SOME METHODS OF DETECTION OF OUTLIERS IN LINEAR REGRESSION MODEL

RANJIT KUMAR PAUL
M. Sc. (Agricultural Statistics), Roll No. 4405
IASRI, Library Avenue, New Delhi-110012

Chairperson: Dr. L. M. Bhar

Abstract: An outlier is an observation that deviates markedly from the majority of the data. To know which observation has greater influence on parameter estimate, detection of outlier is very important. There are several methods for detection of outliers available in the literature. A good number of test-statistics for detecting outliers have been developed. In contrast to detection, outliers are also tackled through robust regression techniques like, M-estimator, Least Median of Square (LMS). Robust regression provides parameter estimates that are insensitive to the presence of outliers and also helps to detect outlying observations. Recently, Forward Search (FS) method has been developed, in which a small number of observations robustly chosen are used to fit a model through Least Square (LS) method. Then more number of observations are included in the subsequent steps. This forward search procedure provides a wealth of information not only for outlier detection but, much more importantly, on the effect of each observation on aspects of inferences about the model. It also reveals the masking problem, if present, very nicely in the data.

Key words: Outlier, Leverage Point, Least Square (LS), Least Median of Square (LMS), Robust Regression, Forward Search (FS), Masking.

1. Introduction

No observation can be guaranteed to be a totally dependable manifestation of the phenomena under study. The probable reliability of an observation is reflected by its relationship to other observations that were obtained under similar conditions. Observations that in the opinion of the investigator stand apart from the bulk of the data have been called “outliers”, “extreme observations” “discordant observations”, “rouge values”, “contaminants”, “surprising values”, “mavericks” or “dirty data”. An outlier is one that appears to deviate markedly from the other members of the sample in which it occurs. An outlier is a data point that is located far from the rest of the data. Given a mean and standard deviation, a statistical distribution expects data points to fall within a specific range. Those that do not are called outliers and should be investigated. The sources of influential subsets are diverse. First, there is the inevitable occurrence of improperly recorded data, either at their sources or in their transcription to computer readable form. Second, observational errors are often inherent in the data. Although procedures more appropriate for estimation than ordinary least squares exist for this situation, the diagnostics may reveal the unsuspected existence and severity of observational errors. Third outlying data points may be legitimately occurring extreme observations. Such data often contain valuable information that improves estimation efficiency by its presence. Even in this beneficial situation, however it is constructive to isolate extreme points and to determine the extent to which the parameter estimates depend on these desirable data. Fourth, since the data could have been generated by model(s) other than that specified, diagnostics may reveal patterns suggestive of these alternatives.
The fact that a small subset of the data can have a disproportionate influence on the estimated parameters or predictions is of concern to users of regression analysis. It is quite possible that the model-estimates are based primarily on this data subset rather than on the majority of the data.

When a regression model is fitted by least squares, the estimated parameters of the fitted model depend on a few statistics aggregated over all the data. If some of the observations are different in some way from the bulk of the data, the overall conclusion drawn from this data set may be wrong. There are a series of powerful general methods for detecting and investigating observations that differ from the bulk of the data. These may be individual observations that do not belong to the general model, that is, outliers. Or there may be a subset of data that is systematically different from the majority.

2. Detection of Outliers

There are two types of outliers depending on the variable in which it occurs. Outliers in the response variable represent model failure. Outliers with respect to the predictors are called leverage points; they can affect the regression model. Their response variables need not be outliers. However, they may almost uniquely determine regression coefficients. They may also cause the standard errors of regression coefficients to be much smaller than they would be if these observations were excluded. Leverage points do not necessarily correspond to outliers. An observation with sufficiently high leverage might exert enough influence to drag the regression equation close to its response and mask the fact that it might otherwise be an outlier.

The ordinary or simple residuals (observed - predicted values) are the most commonly used measures for detecting outliers. The ordinary residuals sum to zero but do not have the same standard deviation. Many other measures improve on or complement simple residuals. Standardized Residuals are the residuals divided by the estimates of their standard errors. They have mean 0 and standard deviation 1. There are two common ways to calculate the standardized residual for the i^{th} observation. The use of residual mean square error from the model fitted to the full dataset (internally studentized residuals) and the use of residual mean square error from the model fitted to all of the data except the i^{th} observation (externally studentized residuals). The externally studentized residuals follow a t distribution with \(n-p-2\) df, where \(n\) is total number of observations and \(p\) is the number of parameters.

Outlier diagnostics are the statistics that focus attention on observations having a large influence on the Least Squares (LS) estimator. Several diagnostic measures have been designed to detect individual cases or groups of cases that may differ from the bulk of the data. The field of diagnostics consists of a combination of numerical and graphical tools. Some commonly used statistics in detection of outliers are described now.

