Chapter 5
Proposed Regression Methodology

Introduction

In robust regression, we simplistically view each of the *n* observations as either "good" if the particular observation follows the underlying trend of the data, or "bad" when the observation does not follow the general trend. Thus, the term "bad" implies that the observation is extreme in the response variable, conditional on its regressor space location. This includes both low leverage outliers as well as high influence points. While detecting high leverage is important, it is important to remember that good leverage points are beneficial to the analysis as they can reduce an estimator's standard error.

Whenever a simple linear regression analysis is performed, addressing issues of outliers and high influence points is not a very daunting task. A simple scatterplot of the response variable versus the regressor variable illustrates graphically the nature of the data. The reliance on having a regression estimator that is able to sift out and detect potentially damaging observations is not at all crucial in this scenario. How one deals with these special observations is, off course, important, but their mere existence can be easily determined. When the regression analysis involves multiple regressor variables, the existence of outliers and high influence points is generally less evident (or even not evident at all) by a casual viewing of the data. The level of sophistication of an analysis method needs to be raised considerably in order to avoid an inferior regression analysis.

The proposed regression methodology offers a new philosophical approach to the robust regression arena. First, an initial high-breakdown regression estimator is produced via a sophisticated clustering algorithm. Second, refinement of this initial regression estimator is

investigated and possibly implemented under a carefully structured use of bounded influence (BI) regression. The rationale behind this second phase is to allow for a possible improvement in efficiency, especially when the level of data contamination does not come close to approaching 50%. The proposed method has been named *cluster-based bounded influence regression*, or CBI for short, to reflect the nature of its two-phase computation process. A typical CBI regression analysis will produce

- A high-breakdown regression estimator,
- A high-breakdown estimate of scale,
- Observation weights,
- A robust analysis of variance summary, including significance tests for the model and for each parameter,
- Classification of observations into various clusters,
    - General regression trend as main cluster; others as minor clusters,
    - Similarity matrix based on regression evaluation, not data space location,
    - Dendogram (graphical summary) of the cluster history.

Accordingly, the CBI regression analysis is offered into the robust regression arena with two primary objectives. First, the regression estimator itself is competitive with current state-of-the-art methods (to come in Chapters 7 and 8), even avoiding certain annoyances or drawbacks associated with these competing methods. Second, the nature of the CBI algorithm directly allows for the creation of a parsimonious and instructive summary regarding the data structure. Multiple outlier detection and coefficient estimation are addressed together in a single analysis. This allows for less statistically literate researchers to more easily understand the data structure and reach appropriate conclusions, beyond basic coefficient estimation and into a higher level of detailed summary.

During the development of the proposed algorithm theoretical issues were also considered. In particular, attention towards maintaining scale equivariance, regression equivariance and affine equivariance properties was made. These properties are in addition to

the 50% high-breakdown point and become topics for discussion in Chapter 6.

Before this new regression methodology is presented, however, a cursory discussion on cluster analysis as it pertains to the application at hand is offered. The CBI algorithm involves a sophisticated cluster phase where certain items and issues need to be fully understood.

§5.1 Cluster Analysis

The cluster analysis of a set of data will ultimately divide the n observations into several subsets. Every observation will be included in exactly one subset, with the goal being that each subset is homogeneous in nature, while subsets themselves are heterogeneous. In order to determine which observations are alike, and which are different, a nonnegative statistic referred to as a *similarity measure* is utilized. Each pair of observations produces a similarity measure between these two particular observations. A value of zero indicates identical observations while larger values indicate a more dissimilar relationship. Euclidean distances are often employed as similarity measures for a set of data, for example. Because the similarity measures are defined on a pairwise basis, it is convenient to let each individual measure represent one element in an $n \times n$ symmetric *similarity matrix*, whose ij^{th} element is the similarity measure between the i^{th} and j^{th} observations.

Once the similarity matrix has been established, the focus turns to the actual method of grouping (or separating) the n observations. There are two main categories of clustering that incorporate a similarity matrix:

- Agglomerative Hierarchical Clustering
- Divisive Hierarchical Clustering

In the first category, observations are initially labeled as individual clusters of size one. The most alike, i.e. having the smallest similarity measure, pair of observations is merged together. Since the elements of the similarity matrix are similarity measures for observations on a pairwise basis, there becomes a specific need to define the similarity between two clusters of observations when at least one cluster consists of multiple observations. Various cluster methods are

available, with the difference lying in the definition of the similarity measure between two clusters, referred to as the *linkage*. One method, for example, is to define the linkage as the average of the similarity measures of all pairwise comparisons across the two clusters, hence called *average-linkage*. Another method, referred to as *single-linkage*, uses the rule that the minimum similarity measure between two points in different clusters defines the similarity measure between those two clusters. A third method, referred to as *complete-linkage*, uses the rule that the maximum similarity measure between two points in different clusters defines the similarity measure between those two clusters. Generally, those involved in cluster analysis want the data described in nice, compact sphere-like clusters, with clusters distinctly different from one another. Single-linkage is not a very popular clustering algorithm in the field of cluster analysis due to the fact that single-linkage has a *chaining* property. Because single-linkage uses the minimum function, this tends to allow the growth of clusters that are elongated (referred to as a *chaining effect*) and too unusual in shape for use in cluster analysis. In regression, however, this could perhaps be viewed as a desirable property because observations at different ends of the regressor space may be part of the same general trend specified by the regression model. Elongated trends could parallel the regression itself. Complete-linkage falls at the other end of the linkage spectrum, and sometimes considered too severe. This leaves average-linkage as one possible compromise, which incorporates more similarity information between two sets of observations than does either of the other two methods mentioned here. There are several other linkage definitions available in the literature (see Johnson and Wichern (1992), Hartigan (1975)), but these three linkages are sufficient for potential use within the proposed regression algorithm.

Continuing the description of agglomerative hierarchical procedures, generally the merging continues until finally one cluster contains all n observations. The cluster identifications for each of the n observations across all n merging steps are referred to as the *cluster history*, and can be presented graphically in a *dendrogram*. Illustrations of both are provided later during the regression analysis of various case studies.

The second category, divisive hierarchical clustering, is a reversal of agglomerative hierarchical clustering. Here, initially every observation is placed in one set. The first step separates the observations into two subsets such that the observations in the first subset are different (dissimilar) from those observations in the second subset. This concept continues until all n observations are separate, i.e. clusters of size one. Again, a dendrogram may be obtained for the analysis.

One characteristic of the hierarchical procedures is that once a pair of observations is joined (or separated, in second category), they remain joined (separated). This is not the situation for all clustering methods, as the next clustering family demonstrates.

Many other clustering methods are members of the *nonhierarchical clustering* family. Here, the goal is to produce K different clusters of observations, where the scalar K is either given in advance, or somehow data-driven. There is no need for a similarity matrix. Observations are randomly separated into K clusters. Movement between clusters is allowed (necessary), where the focus is essentially on how similar a particular observation is to the cluster in which it currently resides when compared to its similarity with each of the other K-1 clusters. *K-means* is one such procedure; see Johnson and Wichern (1992) for details.

In terms of the regression problem at hand, emphasis will be placed solely on agglomerative hierarchical techniques. Nonhierarchical techniques are not practical here since the statistical model is not involved, K is unknown, and robustness to outliers is in question. By utilizing a robust multivariate location and scale methodology from among those mentioned in Chapter 3, a robust similarity matrix will be constructed. Building up to a main cluster, based on robust estimates, seems natural. Atkinson (1994) shows that, in the context of hat diagonals and leverage diagnostics, deleting observations provides no more information than that obtained by the entire set of data. Additionally, the stalactite plot also increases the active size of its subset to identify multiple outliers. Therefore, it is reasonable to keep the clustering algorithm straightforward and limit further discussion to agglomerative hierarchical techniques. This

means that regarding the cluster phase of the CBI algorithm, there are two key decisions that need to be made: the definition of the similarity matrix and the selection of the linkage type.

§5.2 The Proposed Method

As indicated previously, the CBI regression estimator is computed via an algorithm consisting of two phases, a cluster phase and a regression phase. The entire process is a compilation of a wide variety of computations assembled together to form a methodology for obtaining an efficient, high-breakdown regression estimator. A brief overview of the CBI algorithm will be provided first, followed by a more detailed-oriented computational outline. Subsequently, a philosophical discussion will be offered regarding the details of the algorithm, available options and rationale for implementing certain techniques over other alternatives.

The cluster phase begins with high-breakdown location and scale estimation of the p-dimensional regressor-response space. A special set of points, referred to as the set of *anchor points*, is computed that together represent the general trend of the data. Each observation is then characterized by the OLS regression fit that would occur if this individual observation augmented the anchor points. High-breakdown location and scale estimation of this set of OLS coefficients provides the foundation for the construction of the similarity matrix (technically, a distance matrix). Desiring a tight, compact sphere of similar coefficients exhibiting a common trend description, complete-linkage is selected and hierarchical clustering performed until an initial *main cluster* of at least *half* of the data is formed.

A simple OLS fit to this main cluster is used as the basis for the possible adjustment of the anchor set metric to more directly relate to the general trend. A revised similarity matrix is constructed, with a second cluster analysis yielding a revised, final main cluster and g *minor clusters*. The determination of this cluster classification structure completes the cluster phase.

To begin the regression phase, the initial CBI estimator is simply the OLS estimate of the main cluster observations. A high breakdown scale estimate is then computed. High breakdown

BI leverage weights are computed from the regressor space only. Using only the main cluster, a BI regression updates the initial CBI estimator. To this point, the minor clusters have not been utilized in the computation of the CBI regression estimator and their observations are said to be *inactive*. The activation process for these remaining observations has two primary stages. First, a $DFFITS_{+I}^2$ statistic is computed for each of the minor clusters, where $I = 1, 2, \ldots, g$. All minor clusters such that $DFFITS_{+I}^2$ is "small enough" are *candidates* for activation (and, furthermore, the observations comprising the remaining minor clusters will each receive a zero weight). Secondly, a single $DFFITS_{+J}^2$ statistic is computed for the entire collection, *J*, of candidate minor clusters. If $DFFITS_{+J}^2$ is "small enough", then the final CBI estimator is determined from this activation process (provided at least one minor cluster observation obtained a nonzero weight). Otherwise, the minor clusters do not play an active role (i.e. all observations possess a zero weight) and there is no further update to the current CBI regression estimator. A final CBI scale estimate is computed once the final CBI regression estimator has been determined.

Granted, this synopsis is overly simplistic. Details have been smoothed over in an effort to convey the general flow and connectivity of the various computations. The proposed CBI algorithm is now outlined in much greater detail, with a simplistic example to come in the next section that will aid in the understanding of the whole process. Justification for each step follows the example.

The Cluster Phase of the CBI algorithm

Step 1: Perform an MVE estimation of $\mathbf{Z}_{\mathbf{y}}$ (the regressor-response space). The MVE one-step improvement computation is employed (see Appendix B.4 for details).

Step 2: Determine the $(2p+1)\times p$ anchor point matrix, $\boldsymbol{\Omega}$. These points include $\mathbf{MVE}_1\left(\mathbf{Z}_{\mathbf{y}}\right)$ and the endpoints of the ellipsoid of constant distance $\chi^2_{0.975,p}$ from

$\mathbf{MVE}_1\left(\mathbf{Z}_y\right)$ based on an $\mathbf{MVE}_2\left(\mathbf{Z}_y\right)$ metric. For the i[th] axis of the ellipse, the pair of endpoints is determined by the expression $\mathbf{MVE}_1\left(\mathbf{Z}_y\right) \pm \sqrt{\lambda_i \ \chi^2_{0.975,p}} \ \mathbf{e}_i$, where λ_i is the eigenvalue that corresponds to $\mathbf{e}_i$, the i[th] eigenvector of $\mathbf{MVE}_2\left(\mathbf{Z}_y\right)$. See Johnson and Wichern (1992) for more on these details.

