $\hat{y} = Xb = Hy$ which is how the hat matrix gets its name. We can also describe the predicted values as $\hat{y}_i = \sum_{k=1}^{n} h_{ik}y_k$. Using the above results the last expression can be rewritten as

[3] This section is a condensed summary of material in chapter 2 of Belsley, Kuh and Welsch (1980).

$\hat{y}_i = \sum_{k \neq i} h_{ik} y_k + h_{ii} y_i$. Thus, the impact of y_i on $\hat{y}_i$ is controlled by the corresponding diagonal element $h_{ii} (\equiv h_i)$.

There is also a direct distance interpretation of h_i. Let $\tilde{X}$ be the X-matrix centered by column means $\bar{x}$, so that $\tilde{h}_i = (x_i - \bar{x})(\tilde{X}^T \tilde{X})^{-1}(x_i - \bar{x})^T$. It is apparent that large $\tilde{h}_i$ are relatively far removed from the center of the data measured in the $(\tilde{X}^T \tilde{X})$ coordinate system. For simple bivariate regression

$$\tilde{h}_i = \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2},$$

so that a large hat matrix diagonal corresponds in the most transparent fashion to a point removed from the center of the data.

Since H is a projection matrix, it has the following properties:

(i) $0 \leqslant h_i \leqslant 1$
(ii) $\sum h_i = p$.

Thus a perfectly balanced X-matrix – one with equal leverage for all observations – is one for which $h_i = p/n$. As further elaborated in Belsley, Kuh and Welsch (1980), when h_i exceeds $2\,p/n$ (and certainly when it exceeds $3\,p/n$), we are inclined to consider that row of data as potentially influential.

Relatively large residuals have long been viewed as indicators of regression difficulties. Since for spherically distributed errors the least-squares error variance for the ith observation is $\sigma^2(1 - h_i)$, we will scale residuals by $\hat{\sigma}^2(1 - h_i)$. Instead of using s (the sample standard deviation) estimated from all the data to estimate σ, we prefer to use $s(i)$ (the sample standard deviation excluding the ith row) so that the denominator is stochastically independent of the numerator. We thus obtain the *studentized residual*:

$$e_i^* \equiv \frac{e_i}{s(i)\sqrt{1 - h_i}}. \tag{3.1}$$

This has the t-distribution when ε is normally distributed and, interestingly, $1 - h_i$ provides a link between the regular OLS residual, e_i, and the predicted residual:

$$y_i - x_i b(i) = \frac{e_i}{1 - h_i}. \tag{3.2}$$

Furthermore, the standardized predicted residual is just the studentized residual (3.1). Since the least-squares estimator works to reduce large observed residuals, especially at leverage points, residuals (however scaled) need to be augmented by leverage information.

One can observe the influence of an individual data row on regression estimates by comparing OLS quantities based on the full data set with estimates obtained when one row of data at a time has been deleted. The two basic elements, hat matrix diagonals and studentized residuals, reappear in these regression quantities which more directly reflect influential data. We will restrict ourselves here to two such row deletion measures: the predicted response variable, or fitted values, and coefficients. Thus, for fitted values $\hat{y}_i = x_i b$, we have

$$x_i b - x_i b(i) = x_i(b - b(i)) = \frac{h_i e_i}{1 - h_i}.$$

We measure this difference relative to the standard error of the fit here estimated by $s(i)\sqrt{h_i}$, giving a measure we have designated

$$DFFITS_i = \frac{h_i e_i/(1-h_i)}{s(i)\sqrt{h_i}} = \left[\frac{h_i}{1-h_i}\right]^{1/2} e_i^*. \tag{3.3}$$

It is evident that the ith data point:

(i) will have no influence even if $|e_i^*|$ is large provided h_i is small, reinforcing our belief that residuals alone are an inadequate diagnostic, and
(ii) that substantial leverage points can be a major source of influence on the fit even when $|e_i^*|$ is small.

A second direct measure of influence is the vector of estimated regression coefficients when the ith row has been deleted:

$$DFBETA_i \equiv b - b(i) = \frac{(X^{\mathrm{T}}X)^{-1}x_i^{\mathrm{T}}e_i}{1-h_i}. \tag{3.4}$$

This can be scaled by $s(i)\operatorname{diag}\sqrt{(X^{\mathrm{T}}X)^{-1}}$ yielding an expression called *DFBETAS*. The expression for *DFBETA* closely resembles Hampel's definition of the influence function as described subsequently in Section 4. It is clear from inspection that *DFBETA* and the corresponding influence (4.7) are unbounded for OLS. We observe once again that (conditional on X) the absence (presence) of a data row makes a more substantial difference when $|e_i^*|$ is large and/or h_i is large.

There is another way of viewing the influence of the ith data row that is based on fitted values and h_i. If we define $\hat{y}_i(i) = x_i b(i)$, then it can be shown that $\hat{y}_i$ for the full data set is the following weighted average of $\hat{y}_i(i)$ and y_i:

$$\hat{y}_i = (1-h_i)\hat{y}_i(i) + h_i y_i. \tag{3.5}$$

When leverage is substantial for the ith row, the predicted quantity depends heavily on the ith observation. In the example of Section 7, the largest hat matrix diagonal in a sample of 506 observations is 0.29, so that one-fifth of 1 percent of the data has a weight of nearly 1/3 in determining that particular predicted value. Such imbalance is by no means uncommon in our experience.

When several data points in X-space form a relatively tight cluster that is distant from the bulk of the remaining data, the single row deletion methods described here might not work well, since influential subsets could have their effects masked by the presence of nearby points. Then various multiple subset deletion procedures (which can, however, become uncomfortably expensive for large data sets) described in Belsley, Kuh and Welsch (1980) may be used instead. We have also found that *partial regression leverage plots* (a scatter diagram of residuals from y regressed on all but the jth column of X plotted against the residuals of column X_j regressed on all but the jth column of X; its OLS slope is just b_j) contain much highly useful qualitative information about the "masking" problem alluded to here. However, when we turn to bounded-influence regression, we find that the weights provide an alternative and valid source of information about subset influences. This fact enhances the diagnostic appeal of BIF.

4. Bounded-influence regression

In this section we sketch the main ideas behind the Krasker–Welsch bounded-influence estimator. More details may be found in Krasker and Welsch (1982a).

The notation which we will find most useful for our treatment of bounded-influence estimation is

$$y_i = x_i\beta + u_i, \qquad i = 1,\ldots,n. \tag{4.1}$$

For the "central model" we will suppose that the conditional distribution of u_i, given x_i, is $N(0, \sigma^2)$. For reasons which were discussed in detail in earlier sections, one expects small violations of this and all the other assumptions of the model. Our aim is to present an estimator which is not too sensitive to those violations.

To study asymptotic properties such as consistency and asymptotic normality of estimators for β, one usually assumes

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} X^{\mathrm{T}}X = Q \tag{4.2}$$

or, equivalently,

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i^{\mathrm{T}} x_i = Q, \tag{4.3}$$

where Q is non-singular. This will hold provided $E(x^{\mathrm{T}}x)$ exists and is non-singular, where the expectation is over the marginal distribution of x.

When these assumptions hold, the OLS estimator b has many desirable properties. In particular, b is asymptotically efficient, with

$$\sqrt{n}\,(b-\beta) \to N(0, \sigma^2 Q^{-1}) \tag{4.4}$$

in distribution.

As we have seen, least squares is quite sensitive to certain violations of the assumptions. Though formal, less sensitive alternatives have not been widely used in applied work, econometricians often do attempt to protect themselves against model or data failures. For example, a common practice in applied work is to:

(i) run an OLS regression,
(ii) examine the observations with large residuals to determine whether they should be treated separately from the bulk of the data, and
(iii) run another OLS regression with observations deleted, or dummy variables added, etc.

In Section 3 we learned that this practice is not fully satisfactory, since influential observations do not always have large least-squares residuals. Conversely, a large residual does not necessarily imply an influential observation. If we replace the word "residuals" in (ii) by "$|DFFITS|$", the three-step procedure is much improved; and one might ask whether there is any real payoff to using a more formal procedure. The answer is that the simple procedure of examining those observations with large $|DFFITS|$ is not too bad in small samples, but one can do considerably better in large samples. We can explain this as follows: for any reasonable estimator, the variability goes to zero as the sample size goes to infinity. On the other hand, a process which generates gross errors will often generate them as a certain *proportion* of the data, so that the bias caused by gross errors will *not* go to zero as the sample size increases. In these circumstances, *bias will often dominate variability in large samples.* If the concern is with mean squared error, one must therefore focus more on limiting bias as the sample size increases. In small samples it suffices to examine only highly influential observations, since gross errors which are not too influential will cause only a small bias relative to the variability. In large samples, where the variability is very small, we must be suspicious of even moderately influential observations, since even a small bias will be a large part of the mean squared error. If one used the informal three-step procedure outlined above, these considerations would lead us to delete a larger and larger fraction of the data as the sample size increased. As stated in the introduction, it is better to have a formal procedure which smoothly down-weights observations according to how influential they are.

We will now introduce two concepts, the influence function Ω and the sensitivity γ, which are applicable to an arbitrary estimator $\hat{\beta}$. Essentially, the influence $\Omega(y_i, x_i)$ of an observation (y_i, x_i) approximates its effect (suitably normalized) on the estimator $\hat{\beta}$, and γ is the maximum possible influence of a single observation. Our formal definition of influence is based on what is called the "gross error model".

Consider a process which, with probability $1-\varepsilon$, generates a "good" data point (y_i, x_i) from the hypothesized joint distribution. However, with probability ε, the process breaks down and generates an observation identically equal to some fixed (y_0, x_0) [a $(p+1)$-vector which might have nothing to do with the hypothesized joint distribution]. That is, with probability ε, the process generates a "gross error" which is always equal to (y_0, x_0). Under these circumstances the estimator $\hat{\beta}$ will have an asymptotic bias, which we can denote by $C(\varepsilon, y_0, x_0)$. We are interested mainly in how $C(\varepsilon, y_0, x_0)$ varies as a function of (y_0, x_0) for small levels of contamination, ε. Therefore, we define

$$\Omega(y_0, x_0) = \lim_{\varepsilon \downarrow 0} \frac{C(\varepsilon, y_0, x_0)}{\varepsilon}. \tag{4.5}$$

Note that $C(\varepsilon, y_0, x_0)$ is approximately $\varepsilon\Omega(y_0, x_0)$ when ε is small, so that $\varepsilon\Omega(y_0, x_0)$ approximates the bias caused by ε-contamination at (y_0, x_0). Ω is called the *influence function* of the estimator $\hat{\beta}$. If Ω is a bounded function, $\hat{\beta}$ is called a *bounded-influence* estimator.

For the least-squares estimator b, one can show that the influence function for b is

$$\Omega(y, x) = (y - x\beta)Q^{-1}x^{\mathrm{T}}, \tag{4.6}$$

where Q was defined in (4.2). Note that b is *not* a bounded-influence estimator.

The next thing we will do is define the estimator's sensitivity, which we want to think of as the maximum possible influence (suitably normalized) of a single observation in a large sample. The most natural definition (and the one introduced by Hampel) is

$$\max_{y,x} \|\Omega(y, x)\|, \tag{4.7}$$

where $\|\cdot\|$ is the Euclidean norm. The problem with this definition is that it depends on the units of measurement of the explanatory variables. If we change the units in which the explanatory variables are measured, we trivially, but necessarily, redefine the parameters; and the new influence function will generally not have the same maximum as the original one.