2.1 Row Deletion Methods

There are many methods for detection of outliers available in the literature. Some statistics that are obtained through row deletion method are considered here. It is examined in turn how the deletion of each row affects the estimated coefficients, the predicted values (fitted values), the residuals, and the estimated covariance structure of the coefficients.
Consider the following linear regression model

\[ Y = X\beta + \varepsilon \]

where \( Y \) is a \( n \times 1 \) vector of observations, \( X \) is a \( n \times p \) matrix of explanatory variables and \( \beta \) is \( p \times 1 \) vector of parameters. \( \varepsilon \) is a \( n \times 1 \) vector of errors such that

\[ E(\varepsilon) = 0 \quad \text{and} \quad E(\varepsilon\varepsilon') = \sigma^2 I. \]

(i) **DFBETA**

Since the estimated coefficients are often of primary interest in regression models, we look first at the change in the estimate of regression coefficients that would occur if the \( i \)th row were deleted. Denoting the coefficients estimated with the \( i \)th row deleted by \( \beta_{(i)} \), this change is computed by the formula,

\[
\text{DFBETA}_i = \hat{\beta} - \hat{\beta}_{(i)} = \frac{(X'X)^{-1}x_i e_i}{1 - h_{ii}}
\]

where \( x_i \) is the \( i \)th row of the \( X \) matrix, \( e_i \) is the \( i \)th residual and \( h_{ii} \) is the \( i \)th diagonal element of the matrix \( X(X'X)^{-1}X' \). The cut off value for DFBETA it is \( 2/\sqrt{n} \).

(ii) **Cook’s Distance**

Cook (1977) proposed a statistic for detection of outlier as follows:

\[
D_i = \left( \hat{\beta}_{(i)} - \hat{\beta} \right)'X'X(\hat{\beta}_{(i)} - \hat{\beta})/s^2
\]

where \( s^2 \) is the estimate of \( \sigma^2 \).

Large values of \( D_i \) indicate observations that are influential on joint inferences about all the linear parameters in the model. A suggestive alternative form of \( D_i \) is

\[
D_i = \left( \hat{Y}_{(i)} - \hat{Y} \right)'\left( \hat{Y}_{(i)} - \hat{Y} \right)/s^2,
\]

where the \( \hat{Y}_{(i)} = X\hat{\beta}_{(i)} \). An interpretation is that \( D_i \) measures the sum of squared changes in the predictions when observation \( i \) is not used in estimating \( \beta \). \( D_i \) approximately follows \( F(p, n-p) \) distribution. The cut off value of Cook-Statistic is \( 4n \).

(iii) **DFFIT**

It is the difference between the predicted responses from the model constructed using complete data and the predicted responses from the model constructed by setting the \( i \)th observation aside. It is similar to Cook's distance. Unlike Cook's distance, it does not look at all of the predicted values with the \( i \)th observation set aside. It looks only at the predicted values for the \( i \)th observation. \( \text{DFFIT}_i \) is computed as follows:

\[
\text{DFFIT}_i = \hat{Y}_i - \hat{Y}_{(i)} = x_i(\hat{\beta} - \hat{\beta}_{(i)}) = \frac{h_{ii}e_i}{1 - h_{ii}}. \quad \text{The cut off value of DFFIT is} \quad 2\sqrt{\frac{p}{n}}.
\]

(iv) **Covariance Matrix**

Another major aspect of regression is the covariance matrix of the estimated coefficients. We again consider the diagnostic technique of row deletion, this time in a comparison of the covariance matrix using entire data, \( \sigma^2(X'X)^{-1} \), with the covariance matrix that results when \( i \)th row has been deleted, \( \sigma^2[X'(i)X(i)]^{-1} \). Of the various alternative means for
Some Methods of Detection of Outliers in Linear Regression Model

comparing two such positive definite symmetric matrices, the ratio of their determinants \( \frac{\det \sigma^2 \left[ X'(i)X(i) \right]^{-1}}{\det \sigma^2 (X'X)^{-1}} \) is one of the simplest and is quite appealing. Since these two matrices differ only by the inclusion of the \( i^{th} \) row in the sum of squares and cross products, values of this ratio near unity can be taken to indicate that the two covariance matrix is insensitive to the deletion of row \( i \). Of course the estimator \( s^2 \) of \( \sigma^2 \) also changes with the deletion of the \( i^{th} \) observation. We can bring the \( y \) data into consideration by comparing the two matrices \( s^2 (X'X)^{-1} \) and \( s^2 (i) \left[ X'(i)X(i) \right]^{-1} \) in the determinantal ratio

\[
\text{COVRATIO} = \frac{\det s^2 (i) \left[ X'(i)X(i) \right]^{-1}}{\det s^2 (X'X)^{-1}}
\]

or,

\[
\text{COVRATIO} = \frac{1}{\frac{n - p - 1}{n - p} \left[ \frac{\epsilon_i^*}{n - p} \left( 1 - h_{ii} \right) \right]^p + \frac{\epsilon_i^*}{n - p} \left( 1 - h_{ii} \right)}
\]

where \( \epsilon_i^* = \frac{\epsilon_i}{s(i)\sqrt{1 - h_{ii}}} \) is the studentized residual.

For COVRARIO, cut of value is \( 1 \pm 3 \left( \frac{p}{n} \right) \).

2.2 Hat Matrix

Besides the above statistics, the Hat matrix is sometimes used to detect the influential observations. The \( h_{ii} \) are the diagonal elements of the least squares projection matrix, also called hat matrix,

\[ H = X (X'X)^{-1} X' \]

This determines the fitted or predicted values, since

\[ \hat{Y} = X\hat{\beta} = HY \]

The influence of the response value, \( Y_i \), on the fit is most directly reflected in its impact on the corresponding fitted value, \( \hat{Y}_i \), and this information is to be contained in \( h_{ii} \). Where there are two or fewer explanatory variables, scatter plots will quickly reveal any \( x \) outliers, and it is not hard to verify that they have relatively large \( h_{ii} \) values. Here the cutoff value is \( 2p/n \), where \( p \) is the rank of the \( X \) matrix. The \( i^{th} \) observation is called a leverage point when \( h_{ii} \) exceeds \( 2p/n \).