Step 3: Determine the $n \times p$ base regression estimator matrix, **B**. The i[th] row of **B**, denoted by the $1 \times p$ vector $\mathbf{b}_i'$, is defined as the estimator that results from an OLS regression analysis of the set of anchor points supplemented by the addition of the i[th] observation in the dataset. Perform an MVE estimation of **B**, treating each row of **B** as an observation in p dimensions.

Step 4: Using $\mathbf{MVE}_2\left(\mathbf{B}\right)$ as the distance metric, compute an $n \times n$ similarity matrix **S** whose elements are defined to be

$$s_{ij} = \left(\mathbf{b}_i - \mathbf{b}_j\right)' \left(\mathbf{MVE}_2\left(\mathbf{B}\right)\right)^{-1} \left(\mathbf{b}_i - \mathbf{b}_j\right).$$

Perform a cluster analysis on the dataset given the similarity matrix **S** and using complete-linkage to obtain the tightest cluster of $\mathbf{b}_i'$ vectors. The initial main cluster, C_0, is defined at the first instance of which a single cluster consists of at least $h = \left[(n + p + 1)/2\right]$ observations. Continued clustering after this is achieved is not necessary.

Step 5: Define the $n \times 1$ weight vector **w** as

$$w_i = \begin{cases} 1, & i \in C_0, \\ 0, & i \notin C_0. \end{cases}$$

Compute the WLS regression estimate, $\hat{\boldsymbol{\beta}}_0$. A preliminary estimate of scale, $\hat{\sigma}_0$, is defined to be the MAD *of all n residuals*, or

$$\hat{\sigma}_0 = 1.4826 \underset{\forall i}{med} \left| r_i\left(\hat{\boldsymbol{\beta}}_0\right) - \underset{\forall i}{med} \; r_i\left(\hat{\boldsymbol{\beta}}_0\right) \right|.$$

Step 6: Determine the set of observations, H, such that

$$H = \left\{ i : \left| r_i\left(\hat{\boldsymbol{\beta}}_0\right) \right| \le \hat{\sigma}_0 \, 4.685 \sqrt{2pn} \,/ (n - 2p) \right\}.$$

Define the $n \times 1$ weight vector $\boldsymbol{\omega}$ as

$$\omega_i = \begin{cases} 1, & i \in H, \\ 0, & i \notin H. \end{cases}$$

Aside: H is used as the basis for adjusting the anchor set metric. *Steps 7* to *10* mimic *Steps 1* to *4* in terms of their primary purpose. If H equals the set of observations included (i.e. those possessing a weight equal to one) during the one-step improvement for the MVE estimation (for Step 1), then the second cluster phase becomes identical to the first cluster phase.

Step 7: Using $\boldsymbol{\omega}$ and just the regressor space, compute the $p \times 1$ weighted mean vector $\mathbf{m}_H\left(\mathbf{Z}_{\mathbf{y}}\right)$ and $p \times p$ weighted covariance matrix $\mathbf{C}_H\left(\mathbf{Z}_{\mathbf{y}}\right)$. Also compute the $p \times 1$ weighted mean vector $\mathbf{m}_H\left(\mathbf{Z}\right)$ and $p \times p$ weighted covariance matrix $\mathbf{C}_H\left(\mathbf{Z}\right)$ in order to calculate (for use later in *Step 12*) the $p \times 1$ squared robust regressor distance vector, with elements defined by

$$d_i^2 = \left(\mathbf{z}_i - \mathbf{m}_H\left(\mathbf{Z}\right)\right)' \left(\mathbf{C}_H\left(\mathbf{Z}\right)\right)^{-1} \left(\mathbf{z}_i - \mathbf{m}_H\left(\mathbf{Z}\right)\right).$$

Step 8: Update the $(2p+1) \times p$ anchor point matrix, $\boldsymbol{\Omega}$ based on $\mathbf{m}_H\left(\mathbf{Z}_{\mathbf{y}}\right)$ and $\mathbf{C}_H\left(\mathbf{Z}_{\mathbf{y}}\right)$. These points include $\mathbf{m}_H\left(\mathbf{Z}_{\mathbf{y}}\right)$ and the endpoints of the ellipsoid of constant distance $\chi^2_{0.975,p}$ from $\mathbf{m}_H\left(\mathbf{Z}_{\mathbf{y}}\right)$ based on a $\mathbf{C}_H\left(\mathbf{Z}_{\mathbf{y}}\right)$ metric.

Step 9: Update the $n \times p$ base regression estimator matrix, **B**, then compute the $p \times 1$ weighted mean vector $\mathbf{m}_H(\mathbf{B})$ and $p \times p$ weighted covariance matrix $\mathbf{C}_H(\mathbf{B})$ based on the weight vector $\boldsymbol{\omega}$.

Step 10: Using $\mathbf{C}_H(\mathbf{B})$ as the distance metric, compute an $n \times n$ similarity matrix **S** whose elements are defined to be

$$s_{ij} = (\mathbf{b}_i - \mathbf{b}_j)' (\mathbf{C}_H(\mathbf{B}))^{-1} (\mathbf{b}_i - \mathbf{b}_j).$$

Perform a cluster analysis on the dataset given the similarity matrix **S** and using complete linkage. The final main cluster, C_0, is defined at the first instance of which a single cluster consists of at least h observations. The clustering process ceases at this point. The remaining observations fall into one of g minor clusters that are labeled as $C_1, C_2, \ldots, C_g$.

Step 11: Redefine the $n \times 1$ weight vector **w** as

$$w_i = \begin{cases} 1, & i \in C_0, \\ 0, & i \notin C_0. \end{cases}$$

Compute the WLS regression estimate, which becomes the initial CBI estimator, $\hat{\boldsymbol{\beta}}_1$ and subsequently the updated scale estimator

$$\hat{\sigma}_1 = 1.4826 \underset{\forall i}{med} \left| r_i(\hat{\boldsymbol{\beta}}_1) - \underset{\forall i}{med} \; r_i(\hat{\boldsymbol{\beta}}_1) \right|.$$

Step 12: Determine the $p \times 1$ BI leverage weight vector, $\boldsymbol{\pi}$, whose elements are defined as

$$\pi_i = \begin{cases} 1, & i \in C_0, \\ \min\left(1, \dfrac{\chi^2_{0.95, p-1}}{d_i^2}\right), & i \notin C_0. \end{cases}$$

Employing a bisquare ψ-function, another necessary element in the BI regression is the tuning constant, c, which is based on the entire dataset and is given by

$$c = 4.685\sqrt{2pn}\big/(n-2p).$$

Step 13: Using only the main cluster, update the CBI regression estimator. Specifically, $\hat{\boldsymbol{\beta}}_1$, $\hat{\sigma}_1$ and $\boldsymbol{\pi}_{C_0}$ (defined as the sub-vector of $\boldsymbol{\pi}$ that corresponds to C_0 observations only) are inputs into an IRLS computation (using a non-iterated scale estimate) of a BI regression estimator, producing $\hat{\boldsymbol{\beta}}_2$ at convergence. Note that the $n\times 1$ weight vector, $\boldsymbol{w}$, is updated as a result of this step.

Step 14: Let I represent any minor cluster and m_I be the size of I. Let $\boldsymbol{\pi}_{(C_0,C_I)}$ be the sub-vector set of $\boldsymbol{\pi}$ that corresponds only to C_0 and C_I observations. With $\hat{\boldsymbol{\beta}}_2$, $\hat{\sigma}_1$ and $\boldsymbol{\pi}_{(C_0,C_I)}$ as inputs into an IRLS computation (using a non-iterated scale estimate) of a BI regression estimator, $\hat{\boldsymbol{\beta}}_{+I}$ is obtained at convergence. A $DFFITS^2_{+I}$ statistic is then computed via

$$DFFITS^2_{+I} = \frac{\sum_{i=1}^{n}\left(\hat{y}_{i,+I}\left(\hat{\boldsymbol{\beta}}_{+I}\right)-\hat{y}_i\left(\hat{\boldsymbol{\beta}}_2\right)\right)^2}{m_I\ \ \hat{\sigma}_1^2},$$

where $\hat{y}_{i,+I}\left(\hat{\boldsymbol{\beta}}_{+I}\right)$ represents fits when using both C_0 and C_I observations and $\hat{y}_i\left(\hat{\boldsymbol{\beta}}_2\right)$ represents fits when using just C_0 observations (from *Step 13*). This statistic is computed for each of the *g* minor clusters.

Step 15: Define the scalar δ that represents the maximum allowable $DFFITS^2_{+I}$ statistic. To be consistent with the usual critical value for a difference in fits statistic, the recommended (default) value is $\delta = 4$. Then, let *J* represent the union of all activation candidate minor sets, i.e.

$$J = \bigcup_{\forall I} C_I \,\Big|\, \left(DFFITS^2_{+I} \le \delta \text{ and } \exists_{i\in I}\,|\, w_i > 0\right),$$

where the condition $\exists_{i \in I} \left| w_i > 0 \right.$ implies that C_I possesses at least one observation of some value to the regression. This rule is used to offset the effect of essentially iterating the scale and altering $\hat{\boldsymbol{\beta}}_2$ (as IRLS would likely alter $\boldsymbol{w}$) by the activation of non-contributing (zero weight) observations. Provided that $J \neq \varnothing$, then with $\hat{\boldsymbol{\beta}}_2$, $\hat{\sigma}_1$ and $\boldsymbol{\pi}_{(C_0, C_J)}$ as inputs into an IRLS computation (using a non-iterated scale estimate) of a BI regression estimator $\hat{\boldsymbol{\beta}}_{+J}$ is obtained at convergence. Subsequently, the $DFFITS^2_{+J}$ statistic is computed.

Step 16: The final cluster-based bounded influence (CBI) regression estimator is defined as

$$\hat{\boldsymbol{\beta}}_{CBI} = \begin{cases} \hat{\boldsymbol{\beta}}_{+J}, & \text{if } \left(DFFITS^2_{+J} \leq \delta \text{ and } \exists_{j \in J} \left| w_j > 0 \right. \right) \left| J \neq \varnothing, \right. \\ \hat{\boldsymbol{\beta}}_2, & \text{otherwise.} \end{cases}$$

As in *Step 15*, the condition $\exists_{j \in J} \left| w_j > 0 \right.$ protects against estimator drift due to unnecessary scale iteration. The CBI scale estimate is then updated as

$$\hat{\sigma}_{CBI} = 1.4826 \underset{\forall i}{med} \left| r_i \left(\hat{\boldsymbol{\beta}}_{CBI} \right) - \underset{\forall i}{med} \; r_i \left(\hat{\boldsymbol{\beta}}_{CBI} \right) \right|.$$

Section 5.6 will provide a discussion on the recommended choice regarding the final CBI scale estimate for use in inferential statistical analyses, which is based on $\hat{\sigma}_{CBI}$. The final CBI weights for the individual observations are simply the observation weights at convergence of the IRLS computation used to compute $\hat{\boldsymbol{\beta}}_{CBI}$, arising from either $\hat{\boldsymbol{\beta}}_2$ or $\hat{\boldsymbol{\beta}}_{+J}$, as the case may be.

§5.3 Detailed Simple Example

To assist in the understanding of the CBI algorithm and the working mechanisms contained within both phases, consider the generic simple linear regression example dataset in Table 5.1, also shown in Figure 5.1. The regressor variable was randomly generated as

Table 5.1: Simple linear regression example.