Actually, we want more than invariance to the units of measurement. When we work with dummy variables, for example, there are always many equivalent formulations. We can obtain one from another by taking linear combinations of the dummy variables. The list of explanatory variables changes, but the p-dimensional subspace spanned by the explanatory variables stays the same. This suggests that the definition of an estimator's sensitivity should depend only on the p-dimensional subspace spanned by the explanatory variables and not on the particular choice of explanatory variables which appears in the regression.

We can gain some insight into a more reasonable definition of sensitivity by considering the change in the fitted values $\hat{y} = X\hat{\beta}$. The effect on $\hat{y}$ of a gross error (y, x) will be approximately $X\Omega(y, x)$. The norm of this quantity is

$$\|X\Omega(y,x)\| = \left\{\Omega(y,x)^{\mathrm{T}} X^{\mathrm{T}} X \Omega(y,x)\right\}^{1/2}. \tag{4.8}$$

When $\hat{\beta}$ is invariant (so that $\hat{y}$ depends only on the subspace spanned by the p explanatory variables), expression (4.8) will also be invariant.

While (4.8) provides invariance, it only considers the effects of the gross error (y, x) on the fitted value $x\hat{\beta}$. If we are interested in estimating what would happen for new observations on the explanatory variables x_* we would want to consider the effect of the gross error on the estimated value, $x_*\hat{\beta}$.

We will be concerned when the effect of the gross error, $x_*\Omega(y, x)$, is large relative to the standard error $(x_* V x_*^{\mathrm{T}})^{1/2}$ of $x_*\hat{\beta}$, where V denotes the asymptotic covariance matrix of $\hat{\beta}$ and $\Omega(y, x)$ is its influence function. These considerations lead us to consider

$$\max_{y,x} \frac{|x_*\Omega(y,x)|}{\left(x_* V x_*^{\mathrm{T}}\right)^{1/2}} \tag{4.9}$$

as our measure of sensitivity for the particular explanatory variable observations, x_*. However, we often do not know in advance what x_* will be, so we consider the worst possible case and use

$$\max_{y,x} \max_{x_*} \frac{|x_*\Omega(y,x)|}{\left(x_* V x_*^{\mathrm{T}}\right)^{1/2}} = \max_{y,x} \left\{\Omega^{\mathrm{T}}(y,x) V^{-1} \Omega(y,x)\right\}^{1/2} \equiv \gamma \tag{4.10}$$

as our definition of sensitivity. This definition of sensitivity is also invariant to the coordinate system.

An estimator $\hat{\beta}$ is called a bounded-influence estimator if its influence function, Ω, is a bounded function; or, equivalently, if its sensitivity, γ, is finite. The

bounded-influence property is obviously desirable when gross errors or other departures from the assumptions of the model are possible. In this section we will study weighted least-squares (WLS) estimators with the bounded-influence property.

Though OLS is usually expressed in matrix notation:

$$b = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, \tag{4.11}$$

it is more convenient for our purposes to use an earlier notation, the "normal equations":

$$0 = \sum_{i=1}^{n} (y_i - x_i b)x_i^{\mathrm{T}}. \tag{4.12}$$

A WLS estimator $\hat{\beta}$ is an estimator of the form

$$0 = \sum_{i=1}^{n} w_i \cdot (y_i - x_i\hat{\beta})x_i^{\mathrm{T}}. \tag{4.13}$$

(This could be expressed in matrix form as $\hat{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}WY$, where W is a diagonal matrix.) The weight w_i will depend on y_i, x_i, and $\hat{\beta}$ and will also depend on the estimated scale $\hat{\sigma}$ (see Section 6). The $w_i = w(y_i, x_i, \hat{\beta})$ will usually be equal to one, although certain observations may have to be downweighted if the estimator is to have the bounded-influence property.

One can show that under general conditions the influence function of a weighted least squares estimator is

$$\Omega(y, x) = w(y, x, \beta)(y - x\beta)B^{-1}x^{\mathrm{T}} \tag{4.14}$$

for a certain $p \times p$ matrix B, and the estimator's asymptotic covariance matrix will be

$$\begin{aligned} V &= \mathrm{E}\Omega\Omega^{\mathrm{T}} \\ &= \sigma^2 B^{-1}\left[Ew(y, x, \beta)^2\left(\frac{y - x\beta}{\sigma}\right)^2 x^{\mathrm{T}}x\right](B^{-1})^{\mathrm{T}} \\ &= \sigma^2 B^{-1}A(B^{-1})^{\mathrm{T}}, \end{aligned} \tag{4.15}$$

where the expectation is over the joint distribution of (y, x) and A is defined as

the $p \times p$ matrix in the square brackets. It follows that

$$\begin{aligned}\gamma &= \max_{y,x} \{\Omega(y,x)^{\mathrm{T}} V^{-1} \Omega(y,x)\}^{1/2} \\ &= \max_{y,x} \left\{ w(y,x,\beta)^2 (y - x\beta)^2 x (B^{-1})^{\mathrm{T}} \frac{1}{\sigma^2} B^{\mathrm{T}} A^{-1} B B^{-1} x^{\mathrm{T}} \right\}^{1/2} \\ &= \max_{y,x} w(y,x,\beta) \left| \frac{y - x\beta}{\sigma} \right| \{x A^{-1} x^{\mathrm{T}}\}^{1/2}. \end{aligned} \tag{4.16}$$

This is a good point at which to recapitulate. First of all, we adopted a definition of sensitivity which essentially reflects the maximum possible influence, on linear combinations of the estimator $\hat{\beta}$, of a single observation. Since ordinary least squares has infinite sensitivity, we considered the more general class of weighted least-squares (WLS) estimators. We then derived an expression for the sensitivity of an arbitrary WLS estimator, which has a nice interpretation. We see that, apart from the weights, the influence of (y, x) has two components. The first is the normalized residual $(y - x\beta)/\sigma$. The second is the quadratic expression $xA^{-1}x^{\mathrm{T}}$, which should be thought of as the square of a robust measure of the distance of x from the origin.

Suppose that we desire an estimator whose sensitivity γ is $\leqslant a$, where a is some positive number. One reasonable way to choose from among the various candidate estimators would be to find that estimator which is "as close as possible" to least squares, subject to the constraint $\gamma \leqslant a$. By this we mean that we will downweight an observation only if its influence would otherwise exceed the maximum allowable influence. An observation whose influence is below the maximum will be given a weight of one, as would *all* the observations under least squares. In this way we might hope to preserve much of the "central-model" efficiency of OLS, while at the same time protecting ourselves against gross errors. Formally, suppose we require $\gamma \leqslant a$ for $a > 0$. If, for a given observation (y_i, x_i), we have

$$\left| \frac{y_i - x_i \hat{\beta}}{\hat{\sigma}} \right| \{x_i A^{-1} x_i^{\mathrm{T}}\}^{1/2} \leqslant a, \tag{4.17}$$

then we want $w(y_i, x_i, \hat{\beta}) = 1$. Otherwise, we will downweight this observation just enough so that its influence equals the maximum allowable influence, i.e. we set $w(y_i, x_i, \hat{\beta})$ so that

$$w(y_i, x_i, \hat{\beta}) \left| \frac{y_i - x_i \hat{\beta}}{\hat{\sigma}} \right| \{x_i A^{-1} x_i^{\mathrm{T}}\}^{1/2} = a. \tag{4.18}$$

The weight function must therefore satisfy

$$w(y_i, x_i, \hat{\beta}) = \min\left\{1, \frac{a}{\left|\frac{y_i - x_i\hat{\beta}}{\hat{\sigma}}\right|\{x_i A^{-1} x_i^{\mathrm{T}}\}^{1/2}}\right\}. \tag{4.19}$$

Recall that under our "central model", the conditional distribution of $(y - x\beta)/\sigma$, given x, is $N(0,1)$. Let η denote a random variable whose distribution, given x, is $N(0,1)$. Plugging (4.19) into the expression for A, we find

$$\begin{aligned} A &= \mathrm{E}\min\left\{1, \frac{a}{\left|\frac{y - x\beta}{\sigma}\right|\{xA^{-1}x^{\mathrm{T}}\}^{1/2}}\right\}^2 \left(\frac{y - x\beta}{\sigma}\right)^2 x^{\mathrm{T}}x \\ &= \mathrm{E}_x\left[\mathrm{E}_{\eta|x}\min\left\{\eta^2, \frac{a^2}{xA^{-1}x^{\mathrm{T}}}\right\}\right] x^{\mathrm{T}}x \\ &= \mathrm{E}_x r\left(\frac{a}{(xA^{-1}x^{\mathrm{T}})^{1/2}}\right) x^{\mathrm{T}}x, \end{aligned} \tag{4.20}$$

where

$$r(t) = \mathrm{E}_{\eta|x}\min\{\eta^2, t^2\}. \tag{4.21}$$

One can show that the matrix A satisfying (4.20) will exist only if $a > \sqrt{p}$. This suggests the following estimator for β. Choose $a > \sqrt{p}$. Find A to satisfy

$$A = \frac{1}{n}\sum_{i=1}^{n} r\left(\frac{a}{(x_i A^{-1} x_i^{\mathrm{T}})^{1/2}}\right) x_i^{\mathrm{T}} x_i; \tag{4.22}$$

then find b^* to satisfy

$$0 = \sum_{i=1}^{n} \min\left\{1, \frac{a}{\left|\frac{y_i - x_i b^*}{\hat{\sigma}}\right|\{x_i A^{-1} x_i^{\mathrm{T}}\}^{1/2}}\right\}(y_i - x_i b^*) x_i^{\mathrm{T}}. \tag{4.23}$$

One can show that b^*, which is called the *Krasker–Welsch estimator*, has the following properties.

(1) b^* is consistent and asymptotically normal when the assumptions of the model hold.

(2) The sensitivity γ of b^* equals a.
(3) Among all weighted least-squares estimators for β with sensitivity $\leqslant a$, b^* satisfies a necessary condition for minimizing asymptotic variance (in the strong sense that its asymptotic covariance matrix differs from all others by a non-negative definite matrix).

To fully define this estimator we need to specify a. We know that $a > \sqrt{p}$, providing a lower bound. Clearly when $a = \infty$, the bounded-influence estimator reduces to least squares. In practice we want to choose the bound a so that the efficiency of BIF would not be too much lower than the least-squares efficiency if we had data ideal for the use of least squares. This usually means that X is taken as given and the error structure is normal. The relative efficiency then would be obtained by comparing the asymptotic variances $\sigma^2(X^TX)^{-1}$ and $\sigma^2 V(a)$ where $V(a) = n^{-1}\sigma^2 B^{-1}(a)A(a)B^{-1}(a)$.

There is no canonical way to compare two matrices. The trace, determinant, or largest eigenvalue could be used. For example, the relative efficiency could be defined as

$$e(a) = \left\{ \frac{\det\left[\sigma^2(X^TX)^{-1}\right]}{\det\left[\sigma^2 V(a)\right]} \right\}^{1/p} \tag{4.24}$$

and then a found so that $e(a)$ equals, say, 0.95. This means we would be paying about a 5 percent insurance premium by using BIF in ideal situations for least-squares estimation. In return, we obtain protection in non-ideal situations.

The computations involved in obtaining a for a given relative efficiency are complex. Two approximations are available. The first assumes that the X data comes from a spherically symmetric distribution which implies that asymptotically both the OLS covariance matrix and $V(a)$ will be diagonal, say $\alpha(a)I$. Then we need only compare $\sigma^2 I$ to $\sigma^2\alpha(a)I$ which means the relative efficiency is just $e(a) = \alpha^{-1}(a)$. This is much easier to work with than (4.25) but makes unrealistic assumptions about the distribution of X.