2.3 Outlier Detection Based on Robust Regression

Many diagnostics are based on the residuals resulting from LS. However, this starting point may lead to useless results because of the following reason. By definition, LS tries to avoid the large residuals. Consequently, one outlying case may cause a poor fit for the majority of the data because the LS estimator tries to accommodate this case at the expense
of the remaining observations. Therefore an outlier may have a small residual, especially when it is a leverage point. As a consequence, diagnostics based on LS residuals often fail to reveal such points.

A least squares model can be distorted by a single observation. The fitted line or surface might be tipped so that it no longer passes through the bulk of the data. In order to reduce the effect of a very large error it will introduce many small or moderate errors. For example, if a large error is reduced from 200 to 50, its square is reduced from 40,000 to 2,500. Increasing an error from 5 to 15 increases its square from 25 to 225. Thus, a least squares fit might introduce many small errors in order to reduce a large one.

In Figure 2.1 the line A denotes the regression line and passes through the bulk of the data. But in presence of the outlier the line is dragged out to the outlier because the principle of the least square method says that the residuals sum of squares should be minimum. The line B indicates the regression line in presence of the outlier. Again in some cases where the LS estimate is used to detect outlier may diagnose a clean point as an outlier and an outlier as a clean point.
In Figure 2.2, the point 1 is actually an outlier and in presence of this point the regression line is dragged out to that point resulting the point 2 as an outlier, though it is a clean point, because now the distance of this point from the line is longest.

Moreover, all those statistics described above involved LS residuals. Therefore, if these statistics are applied in detection of outliers, it may wrongly detect some clean observations as outliers. So it is suggested to use robust regression in place of LS. Robust regression not only provides parameter estimation that is insensitive to the presence of outliers, but also helps in detecting outliers.

Robust regression is a term used to describe model fitting procedures that are insensitive to the effects of outlier observations. Many robust regression methods have been developed, out of which M estimator is most popular.

**M-estimator**

In general, a class of robust estimators that minimize a function $f$ of the residuals is defined as,

$$\beta = \text{Minimize} \sum_{i=1}^{n} f(e_i) = \text{Minimize} \sum_{i=1}^{n} f(y_i - x_i \beta) \quad \ldots(2.1)$$

where $x_i$ denotes the $i$th row of $X$.

Generally the following equation is solved:

$$\beta \text{Minimize} \sum_{i=1}^{n} f\left(\frac{e_i}{s}\right) = \text{Minimize} \sum_{i=1}^{n} f\left(\frac{y_i - x_i \beta}{s}\right) \quad \ldots(2.2)$$

where $s = \text{median}|e_i - \text{median}(e_i)|/0.6745 \quad \ldots(2.3)$

$s$ is an approximately unbiased estimator of $\sigma$ if $n$ is large and the error distribution is normal.

An estimator of this type is called an **M-estimator**, where $M$ stands for maximum likelihood. That is, the function $f$ is related to the likelihood function for an appropriate choice of the error distribution. For example, if the method of least squares is used (implying the error distribution is normal), then $f(z) = (1/2)z^2$.

To minimize Eq. 2.2, equate the first partial derivatives of $f$ with respect to $\beta_j (j = 1, 2, \ldots, p)$ to zero, yielding a necessary condition for a minimum. This gives the system of $p$ equations

$$\sum_{i=1}^{n} x_{ij} \psi\left(\frac{y_i - x_i \beta}{s}\right) = 0, \quad (j = 1, \ldots, p) \quad \ldots(2.4)$$

where $s$ is the robust estimate of scale, $\psi = f'$ and $x_{ij}$ is the $i$th observation on the $j$th regressor and $x_{i0} = 1$. In general the $\psi$ function is non linear and the Eq. 2.4 must be
Some Methods of Detection of Outliers in Linear Regression Model

solved by iterative methods. Iterative reweighted least squares is most widely used. This approach is usually attributed to Beaton and Tukey (1974).

To use iteratively reweighted least squares, suppose that an initial estimate \( \hat{\beta}_0 \) is available and that \( s \) is an estimate of scale. Then write \( p \) equations in (2.4),

\[
\sum_{i=1}^{n} x_{ij} w_0 \left( \frac{y_i - x_i^t \hat{\beta}_0}{s} \right) = 0
\]

\[
\sum_{i=1}^{n} x_{ij} \left[ \psi \left( \frac{(y_i - x_i^t \hat{\beta})}{s} \right) / \psi \left( \frac{(y_i - x_i^t \hat{\beta}_0)}{s} \right) \right] \left( y_i - x_i^t \hat{\beta} \right) = 0
\]

as

\[
\sum_{i=1}^{n} x_{ij} w_{i0} (y_i - x_i^t \hat{\beta}_0) = 0
\]

where

\[
w_{i0} = \frac{\psi \left( \frac{(y_i - x_i^t \hat{\beta}_0)}{s} \right)}{\psi \left( \frac{(y_i - x_i^t \hat{\beta})}{s} \right)} \quad \text{if} \quad y_i \neq x_i^t \hat{\beta}_0
\]

\[= 1 \quad \text{if} \quad y_i = x_i^t \hat{\beta}_0\]

In matrix notation Eq.2.6 becomes

\[
X^t W_0 X \hat{\beta} = X^t W_0 y
\]

where \( W_0 \) is an \( n \times n \) diagonal matrix of “weights” with diagonal elements \( w_{10}, w_{20}, \ldots, w_{n0} \) given by Eq.2.7. We recognize Eq.2.8 as the usual weighted least-squares normal equations. Consequently the one step estimator is

\[
\hat{\beta}_1 = (X^t W_0 X)^{-1} X^t W_0 y
\]

In the next step we recompute the weights from Eq. 2.7 but using \( \hat{\beta}_1 \) instead of \( \hat{\beta}_0 \).