Ob.	Y	X	Ob.	Y	X
1	167.611	9.4826	7	137.416	14.7705
2	162.155	9.7717	8	141.657	13.5850
3	154.880	11.3759	9	134.214	15.4460
4	144.862	12.3826	10	153.800	29.0000
5	148.578	14.7535	11	155.800	30.0000
6	144.966	15.1601	12	80.932	31.0000

$$x_i = \begin{cases} U[8,18], & i \le 9, \\ 29, & i = 10, \\ 30, & i = 11, \\ 31, & i = 12. \end{cases}$$

Then, the response variable was randomly generated according to

$$y_i = \begin{cases} 153.8, & i = 10, \\ 155.8, & i = 11, \\ 200 - 4\ x_i + \varepsilon_i, & otherwise, \end{cases}$$

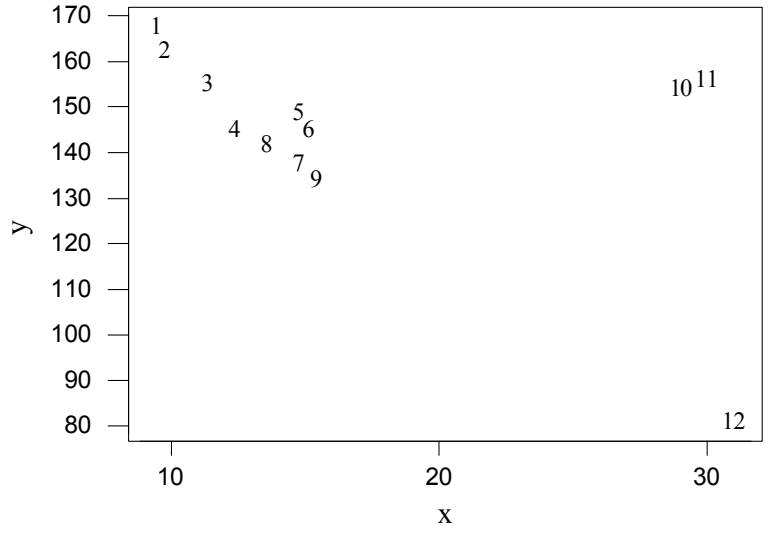

Figure 5.1: Scatterplot of Example 1 data, which contains a high influence cluster of two observations (10 and 11) and also a good leverage point (12).

with $\varepsilon_i \sim N[0,\ 25]$. The data structure clearly indicates the presence of two high influence points (labeled 10 and 11) and one good leverage point (12). The OLS fit using the data subset $\{1:9, 12\}$ is

$$\hat{\boldsymbol{\beta}}'_{OLS} = [198.146 \quad -3.819],$$

with $\hat{\sigma}_{OLS} = \sqrt{MSE} = 5.048$. These subset OLS estimates reflect the true underlying parameters aside from the variability associated due to random generation. The CBI regression analysis of this dataset is presented in step-by-step form.

Step 1: The MVE estimation of $\mathbf{Z}_y$ is

$$\mathbf{MVE}_1(\mathbf{Z}_y) = \begin{bmatrix} 12.970 \\ 148.482 \end{bmatrix}$$

and

$$\mathbf{MVE}_2(\mathbf{Z}_y) = \begin{bmatrix} 5.3756 & -23.1869 \\ -23.1869 & 123.9895 \end{bmatrix}.$$

It is noted that these one-step MVE estimates utilize observations $\{1:9\}$ in their computation.

Step 2: The eigenvectors and eigenvalues for $\mathbf{MVE}_2(\mathbf{Z}_y)$ are

$$\mathbf{E} = [\mathbf{e}_1 \quad \mathbf{e}_2] = \begin{bmatrix} -0.1853 & 0.9827 \\ 0.9827 & 0.1853 \end{bmatrix}$$

and

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 128.361 \\ 1.004 \end{bmatrix},$$

respectively. With $\chi^2_{0.975,2} = 7.378$, the anchor set matrix, $\boldsymbol{\Omega}$, is then computed as

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{MVE}_1\left(\mathbf{Z}_y\right)' \\ \mathbf{MVE}_1\left(\mathbf{Z}_y\right)' - \sqrt{\lambda_1 \ \chi^2_{0.975,2}} \ \mathbf{e}_1' \\ \mathbf{MVE}_1\left(\mathbf{Z}_y\right)' + \sqrt{\lambda_1 \ \chi^2_{0.975,2}} \ \mathbf{e}_1' \\ \mathbf{MVE}_1\left(\mathbf{Z}_y\right)' - \sqrt{\lambda_2 \ \chi^2_{0.975,2}} \ \mathbf{e}_2' \\ \mathbf{MVE}_1\left(\mathbf{Z}_y\right)' + \sqrt{\lambda_2 \ \chi^2_{0.975,2}} \ \mathbf{e}_2' \end{bmatrix} = \begin{bmatrix} 12.9698 & 148.4820 \\ 18.6712 & 118.2411 \\ 7.2684 & 178.7229 \\ 10.2951 & 147.9777 \\ 15.6445 & 148.9863 \end{bmatrix}.$$

Figure 5.2 displays the relationship of the anchor points and the original data. It is seen how the major axis of the ellipse has anchor points that are extended in the direction of the general regression trend.

Step 3: The collection of OLS anchor set regression estimates is shown in Table 5.2, i.e. matrix **B**. It is clear from this table that observations 10 and 11 are similar to each other, yet dramatically different from the rest of the estimates.

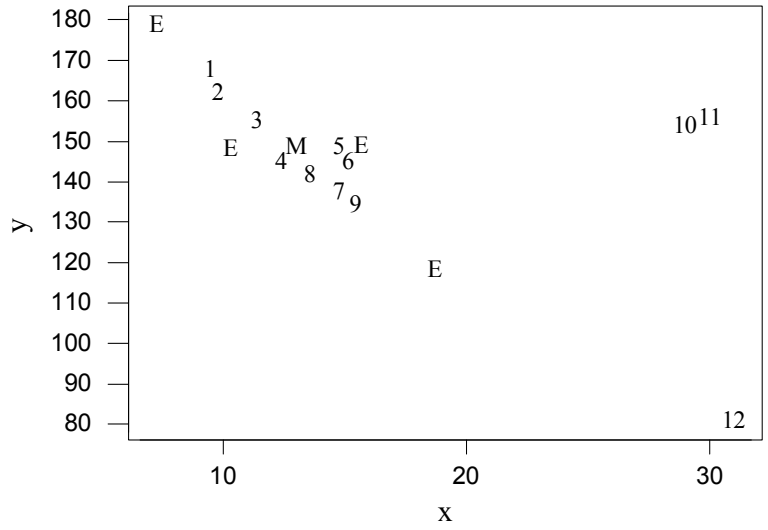

Figure 5.2: Illustration of the anchor point locations (first cluster phase of the CBI algorithm) for the Section 5.3 example dataset.

*Table 5.2: The 12×2 **B** matrix; the OLS regression estimators from anchor set regressions.*

Ob.	Intercept	Slope	Ob.	Intercept	Slope
1	206.751	-4.4461	7	204.676	-4.3737
2	204.359	-4.3096	8	204.081	-4.3402
3	204.247	-4.3055	9	205.000	-4.4010
4	202.913	-4.2755	10	163.818	-0.9238
5	203.849	-4.1721	11	161.436	-0.7423
6	203.681	-4.1834	12	199.124	-3.8748

Table 5.3: Similarity matrix using $s_{ij} = \left(\mathbf{b}_i - \mathbf{b}_j\right)' \left(\mathbf{MVE}_2\left(\mathbf{B}\right)\right)^{-1} \left(\mathbf{b}_i - \mathbf{b}_j\right)$.

s_{ij}	1	2	3	4	5	6	7	8	9	10	11	12
1	0	5.146	5.671	14.840	9.314	9.291	4.948	7.644	4.086	1751	1945	52.320
2	5.146	0	0.014	2.893	3.280	2.281	0.614	0.743	1.040	1584	1769	26.290
3	5.671	0.014	0	2.508	3.388	2.314	0.610	0.611	1.089	1578	1761	25.440
4	14.840	2.893	2.508	0	8.476	6.274	2.818	1.239	3.863	1509	1688	19.010
5	9.314	3.280	3.388	8.476	0	0.167	6.710	6.644	7.775	1509	1690	19.630
6	9.291	2.281	2.314	6.274	0.167	0	5.258	4.944	6.344	1505	1686	18.220
7	4.948	0.614	0.610	2.818	6.710	5.258	0	0.320	0.102	1625	1811	32.000
8	7.644	0.743	0.611	1.239	6.644	4.944	0.320	0	0.741	1585	1769	26.950
9	4.086	1.040	1.089	3.863	7.775	6.344	0.102	0.741	0	1651	1838	35.700
10	1751	1584	1578	1509	1509	1505	1625	1585	1651	0	5.143	1202
11	1945	1769	1761	1688	1690	1686	1811	1769	1838	5.143	0	1364
12	52.320	26.290	25.440	19.010	19.630	18.220	32.000	26.950	35.700	1202	1364	0

Step 4: MVE estimation of **B** produces

$$\mathbf{MVE}_1\left(\mathbf{B}\right) = \begin{bmatrix} 204.3950 \\ -4.3119 \end{bmatrix}$$

and

$$\mathbf{MVE}_2\left(\mathbf{B}\right) = \begin{bmatrix} 1.1406 & -0.0752 \\ -0.0752 & 0.00850 \end{bmatrix}.$$

Using $\mathbf{MVE}_2\left(\mathbf{B}\right)$ as a distance metric, Table 5.3 provides the similarity matrix, **S**. The main cluster is required to contain at least $h = \left[(n + p + 1)/2\right] = 7$ observations. The cluster history for a complete-linkage clustering of **S** is displayed in Table 5.4, with Figure 5.3 providing a dendrogram representation of

Table 5.4: Cluster history for example data.

Step	1	2	3	4	5	6	7	8	9	10	11	12
0	1	2	3	4	5	6	7	8	9	10	11	12
1	1	2	2	4	5	6	7	8	9	10	11	12
2	1	2	2	4	5	6	7	8	7	10	11	12
3	1	2	2	4	5	5	7	8	7	10	11	12
4	1	2	2	4	5	5	7	7	7	10	11	12
5	1	2	2	4	5	5	2	2	2	10	11	12
6	1	2	2	2	5	5	2	2	2	10	11	12
7	1	2	2	2	5	5	2	2	2	10	10	12
8	***1***	***2***	***2***	***2***	***2***	***2***	***2***	***2***	***2***	***10***	***10***	***12***
9	1	1	1	1	1	1	1	1	1	10	10	12
10	1	1	1	1	1	1	1	1	1	10	10	1
11	1	1	1	1	1	1	1	1	1	1	1	1

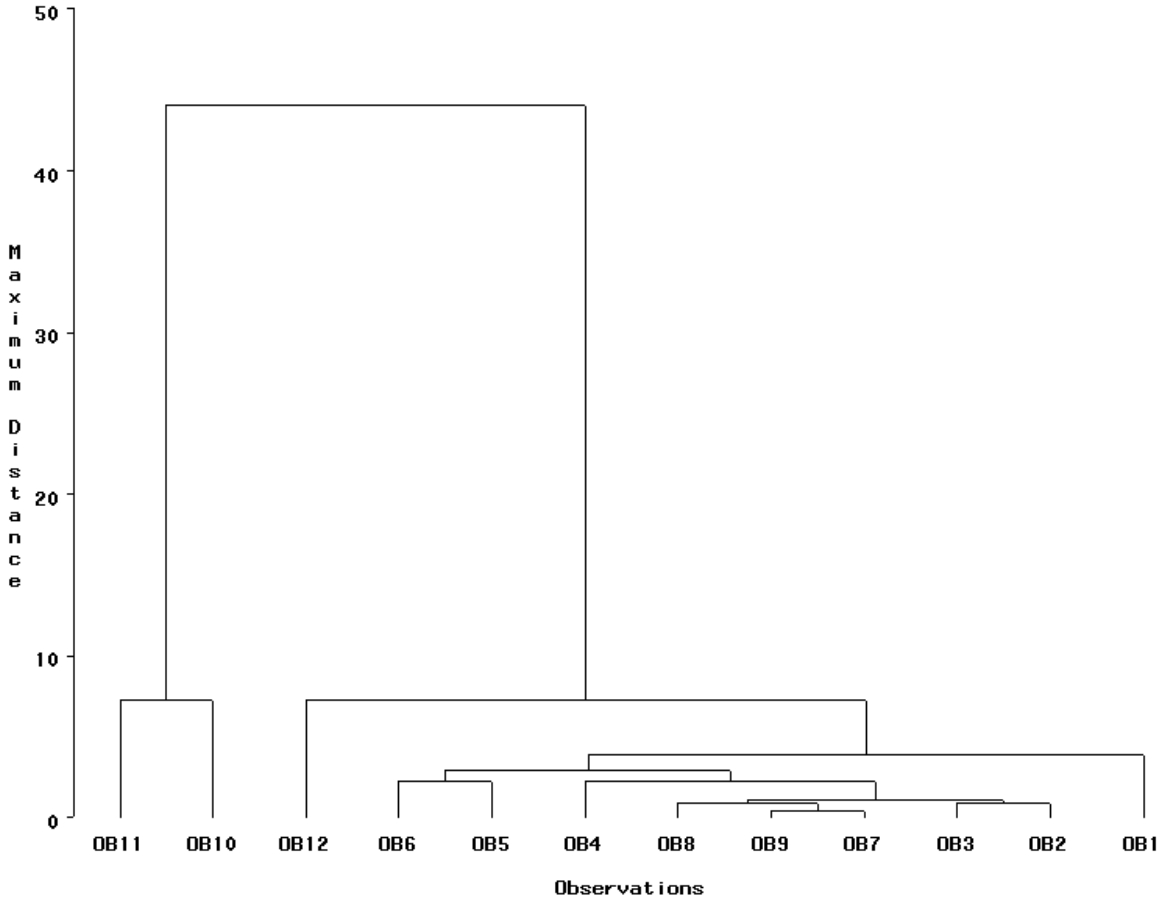

Figure 5.3: Dendrogram for initial clustering of Section 5.3 example dataset.