The simplest approach is to examine the estimator in the location case. Then $V(a)$ and X^TX are scalars. It is then possible to compute the relative efficiencies because the BIF estimator reduces to a simple form. When the a value for location is found, say a_L, we then approximate the bound, a, for higher dimensions by using $a = a_L\sqrt{p}$. Further details may be found in Krasker and Welsch (1982a) and Peters, Samarov and Welsch (1982).

We would now like to show briefly how the concepts of bounded-influence relate to the regression diagnostics of Section 3. Full details may be found in Welsch (1982). Consider again the "gross error model" introduced above. Assume that our "good" data are (x_k, y_k), $k \neq i$, and the suspected bad observation is

(x_i, y_i). Then we can show that the potential influence [what would happen if we decided to use (x_i, y_i)] of the ith observation on $b(i)$ is

$$\Omega(x_i, y_i, b(i)) = (n-1)\left[X^{\mathrm{T}}(i)X(i)\right]^{-1}x_i^{\mathrm{T}}(y_i - x_i b(i)) \tag{4.25}$$

$$= (n-1)(X^{\mathrm{T}}X)^{-1}x_i^{\mathrm{T}}(y_i - x_i b)/(1-h_i)^2. \tag{4.26}$$

Note the presence of the predicted residual (3.2) in (4.25).

The analog to V in (4.16) turns out to be

$$V(i) = (n-1)s^2(i)\left[X^{\mathrm{T}}(i)X(i)\right]^{-1}. \tag{4.27}$$

Therefore, the norm for our measure of sensitivity (4.11) is

$$\left[\Omega^{\mathrm{T}}(y_i, x_i, b(i))V^{-1}(i)\Omega(y_i, x_i, b(i))\right]^{1/2}$$

which, after some matrix algebra, is just

$$\left((n-1)\frac{h_i}{1-h_i}\frac{(y_i - x_i b(i))^2}{s^2(i)}\right)^{1/2} \tag{4.28}$$

or

$$\left((n-1)\frac{h_i}{1-h_i}\frac{(y_i - x_i b)^2}{s^2(i)(1-h_i)^2}\right)^{1/2}. \tag{4.29}$$

Comparing this to (3.3), we obtain that (4.29) is equivalent to

$$(n-1)^{1/2}|DFFITS_i|/(1-h_i)^{1/2}. \tag{4.30}$$

To bound the influence, we require that

$$\max_i (n-1)^{1/2}|DFFITS_i|/(1-h_i)^{1/2} \leqslant a,$$

which clearly implies that

$$(n-1)^{1/2}|DFFITS_i| \leqslant a.$$

The simple choice of a for BIF discussed above was $a_L\sqrt{p}$. For location,

$$DFFITS_i = \frac{n^{1/2}}{n-1}\frac{e_i}{s(i)},$$

and we might consider $(n-1)^{1/2}|DFFITS_i|$ large if it exceeded 2. Hence, a_L around 2 is good for diagnostic purposes.

Clearly (4.30) could have been chosen as our basic diagnostic tool. However, *DFFITS* has a natural interpretation in the context of least squares and therefore we feel it is easier to understand and to use.

5. Aspects of robust inference

When we estimate the coefficient vector β in the linear model $y_i = x_i\beta + u_i$, it is usually because we want to draw inferences about some aspect of the conditional distribution of y given x. In forecasting, for example, we need a probability distribution for the response variable, conditional on a particular x. Alternatively, we might want to know how the conditional expectation of the response variable varies with x.

In this section we analyze the problems that are created for inference by the fact that the linear model will never be exactly correct. To be sure, failures of linearity that occur for *extreme* values of the x-vector will always show up in a bounded-influence regression. However, *gradual* curvature over the entire range of X is much more difficult to detect. Moreover, departures from linearity in extreme regions of the x-space are sometimes very difficult to distinguish from aberrant data. Unfortunately, there are applications in which the distinction can be important.

To illustrate this point, consider the data plotted in Figure 5.1, and suppose that we are trying to predict y, conditional upon $x = 4$. Obviously, the outlier is crucial. If these were known to be good data from a linear model, then the outlier would be allowed to have a large effect on the prediction. On the other hand, if the outliers were known to be erroneous or inapplicable for some reason, one would base inferences about $(y|x = 4)$ on the remaining nine observations. The prediction for y would be substantially higher in the latter case; or, more precisely, the probability distribution would be centered at a larger value.

There is a third possibility: namely that the true regression line is slightly curved. With the data in Figure 5.1 even a small amount of curvature would make the outlier consistent with the rest of the sample. Were such curvature permitted, one would obtain a prediction for $(y|x = 4)$ lying between the two just mentioned.

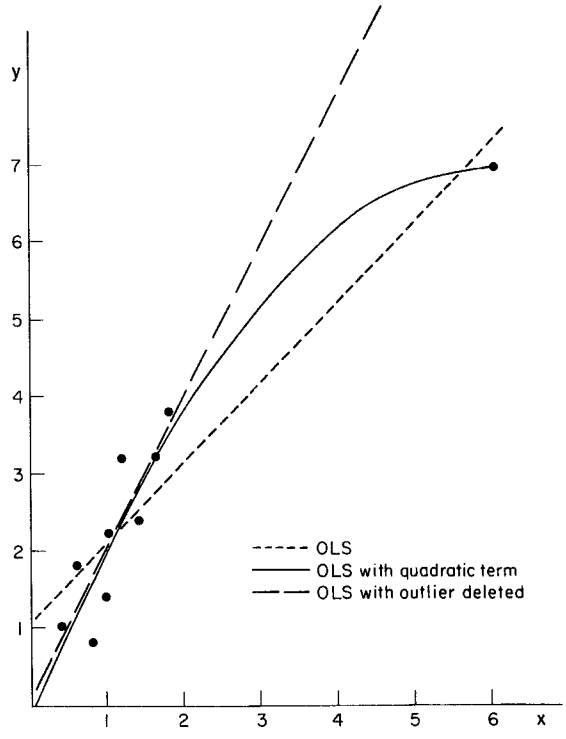

Figure 5.1

With data as simple as those in Figure 5.1, one could do a satisfactory job by computing three different fits and weighting them subjectively. The resulting probability distribution for y would be trimodal. However, this approach will not work with larger, multivariate regression problems. It seems that one has to use a model that builds-in the possibilities of bad data and curvature. As an illustration we will describe an approach proposed by Krasker (1981). Suppose that there are bad data occurring with probability ε, and that the good data are generated by a process satisfying

$$y_i = R(x_i, \theta) + u_i,$$

with $(u_i/\sigma | x_i, \theta, \sigma) \sim N(0, 1)$.

For a bad (y_i, x_i) observation, suppose that $(y_i | x_i)$ has a density $h(y_i | x_i, \alpha)$ for some parameter α. In order to apply this approach one has to make specific

choices for R and h. Krasker proposed a uniform distribution for h:

$$h(y|x,\alpha) \equiv \alpha \quad \text{for all } y.$$

Strictly speaking this is only a probability distribution when limited to an interval of length $1/\alpha$. The practical implication is that the parameter α ought to be small enough for all the y_i to lie in an interval of length $\leq 1/\alpha$. The uniform distribution reflects the notion that a gross error "could be anywhere".

For R, Krasker used

$$R(x,\beta,\gamma) = \beta_1 + \beta_2\left[\frac{\exp\{\gamma_2 x_2\}-1}{\gamma_2}\right] + \cdots + \beta_p\left[\frac{\exp\{\gamma_p x_p\}-1}{\gamma_p}\right].$$

This makes sense provided $\gamma_j \neq 0$ for all j. However.

$$\lim_{\gamma_j \to 0} \frac{\exp\{\gamma_j x_j\}-1}{\gamma_j} = x_j, \tag{5.1}$$

so that one can extend the definition of R even to $\gamma = 0$. Eq. (5.1) shows that when γ is small, $R(x,\beta,\gamma)$ is nearly linear in x.

Given x and the parameters, the density for y is the mixture

$$(1-\varepsilon)\frac{1}{\sigma}\phi\left(\frac{y - R(x,\beta,\gamma)}{\sigma}\right) + \varepsilon\alpha,$$

where ϕ is the density for $N(0,1)$. The likelihood function is

$$\prod_{i=1}^{n}\left\{(1-\varepsilon)\frac{1}{\sigma}\phi\left(\frac{y_i - R(x_i,\beta,\gamma)}{\sigma}\right) + \varepsilon\alpha\right\}.$$

The likelihood function will often be multimodal, for essentially the same reason that the subjective weighting of three fits for the data in figure 5.1 would lead to a trimodal probability distribution for y. Consequently, maximum likelihood is not adequate; one has to work with the entire likelihood function. Krasker's approach is Bayesian. Given a prior $p(\beta,\gamma,\sigma,\varepsilon,\alpha)$ for the parameters, one can find the posterior marginal distribution for any quantity of interest [such as some β_j or $R(x,\beta,\gamma)$] by numerical integration.

In most problems the priors on β and σ would be relatively diffuse. However, the prior information on the other parameters is sometimes crucial. This is particularly true for γ, which in a sense represents the amount of curvature in the regression surface. If, as is usually the case, the variables were transformed beforehand in order to make a linear model plausible, then one would choose a prior under which γ is near zero with high probability.

6. Historical background

In this section we provide a brief overview of the history and structure of bounded-influence estimation. For more background on classical robust estimation see Huber (1981) and Barnett and Lewis (1978, ch. 4). Koenker (1982) provides a good survey of robust procedures.

Huber (1973) proposed a form of robust regression based on a direct generalization of his work on the robust estimation of location parameters. His approach was to define

$$\rho_c(t) = \begin{cases} t^2/2 & |t| < c, \\ c|t| - c^2/2 & |t| \geq c, \end{cases} \tag{6.1}$$

and then minimize

$$d\sigma + \sum_{i=1}^{n} \sigma \rho_c[(y_i - x_i\beta)/\sigma] \tag{6.2}$$

with respect to β and σ. (The constant d is used to make the scale estimate consistent.) The influence function is

$$\Omega(y, x) = \psi_c[(y - x\beta)/\sigma] B^{-1} x^{\mathrm{T}}, \tag{6.3}$$

where $\psi_c(t) = \rho_c'(t)$ and B is a certain $p \times p$ matrix. Even though $\psi_c(\cdot)$ limits the effect of large residuals, the influence of (y, x) can be arbitrarily large because x^{T} multiplies $\psi_c(\cdot)$. This form of robust regression should be used with caution if there is a potential for outlying observations in the x data. Huber, and especially Hampel (1973), also stress this point.

Many other criterion functions like (6.1) can be considered [Holland and Welsch (1977)]. Those that have non-monotone $\psi(\cdot)$ functions are of special interest in regression because these functions are often zero for large residuals and hence remain zero when multiplied by x^{T}. However, they will not be bounded-influence estimators in all regions of the x-space and, because of the possibility of multiple solutions to (6.2), need a bounded-influence start to be effective.

Mallows (1973, 1975) proposed a way to construct bounded-influence estimators by, in essence, modifying (6.2) to read

$$d\sigma + \sum_{i=1}^{n} \sigma u(x_i) \rho_c[(y_i - x_i\beta)/\sigma] \tag{6.4}$$

for certain weights $u(x_i)$ which may depend on the entire X-matrix and not just

x_i. The influence function

$$\Omega(y, x) = u(x)\psi_c[(y - x\beta)/\sigma]B^{-1}x^{\mathrm{T}}, \tag{6.5}$$

where the B-matrix is not the same as in (6.3). If u is appropriately chosen, then Ω will be bounded. Some optimality results for this form are contained in Maronna, Bustos and Yohai (1979).