Least Median of Squares (LMS) Estimator

Least Median of Squares (LMS) regression, developed by Rousseeuw (1984) minimizes the median squared residuals. Since it focuses on the median residual, up to half of the observations can disagree without masking a model that fits the rest of the data.

For the linear regression model \( E(Y) = X\beta \), with \( X \) of rank \( p \), let \( b \) be any estimate of \( \beta \). With \( n \) observations, the residuals from this estimate are \( e_i(b) = y_i - x_i^t b, (i=1,2,\ldots,n) \). The LMS estimate \( \hat{\beta}_p \) is the value of \( b \) minimizing the median of the square residuals \( e_i^2(b) \).

Thus \( \hat{\beta}_p \) minimizes the scale estimate

\[
\sigma^2(b) = e_{[\text{med}]}^2(b)
\]

where \( e_{[k]}(b) \) is the \( k^\text{th} \) ordered squared residual. In order to allow for estimation of the parameters of the linear model the median is taken as

\[
\text{med} = \text{integer part of } [(n+p+1)/2],
\]

... (2.11)
The parameter estimate satisfying (2.10) has asymptotically, a break down point of 50%. Thus, for large $n$, almost half the data can be outliers, or come from some other model and LMS will still provide an unbiased estimate of the regression parameters. This is the maximum break down that can be tolerated. For a higher proportion of outliers there is no longer a model that fits the majority of the data. The very robust behavior of the LMS estimate is in contrast to that of the least squares estimate $\hat{\beta}$ minimizing

$$S(\beta) = (y - X\beta)'(y - X\beta),$$

which can be written as

$$S(b) = \sum_{i=1}^{n} e_i^2(b).$$

Only one observation needs to be moved towards infinity to cause an arbitrarily large change in the estimate $\hat{\beta}$: the breakdown point of $\hat{\beta}$ is zero.

The definition of $\hat{\beta}_p$ in (2.10) gives no indication of how to find such a parameter estimate. Since the surface to be minimized has many local minima, approximate methods are used. Roussaeuw (1984) finds an approximation to $\hat{\beta}_p$ by searching only over elemental sets that is subsets of $p$ observations, taken at random.

Fitting an LMS regression model possesses some difficulties. The first is computational. Unlike least squares regression, there is no formula that can be used to calculate the coefficients for an LMS regression. Roussaeuw (1984) has proposed an algorithm to obtain LMS estimator. According to this algorithm a random sample of size $p$, is drawn. A regression surface is fitted to each set of observations and the median squared residual is calculated. The model that has the smallest median squared residual is used. Once a robust fit of the model is obtained, residuals from this model are used for detecting outlying observations.

The following Figure 2.3 describes that the LMS fit is not affected by the outlying observation.

![Figure 2.3](image-url)
One of the drawbacks of the LMS estimate is that it does not consider all the data points for estimation of parameters. Also it does not reveal the masking effect of outliers if any. So the forward search method of detection of outliers has been used.

2.4 Forward Search
If the values of the parameters of the model were known, there would be no difficulty in detecting the outliers, which would have large residuals. The difficulty arises because the outliers are included in the data used to estimate the parameters, which can then be badly biased. Most methods for outlier detection therefore seek to divide the data into two parts, a larger “clean” part and the outliers.

For detecting multiple outliers, some apply single row deletion method repeatedly. But this method fails when there is a problem of masking. Multiple row deletion technique has also been suggested. But the difficulty here is the explosion of the number combination to be considered. To overcome such problems, forward search has been evolved by Atkinson (1994).

The basic idea of this method is to order the observations by their closeness to the fitted model. It starts with a fit to very few observations and then successively fit to larger subsets. The starting point is found by fitting to a large number of small subsets, using methods from robust statistics to determine which subset fits best. Then all the observations are ordered by closeness to this fitted model; for regression models the residuals determine closeness. For multivariate models, the subset size is increased by one and the model refitted to the observations with the smallest residuals for the increased subset size. Usually one observation enters, but some times two or more enter the subset as one or more leave. The process continues with increasing subset sizes until; finally, all the data are fitted. As a result of this forward search we have an ordering of the observations by closeness to the assumed model.

If the model and the data agree, the robust and least squares fit will be similar, as will be the parameter estimates and residuals from the two fits. But often the estimates and the residuals of the fitted model change appreciably during the forward search. The changes in these quantities and in various statistics are monitored, such as score tests for transformation, as the process moves forward through the data, adding one observation at a time. This forward procedure provides a wealth of information not only for outlier detection but, much more importantly, on the effect of each observation on aspects of inference about the model.