the clustering process. After a total of 8 steps the main cluster is defined as $C_0 = \{2:9\}$. The classification of other observations is immaterial at this time.

Step 5: The OLS estimate based solely on the seven main cluster observations is

$$\hat{\boldsymbol{\beta}}_0 = \begin{bmatrix} 196.5825 & -3.7664 \end{bmatrix}'.$$

The scale estimate from this regression fit is computed across all eleven residuals and becomes $\hat{\sigma}_0 = 8.0168$.

Step 6: $H = \{1:9, 12\}$ and the weight vector $\boldsymbol{\omega}$ becomes

$$\boldsymbol{\omega}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

Step 7: Being essentially an update to *Step 1*, the data are now characterized by

$$\mathbf{m}_H\left(\mathbf{Z}_{\mathrm{y}}\right) = \begin{bmatrix} 14.7728 \\ 141.7270 \end{bmatrix}$$

and

$$\mathbf{C}_H\left(\mathbf{Z}_{\mathrm{y}}\right) = \begin{bmatrix} 37.2873 & -142.4039 \\ -142.4039 & 566.5070 \end{bmatrix}.$$

Step 8: Being essentially an update to *Step 2*, the eigenvectors and eigenvalues for $\mathbf{C}_H\left(\mathbf{Z}_{\mathrm{y}}\right)$ are

$$\mathbf{E} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix} = \begin{bmatrix} -0.2444 & 0.9697 \\ 0.9697 & 0.2444 \end{bmatrix}$$

and

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 602.3921 \\ 1.4021 \end{bmatrix},$$

respectively. The updated anchor set matrix, $\boldsymbol{\Omega}$, becomes

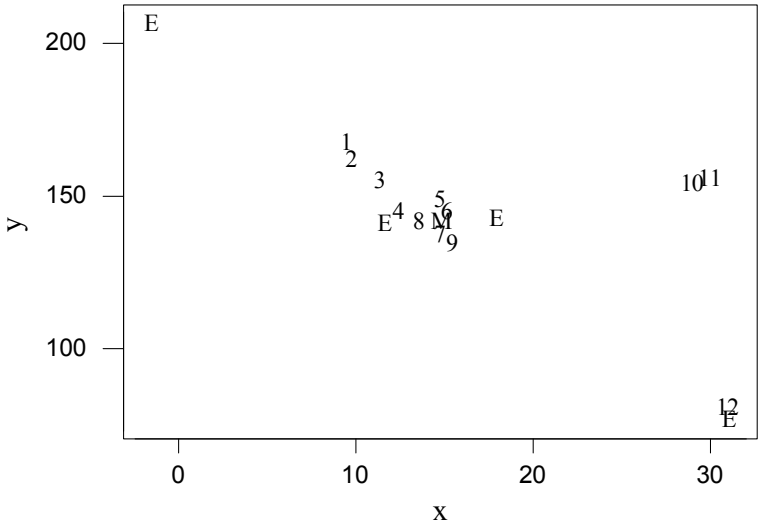

Figure 5.4: Updated anchor points in relation to the original dataset.

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{m}_H\left(\mathbf{Z}_\mathbf{y}\right)' \\ \mathbf{m}_H\left(\mathbf{Z}_\mathbf{y}\right)' - \sqrt{\lambda_1\ \chi^2_{0.975,2}}\ \mathbf{e}_1' \\ \mathbf{m}_H\left(\mathbf{Z}_\mathbf{y}\right)' + \sqrt{\lambda_1\ \chi^2_{0.975,2}}\ \mathbf{e}_1' \\ \mathbf{m}_H\left(\mathbf{Z}_\mathbf{y}\right)' - \sqrt{\lambda_2\ \chi^2_{0.975,2}}\ \mathbf{e}_2' \\ \mathbf{m}_H\left(\mathbf{Z}_\mathbf{y}\right)' + \sqrt{\lambda_2\ \chi^2_{0.975,2}}\ \mathbf{e}_2' \end{bmatrix} = \begin{bmatrix} 14.7728 & 141.7270 \\ 31.0630 & 77.0824 \\ -1.5174 & 206.3717 \\ 11.6540 & 140.9411 \\ 17.8916 & 142.5130 \end{bmatrix}.$$

Figure 5.4 displays the relationship of the updated anchor points and the original data. The general trend is more pronounced here than it was in Figure 5.2.

Step 9: Being essentially an update to *Step 3*, the updated **B** matrix is displayed in Table 5.5. Observations 10 and 11 still distinguish themselves from the remaining observations.

Step 10: Being essentially an update to *Step 4*, the weighted mean vector and covariance matrix for **B** are

$$\mathbf{m}_H(\mathbf{B}) = \begin{bmatrix} 198.1458 \\ -3.8197 \end{bmatrix}$$

and

$$\mathbf{C}_H(\mathbf{B}) = \begin{bmatrix} 0.8873 & -0.0116 \\ -0.0116 & 0.00033 \end{bmatrix},$$

respectively. Using $\mathbf{C}_H(\mathbf{B})$ as a distance metric, Table 5.6 provides the updated similarity matrix, **S**. The main cluster is required to contain at least $h = 7$ observations. The cluster history for a complete-linkage clustering of **S** is displayed in Table 5.7, dendrogram in Figure 5.5. After a total of 8 steps the main cluster is defined as $C_0 = \{2-9, 12\}$. There are $g = 3$ minor clusters, $C_1 = \{1\}$, $C_2 = \{10\}$ and $C_3 = \{11\}$. Observations 10 and 11 have not yet merged primarily due to the fact that the difference in their slopes (**B** matrix) is much greater than any difference amongst slopes corresponding to observations in C_0.

*Table 5.5: The updated 12×2 **B** matrix; the OLS regression estimators from updated anchor set regressions.*

Ob.	Intercept	Slope	Ob.	Intercept	Slope
1	199.699	-3.8628	7	197.426	-3.8191
2	198.502	-3.8288	8	197.258	-3.8108
3	198.189	-3.8200	9	197.397	-3.8241
4	196.838	-3.7976	10	190.437	-2.7239
5	199.278	-3.8193	11	188.842	-2.5863
6	198.891	-3.8163	12	197.980	-3.7984

Table 5.6: The updated 12×12 similarity matrix using $s_{ij}=\left(\mathbf{b}_i-\mathbf{b}_j\right)'\left(\mathbf{C}_H\left(\mathbf{B}\right)\right)^{-1}\left(\mathbf{b}_i-\mathbf{b}_j\right)$.

s_{ij}	1	2	3	4	5	6	7	8	9	10	11	12
1	0	3.473	5.501	13.490	8.203	7.875	6.897	8.928	6.370	5850	7281	13.060
2	3.473	0	0.232	3.613	2.827	1.882	1.414	1.762	1.913	5621	7025	3.393
3	5.501	0.232	0	2.181	2.586	1.477	1.114	1.028	1.874	5564	6960	2.032
4	13.490	3.613	2.181	0	7.298	5.121	1.447	0.532	2.405	5490	6876	2.587
5	8.203	2.827	2.586	7.298	0	0.194	7.085	6.403	8.818	5420	6801	1.983
6	7.875	1.882	1.477	5.121	0.194	0	5.098	4.412	6.680	5436	6818	1.136
7	6.897	1.414	1.114	1.447	7.085	5.098	0	0.235	0.164	5652	7058	4.692
8	8.928	1.762	1.028	0.532	6.403	4.412	0.235	0	0.754	5582	6979	3.256
9	6.370	1.913	1.874	2.405	8.818	6.680	0.164	0.754	0	5712	7125	6.578
10	5850	5621	5564	5490	5420	5436	5652	5582	5712	0	78.550	5354
11	7281	7025	6960	6876	6801	6818	7058	6979	7125	78.550	0	6725
12	13.060	3.393	2.032	2.587	1.983	1.136	4.692	3.256	6.578	5354	6725	0

Table 5.7: The updated cluster history for example data.

Step	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
0	1	2	3	4	5	6	7	8	9	10	11	12
1	1	2	3	4	5	6	7	8	7	10	11	12
2	1	2	3	4	5	5	7	8	7	10	11	12
3	1	2	2	4	5	5	7	8	7	10	11	12
4	1	2	2	4	5	5	7	4	7	10	11	12
5	1	2	2	4	5	5	2	4	2	10	11	12
6	1	2	2	4	5	5	2	4	2	10	11	5
7	1	2	2	2	5	5	2	2	2	10	11	5
8	***1***	***2***	***2***	***2***	***2***	***2***	***2***	***2***	***2***	***10***	***11***	***2***
9	1	1	1	1	1	1	1	1	1	10	11	1
10	1	1	1	1	1	1	1	1	1	10	10	1
11	1	1	1	1	1	1	1	1	1	1	1	1