There is one problem that is immediately apparent. Outlying points in the X-space increase the efficiency of most estimation procedures. Any downweighting in X-space that does not include some consideration of how the y-values at these outlying observations fit the pattern set by the bulk of the data cannot be efficient.

In 1975, Schweppe [Handschin et al. (1975)] proposed essentially the form

$$d\sigma + \sum_{i=1}^{n} \sigma v^2(x_i)\rho_c[(y_i - x_i\beta_n)/\sigma v(x_i)], \tag{6.6}$$

with $v(x_i) = (1 - h_i)^{1/2}$ and $h_i = x_i(X^tX)^{-1}x_i^{\mathrm{T}}$. Again, (6.6) can provide bounded influence but with the additional property that if $(y - x\beta)/\sigma v(x)$ is small, the effect of $v(x)$ will be cancelled out. This has the potential to help overcome some of the efficiency problems outlined for the Mallows approach. Hill (1977) compared, via Monte Carlo, the Mallows and Schweppe forms along with several others and found that these two dominate, with the Schweppe form having an advantage.

Welsch (1977) tried a more direct approach to overcoming these efficiency problems. If the ith observation is an outlier in X-space (perhaps indicated by a large value of h_i), but (y_i, x_i) is consistent with the fit obtained from the rest of the data, $DFFITS_i$ [see (3.3)] would not be unduly large. Thus, Welsch proposed solving

$$\sum_{i=1}^{n} w_i(y_i - x_i\beta)x_i^{\mathrm{T}} = 0,$$

where $w_i = w(DFFITS_i)$ and $w(\cdot)$ is a weight function such as

$$w_c(t) = \psi_c(t)/t. \tag{6.7}$$

This is just one iteratively-reweighted least-squares step of (6.6) with $v(x_i) = (1 - h_i)/h_i^{1/2}$. Hinkley (1977), motivated by the jackknife, has proposed a similar class of estimators.

Without some additional criteria there is no way to choose among these approaches. A natural criterion, suggested by Hampel (1968), is to minimize the asymptotic variance subject to a bound on the influence function.

In the single parameter case this problem was solved by Hampel (1968). The multiparameter case was first considered in 1976 in several talks at Bell Laboratories by Hampel [see Hampel (1978) for further discussion] and at about the same time by Krasker during his doctoral research at M.I.T. [Krasker (1978, 1980)].

Krasker examined the very general class of estimators

$$0 = \sum_{i=1}^{n} \phi(y_i, x_i, \beta_n) \tag{6.8}$$

for some function ϕ: $\mathbf{R} \times \mathbf{R}^p \times \mathbf{R}^p \to \mathbf{R}^p$ and showed that the estimator defined by

$$0 = \sum_{i=1}^{n} \frac{\psi_c(|(y_i - x_i\beta_n)B^{-1}x_i^{\mathrm{T}}|)}{|(y_i - x_i\beta_n)B^{-1}x_i^{\mathrm{T}}|}(y_i - x_i\beta_n)x_i^{\mathrm{T}}, \tag{6.9}$$

with the side condition

$$B = \frac{-\partial}{\partial\beta} E \frac{\psi_c(|(y - x\beta)B^{-1}x^{\mathrm{T}}|)}{|(y - x\beta)B^{-1}x^{\mathrm{T}}|}(y - x\beta)x^{\mathrm{T}},$$

minimizes the trace of the asymptotic covariance matrix when a bound is placed on γ.

In Section 4 we restricted attention to estimators of the form

$$0 = \sum_{i=1}^{n} w(y_i, x_i, \beta_n)(y_i - x_i\beta_n)x_i^{\mathrm{T}}, \tag{6.10}$$

where β_n is the estimate of β and the weight function w is non-negative, bounded, and continuous.

7. Bounded-influence estimates for a hedonic price index

Our example is drawn from a study by Harrison and Rubinfeld (1978), in which a hedonic price index for housing is estimated for use in a subsequent calculation of the marginal-willingness-to-pay for clean air.[4] Hedonic price indexes were introduced into recent econometrics literature by Griliches (1968). In essence, a hedonic price index is obtained from the fitted values in a regression where price is the response variable and the explanatory variables represent its qualitative

[4]Section 4.4 of Belsley, Kuh and Welsch (1980) provides a description of the Harrison–Rubinfeld problem as well as detailed regression diagnostics for it.

Table 7.1
Definition of model variables

Symbol	Definition
LMV	logarithm of the median value of owner-occupied homes
CRIM	per capita crime rate by town
ZN	proportion of a town's residential land zoned for lots greater than 25 000 square feet
INDUS	proportion of nonretail business acres per town
CHAS	Charles River dummy variable with value 1 if tract bounds on the Charles River
NOXSQ	nitrogen oxide concentration (parts per hundred million) squared
RM	average number of rooms squared
AGE	proportion of owner-occupied units built prior to 1940
DIS	logarithm of the weighted distances to five employment centers in the Boston region
RAD	logarithm of index of accessibility to radial highways
TAX	full-value property-tax rate (per $10 000)
PTRATIO	pupil–teacher ratio by town
B	$(Bk - 0.63)^2$ where Bk is the proportion of blacks in the population
LSTAT	logarithm of the proportion of the population that is lower status

Table 7.2
OLS estimates: Housing-price equation

Variable	Coefficient estimate	Standard error	*t*-Statistic
INTERCEPT	9.758	0.150	65.23
CRIM	−0.0119	0.00124	−9.53
ZN	7.94×10^{-5}	5.06×10^{-4}	0.16
INDUS	2.36×10^{-4}	2.36×10^{-3}	0.10
CHAS	0.0914	0.0332	2.75
NOXSQ	−0.00639	0.00113	−5.64
RM	0.00633	0.00131	4.82
AGE	8.86×10^{-5}	5.26×10^{-4}	0.17
DIS	−0.191	0.0334	−5.73
RAD	0.0957	0.0191	5.00
TAX	-4.20×10^{-4}	1.23×10^{-4}	−3.42
PTRATIO	−0.0311	0.00501	−6.21
B	0.364	0.103	3.53
LSTAT	−0.371	0.0250	−14.83
$R^2 = 0.806$	$SER = 0.182$		

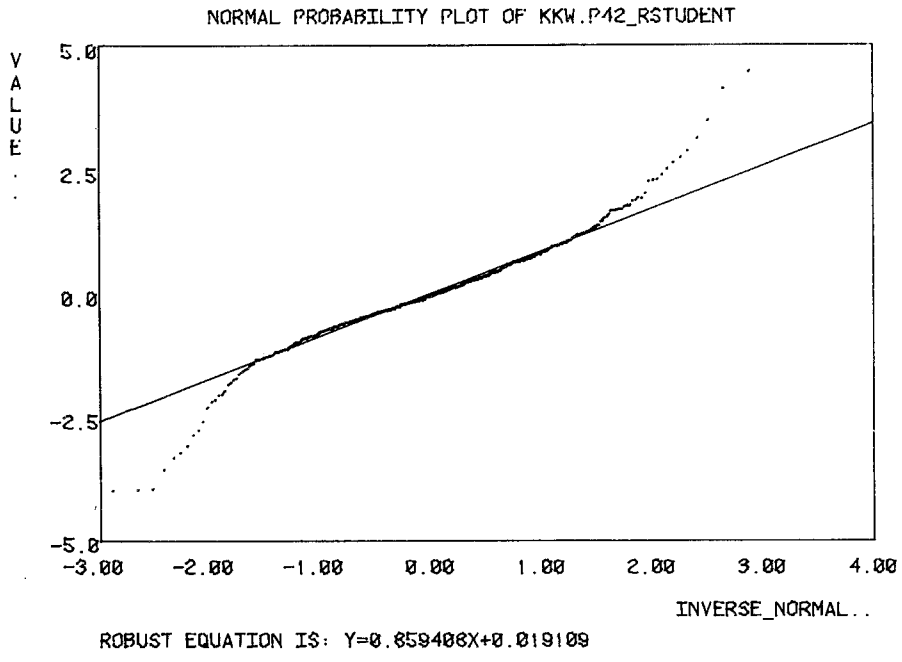

Figure 7.1. Normal probability plot for studentized residuals from OLS estimates; housing-price equation.

determinants. Harrison and Rubinfeld are principally interested in examining the impact of air pollution (as measured by the square of nitrogen oxide concentration, NOXSQ) on the price of owner-occupied homes. Thus, their hedonic price equation includes NOXSQ and thirteen other explanatory variables as indicators of qualities that affect the price of houses.

The basic data are a sample of 506 observations on census tracts in the Boston Standard Metropolitan Statistical Area (SMSA) in 1970[5] and the variables used are defined in Table 7.1. This study includes many socio-demographic variables at a relatively disaggregated level in common with many other current cross-section studies.

Table 7.2 reports least-squares estimates of eq. (7.1). The normal probability plot of studentized residuals in Figure 7.1 provides an instructive first diagnostic step. Substantial departures from normality are clearly evident, since large residuals are disproportionately present. Thus, at least for these observations, the potential exists for strongly influential observations.

[5]The original data together with a list of the census tracts appear in Belsley, Kuh and Welsch (1980).

7.1. The model

The hedonic housing-price model used by Harrison and Rubinfeld is

$$LMV = \beta_1 + \beta_2 CRIM + \beta_3 ZN + \beta_4 INDUS + \beta_5 CHAS + \beta_6 NOXSQ + \beta_7 RM + \beta_8 AGE + \beta_9 DIS + \beta_{10} RAD + \beta_{11} TAX + \beta_{12} PTRATIO + \beta_{13} B + \beta_{14} LSTAT + \varepsilon. \tag{7.1}$$

A brief description of each variable is given in Table 7.1. Further details may be found in Harrison and Rubinfeld (1978).

7.2. Partial plots

Two partial-regression leverage plots (see the end of Section 3 for their description) reveal helpful information of both positive and negative import. Figure 7.2 for *NOXSQ*, a variable of major concern, reveals a scatter which is not obviously

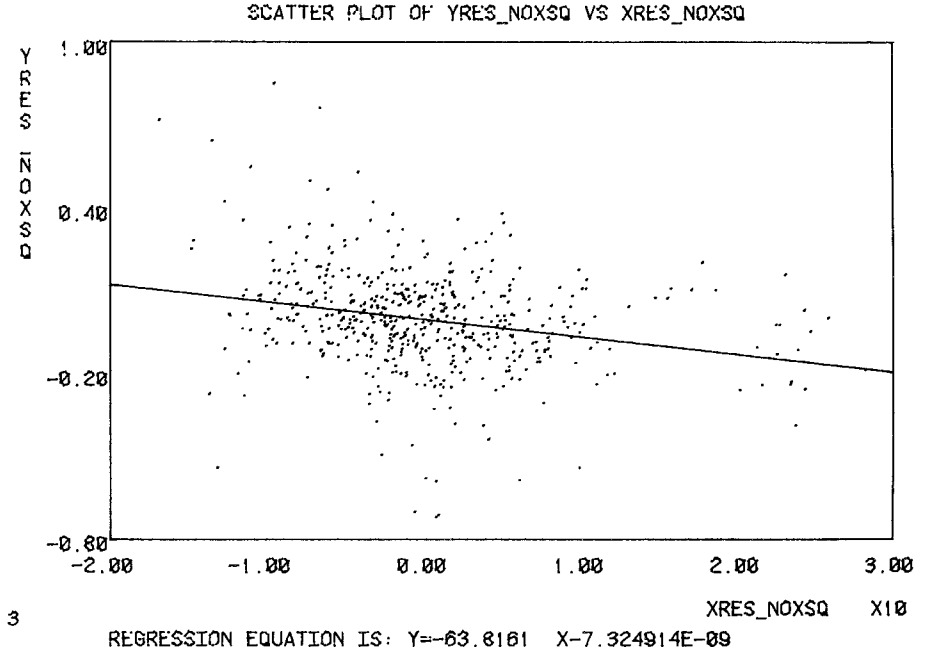

Figure 7.2. Partial-regression leverage plot for b_6 (*NOXSQ*), SE = 0.00113; housing-price equation.