In the forward search, such larger sub-samples of outlier free observations are found by starting from small subsets and incrementing them with observations that have small residuals, and so are unlikely to be outliers. The method was introduced by Hadi (1992) for the detection of outliers from a fit using approximately half the observations. Different versions of method are described by Hadi and Simonoff (1993), Hadi and Simonoff (1994) and by Atkinson (1994).

Suppose at some stage in the forward search the set of \( m \) observations used in fitting is \( S^{(m)} \). Fitting to this subset is by least squares (for regression models) yielding the parameter estimates \( \hat{\beta}_m \). From these parameter estimates a set of \( n \) residuals \( \hat{e}_n \) are
calculated and also estimate $\sigma^2$. Suppose the subset $S_{*}^{(m)}$ is clear of outliers. There will then be $n-m$ observations not used in fitting that may contain outliers. The interest is in the evolution, as $m$ goes from $p$ to $n$, of quantities such as residuals, and test statistics, together with Cook’s distance and other diagnostic quantities. The sequence of the parameter estimates $\hat{\beta}_m$ and related $t$ statistics are also monitored. The changes that occur, which can always be associated with the introduction of a particular group of observations are monitored. In practice almost always one observation, into the subset $m$ used for fitting.

Interpretation of these changes is complemented by examination of changes in the forward plot of residuals.

**Remark 1**
The search starts with the approximate LMS estimator found by sampling subsets of size $p$.

Let this be $\hat{\beta}_p^*$ and the Least Square estimator at the end of the search be $\hat{\beta}_n^* = \hat{\beta}$. In the absence of outliers and systematic departures from the model

$$E (\hat{\beta}_p^*) = E (\hat{\beta}) = \beta;$$

that is, both the parameter estimates are unbiased estimators of the same quantity. The same property holds for the sequence of estimates $\hat{\beta}_m^*$ produced in the forward search. Therefore, in the absence of outliers, it is expected that both parameter estimates and residuals would remain sensibly constant during the forward search.

**Remark 2**
Now suppose there are $k$ outliers. Starting from a clean subset, the forward procedure will include these towards the end of the search, usually in the last $k$ steps. Until these outliers are included, the condition of Remark 1 will hold and that residuals plots and parameter estimates will remain sensibly constant until the outliers are incorporated in the subset used for fitting.

**Remark 3**
If there are indications that the regression data should be transformed, it is important to remember that outliers in the transformed scale may not be outliers in another scale. If the data are analyzed using the wrong transformation, the $k$ outliers may enter the search well before the end.

The forward algorithm is made up of three steps: the first concerns the choice of an initial subset the second refers to the way in which the forward search is progressed and the third relates to the monitoring of the statistics during the progress of the search.

**Step 1: Choice of the Initial Subset**
A formal definition of the algorithm used to find the LMS estimator is now given. If the model contains $p$ parameters, the forward search algorithm starts with the selection of a subset of $p$ units. Observations in this subset are intended to be outlier free. If $n$ is moderate and $p \ll n$, the choice of the initial subset can be performed by exhaustive
enumeration of all \( \binom{n}{p} \) distinct \( p \) tuples set. Let \( e_{i,S(p)} \) be the least squares residual for unit \( i \) given observations in \( S_i^{(p)} \). The initial subset is taken as the \( p \) tuple \( S_i^{(p)} \) which satisfies \( e_{\text{med},S_i^{(p)}}^2 = \min_{1} \left[ e_{\text{med},S_i^{(p)}}^2 \right] \), where \( e_{[k]}^{(p)} S_i^{(p)} \) is the \( k \)th ordered squared residual among \( e_{i,S_i^{(p)}}^2 \), \( i = 1, \ldots, n \), and, med is the integer part of \( (n + p + 1)/2 \). If \( \binom{n}{p} \) is too large, 1,000 samples are taken.

**Step 2: Adding Observations during the Forward Search**

Given a subset of dimension \( m \geq p \), say \( S_i^{(m)} \), the forward search moves to dimension \( m + 1 \) by selecting the \( m + 1 \) units being chosen by ordering all squared residuals \( e_{i,S_i^{(m)}}^2 \), \( i = 1, \ldots, n \). The forward search estimator \( \hat{\beta}_{FS} \) is defined as a collection of least squares estimators in each step of forward search; that is,

\[
\hat{\beta}_{FS} = (\hat{\beta}_p^*, \ldots, \hat{\beta}_n^*).
\]

In most moves from \( m \) to \( m + 1 \) just one new unit joins the subset. It may also happen that two or more units join \( S_i^{(m)} \) as one or more leave. However, such an event is quite unusual, only occurring when the search includes one unit that belongs to a cluster of outliers. At the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. Of course, several other units then have to leave the subset.

The search avoids, in the first steps, the inclusion of outliers and provides a natural ordering of the data according to the specified null model. In this approach a highly robust method and at the same time least squares estimators are used. The zero breakdown point of least square estimators, in the context of the forward search, does not turn out to be disadvantageous. The introduction of typical (influential) observations is signaled by sharp changes in the curves that monitor parameter estimates, t tests, or any other statistic at every step. In this context the robustness of the method does not derive from the choice of a particular estimator with a high breakdown point, but from the progressive inclusion of the units into a subset that are outlier free. As a bonus of suggestive procedure, the observations can be naturally ordered according to the specified null model and it is possible to know how many of the m are compatible with a particular specification. Furthermore the suggested approach enables to analyze the influential effect of the atypical units (outliers) on the results of the statistical analyses.