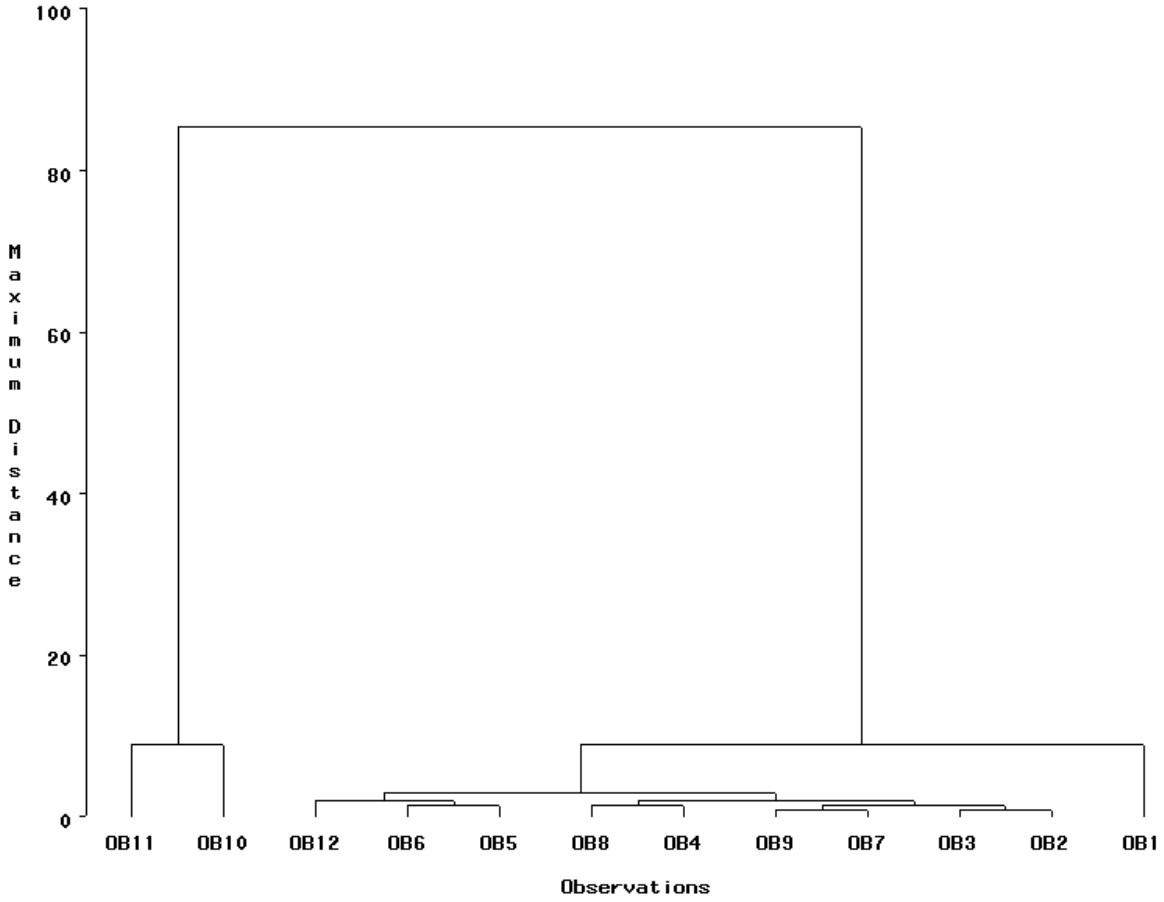

Figure 5.5: Dendrogram for final clustering of Section 5.3 dataset.

Step 11: Based on C_0, the weight vector **w** is

$$\mathbf{w}' = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

The OLS estimator based solely on the nine main cluster observations is

$$\hat{\boldsymbol{\beta}}_1' = \begin{bmatrix} 195.8304 & -3.7094 \end{bmatrix}.$$

The scale estimate from this regression fit is computed across all twelve residuals and is $\hat{\sigma} = 8.2277$.

Step 12: The regressor space is characterized by

$$\mathbf{m}_H(\mathbf{Z}) = [14.7728]$$

and

$$\mathbf{C}_H(\mathbf{Z}) = [37.2873],$$

which are used to determine the BI leverage weight vector $\boldsymbol{\pi}$, where

$$\boldsymbol{\pi}' = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0.708 \quad 0.618 \quad 1].$$

Also, the BI regression tuning constant is $c = 4.685\sqrt{2pn}/(n-2p) = 4.057$.

Step 13: At convergence, the IRLS solution (using a non-iterated scale estimate) of the BI regression estimator for C_0 is

$$\hat{\boldsymbol{\beta}}_2' = [195.7777 \quad -3.7099].$$

Step 14: To judge the merit of each minor cluster with respect to the general trend, computations yield $DFFITS_{+1}^2 = 0.1329$ (with $w_1 = 0.9424$), $DFFITS_{+2}^2 = 0$ (with $w_{10} = 0$) and $DFFITS_{+3}^2 = 0$ (with $w_{11} = 0$).

Step 15: Let $\delta = 4$. It is clear that $J = C_1 = \{1\}$ as only $DFFITS_{+1}^2 \le \delta$. The activation of a single minor cluster yields $DFFITS_{+J}^2 = DFFITS_{+1}^2 \le \delta$. Since $w_1 = 0.9424$ (from *Step 14*), the CBI estimator is obtained from this activation step.

Table 5.8: Final observation weights for the CBI regression estimator.

Ob.	w_i	Ob.	w_i
1	0.942	7	0.967
2	0.997	8	0.963
3	1.000	9	0.957
4	0.937	10	0
5	0.919	11	0
6	0.960	12	0.998

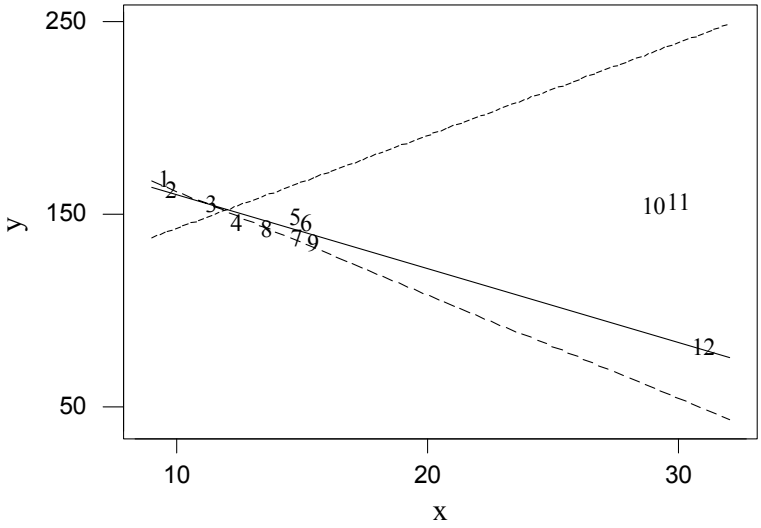

Figure 5.6: CBI regression fit (solid line) for the Section 5.3 example dataset. Also shown are the LTS fit (dashed line) and the S1S fit (dotted line). The M1S fit is identical to the S1S fit. The OLS fit (omitting observations 10 and 11) is virtually identical (graphically indistinguishable) to the CBI fit.

Step 16: The final cluster-based bounded influence (CBI) regression estimator is

$$\hat{\boldsymbol{\beta}}'_{CBI} = \hat{\boldsymbol{\beta}}'_2 = \begin{bmatrix} 198.0978 & -3.8170 \end{bmatrix}.$$

Based on $\hat{\boldsymbol{\beta}}_{CBI}$, the final CBI scale estimate is computed using all twelve residuals as $\hat{\sigma}_{CBI} = 8.2277$. The final CBI weights for the individual observations (see Table 5.8) are the weights at convergence of the IRLS computation used to compute $\hat{\boldsymbol{\beta}}_2$ (for this example).

The scatterplot of Figure 5.6 illustrates the CBI regression fit (solid line) with respect to the original data. This simplistic example has displayed the ability of the CBI algorithm to detect a high influence cluster of size 2 and eliminate its effect on the quality of the regression fit. In addition, observation 12 (and its good leverage) was utilized nicely. Figure 5.6 also illustrates the LTS (dashed line) and S1S (dotted line) regression fits for this dataset, noting that M1S was identical to S1S. The actual regression estimators are given by

$$\hat{\boldsymbol{\beta}}'_{LTS} = \begin{bmatrix} 215.6378 & -5.3755 \end{bmatrix}$$

and

$$\hat{\boldsymbol{\beta}}'_{SIS} = \hat{\boldsymbol{\beta}}'_{MIS} = \begin{bmatrix} 94.6080 & 4.8136 \end{bmatrix}.$$

While LTS followed the tightness of the trend given by seven observations $\{1:4, 7:9\}$, S1S became misguided altogether. CBI was able to offer an improvement over LTS by recognizing that more than half of the observations would follow a similar general trend.

The inferential side of the CBI regression analysis for this example will be provided later in Section 5.6.

§5.4 The CBI Algorithm Philosophy

The technical and computational details of the CBI algorithm evolved during the development process. However, the general philosophy and intent have remained steadfast. The goal was to take an efficient low-breakdown point method, BI regression, and improve the breakdown point while not making a huge sacrifice regarding efficiency. The purpose was to offer an alternative to the currently available high-breakdown point methods, methods that do have drawbacks that were previously discussed in Chapter 4.

The CBI philosophy has always been to attack the problem in two stages. The first stage would involve the use of clustering observations to determine a main cluster that would ultimately produce initial estimates for the regression coefficients as well as for scale. To improve efficiency, the remaining observations would be added "intelligently" and the regression fit updated. A wide variety of options are available to formulate an algorithm that meets these broad, general guidelines. However, there are critical details and important nuances contained within the proposed algorithm that due to their apparent subtlety might not be recognized as terribly important. For this reason, the development of the algorithm and a brief discussion on possible options is offered.

The primary points of concern regarding the development of the CBI algorithm are

- The definition of the similarity matrix,
- The clustering procedure to invoke and
- The rules governing the sequential addition of minor clusters.

The foundation of the CBI algorithm is the definition of the similarity matrix. The performance of the CBI estimator hinges on the quality of this matrix and its ability to discriminate observations according to general trend tendencies.

§5.4.1 The Cluster Phase of the CBI Algorithm

Originally, the CBI algorithm was drafted by noting that multiple outliers and joint influence appeared to have a structure within the altered hat matrix that could be exploited by clustering. Namely, the columns (or rows, by symmetry) corresponding to a group of similarly outlying observations (a multiple outlier set) were themselves similar while simultaneously being dramatically different from the general trend data. However, observations in vastly different regions of the response-regressor space may be jointly influential, but this fact is not detectable by this clustering procedure. In such cases, joint influence would be combated during the regression phase by viewing clusters individually or sequentially.

Of course, the altered hat matrix is a mean-based numerical entity, with the inherent low-breakdown concerns that multiple outliers could potentially overwhelm the matrix and destroy the structure that allows for the reasonable discrimination of the observations. Even though the use of this matrix produced nice results in a variety of case studies, a truly robust modification of the clustering procedure was investigated and eventually incorporated into the CBI algorithm.

Another subtlety of the cluster phase is that under the original altered hat matrix based procedure, the single-linkage method was employed for a very important reason. Clustering the altered hat matrix means clustering observations based upon their location in the response-regressor space, NOT by a measure of goodness towards some general trend. As a consequence,

good leverage points are at a distinct disadvantage. The chaining property of single-linkage was to be utilized to enable the formation of the main cluster to elongate and follow the general trend of the data, pulling in good leverage points in the process. Yet, often, good leverage points remained outside the main cluster and needed to rely on the sequential cluster phase to become involved in a useful manner. Without good leverage points, an initial regression estimator is left vulnerable to internal instability if there is a large bulk of the data in a small region of the response-regressor space (due to higher variances in the estimated coefficients). Thus, a similarity matrix that fairly treats good leverage points would also lead to an improvement in the overall performance of the CBI estimator.

§5.4.1.1 Clustering Foundation

The discrimination of observations according to how well each follows the general trend is the primary purpose of the similarity matrix. The ability to provide insight into subsets of data that do not follow the general trend, but themselves each suggest a different trend, is also a valuable trait. There were two general approaches taken towards the development of the similarity matrix.

The first (and original) approach dealt with an observation's location in the response-regressor space and its relationship towards the other observations. The regression model is not explicitly used to draw a similarity structure. Instead, this approach relies on the distance metric used to mimic the general trend. Linkage is selected in order to follow a potentially elongated set of observations defining the general trend.

The classical dispersion matrix is the sample covariance matrix, or in terms of distances, the sample SSCP (sums of squares and cross-products) matrix. Its relevance to this discussion is seen by the following result. As stated previously in Section 1.3.3, the $n \times n$ *altered hat matrix* is found by $\mathbf{H}_y = \mathbf{X}_y(\mathbf{X}_y'\mathbf{X}_y)^{-1}\mathbf{X}_y'$. In very well behaved data, the diagonal elements are leverage measures, with off-diagonals representing joint position. As noted by Gray and Ling (1984), patterns of diagonal blocks of large values indicate separate clusters of observations.