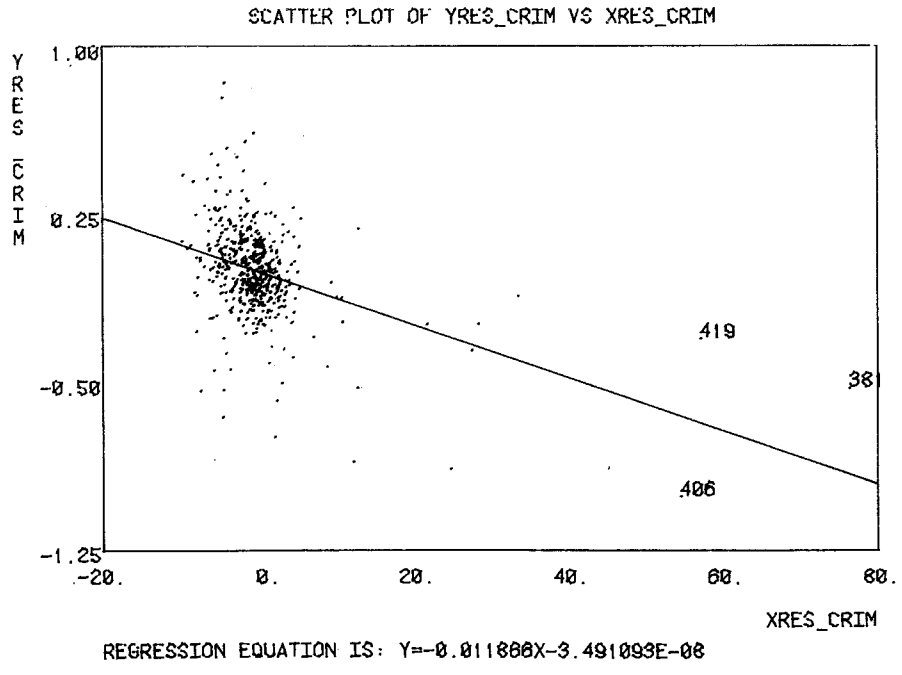

Figure 7.3. Partial-regression leverage plot for b_2 (*CRIM*), SE = 0.00124; housing-price equation.

dominated by extreme points. *CRIM*, shown in Figure 7.3 is another story. Three points near the center of Boston – 381 (Charlestown), 406 (South Boston), and 419 (Roxbury) – dominate this partial plot in terms of leverage.[6] Further investigation with deletion diagnostics, and then bounded-influence estimation, is clearly worthwhile. The willingness to pay for clean air function, W, mentioned in the first paragraph of this section, includes terms in both the predicted response variable and *NOXSQ*: $W_i = e^{\hat{y}_i}(-2b_6 NOX_i)$. We therefore ought to pay especially close attention to *DFFITS* and *DFBETAS*.

Table 7.3 contains row-deletion diagnostic information on 30 census tracts that exceeded stringent cutoffs for any of the following measures: hat-matrix diagonals, studentized residuals, *DFFITS*, or the *DFBETAS* for the two explanatory variables *NOXSQ* and *CRIM*.[7] *NOXSQ* is of special interest because its presence along with $\hat{y}$ (in this instance LMV) in the equation serves an important purpose

[6] The remaining plots, which were omitted to conserve space, do not portray strikingly influential observations comparable to that for *CRIM*.

[7] The cutoffs are more severe than those in Belsley, Kuh and Welsch (1980) since our objective here is to locate the most extreme influential points to compare them with related diagnostic information from bounded-influence estimation.

Table 7.3
Hat-matrix diagonals, studentized residuals, *DFFITS*, and *DFBETAS* for selected census tracts: housing-price equation[a]

Census tract	h_i	*RSTUDENT*	*DFFITS*	*DFBETAS*	
				NOXSQ	*CRIM*
149	0.0485	1.773	0.4006	0.2609*	−0.0144
151	0.0456	0.980	0.2142	0.1612*	0.0026
153	0.0737	−1.312	−0.3701	−0.2013*	−0.0028
156	0.0840*	−0.427	−0.1292	−0.0650	0.0007
215	0.0579	2.910*	0.7214*	−0.0776	−0.0751
359	0.0520	1.250	0.2930	0.1364*	−0.0056
365	0.0891*	−2.747*	−0.8590*	−0.1780*	0.0649
366	0.0773	1.960	0.5671*	0.0549	−0.0875
368	0.0607	2.764*	0.7026*	−0.1952*	−0.0468
369	0.0982*	2.663*	0.8791*	−0.1256	−0.1105
372	0.0242	4.512*	0.7110*	−0.2709*	−0.1293
373	0.0532	4.160*	0.9856*	−0.1751*	−0.1264
381	0.2949*	2.559*	1.6551*	0.0975	1.5914*
386	0.0183	−2.564*	−0.3499	0.0113	−0.0567
398	0.0125	−3.212*	−0.3617	−0.0025	0.0928
399	0.0458	−3.301*	−0.7235*	−0.0168	−0.5774*
400	0.0213	−3.936*	−0.5808*	0.0132	0.1176
401	0.0225	−3.954*	−0.5995*	−0.0242	−0.3424*
402	0.0141	−3.988*	0.4766	−0.0204	−0.0636
406	0.1533*	2.141	−0.9112*	−0.0552	−0.8699*
410	0.0252	3.162*	0.5079*	−0.2685*	−0.0308
411	0.1116*	1.688	0.5983*	−0.1278	0.4130*
413	0.0477	3.520*	0.7878*	−0.3782*	−0.0060
414	0.0307	1.947	0.3470	−0.1554*	0.1783*
417	0.0387	−2.852*	−0.5724*	0.0473	0.1098
419	0.1843*	2.316	1.1009*	0.0352	1.0041*
427	0.0410	−1.956	−0.4012	0.1416*	0.0582
490	0.0514	−3.534*	−0.8225*	0.2957*	0.1797*
491	0.0527	−2.019	−0.4763	0.1760*	0.1107
506	0.0357	−3.070*	−0.5906*	−0.1193	−0.0547

[a]Starred values which exceed cutoff values: $h_i = 3(p/n) = 0.083$; $RSTUDENT = 2.5$; $DFFITS = 3(\sqrt{p/n}) = 0.50$; $DFBETAS = 3/\sqrt{n} = 0.133$.

of the authors, while *CRIM* is affected by some quite exceptional leverage points. The most influential point is clearly census tract 381 with a large hat-matrix diagonal and studentized residual. Its deletion alone causes $\hat{y}$ to shift by 1.66 standard deviations according to *DFFITS*.

Furthermore, the explanatory variables show severe imbalance in several instances. Since $p/n = 0.0275$ in the perfectly balanced case, the largest h_i of 0.2949 is more than ten times greater. Some other h_i are also severely out of balance. An excessive (relative to the normal distribution) number of large studentized residuals, first observed in the normal probability plot, are also apparent. This is a

Table 7.4
Distributions of diagnostic information from bounded-influence estimation of hedonic price index – centered and scaled robust distance based on A-matrix for X data

	99 percent EM	95 percent EM
13–14	0	1
12–13	0	0
11–12	1	0
10–11	0	1
9–10	0	1
8–9	0	0
7–8	2	0
6–7	0	1
5–6	1	0
4–5	0	1
3–4	4	6
2–3	18	16
1–2	129	128
0–1	351	351

situation where bounded-influence regression seems a natural alternative to least squares – initially as a diagnostic procedure and perhaps later to provide better regression estimates.

Two sets of bounded-influence estimates are presented next, for approximately 99 and 95 percent efficiency at the normal.[8] In the latter case, the limited influence allowed a given row of data would "cost" 5 percent efficiency if the normal error model prevails and leverage points correspond to completely valid elements of the model. The sensitivity bounds were $a_{0.99} = 12.0$ and $a_{0.95} = 8.0$, respectively. The 95 percent efficient model (hereafter 95 percent EM) will downweight a larger number of observations more heavily than the 99 percent EM, but if there are only a few very large influential points, even the less stringent 99 percent EM estimates will differ substantially from the OLS estimates.

We first examine the diagnostic content of the bounded-influence estimates. Table 7.4 describes the frequency distribution of robust distances. These are centered by the median distance and divided by 1.48 times the median absolute deviation from the median,[9] since the robust distances alone have no direct

[8]The TROLL program BIFMOD, written by Stephen Peters and Alexander Samarov, took a total of 19.5 CPU seconds on an IBM 370/168 for the more computer-intensive 95 percent efficiency model, at a cost well under $10. This total breaks down as follows: 1.9 CPU seconds for program loading and the OLS Initial start; 10.4 CPU seconds for 10 iterations to estimate the A-matrix; 7.2 CPU seconds for 54 iterations to estimate coefficients, scale, asymptotic covariance matrix and, of course, the weights. More details on the program and methods for computing efficiency may be found in Peters, Samarov and Welsch (1981).

[9]The 1.48 insures a consistent scale estimate for Gaussian data.

interpretation. Restricting comment to the 95 percent EM case we note that 479 observations, or 95 percent of the observations in all, are less than two robust standard errors from the median distance. Five observations are greatly distant from the median value of the robust distance; all of these correspond to large hat-matrix diagonals in the regression diagnostics with the exception of 415 which has a hat matrix diagonal of 0.067.

Table 7.5 presents individual diagnostic information for those observations with a final BIF weight less than one. Comment will be limited to the 95 percent EM and to a brief comparison with the regression diagnostics. First, the rank orderings of the largest hat-matrix diagonals and standardized robust distances are similar, but 415, 405, and 428 are now brought to our attention more forcefully. Second, while the two most influential bounded-influence observations (nos. 381 and 419) are also most influential according to *DFFITS* and the rank orderings are roughly alike, there are a few striking differences. The third most influential for BIF (no. 411) ranks 14th according to *DFFITS* and the third largest for *DFFITS* (no. 406) ranks 28th for BIF. Note that 406 has the third largest robust distance. From the structure of the bounded influence algorithm, the residuals associated with the BIF estimates must be extremely small for downweighting to be small in the presence of large leverage. It also appears that observations 370 and 415 were missed by the single row deletion diagnostics. This is especially true for 415 which does not appear in the first 100 rank ordered *DFFITS*.[10]

Regression coefficients for the bounded-influence estimates, their estimated standard errors, and t statistics (for the null hypothesis H_0: $\beta = 0$) appear in Table 7.6. The standard errors come from an asymptotic distribution, analogous to the way one finds standard errors for two-stage least squares or any non-linear estimator, or even for OLS when the disturbances are not exactly normal. Since the convergence to the asymptotic distribution can be slow when there are high-leverage X-rows, one must interpret these standard errors with some care. Of course, this comment applies also to OLS.