**Remark 4**

The method is not sensitive to the method used to select an initial subset, provided unmasked outliers are not included at the start. For example the least median of squares criterion for regression can be replaced by that of Least Trimmed Squares (LTS). This criterion provides estimators with better properties than LMS estimators, found by minimizing the sum of the smallest \( h \) squared residuals.
Some Methods of Detection of Outliers in Linear Regression Model

$$S_h(b) = \sum_{i=1}^{n} e_i^2(b),$$ for some $h$ with $[(n + p + 1) / 2] \leq h < n$.

What is important in this procedure is that the initial subset is either free of outliers or contains unmasked outliers which are immediately removed by the forward procedure. The search is often able to recover from a start that is not very robust. An example, for regression, is given by Atkinson and Mulira (1993) and for spatial data by Cerioli and Riani (1999).

**Step 3: Monitoring the Search**

Step 2 of the forward search is repeated until all units are included in the subset. If just one observation enters $S_{(m)}$ at each move, the algorithm provides an ordering of the data according to the specified null model, with observations furthest from it joining the subset at the last stages of the procedure.

**Example 2.1**: A data set used by Weisberg (1985) is considered here to introduce the ideas of regression analysis. There are 17 observations on the boiling point of water in °F at different pressures, obtained from measurements at a variety of elevations in the Alps. The purpose of the original experiment was to allow prediction of pressure from boiling point, which is easily measured, and so to provide an estimate of altitude. The higher the altitude, the lower the pressure and the consequent boiling point. Weisberg (1985) gives values of both pressure and 100 × log (pressure) as possible response. The variables are:

- $x$: boiling point, °F and $y$: 100 × log(pressure).

**Table 2.1**: Data on air pressure in the Alps and the boiling point of water

<table>
<thead>
<tr>
<th>Observation Number</th>
<th>Boiling Point</th>
<th>100*Log Pressure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>194.5</td>
<td>131.79</td>
</tr>
<tr>
<td>2</td>
<td>194.3</td>
<td>131.79</td>
</tr>
<tr>
<td>3</td>
<td>197.9</td>
<td>135.02</td>
</tr>
<tr>
<td>4</td>
<td>198.4</td>
<td>135.55</td>
</tr>
<tr>
<td>5</td>
<td>199.4</td>
<td>136.46</td>
</tr>
<tr>
<td>6</td>
<td>199.9</td>
<td>136.83</td>
</tr>
<tr>
<td>7</td>
<td>200.9</td>
<td>137.82</td>
</tr>
<tr>
<td>8</td>
<td>201.1</td>
<td>138.00</td>
</tr>
<tr>
<td>9</td>
<td>201.4</td>
<td>138.06</td>
</tr>
<tr>
<td>10</td>
<td>201.3</td>
<td>138.05</td>
</tr>
<tr>
<td>11</td>
<td>203.6</td>
<td>140.04</td>
</tr>
<tr>
<td>12</td>
<td>204.6</td>
<td>142.44</td>
</tr>
<tr>
<td>13</td>
<td>209.5</td>
<td>145.47</td>
</tr>
<tr>
<td>14</td>
<td>208.6</td>
<td>144.34</td>
</tr>
<tr>
<td>15</td>
<td>210.7</td>
<td>146.3</td>
</tr>
<tr>
<td>16</td>
<td>211.9</td>
<td>147.54</td>
</tr>
<tr>
<td>17</td>
<td>212.2</td>
<td>147.80</td>
</tr>
</tbody>
</table>

The data are plotted in Figure 2.4. A quick glance at the plot shows there is a strong linear relationship between log(pressure) and boiling point. A slightly longer glance reveals that one of the points lies slightly off the line. Linear regression of $y$ on $x$ yields a $t$ value for the regression of $54.45$, clear evidence of the significance of the relationship.
Two plots of the least square residuals $e$ are often used to check fitted models. Figure 2.5 shows a plot of residuals against fitted values $y$. This clearly shows one outlier, observation 12. The normal plot of the studentized residuals is an almost straight line from which the large residual for observation 12 is clearly distanced. It is clear that observation 12 is an outlier.

Now it is shown that how forward search can reveal this point as an outlier.

It is started with a least squares fit to two observations, robustly chosen. From this fit residuals for all 17 observations are calculated and next fit to the three observations with smallest residuals. In general we fit to a subset of size $m$, order the residuals and take as the next subset the $m+1$ case with smallest residuals. This gives a forward search through the data, order by closeness to the model. It is expected that the last observations to enter the search will be those which are furthest from the model and so may cause changes once
they are included in the subset used for fitting. In the search through the data, the outlying observation 12 was the last to enter the search.

For each value of m from 2 to 17 quantities such as the residuals and the parameter estimates are calculated and see how they change. Figure 2.6 is a plot of the values of the parameter estimates during the forward search. The values are extremely stable, reflecting the closeness of all observations to the straight line. The introduction of observation 12 at the end of the search causes virtually no change in the position of the line. However, Figure 2.7 shows that introduction of observation 12 causes a huge increase in \( s^2 \), the residuals mean square estimate of the error variance (\( \sigma^2 \)). The information from these plots about observation 12 confirms and quantifies that from the scatter plot of Figure 2.4, observation 12 is an outlier, but the observation is in the centre of the data, so that its exclusion or inclusion has a small effect on the estimated parameters. The plots also show that all other observations agree with the overall model. Throughout the search, all cases have small residuals, apart from case 12 which is outlying from all fitted subsets. Even when it is included in the last step of the search, its residual only decreases slightly.