Additionally, off-diagonal elements between rows defined by some block diagonal are similar in magnitude. Therefore, similar observations would have similar rows, and their differences would hover around zero. Dissimilar observations would have a larger difference, in magnitude. So, consider each row of $\mathbf{H}_y$ as a coordinate vector, or as a transformed observation. Then, use the squared Euclidean distance between pairs of rows to define the similarity matrix, **S**, whose elements are defined as

$$s_{ij} = \sum_{l=1}^{n} (h_{y,il} - h_{y,jl})^2 .$$

This similarity measure simplifies to

$$s_{ij} = (\mathbf{x}_{y,i} - \mathbf{x}_{y,j})'(\mathbf{X}_y'\mathbf{X}_y)^{-1}(\mathbf{x}_{y,i} - \mathbf{x}_{y,j}),$$

which shows that the Euclidean distance needs to be rescaled by the data metric, or SSCP matrix, in order to be appropriate for linear models. The problem is that the SSCP matrix is not resistant to unusual observations. Now the summation-based statistic for s_{ij} could be replaced by a variety of other statistics, such as the median squared difference, minimum squared difference, maximum squared difference, or sum of smallest *h* squared differences. Of course, the Euclidean distance would be proportional to the average squared difference so no practical difference exists here. Some preliminary simulation studies of these various similarity measures indicated that very poor results could be obtained when either the minimum squared difference or the maximum squared difference is used. It appears that not enough information is being used by these criteria to be of practical use. The median squared difference is also eliminated from consideration due to mediocre performance. The Euclidean distance worked fairly well, especially in conjunction with the non-robust altered hat matrix, in certain case studies. When using a robust distance matrix, however, low leverage outliers can be classified with good observations due to their central location. An improvement was reached when the sum of the smallest *h* squared distances was used instead. Low leverage outliers may not be extreme in any of the *p* dimensions. Due to this central location, it becomes easier for these observations to blend in with the good data when all *n* pairwise differences are incorporated. This is because good leverage observations have larger distances to observations residing in "opposite sides" of the regressor space. Using only the smallest *h* of the squared differences keeps good

observations from being unduly penalized because of their location in the regressor space. Furthermore, Rousseeuw and van Zomeren (1990) state that the hat diagonals really should not be considered as leverage statistics, given their high sensitivity to unusual data, and recommend the use of robust Mahalanobis distances. Therefore, other, more robust, matrices are investigated.

To improve the robustness of the similarity matrix the $\mathbf{X}_y'\mathbf{X}_y$ matrix may be replaced by another $p \times p$ matrix that captures the general shape of the data, but is not affected by extreme observations. Alternatively, one could replace the hat diagonals with robust distances, due to the monotonic relationship that exists between them. This brings the focus back to the various multivariate location and dispersion estimators that were listed and discussed in Chapter 3. One additional procedure, not mentioned earlier, is now considered.

First, consider the usual altered hat matrix. It is computationally simple and has nice equivariance properties, a topic of interest in Chapter 6. Suppose that the altered hat diagonals are ranked, and the median determined. A new, trimmed version of the altered hat matrix can then be formed by using those observations whose diagonals were less than or equal to the median diagonal. Given this new altered hat matrix, a new median diagonal can be found, and so on. When the set of active observations remains unchanged on consecutive iterations, convergence is achieved. This last trimmed altered hat matrix becomes the shape matrix used in the computation of the similarity matrix. It turns out that this procedure was previously investigated in the literature. Rousseeuw and Leroy (1987) refer to it as *Iterative Trimming*, and give the references Gnanadesikan and Kettering (1972) and Devlin, et. al. (1975). Furthermore, they report that Donoho (1982) reasons that the breakdown point for the location estimator "is at most about $1/p$". Without a high breakdown point, this procedure is no longer considered.

A straightforward adjustment to the Iterative Trimming procedure would be incorporate robust Mahalanobis distances that are based on coordinatewise medians. Proceed with the trimming based on the median of this set of robust distances. This method would almost

assuredly attain some low breakdown point, but whether large quantities of extreme data could be handled is not clear. This then becomes very similar to the forward search of Hadi (1992), except that the active subset remains the same size. The breakdown point of this procedure is very much in question; it is probably moderate at best. Regardless, another drawback is that equivariance problems occur when a multivariate problem is viewed coordinatewise. Another potential source of difficulty involves the convergence of this iterative scheme. For example, is it possible for one subset to provide trimming that produces a second subset, but the trimming from this subset leads back to the first subset? Perhaps this cycle involves more than two subsets. This procedure, having several potentially damaging aspects, is eliminated from further discussion.

Creating an outlier resistant similarity matrix could also be developed with either of two high breakdown point methodologies introduced in Chapter 3. The first method is the Stahel-Donoho estimator. This projection-based estimator is very computationally cheap when using the set of n directional vectors suggested by Rousseeuw and van Zomeren (1990). Regression equivariance is not attained, although it is likely that clustering will often smooth out these departures from true regression equivariance. The second method from Chapter 3 considered for use in the CBI cluster phase is the MCD estimator. Recall that the exact MCD algorithm is too computationally exhaustive for consideration. However, the FSA approach for obtaining the exact MCD with "high probability" is available. More importantly, unlike the subsampling methods used for the calculation of LTS, subsampling variability has not been observed when using the FSA for the MCD computation. This means that while random subsampling is used, there is virtually no effect on repeatability provided enough random starts are employed. Furthermore, this number of starts is significantly smaller than the number of random subsets needed to obtain LTS.

Therefore, there are viable options to attain a high breakdown property with a similarity matrix based on response-regressor space location. The major drawback to such an approach, however, lies squarely on the handling of good leverage points. Cluster formation that extends

to reach such observations may not occur, which allows more variability to drift into the formation of the initial estimator. For a more stable initial estimator, good leverage points should not be unduly penalized. They are invaluable to the regression analysis.

The second approach in the development of a similarity matrix dealt with a direct use of the regression model to discriminate between observations. The idea is that a manageable number of trial regressions are performed and the results of these regressions are then formulated into some measure of similarity. The simplest method (conceptually) has been touched on before. For a simple linear regression problem, all pairwise OLS regressions could be performed. In multiple linear regression this expands to OLS regressions on all possible *p*-subsets. Clearly, this is not computationally feasible as either *n* or *p* increases. Instead, the idea offered here is to alleviate the combinatorics issue by creating a small set of special points (referred to as *anchor points*) that describe the general trend of the data.

Two approaches towards the use of the anchor points are considered here. First, an OLS regression for each observation augmented with the anchor set could be made, with the resulting collection of regression coefficients forming the basis for the similarity measures. This approach requires just *n* regressions, but also requires a method for defining similarity between pairs of observations. Second, an OLS regression for each pair of observations augmented with the anchor set could be made, with the resulting collection of regression coefficients forming the basis for the similarity measures. This approach requires a larger C_2^n regressions, although similarity could be directly forthcoming via use of robust Mahalanobis distances.

Using a single point augmentation approach has been recommended and implemented into the proposed CBI methodology. The rationale is that

- the reduction in the number of necessary OLS regressions is substantial and
- With elements $s_{ij} = \left(\mathbf{b}_i - \mathbf{b}_j\right)' \left(\mathbf{MVE}_2\left(\mathbf{B}\right)\right)^{-1} \left(\mathbf{b}_i - \mathbf{b}_j\right)$, the similarity matrix **S** is computationally the same expense as obtaining robust Mahalanobis distances from **B**. Both cases require MCD estimation of **B**.

The determination of these special anchor points is, obviously, the basis of such a methodology. Via a MVE estimation of the response-regressor space, an ellipsoid of constant distance $\chi^2_{0.975,p}$ from the center $\mathbf{MVE}_1\left(\mathbf{Z}_\mathbf{y}\right)$ and with dispersion matrix $\mathbf{MVE}_2\left(\mathbf{Z}_\mathbf{y}\right)$ describes the general data cloud in a high-breakdown sense. In order to perform the series of *n* OLS regressions, the anchor set must contain at least $p-1$ points (as the data is assumed to be in general position). Recall that in Figure 5.4 the endpoints of major axis of the ellipsoid along with the center together form an accurate depiction of the general trend. The minor axis seems unnecessary, perhaps useless or even detrimental. The argument *could* be made that only the $INT[(p-1)/2]$ major axes (as defined by the ordering of the eigenvalues corresponding to the eigenvectors representing each axis) are used to create the anchor set, along with $\mathbf{MVE}_1\left(\mathbf{Z}_\mathbf{y}\right)$. The rebuttal is that this restriction leads to the undesirable condition that certain equivariance properties no longer are achieved by the resulting set of *n* OLS regression estimators. Furthermore, the level of accuracy of this collection of OLS regression estimators with respect to the true, underlying parameter set is not really of much concern. What really matters is that the structure of this set of *n* OLS regression estimators can be used to discriminate the observations with respect to the regression model. Therefore, the anchor set formation *must* utilize all *p* axes.

A second MVE estimation, this time on the collection of *n* estimated OLS estimators, offers a high breakdown metric on which to compute similarities between pairs of observations. The underlying philosophy states that the observations that follow the general regression trend will possess similar OLS estimators. This includes good leverage points, which are NOT penalized due to their location in the regressor space, a MAJOR advantage not to be lightly dismissed. This feature lends itself towards the strengthening of the eventual initial CBI estimator by lowering its variance. In addition, low leverage outliers will typically produce pronounced deviations in the OLS intercept estimate. Any high influence point, meanwhile, will produce an OLS estimator that is dramatically different from the general regression trend.

§5.4.1.2 Linkage Selection

Once the similarity matrix, **S**, has been calculated, agglomerative hierarchical clustering begins. Three linkages are under consideration: single, complete, and average. The selection of appropriate linkage begins with the philosophical development of the similarity measure. When response-regressor space location is the basis for similarity, single linkage is preferred due to its chaining property. This formation of an elongated main cluster simulates fitting the prescribed linear model with actually performing a regression analysis. By analyzing various case studies, as well as other simulated data, it was determined that both average and complete linkage have problems with low leverage outliers blending in with the good data. This is again due to the fact that their central location is able to overcome the outlier nature in just one (the response variable) direction.

When the basis for similarity is based upon OLS regressions involving the anchor set, the objective criterion for linkage performance changes dramatically. As the prescribed model is now involved in the discrimination of the observations, the chaining property of single-linkage is not only not desired, it is to be avoided. The main cluster should consist of observations whose regression characteristics are tightly bound together. In this light, complete-linkage is the preferred method, essentially be default. Single-linkage and average-linkage could produce main clusters that are less representative of the general trend than that defined by complete-linkage.

§5.4.1.3 Clustering Stopping Rule

The underlying principle of the CBI algorithm is that the main cluster will be sufficient to produce a high-breakdown regression estimator. As the assumption is that there are at least h good observations in the data, clustering may continue until a single cluster contains at least h observations, provided that the similarity matrix possesses a high-breakdown point. However, consider the case where after the formation of the (final) main cluster is made, minor cluster merging takes place. Three options exist for the final cluster classification scheme:

- cease the cluster merging process immediately upon the defining of the main cluster,

- upon defining the main cluster, allow the clustering to continue as long as the main cluster is not involved in the merger, or
- upon defining the main cluster, allow the clustering to continue as long as the main cluster is not involved in the merger and the similarity measure used to define the merger does not exceed some threshold value.

The CBI algorithm employs the first option: clustering ceases upon formation of the main cluster. Minor clusters are required to have the same or better similarity as does the main cluster. With the use of the anchor set to create similarity measures that are regression oriented, minor cluster formation are not penalized due to response-regressor space location. Furthermore, two minor clusters of differing regression traits might become merged before the main cluster is involved in another merge. The third option could combat this potential difficulty; however, the determination of the threshold value (the "how large is large" statistical problem) is unclear. Regarding CBI coefficient estimation, no further clustering is employed once the main cluster is defined. The description of the minor clusters will be illustrated to the user via the final dendrogram. Minor cluster performance will be numerically displayed with a summary of the activation process. Together, this information is sufficient to provide the user with enough information regarding the nature of any multiple outlier issues.