With a perfect model (including normal disturbances) one would expect the bounded-influence standard errors to exceed those of OLS, because bounded-influence will downweight some of the high-leverage observations which give OLS its central-model efficiency. However, bounded-influence can be *more* efficient than OLS if the disturbance distribution is heavy-tailed. We note that all coefficients change monotonically in the progression from least to most severe bounding of influence, i.e. from OLS to 99 percent EM to 95 percent EM. There is no certainty that monotonicity holds more generally. The coefficient of the key

[10]Multiple deletion methods, however, did turn up no. 415 as a potentially masked point [see Belsley, Kuh and Welsch (1980, p. 242)] thus confirming our expectation that BIF diagnostics offer effective alternatives to multiple deletion procedures.

Table 7.5
Bounded-influence diagnostic information about the most influential data tows: hedonic price index

99 percent EM				95 percent EM			
Standardized robust distance		Final Weights		Standardized robust distance		Final Weights	
Index	Value	Index	Value	Index	Value	Index	Value
381	11.14	381	0.231	381	13.68	381	0.086
419	7.98	419	0.301	419	10.07	419	0.103
406	7.10	373	0.489	406	9.02	411	0.165
411	5.01	411	0.511	411	6.28	369	0.208
369	3.68	369	0.517	415	4.42	373	0.228
365	3.30	365	0.558	369	3.73	365	0.252
415	3.26	413	0.579	405	3.57	368	0.254
156	3.12	490	0.591	428	3.36	413	0.264
343	2.99	368	0.618	365	3.33	490	0.298
366	2.90	399	0.670	156	3.30	366	0.307
163	2.84	372	0.673	399	3.13	399	0.317
153	2.72	215	0.736	163	2.97	372	0.334
371	2.70	401	0.771	366	2.97	401	0.360
284	2.49	366	0.777	153	2.94	370	0.397
405	2.47	400	0.836	343	2.89	414	0.399
428	2.47	406	0.853	371	2.84	415	0.403
164	2.46	506	0.856	143	2.65	400	0.433
162	2.45	417	0.865	164	2.57	215	0.437
143	2.44	410	0.940	284	2.56	417	0.440
157	2.42	370	0.978	370	2.56	506	0.464
370	2.39	491	0.989	157	2.55	410	0.481
				155	2.51	491	0.488
				162	2.45	402	0.520
				368	2.30	420	0.536
				161	2.06	371	0.548
				127	2.05	408	0.599
				124	2.00	467	0.635
				160	1.99	406	0.638
				373	1.94	375	0.661
				146	1.90	416	0.667
				258	1.89	8	0.673
				215	1.89	343	0.696
				147	1.88	398	0.725
				359	1.87	386	0.742
				148	1.82	149	0.748
				121	1.82	428	0.750
				123	1.81	367	0.786
				145	1.80	153	0.804
				125	1.80	427	0.829
				126	1.78	404	0.880
				319	1.77	182	0.896
				152	1.77	359	0.903
				122	1.76	412	0.923
				358	1.75	388	0.957

Table 7.6
Hedonic price index regression estimates

Estimated coefficients				Coefficient standard errors and t-statistics					
						Bounded-influence			
	Bounded-influence			OLS		99 percent EM		95 percent EM	
OLS	99 percent EM	95 percent EM	RHS variable	SE(b)	t-stat	SE(b)	t-stat	SE(b)	t-stat
-0.0119	-0.0143	-0.0158	*CRIM*	1.24×10^{-3}	-9.53	4.33×10^{-3}	-3.31	4.34×10^{-3}	-3.65
7.94×10^{-5}	7.52×10^{-5}	-2.39×10^{-5}	*ZN*	5.06×10^{-4}	0.16	3.63×10^{-4}	0.21	3.26×10^{-4}	-0.07
2.36×10^{-4}	3.98×10^{-4}	7.25×10^{-4}	*INDUS*	2.36×10^{-3}	0.10	1.68×10^{-3}	0.24	1.50×10^{-3}	0.48
0.0914	0.0863	0.0768	*CHAS*	0.0332	2.75	0.0301	2.87	0.0251	3.06
-6.39×10^{-3}	-5.86×10^{-3}	-4.84×10^{-3}	*NOXSQ*	1.13×10^{-3}	-5.64	1.18×10^{-3}	-4.97	1.04×10^{-3}	-4.67
6.33×10^{-3}	7.87×10^{-3}	0.0110	*RM*	1.31×10^{-3}	4.82	2.22×10^{-3}	3.55	1.67×10^{-3}	6.57
8.86×10^{-5}	-1.26×10^{-4}	-6.84×10^{-4}	*AGE*	5.26×10^{-4}	0.17	5.87×10^{-4}	-0.22	4.53×10^{-4}	-1.51
-0.191	-0.182	-0.165	*DIS*	0.0334	-5.73	0.0381	-4.78	0.0316	-5.21
0.0957	0.0922	0.0785	*RAD*	0.0191	5.00	0.0187	4.93	0.0152	5.15
-4.20×10^{-4}	-3.76×10^{-4}	-3.25×10^{-4}	*TAX*	1.23×10^{-4}	-3.42	1.14×10^{-4}	-3.30	9.56×10^{-5}	-3.40
-0.0311	-0.0305	-0.0290	*PTRATIO*	5.01×10^{-3}	-6.21	3.76×10^{-3}	-8.10	3.22×10^{-3}	-9.01
0.364	0.423	0.532	*B*	0.103	3.53	0.146	2.90	0.127	4.18
-0.371	-0.341	-0.284	*LSTAT*	0.0250	-14.83	0.0422	-8.09	0.0319	-8.91
9.76	9.71	9.64	*CONST*	0.150	65.23	0.156	62.35	0.132	73.18

variable *NOXSQ* follows the sequence −63.9; −58.6; −48.4 – differences that are large enough to cause big changes in the hedonic price calculations. However, since the fitted value also enters the willingness-to-pay function and other coefficients change, possibly in offsetting ways, a more detailed analysis would be needed to assess how much the bounded-influence calculations would ultimately affect the Harrison–Rubinfeld analysis.

Table 7.7

A. Difference between OLS and bounded-influence scaled by average of OLS and bounded-influence coefficient standard errors

RHS variable[a]	$\frac{b_{OLS} - b_{99}}{\frac{1}{2}(SE(b)_{OLS} + SE(b)_{99})}$	$\frac{b_{OLS} - b_{95}}{\frac{1}{2}(SE(b)_{OLS} + SE(b)_{95})}$
**CRIM*	0.88	1.42
ZN	0.01	0.25
INDUS	0.08	0.25
CHAS	0.16	0.50
**NOXSQ*	0.46	1.43
**RM*	0.87	3.12
**AGE*	0.39	1.58
DIS	0.27	0.81
**RAD*	0.19	1.00
TAX	0.37	0.87
PTRATIO	0.15	0.52
**B*	0.48	1.46
**LSTAT*	0.89	3.05
CONST	0.32	0.86

B. Percentage difference between OLS and bounded-influence estimates

RHS variable[a]	$\frac{b_{OLS} - b_{99}}{b_{OLS}}$	$\frac{b_{OLS} - b_{95}}{b_{OLS}}$
**CRIM*	21%	33%
ZN	5	130
INDUS	69	207
CHAS	6	16
**NOXSQ*	8	24
**RM*	24	74
AGE	242	872
DIS	5	14
**RAD*	4	18
TAX	10	23
PTRATIO	2	7
**B*	16	46
**LSTAT*	8	23
CONST	0	1

[a]Values are starred when magnitude exceeds 1 for OLS−95 percent EM differences, as explained in the text.

OLS and BIF estimates are compared in two ways in Table 7.7. First, Part A of Table 7.7 has their differences scaled by the average of the OLS and BIF coefficient standard errors. Percent differences, measured as the coefficient difference divided by the OLS coefficient, are shown in Part B of Table 7.7. Seven coefficients, including that for *NOXSQ*, change by one or more standard deviations using the 95 percent EM results; these changes have been starred. Two coefficients for the 95 percent EM model differ by more than three standard deviations. Next, the percent differences in Part B of Table 7.7 show that BIF estimation makes a large quantitative (as well as statistical) difference in the estimates, including those coefficients with the statistically most significant differences between them.

We noted above that $\hat{y}$ is an important part of the willingness-to-pay analysis based on the hedonic price index. More generally, the investigator is interested in comparing *ex-post* predictions among different procedures. As a first step we look at an index plot in Figure 7.4 of the point-by-point predicted value ratio, i.e. the $\hat{y}_i$ for OLS divided by $\hat{y}_i$ for the BIF 95 percent EM *after* conversion from logarithms into original units of dollar per dwelling unit. There are sizable numbers of large relative differences: the combined impact of coefficient changes

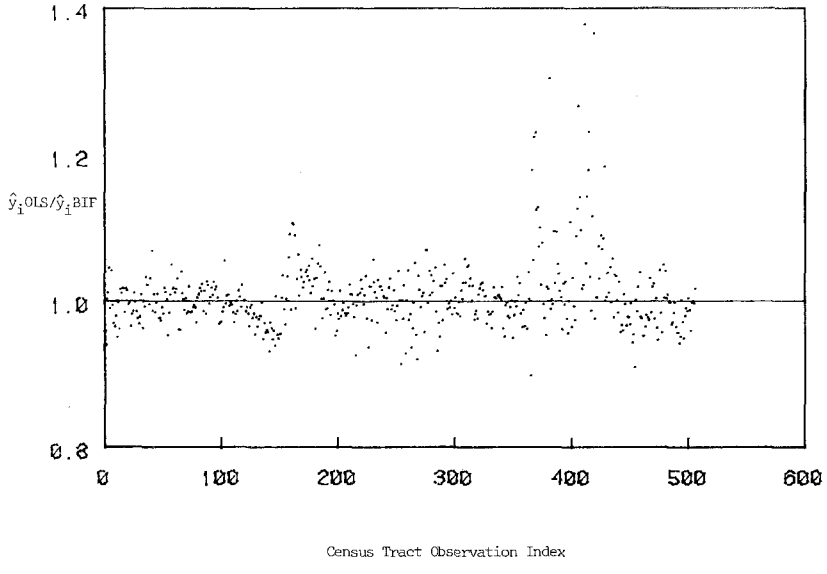

Figure 7.4. Ratio of OLS $\hat{y}_i$ to BIF $\hat{y}_i$: Census Tract Index for hedonic price equation.

sometimes *does* make a big difference to the predicted outcome. We also not that the largest differences occur when the BIF prediction is smaller than the OLS prediction which shows up in ratios well above unity; a number of predictions differ by more than 25 percent. Large differences are concentrated in the vicinity of Census Tracts 350–450. This points toward geographic concentration of large OLS–BIF differences, a fact on which we elaborate next in a related context.

A primary objective of bounding influence, we recall, is to limit bias, which tends to dominate variance in large samples. Upon noting that differences between OLS and BIF predictions are sometimes uncomfortably large, we ought to look for indications of systematic divergences between OLS and BIF estimates indicative of potential bias. This is readily achieved in the present instance by the scatter diagram in Figure 7.5 of BIF $\hat{y}_i$ against OLS $\hat{y}_i$ in logarithms that were used in the estimating equations. If the scatter is approximately uniformly distributed around a 45° line, bias can be treated as unimportant for practical purposes.

It is striking to observe in Figure 7.5 that the eight largest divergences all lie below the 45° line. It is furthermore clear that all but one of the most extreme points are below the average value of the response variable. Both effects are strongly suggestive of distortions or systematic differences indicative of potential bias in the OLS estimates. All the extreme points are in the center city, Boston.