**Remark 5**
The estimate of \( \sigma^2 \) does not remain constant during the forward search as observations are sequentially selected that have small residuals. Thus, even in the absence of outliers, the
Some Methods of Detection of Outliers in Linear Regression Model

The residuals mean square estimate $s^2_{n(m)} < s^2_{n(n)} = s^2$ for $m < n$. The smooth increase of $s^2_{n(m)}$ with $m$ for the transformed data is typical of what is expected when the data agree with the model and are correctly ordered by the forward search.

**Example 2.2:** Table 2.2 gives 60 observations on a response $y$ with the value of three explanatory variables [Atkinson and Riani (2000)].

<table>
<thead>
<tr>
<th>obs</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2133</td>
<td>-2.308</td>
<td>-8.5132</td>
<td>-10.6651</td>
</tr>
<tr>
<td>2</td>
<td>0.9413</td>
<td>1.2048</td>
<td>-4.1313</td>
<td>5.2739</td>
</tr>
<tr>
<td>3</td>
<td>-0.6165</td>
<td>-1.0754</td>
<td>-5.4988</td>
<td>-2.2908</td>
</tr>
<tr>
<td>4</td>
<td>-1.2154</td>
<td>1.4384</td>
<td>-4.7902</td>
<td>2.805</td>
</tr>
<tr>
<td>5</td>
<td>0.3678</td>
<td>1.6125</td>
<td>-2.8339</td>
<td>7.5527</td>
</tr>
<tr>
<td>6</td>
<td>0.0366</td>
<td>0.584</td>
<td>-5.8016</td>
<td>0.4213</td>
</tr>
<tr>
<td>7</td>
<td>0.1636</td>
<td>0.7677</td>
<td>-3.9292</td>
<td>1.8518</td>
</tr>
<tr>
<td>8</td>
<td>-1.4562</td>
<td>0.3064</td>
<td>-7.1359</td>
<td>-4.0885</td>
</tr>
<tr>
<td>9</td>
<td>-2.6535</td>
<td>-2.142</td>
<td>-7.6303</td>
<td>-7.0537</td>
</tr>
<tr>
<td>10</td>
<td>-1.2722</td>
<td>0.2899</td>
<td>-5.0166</td>
<td>-0.8915</td>
</tr>
<tr>
<td>11</td>
<td>-1.2276</td>
<td>-1.4133</td>
<td>-7.6792</td>
<td>-7.0021</td>
</tr>
<tr>
<td>12</td>
<td>1.3087</td>
<td>0.5436</td>
<td>-2.9362</td>
<td>4.4775</td>
</tr>
<tr>
<td>13</td>
<td>-0.1036</td>
<td>-2.227</td>
<td>-7.7164</td>
<td>-10.7953</td>
</tr>
<tr>
<td>14</td>
<td>-0.2724</td>
<td>0.2107</td>
<td>-3.156</td>
<td>5.3512</td>
</tr>
<tr>
<td>15</td>
<td>-0.3896</td>
<td>-3.1353</td>
<td>-8.7172</td>
<td>-14.5927</td>
</tr>
<tr>
<td>16</td>
<td>0.6659</td>
<td>0.8877</td>
<td>-3.2162</td>
<td>5.8618</td>
</tr>
<tr>
<td>17</td>
<td>2.121</td>
<td>1.652</td>
<td>-1.1933</td>
<td>11.1076</td>
</tr>
<tr>
<td>18</td>
<td>-0.9035</td>
<td>-1.1082</td>
<td>-5.7122</td>
<td>-4.6404</td>
</tr>
<tr>
<td>19</td>
<td>-0.88</td>
<td>-1.703</td>
<td>-7.0756</td>
<td>-8.3006</td>
</tr>
<tr>
<td>20</td>
<td>-0.4792</td>
<td>-1.4409</td>
<td>-7.1334</td>
<td>-7.3699</td>
</tr>
<tr>
<td>21</td>
<td>-1.9608</td>
<td>-1.9674</td>
<td>-7.0908</td>
<td>-6.2175</td>
</tr>
<tr>
<td>22</td>
<td>-1.9271</td>
<td>0.5333</td>
<td>-5.8178</td>
<td>-1.5325</td>
</tr>
<tr>
<td>23</td>
<td>-0.497</td>
<td>-1.2314</td>
<td>-7.8755</td>
<td>-9.2648</td>
</tr>
<tr>
<td>24</td>
<td>-1.0759</td>
<td>-1.0643</td>
<td>-5.4215</td>
<td>-2.9707</td>
</tr>
<tr>
<td>25</td>
<td>-1.3846</td>
<td>0.8026</td>
<td>-5.1148</td>
<td>0.1859</td>
</tr>
<tr>
<td>26</td>
<td>2.165</td>
<td>-1.6213</td>
<td>-3.0335</td>
<td>0.5601</td>
</tr>
<tr>
<td>27</td>
<td>-0.9351</td>
<td>0.197</td>
<td>-4.5964</td>
<td>2.2223</td>
</tr>
<tr>
<td>28</td>
<td>0.1378</td>
<td>-1.1782</td>
<td>-5.888</td>
<td>-3.6547</td>
</tr>
<tr>
<td>29</td>
<td>-2.5223</td>
<td>-1.6597</td>
<td>-9.4292</td>
<td>-13.6328</td>
</tr>
<tr>
<td>30</td>
<td>-0.6787</td>
<td>-1.9986</td>
<td>-4.9259</td>
<td>-1.4192</td>
</tr>
<tr>
<td>31</td>
<td>-1.8665</td>
<td>-1.7153</td>
<td>-6.511</td>
<td>-3.9751</td>
</tr>
<tr>
<td>32</td>
<td>0.2733</td>
<td>0.4637</td>
<td>-3.937</td>
<td>2.207</td>
</tr>
<tr>
<td>33</td>
<td>2.5053</td>
<td>-0.2927</td>
<td>-3.3827</td>
<td>2.9221</td>
</tr>
<tr>
<td>34</td>
<td>-0.6474</td>
<td>0.8221</td>
<td>-4.1466</td>
<td>2.9272</td>
</tr>
<tr>
<td>35</td>
<td>-0.2099</td>
<td>-0.785</td>
<td>-6.1856</td>
<td>-3.8248</td>
</tr>
<tr>
<td>36</td>
<td>-0.1594</td>
<td>-2.4325</td>
<td>-6.8908</td>
<td>-7.3217</td>
</tr>
<tr>
<td>37</td>
<td>-0.3642</td>
<td>-0.8784</td>
<td>-4.4361</td>
<td>-1.0826</td>
</tr>
<tr>
<td>38</td>
<td>-1.7083</td>
<td>-2.0538</td>
<td>-6.9914</td>
<td>-5.7308</td>
</tr>
<tr>
<td>39</td>
<td>0.0828</td>
<td>1.6929</td>
<td>-3.0457</td>
<td>5.2853</td>
</tr>
</tbody>
</table>
Some Methods of Detection of Outliers in Linear Regression Model