§5.4.1.4 Revised Similarity Matrix

Upon the completion of *Step 4* of the CBI algorithm, a robust similarity matrix has been constructed. However, *Steps 7* to *10* are involved in the creation of a revised similarity matrix, based on a preliminary OLS regression fit to the initial main cluster (*Steps 5* and *6*). The rationale behind this apparent redundancy lies in the metric used to compute the anchor set. The first anchor set (*Step 2*) is based solely on a compact ellipsoid region argument (with expansion via a 1-step improvement based on robust Mahalanobis distances). Utilizing a regression fit to the general trend (*Step 5*), however, the second anchor set now directly incorporates the regression model. When the two anchor sets differ, resulting similarities, particularly those involving minor cluster formation, are refined so that they relate better to this general trend fit.

By being computationally viable, *Steps 7* to *10* represent a worthwhile fine-tuning of the similarity matrix.

§5.4.2 The Sequential Regression Phase of the CBI Algorithm

The cluster phase emphasized the determination of a main cluster of observations that are most similar to a common regression trend, with a further classification of g minor clusters such that each minor cluster suggests a different regression trend. Without the issues of outliers and high influence, an OLS regression fit to the main cluster becomes the initial CBI estimator. A scale estimate is then computed from this estimator, but across all n observations to avoid it from becoming a gross underestimate of the exhibited variability.

The handling of the minor clusters has evolved during the course of study. Originally, the philosophy intended that there be a truly sequential nature to the activation of minor clusters. Minor clusters were ranked according to some fit criterion, then added one at a time with the hope that efficiency would be increased via this use of more than half of the data while maintaining a 50% breakdown point. Activating one minor cluster at a time would allow the algorithm to handle jointly influential minor clusters in that these clusters are not evaluated together, where their joint influence could become problematic. The drawback to such an approach has to do with possible drifting of the BI estimator. For example, a first minor cluster activation could pull the estimator away from the minor cluster to be added next, perhaps to the point where this second minor cluster is poorly weighted. Together, the first and second minor clusters would solidify the initial main cluster-based estimator, but that opportunity does not arise. Complex alternatives to add multiple minor clusters in a sequential fashion were investigated, but later abandoned. A more parsimonious solution was then sought. To this end, the CBI regression phase is reduced to essentially three steps:

- initial estimator from the main cluster: an OLS estimator fed into the initial CBI estimator,
- a fit criterion for each minor cluster (based on fully iterated BI estimation) and the creation of a set of candidate minor clusters and

- a fit criterion for the candidate set (based on fully iterated BI estimation), with a possible final update to the CBI estimator.

Joint influence is addressed by viewing each minor cluster separately first. It also is addressed by viewing the fit criterion after IRLS convergence of the BI estimator. If certain observations together produce a dramatically different BI estimator, the result is simply dismissed. The creation of the candidate set, along with the fit criterion imposed on it, will provide protection from a drift away from the initial CBI estimator. Yet the intermediate BI estimators are allowed to iterate to convergence, a computation that ordinarily would allow for the possible breakdown of the estimator – and the reason that one-step methods are employed by the current state-of-the-art procedures.

A few details within the regression phase are now further addressed.

§5.4.2.1 The Scale Estimate

While the CBI algorithm is focused on producing a high-breakdown regression estimator, the scale estimate plays a vital role in the computation process. After all, bounded influence regression requires a good estimate of scale in order to determine appropriate downweighting of observations. There are only three times within the CBI algorithm where the scale estimate is computed:

- after the OLS fit to the initial main cluster, $\hat{\sigma}_0$,
- after the OLS fit to the final main cluster, $\hat{\sigma}_1$, and
- after the final CBI regression estimator has been determined, $\hat{\sigma}_{CBI}$.

In all three cases, the scale estimate is the MAD, based on residuals from the current fit across all n observations, not just those observations with non-zero weights that were used to create the fit. The estimate for scale is never updated during the IRLS computation of a BI regression estimator, thereby increasing the likelihood of convergence of the IRLS algorithm.

The second scale estimate, $\hat{\sigma}_1$, is not updated after $\hat{\boldsymbol{\beta}}_2$ is determined, but before the minor cluster activation stage, so that the effect of an activation is clear and not confounded with differences in **w** due to scale iteration. By forbidding the scale update at this point, a minor cluster with zero weight must yield $\hat{\boldsymbol{\beta}}_{+1} = \hat{\boldsymbol{\beta}}_2$, a condition that likely would not hold if the scale were updated prior to the activation stage. This is in agreement with the principle that observations with no weight should not alter the regression estimator.

§5.4.2.2 Weighting Schemes

There are benefits gained from the inclusion of good leverage points, from coefficient stability to reduced standard errors. The basis of the philosophy used in developing the cluster phase of the CBI algorithm is aimed directly at exploiting this actuality. However, it should also be noted that a decision regarding the form of the IRLS weights used in the CBI algorithm (BI regression) also has consequences relating to the treatment of good leverage points. Consider

- the Mallows form of $\pi_i \left(1 - \left(\frac{r_i}{\hat{\sigma}} \right)^2 \right)^2$ and
- the Schweppe form of $\left(1 - \left(\frac{r_i}{\pi_i \hat{\sigma}} \right)^2 \right)^2$.

Figure 5.7 illustrates the nature of the Mallows form. Beyond a critical value in the direction of the robust distance, the weight becomes smaller rapidly. Basically, the shape of the bisquare weight function is seen on the plane defined by any robust distance. However, the maximum weight, occurring when the scaled residual is zero, becomes smaller as the robust distance increases. Therefore, even a perfectly fitting good leverage point is penalized due to its high leverage. Meanwhile, Figure 5.8 shows the nature of the Schweppe form. Beyond a critical value in the direction of the robust distance, the critical value for the scaled residual becomes smaller in magnitude. The general shape of the bisquare weight function is still seen on the plane defined by any robust distance. The maximum weight remains at one as the robust distance increases. However, it becomes increasingly difficult for good leverage points to receive positive weight if the fit is not quite perfect. Therefore, a perfectly fitting good leverage

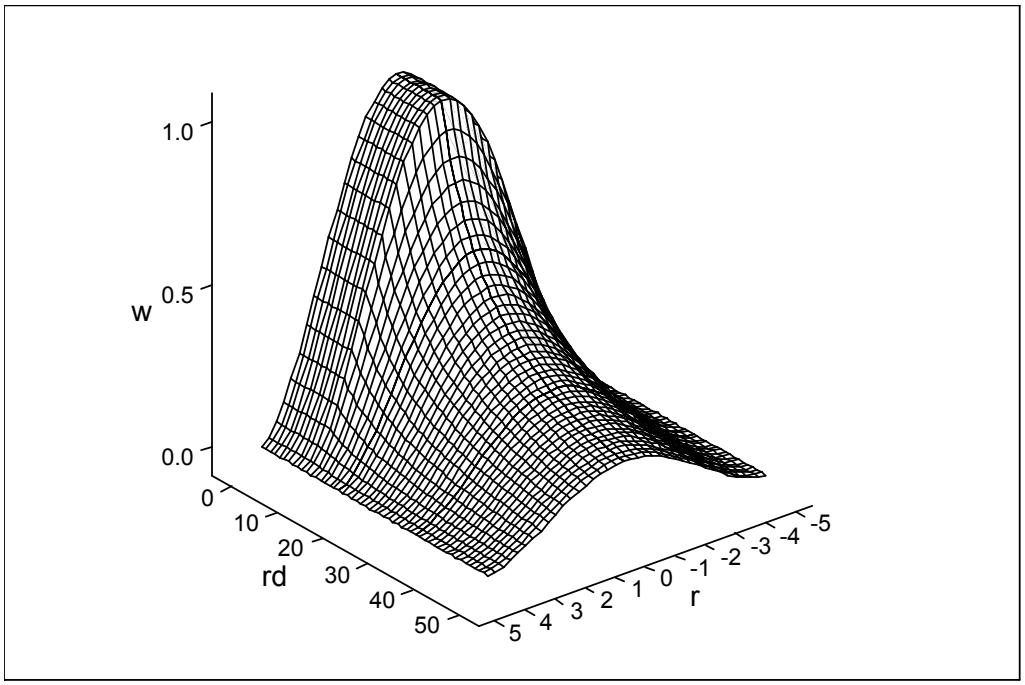

Figure 5.7: IRLS weights via Mallows weighting philosophy. Illustration uses a bisquare ψ-function, $\sigma = 1$, and the robust distance (rd) critical value is 10.

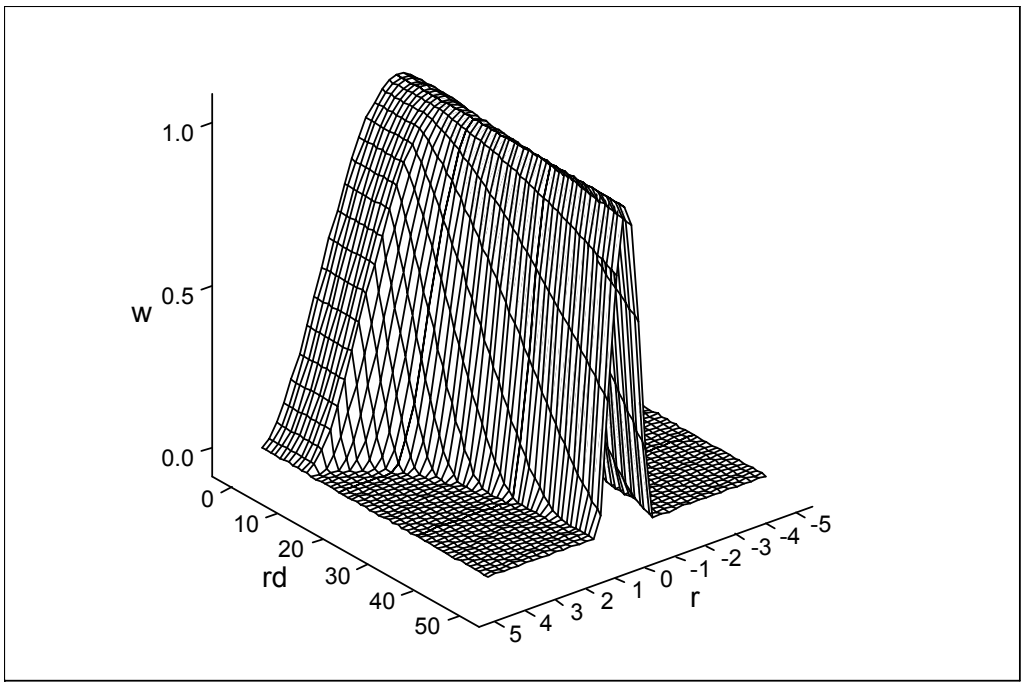

Figure 5.8: IRLS weights via Schweppe weighting philosophy. Illustration uses a bisquare ψ-function, $\sigma = 1$, and the robust distance (rd) critical value is 10.

point is not penalized due to its high leverage, but the window of opportunity for good leverage points to play a role in the estimation becomes narrower with respect to its leverage.

The CBI algorithm incorporates the Schweppe form for IRLS weights. Minor clusters must then follow the current fit rather well, especially as their leverage increases, in order to become active and possess non-zero weights.

§5.5 Case Study: Stackloss Data

To further the illustration of the proposed methodology, the stackloss data case study is now analyzed, thereby presenting a multiple linear regression analysis to supplement the simple linear regression example of Section 5.3. As before, step-by-step details are provided to illustrate the workings of the CBI algorithm.