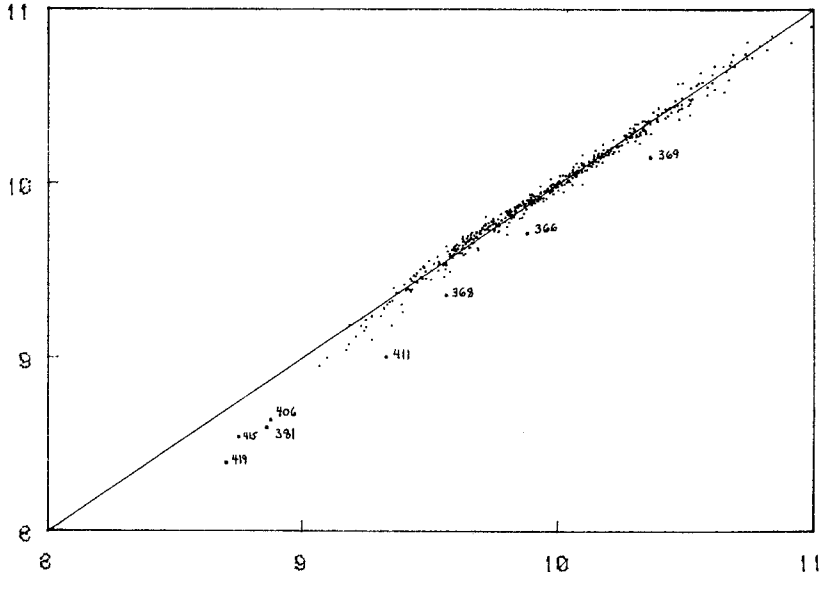

Figure 7.5. Scatter diagram of BIF $\hat{y}_i$ versus OLS $\hat{y}_i$: hedonic price equation.(*Note*: Both magnitudes are in natural logarithms.)

Observations 366, 368, and 369 are in Back Bay; Census Tract 381 is in Charlestown; 406 is in South Boston, 411 is in South Bay; 415 and 419 are in Roxbury. The close geographical proximity of these tracts is consistent with the hypothesis that different specifications hold for inner and outer city areas.[11]

We must emphasize that downweighted observations are not necessarily wrong, although some of them might be. Moreover, it is only under certain stringent assumptions that one could say that BIF is "better" than OLS. From the analysis done so far one can conclude only that certain observations seem to depart from the stochastic pattern of the bulk of the sample; moreover, those observations seem to be concentrated in the central city. Whether one should modify the model, or analyze the center-city observations separately, or simply assume that the anomalies are sampling error, is a question that cannot be decided by estimators alone. However, we feel it is highly advantageous to have an alternative fit that at least *reveals* the anomalies and the impact on estimated coefficients when their influence has been diminished.

How one chooses to treat influential observations, including the benign neglect of OLS, makes a sizable difference in this particular instance. Since many data sets that econometricians now prefer have related characteristics, this issue needs to be confronted.

8. Bounded-influence estimation with endogenous explanatory variables

So far we have dealt solely with cases in which, under the assumptions of the "central model", the conditional distribution of y_i given x_i is $N(x_i\beta, \sigma^2)$. In other words, we have assumed that there is a linear regression function whose parameters we wish to estimate. Quite commonly in economics, however, one can assume only that the conditional distribution of $y_i - x_i\beta$, given the values of the exogenous variables, is $N(0, \sigma^2)$. These assumptions are identical only if all the explanatory variables are exogenous, which is generally not true in simultaneous-equations models.

The most widely used estimator in simultaneous-equations models is two-stage least-squares (2SLS), which is a particular example of the instrumental variables (IV) approach which we will develop below. As we will see, 2SLS (or any IV estimator, for that matter) shares with ordinary least squares a very high sensitivity to failures of the underlying assumptions; and so we would like to have available a bounded-influence estimator for β. It turns out that the Krasker–Welsch estimator extends naturally, within the IV framework, to models with endogenous explanatory variables.

[11] The regression diagnostics in Belsley, Kuh and Welsch (1980, pp. 239 and 243) also pointed out systematic geographic concentrations of influential data.

We will suppose that we have an $n \times p$ matrix Z, whose columns are called instruments, and which (under the central model) satisfies the property mentioned above that the distribution of $y_i - x_i\beta$, given z_i, is $N(0, \sigma^2)$. The ordinary instrumental variables (IV) estimator for β is then defined by

$$b = (Z^{\mathrm{T}}X)^{-1}Z^{\mathrm{T}}y, \tag{8.1}$$

which can also be written as

$$0 = \sum_{i=1}^{n} (y_i - x_i b) z_i^{\mathrm{T}}. \tag{8.2}$$

2SLS is a special case of IV in which the instruments are formed by projecting the explanatory variables onto the space spanned by the model's exogenous variables. Let M be the $n \times p'$ matrix ($p' \geq p$) of all the exogenous variables in the model, and define $Z = M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}}X$. The 2SLS estimate is

$$b_{\mathrm{2SLS}} = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y, \tag{8.3}$$

which one can show is identical to the IV estimate in eq. (8.1) since $M(M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}}$ is a projection matrix [see Theil (1971, p. 459)].

The influence function for the IV estimator turns out to be

$$\Omega(y, x, z) = (y - x\beta)Q^{-1}z^{\mathrm{T}}, \tag{8.4}$$

where

$$Q = Ez^{\mathrm{T}}x. \tag{8.5}$$

It is interesting to compare this influence function with the least-squares influence function in eq. (4.6). There, we noticed that the influence was a product of the disturbance $y - x\beta$ and a term $Q^{-1}x^{\mathrm{T}}$ which depended only on x and represented the "leverage" of that row of the X-matrix. For IV, the influence of a particular observation still depends on the disturbance. However, x now has no "independent" effect, but rather affects b *only* through the disturbance. The "leverage" term depends on z, not x (though some of the columns of X are also columns of Z). Even if the X-matrix were highly unbalanced, so that certain observations would have considerable least-squares leverage, it is possible for Z to be relatively balanced. This could happen, for example, if a "Wald Instrument" (consisting entirely of zeros and ones) were used as a column of Z.

A logical way to construct a bounded-influence estimator for β is to restrict attention to "weighted instrumental variables" (WIV) estimators of the form

$$0 = \sum_{i=1}^{n} w_i \cdot (y_i - x_i\hat{\beta}) z_i^{\mathrm{T}}. \tag{8.6}$$

The weights w_i will depend on y_i, x_i, z_i, $\hat{\beta}$, and $\hat{\sigma}$. One can show that the influence function is

$$\Omega(y, x, z) = w(y, x, z, \beta, \sigma)(y - x\beta)B^{-1}z^{\mathrm{T}} \tag{8.7}$$

for a certain non-singular $p \times p$ matrix B. The same argument which justified the Krasker–Welsch estimator can be used to suggest the following WIV estimator b^* for β. Choose $a > \sqrt{p}$. Find A to satisfy

$$A = \frac{1}{n} \sum_{i=1}^{n} r\left(\frac{a}{(z_i A^{-1} z_i^{\mathrm{T}})^{1/2}} \right) z_i^{\mathrm{T}} z_i. \tag{8.8}$$

Finally, find b^* to satisfy

$$0 = \sum_{i=1}^{n} \min\left\{ 1, \frac{a}{\left| \frac{y_i - x_i b^*}{\hat{\sigma}} \right| (z_i A^{-1} z_i^{\mathrm{T}})^{1/2}} \right\} (y_i - x_i b^*) z_i^{\mathrm{T}}. \tag{8.9}$$

This estimator will have sensitivity equal to a, and analogous to the Krasker–Welsch estimator presented earlier, is in a sense as close to IV as possible subject to the constraint that the sensitivity be $\leq a$.

This bounded-influence weighted IV estimator provides the same powerful diagnostic information as the ordinary Krasker–Welsch estimator. However, many of its properties are not as well understood. We conjecture that it has maximum efficiency among all WIV estimators with the same sensitivity, though we have no proof as yet. Moreover, the process of approximating the distribution of b^* presents even more difficulties in the IV context than in ordinary regression. Further details may be found in Krasker and Welsch (1982b). An example of 2SLS regression diagnostics is contained in Kuh and Welsch (1980).

9. Resistant time-series estimation

In the previous sections we have said little about any special structure that might be present in the explanatory variable matrix, X. In particular, what are the

consequences of having lagged endogenous and exogenous variables? Formally, the methods described above can be applied and useful diagnostic information obtained. However, potentially far more useful diagnostic information could be obtained if the time-series structure of the data were utilized.

Research on resistant estimation in the time-series context has lagged behind studies in resistant location and regression estimation; understandably so in view of the increased difficulties imposed by dependency between the data points. In the face of such complications, it seems imperative to study resistant time-series methods by first specifying simple outlier-generating models and focusing on single time series.

Martin (1979), generalizing the earlier work of Fox (1972), introduced two kinds of outliers for linear time-series models: innovations outliers (IO) and additive outliers (AO). Roughly speaking, IO correspond to problems with the error distribution, ε, and additive outliers to the gross errors discussed in previous sections.

The observational model considered by Martin (1980b) is of the following general form:

$$z_t = \mu + y_t + \delta_t, \qquad t = 1,2,\ldots,n, \tag{9.1}$$

where μ is a location parameter, y_t is a zero-mean stationary ARMA (p, q) model:

$$y_t = \psi_1 y_{t-1} + \cdots + \psi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

and ε_t has a symmetric distribution.

The IO model is obtained when $\delta_t \equiv 0$ and the ε_t are assumed to have a heavy-tailed non-Gaussian distribution. The AO model results when ε_t are Gaussian, δ_t independent of x_t, and $\text{prob}(\delta_t = 0) = 1 - \delta$ with δ positive and not too large. The problem of robust model selection for p-order autoregressive models was considered in Martin (1980a).

For autoregressive IO and AO models, Martin and Jong (1976), Denby and Martin (1979), and Martin (1980a) have shown that:

(a) The efficiency of the LS estimates in IO models decreases dramatically at "near-Gaussian" heavy-tailed distributions;
(b) Huber-type estimates (see Section 6) of the autoregressive parameters overcome the efficiency problem for finite variance IO models;
(c) both LS and Huber estimates lack resistance toward AO situations: and
(d) the time-series analog of the Mallows estimator (see Section 6) is resistant to both IO and AO situations.

Time-series analogs of the Krasker–Welsch estimator would also be resistant but no detailed comparisons have been made with the Mallows estimator. Thus, the

bounded-influence approach is a good way to proceed for autoregressive models as well as regular regression models.

General ARMA models with additive outliers have been studied by Martin (1980b) and Martin, Samarov and Vandaele (1980). The later paper makes use of approximate conditional mean filters, a generalization of the Kalman filter. The method is, in fact, a special case of the following process [Huber (1980)].

(1) "Clean" the data by pulling outliers toward their fitted values (in the time series case by using a generalized Kalman filter).
(2) Apply LS to this adjusted data.
(3) Iterate (1) and (2) until convergence.

This method is most easily illustrated by using simple linear regression ($y_i = \alpha + \beta x_i + \varepsilon_i$) and the weights given in (4.19).

Assume that we have a fit ($\hat{y}$) to the data, i.e. a tentative regression line, an estimate of scale $\hat{\sigma}$, and an estimate of A [from (4.22)]. Then we "clean the data" by forming adjusted data,

$$\tilde{y}_i = \hat{y}_i + \left[\min\left(1, \frac{a}{\left| \frac{y_i - \hat{y}_i}{\hat{\sigma}} \right| \left(x_i A^{-1} x_i^{\mathrm{T}} \right)^{1/2}} \right) \right] (y_i - \hat{y}_i), \tag{9.2}$$

and use LS on $\tilde{y}_i$(i.e. $\hat{\beta}_{\text{new}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\tilde{y}$) to get the next tentative regression line.