The plot of residuals against fitted values, Figure 2.8 shows no obvious pattern. The largest residual is that of case 43. There is therefore no clear indication that the data are not homogeneous and well behaved. Evidence of the structure of the data is clearly shown in the Fig.9, the scaled squared residuals from the forward search. This fascinating plot reveals the presence of 6 masked outliers. The left hand end of the plot gives the residuals from the LMS estimates found by sampling 1000 subsets of size $p=4$. From the most extreme residual downwards, the cases giving rise to the outliers are 9, 30, 31, 38, 47, and 21. When all the data are fitted the largest residuals belong to, in order cases 43, 51, 2, 47, 31, 9, 38, 29, 7 and 48.

![Figure 2.8](image-url)

Figure 2.8
The assessment of the importance of these outliers can be made by the behavior of the parameter estimates and of the related t statistics. Apart from the $\hat{\beta}_1$, all remain positive with t values around 10 or greater during the course of the forward search. We therefore concentrate on the behavior of $t_1$, the t statistics for $\beta_1$. The values for the last 20 steps of the forward search are plotted in the Fig.10. The general downward trend is typical of plots of t statistics from the forward search. It is caused by the increasing value of $s^2$, Figure 2.11 as observations with larger residuals are entered during the search.

An important feature in the interpretation of Figure 2.10 is the two upward jumps in the value of the statistic. The first results form the inclusion of observation 43 when m=54, giving a t value of 2.25, evidence significant at the 3% level, of a positive value of $\beta_1$. Therefore the outlier enter the subset, with the observation 43 leaving when m=58, as two outliers enter. When m=59 the value of the statistic has decreased to -1.93, close to evidence for a negative value of the parameter. Reintroduction of observation 43 in the last step of the search results in a value of -1.26, indicating that $\beta_1$ may well be zero. It is therefore important that the outliers be identified.
Some Methods of Detection of Outliers in Linear Regression Model

This example shows very clearly the existence of masked outliers, which would not be detected by the LS. However the forward plot of residuals in Figure 2.9 clearly indicates a structure that is hidden in the conventional plot of residuals.

3. Conclusions
The least squares method is very much sensitive to the extreme observations. Only one observation needs to be moved towards infinity to cause an arbitrarily large change in the estimate $\hat{\beta}$. The principle of LS tries to minimize the residual sum of squares. If any outlier is present it will try to minimize the residual of this particular point, at the expense of the other observations. In that sometimes a good observation may be detected as outlier. It has nothing to do when masking is present.

In contrast to LS, robust regression procedure provides parameter estimates that are insensitive to presence of outliers. Least Median of Square (LMS) is one of the popular robust regression methods. It has 50% break down point. It not only estimates the parameters but also detects outliers. But it does not consider all the observations for parameter estimation. Therefore some good observations may not be used for parameters estimation.

Forward search reveals that outliers which were initially masked and not revealed by LS. It starts with $p$ observations by LMS and then subset is increased one by one. Here the observations are arranged according to their residuals. Therefore if any outlier is present in the data, it will enter in the forward search at the end of the search. The search that we use avoids, in the first step, the inclusion of the outliers and provides a natural ordering of the data. In this method we use a highly robust method and the same time least square estimators. The zero break down point of the least square estimators, in the context of forward search, does not turn out to be disadvantageous.

References