Kleiner, Martin and Thompson (1979) treat the problem of robust estimation of power spectrum density. A robust method for dealing with seasonality in fitting ARMA models is presented in the robust seasonal adjustment procedure SABL (Seasonal Adjustment Bell Laboratories) due to Cleveland, Dunn and Terpening (1978a, 1978b).

10. Directions for further research

Economists often confront data and models (structural and stochastic) that are plagued by imperfections that can disproportionately influence estimates. Recent developments have led to model- and data-dependent weights for weighted least squares that bound the maximum permissible influence any row of data is allowed.

These iterative estimators have diagnostic content and, like other recently devised regression diagnostics, can be used to highlight data peculiarities and/or possible model failure. In addition, they can provide alternative estimates and predictions. The residuals from a bounded-influence fit are often more useful for

assessing problems than the residuals from a least-squares fit. Bounded-influence procedures represent a significant advance over arbitrary use of dummy variables and/or judgmental elimination of data.

There are many interesting areas for further research. Bounded-influence estimators now exist for linear single equation models and linear instrumental variables models. Extensions to non-linear models, including logit and probit, remain to be fully worked out.

It is possible to consider bounding the influence of small departures from independent errors or small departures from a specified dependent error model. Work on resistant time-series estimation for full-scale econometric models is in its infancy. Extensions to more complex stochastic situations – variance component models for instance – would also be interesting.

The theoretical foundations for bounded influence are recent. We have focused on estimators that have a strong link to maximum likelihood (with a constraint to bound the influence) but other approaches – quantile estimation, pth power, and non-parametric regression – have their advocates. We feel that a new area of inquiry has been opened with basic assumptions that correspond more closely to the data and model properties encountered in the social sciences.

References

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers and J. W. Tukey (1972) *Robust Estimates of Location*. Princeton University Press.

Annales de l'Insée, The Econometrics of Panel Data (1978) *Special Issue*, Issue 30–31, April–Sept. Paris.

Atkinson, Margaret and Jacques Mairesse (1978), "Length of life equipment in French manufacturing industries", *Annales de l'Insée*, 30–31, 23–48.

Barnett, V. and T. Lewis (1978) *Outliers in Statistical Data*. New York: John Wiley & Sons.

Belsley, David A., Edwin Kuh and Roy E. Welsch (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Chamberlain, Gary (1978) "Omitted Variable Bias in Panel Data: Estimating the Returns to Schooling", *Annales de l'Insée*, 30–31, 49–82.

Charnes, A., W. W. Cooper and R. O. Ferguson (1955) "Optimal Estimation of Executive Compensation by Linear Programming", *Management Science*, 1, 138–151.

Cleveland, W. S., D. M. Dunn and I. J. Terpenning (1978a) "A Resistant Seasonal Adjustment Procedure with Graphical Methods for Interpretation and Diagnosis", in: A. Zellner (ed.), *Seasonal Analysis of Economic Time Series*. U.S. Dept. of Commerce, Bureau of the Census.

Cleveland, W. S., D. M. Dunn and I. J. Terpenning (1978b) "The SABL Seasonal Adjustment Package-Statistical and Graphical Procedures", available from Computing Inf. Library, Bell Laboratories, Murray Hill, N. J.

Denby, L. and R. D. Martin (1979) "Robust Estimation of the First-Order Autoregressive Parameter", *Journal of the American Statistical Association*, 74, 140–146.

Diamond, Peter, Richard Anderson and Yves Balcer (1976) "A Model of Lifetime Earnings Patterns", in: *Report of the Consultant Panel on Social Security to the Congressional Research Service*, Appendix B, pp. 81–119.

Eisner, Robert (1978) "Cross Section and Time Series Estimates of Investment Functions", *Annales de l'Insée*, 30–31, 99–129.

Fair, Ray C. (1974) "On the Robust Estimation of Econometric Models", *Annals of Economic and Social Measurement*, 3.
Fisher, W. D. (1961) "A Note on Curve Fitting with Minimum Deviations by Linear Programming," *Journal of the American Statistical Association*. 56, 359–361.
Fox, A. J. (1972) "Outliers in Time Series", *Journal of the Royal Statistical Society*, B 34, 350–363.
Griliches, Z. (1968) "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change", in: A. Zellner (ed.), *Readings in Economics and Statistics*. Little Brown.
Griliches, Zvi, Bronwyn H. Hall and Jerry A. Hausman (1978) "Missing Data and Self-Selection in Large Panels", *Annales de l'Insée*, 30–31, 137–176.
Hampel, F. R. (1968) "Contributions to the Theory of Robust Estimation", Ph.D. thesis, Berkeley.
Hampel, F. R. (1973) "Robust Estimation: A Condensed Partial Survey", *Zeitschrift für Wahrscheinlichkeitstheorie und Verw. Gebeite*, 27, 87–104.
Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation", *Journal of the American Statistical Association*, 69, 383–393.
Hampel, F. R. (1978) "Optimally Bounding the Gross-Error-Sensitivity and the Influence of Position in Factor Space", in: 1978 *Proceedings of the Statistical Computing Section*. ASA, Washington, D.C. pp. 59–64.
Handschin, E., J. Kohlas, A. Fiechter and F. Schweppe (1975) "Bad Data Analysis for Power System State Estimation", *IEEE Transactions on Power Apparatus and Systems*, PAS-94, 2, 239–337.
Harrison, D. and D. L. Rubinfeld (1978) "Hedonic Prices and the Demand for Clean Air", *Journal of Environmental Economics and Management*, 5, 81–102.
Hill, R. W. (1977) "Robust Regression when there are Outliers in the Carriers", unpublished Ph.D. dissertation, Department of Statistics, Harvard University.
Hinkley, David V. (1977) "On Jackknifing in Unbalanced Situations", *Technometrics*, 19, 285–292.
Holland, P. W. and R. E. Welsch (1977) "Robust Regression Using Iteratively Reweighted Least-Squares", *Communications in Statistics*, A6, 813–827.
Huber, P. J. (1973) "Robust Regression: Asymptotics, Conjectures and Monte Carlo", *Annals of Statistics*, 1, 799–821.
Huber, Peter J. (1977) *Robust Statistical Procedures*. Philadelphia: SIAM.
Huber, P. J. (1981) *Robust Statistics*. New York: John Wiley & Sons.
Karst, O. J. (1958) "Linear Curve Fitting Using Least Deviations", *Journal of the American Statistical Association*, 53, 118–132.
Kleiner, R., R. D. Martin and D. J. Thomson (1979) "Robust Estimation of Power Spectra with Discussion", *Journal of the Royal Statistical Society*, B 41,
Koenker, Roger and Gilbert Basset, Jr. (1978) "Regression Quantiles", *Econometrica*, 46, 33–50.
Koenker, R. (1982) "Robust Methods in Econometrics", *Econometric Reviews* (to appear).
Krasker, William S. and Roy E. Welsch (1982a) "Efficient Bounded-Influence Regression Estimation", *Journal of the American Statistical Association*, 77, 595–604.
Krasker, William S. and Roy E. Welsch (1982b) "Resistant Estimation for Simultaneous-Equations Models Using Weighted Instrumental Variables", Unpublished manuscript, MIT Center for Computational Research in Economics and Management Science, Cambridge, MA.
Krasker, W. S. (1978) "Applications of Robust Estimation to Econometric Problems", unpublished Ph.D. thesis, Department of Economics, M.I.T.
Krasker, W. S. (1980) "Estimation in Linear Regression Models with Disparate Data Points", *Econometrica* 48, 1333–1346.
Krasker, W. (1981) "Robust Regression Inference", Unpublished manuscript, Graduate School of Business, Harvard University, Cambridge, MA.
Kuh, Edwin and Roy E. Welsch (1980), "Econometric Models and Their Assessment for Policy: Some New Diagnostics Applied to the Translog Energy Demand in Manufacturing", in: S. Gass (ed.), *Proceedings of the Workshop on Validation and Assessment Issues of Energy Models*. Washington, D.C.: National Bureau of Standards, pp. 445–475.
Mallows, C. L. (1973) "Influence Functions", talk at NBER Conference on Robust Regression, Cambridge, MA.
Mallows, C. L. (1975) "On Some Topics in Robustness", unpublished memorandum, Bell Telephone Laboratories, Murray Hill, New Jersey.
Marazzi (1980) "Robust Linear Regression Programs in ROBETH", Research Report No. 23,

Fachgruppe für Statistik, ETH, Zurich.
Maronna, R. A. (1976) "Robust M-estimators of Multivariate Location and Scatter", *Annals of Statistics* 4, 51–67.
Maronna, R. A. and V. J. Yohai (1980) "Asymptotic Behavior of General M-estimates for Regression and Scale with Random Carriers", *Zeitschrift für Wahrscheinlichkeitstheorie*, 58, 7–20.
Maronna, R. A., V. J. Yohai and O. Bustos (1979) "Bias- and Efficiency-Robustness of General M-estimators for Regression with Random Carriers", in: T. Gasser and M. Rosenblatt (eds.), *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics no. 757. New York: Springer Verlag, pp. 91–116.
Martin, R. D. (1979) "Robust Estimation for Time Series Autoregressions", in: R. L. Launer and G. Wilkinson (eds.), *Robustness in Statistics*. New York: Academic Press.
Martin, R. D. (1980a) "Robust Estimation of Autoregressive Models", in: Brillinger et al. (eds.), *Directions in Time Series*, Institute of Mathematical Statistics Publication.
Martin, R. D. (1980b) "Robust Methods for Time Series", in: *Proceedings of Nottingham International Time Series Meeting*. Amsterdam: North-Holland, Publishing Co.
Martin, R. D. and J. Jong (1976) "Asymptotic properties of robust generalized M-estimates for the first order autoregressive parameter", Bell Labs., Tech. Memo, Murray Hill, N.J.
Martin, R. D., A. Samarov and W. Vandaele (1981), "Robust Methods for ARIMA Models", M.I.T. Center for Computational Research in Economics and Management Science Technical Report no. 29, Cambridge, Mass.
McFadden, Daniel and Jeff Dubbin (1980) "An Econometric Analysis of Residential Electrical Appliance Holdings and Consumption", Department of Economics, Massachusetts Institute of Technology.
Meyer, J. R. and R. R. Glauber (1964) *Investment Decisions, Economic Forecasting and Public Policy*. Harvard Business School Press, Cambridge, Mass
Peters, Stephen C., Alexander Samarov and Roy E. Welsch (1982) "TROLL PROGRAM: BIF and BIFMOD", MIT Center for Computational Research in Economics and Management Science, Technical Report no. 30, Cambridge, MA.
Taylor, Lester D. (1974) "Estimation by Minimizing the Sum of Absolute Errors", in Paul Zarembka (ed.), *Frontiers in Econometrics*. New York: Academic Press, ch. 6.
Theil, H. (1971) *Principles of Econometrics*. New York: John Wiley & Sons, p. 459.
Wagner, H. M. (1959) "Linear Programming Techniques for Regression Analysis", *Journal of the American Statistical Association* 56, 206–212.
Welsch, R. E. (1977) "Regression Sensitivity Analysis and Bounded-Influence Estimation", Talk at NBER Conference on Criteria for Evaluation of Econometric Models, University of Michigan. Appears in J. Kmenta and J. Ramsey (1980) *Evaluation of Econometric Models*. New York: Academic Press, pp. 153–167.
Welsch, R. E. (1982) "Influence Functions and Regression Diagnostics", in: R. Launer and A. Siegel (eds.), *Modern Data Analysis*. New York: Academic Press, pp. 149–169.