Step 1: The MVE estimation of the $\mathbf{Z}_y$ is

$$\mathbf{MVE}_1\left(\mathbf{Z}_y\right)' = \begin{bmatrix} 59.45 & 20.80 & 86.15 & 16.30 \end{bmatrix}$$

and

$$\mathbf{MVE}_2\left(\mathbf{Z}_y\right) = \begin{bmatrix} 67.3132 & 17.4632 & 22.9289 & 63.8053 \\ 17.4632 & 8.5895 & 6.0842 & 21.6421 \\ 22.9289 & 6.0842 & 29.8184 & 19.2684 \\ 63.8053 & 21.6421 & 19.2684 & 75.8000 \end{bmatrix}.$$

Step 2: The eigenvectors and eigenvalues for $\mathbf{MVE}_2\left(\mathbf{Z}_y\right)$ are

$$\mathbf{E} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_4 \end{bmatrix} = \begin{bmatrix} 0.6506 & 0.0110 & -0.7447 & 0.1484 \\ 0.1980 & -0.0774 & 0.3534 & 0.9110 \\ 0.2468 & 0.9424 & 0.2180 & -0.0582 \\ 0.6904 & -0.3252 & 0.5225 & -0.3804 \end{bmatrix}$$

and

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 149.0283 \\ 22.9392 \\ 7.5431 \\ 2.0105 \end{bmatrix},$$

respectively. The anchor set matrix, $\boldsymbol{\Omega}$, becomes

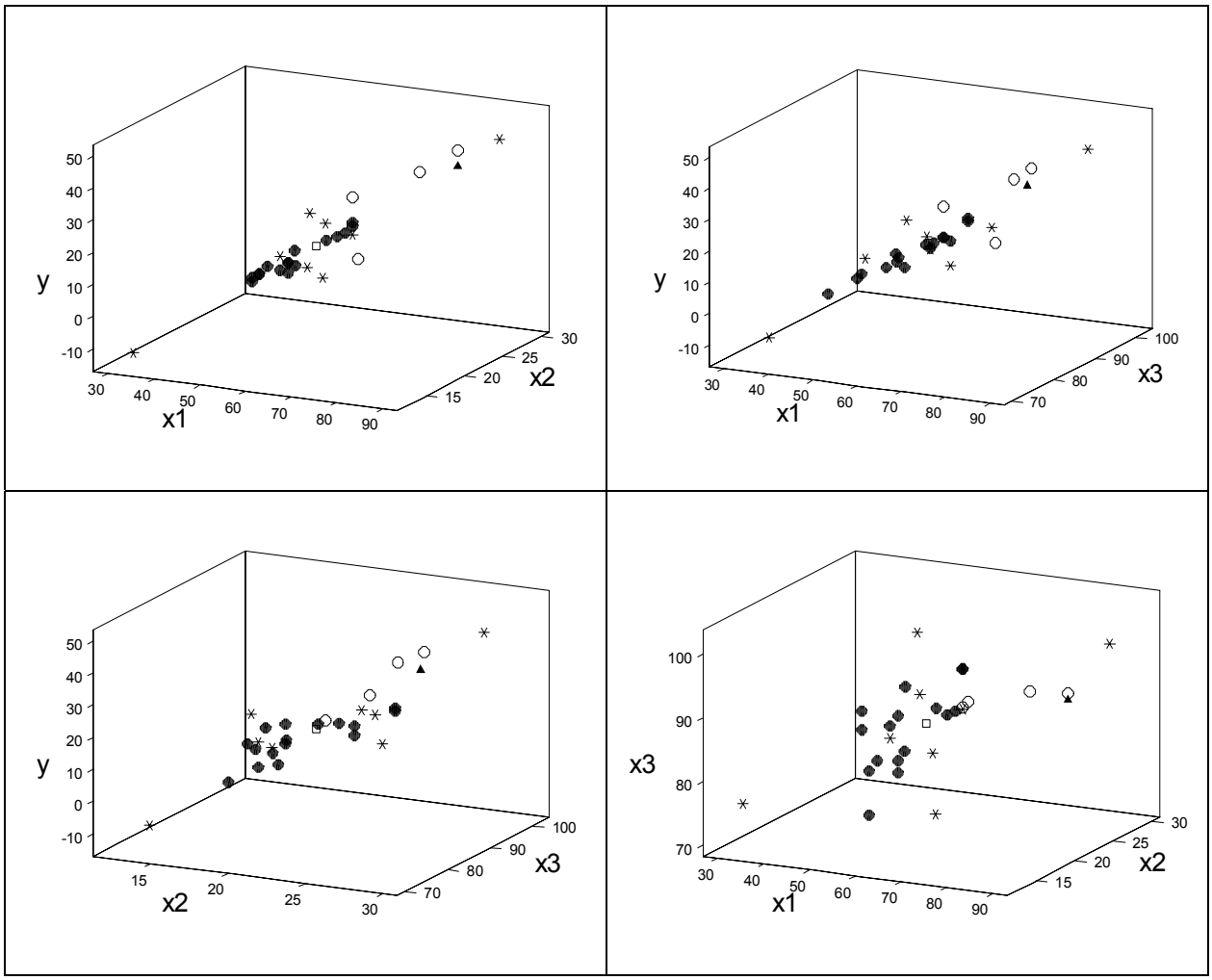

Figure 5.9: Illustration of the anchor point layout for the stackloss data, as well as the visual representation of the high influence points. Center of the anchor points a square, other anchor points are asterisks, observations 1, 3, 4 and 21 are circles, observation 2 a solid triangle, remaining data are solid circles.

$$\boldsymbol{\Omega} = \begin{bmatrix} 59.4000 & 20.8000 & 86.1500 & 16.3000 \\ 32.9367 & 12.7321 & 76.0914 & -11.8326 \\ 85.9633 & 28.8679 & 96.2086 & 44.4326 \\ 59.2735 & 22.0373 & 71.0825 & 21.4987 \\ 59.6265 & 19.5627 & 101.2175 & 11.1013 \\ 66.2773 & 17.5598 & 84.1511 & 11.5096 \\ 52.6227 & 24.0402 & 88.1489 & 21.0904 \\ 58.7474 & 16.4880 & 86.4253 & 18.1004 \\ 60.1526 & 25.1120 & 85.8747 & 14.4996 \end{bmatrix}.$$

Figure 5.9 illustrates the four 3-dimensional representations of the stackloss data along with a graphical illustration of the anchor points. Together, these figures demonstrate how the anchor points represent the general trend, while also highlighting the five outliers.

*Table 5.9: The initial 21×4 **B** matrix for the stackloss case study.*

Ob.	Intercept	X_1	X_2	X_3
1	-39.549	0.708	1.290	-0.147
2	-38.900	0.657	1.254	-0.116
3	-40.320	0.710	1.266	-0.133
4	-39.572	0.634	1.402	-0.121
5	-38.899	0.662	1.250	-0.120
6	-38.753	0.669	1.218	-0.120
7	-37.297	0.679	1.210	-0.141
8	-38.037	0.672	1.232	-0.131
9	-38.534	0.686	1.186	-0.126
10	-37.686	0.674	1.227	-0.133
11	-39.228	0.670	1.201	-0.105
12	-38.830	0.678	1.177	-0.109
13	-39.786	0.653	1.287	-0.112
14	-38.770	0.663	1.261	-0.123
15	-39.372	0.649	1.252	-0.102
16	-38.884	0.660	1.256	-0.117
17	-40.905	0.663	1.250	-0.097
18	-39.433	0.665	1.254	-0.116
19	-39.419	0.668	1.246	-0.116
20	-38.078	0.662	1.259	-0.128
21	-37.502	0.580	1.456	-0.134

Step 3: The collection of OLS anchor set regression estimates, matrix **B**, is shown in Table 5.9. This matrix shows that observations 1, 2, 3, 4 and 21 deviate from the general trend, and observations 13, 14 and 20 are perhaps what might be described as modestly different from the general trend.

Step 4: MVE estimation of **B** produces

$$\mathbf{MVE}_1(\mathbf{B})' = [-38.8712 \quad 0.6664 \quad 1.2365 \quad -0.1183]$$

and

$$\mathbf{MVE}_2(\mathbf{B}) = \begin{bmatrix} 0.7072 & 0.0039 & -0.0089 & -0.0087 \\ 0.0039 & 0.00009 & -0.00024 & -0.00005 \\ -0.0089 & -0.00024 & 0.00087 & 0.00005 \\ -0.0087 & -0.00005 & 0.00005 & 0.00013 \end{bmatrix}.$$

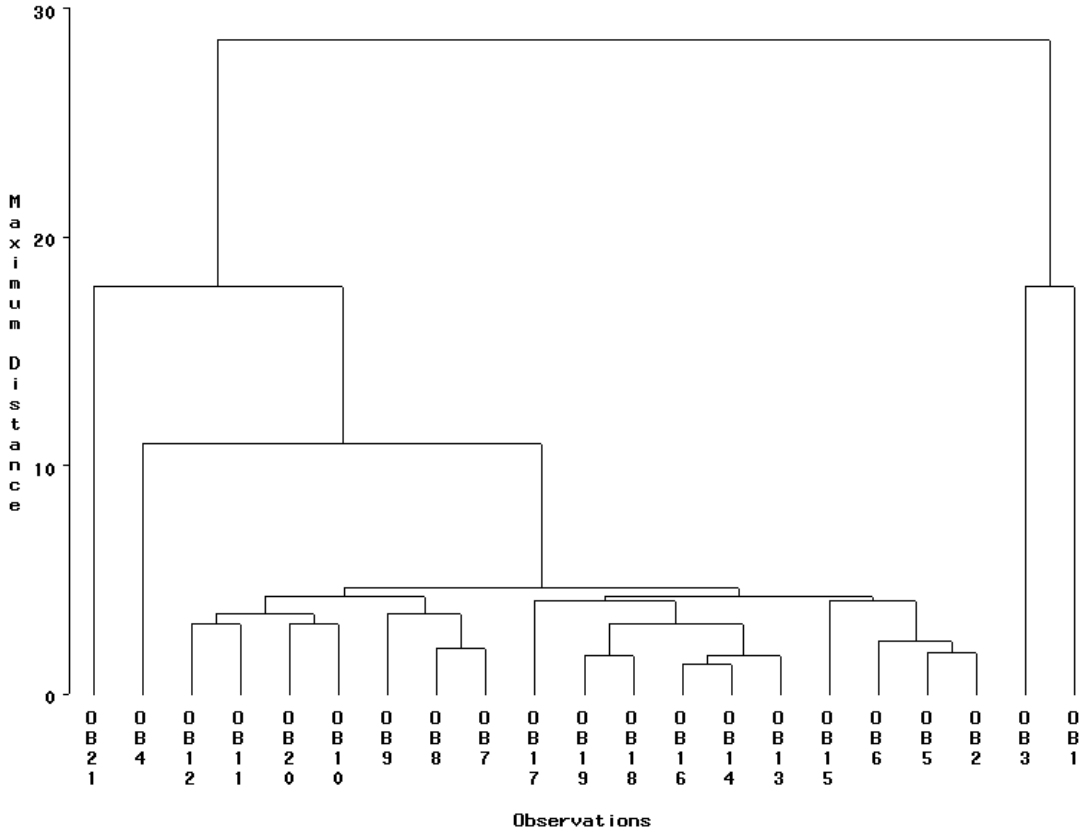

Figure 5.10: Dendrogram for initial clustering of the stackloss data.

Table 5.10: Similarity matrix using $s_{ij} = \left(\mathbf{b}_i - \mathbf{b}_j\right)' \left(\mathbf{MVE}_2\left(\mathbf{B}\right)\right)^{-1} \left(\mathbf{b}_i - \mathbf{b}_j\right)$.

s_{ij}	1	2	3	4	5	6	7	8	9	10	11
1	0	296.40	2.82	130.20	283.00	327.30	278.80	260.40	268.50	224.10	264.00
2	296.40	0	303.20	84.51	0.70	3.96	6.54	3.12	10.72	8.69	8.08
3	2.82	303.20	0	141.40	290.40	333.50	288.40	268.70	272.80	230.90	262.80
4	130.20	84.51	141.40	0	85.29	119.70	98.82	80.85	113.30	61.82	86.73
5	283.00	0.70	290.40	85.29	0	2.94	4.60	1.77	7.01	8.13	9.04
6	327.30	3.96	333.50	119.70	2.94	0	5.54	5.57	5.39	16.25	12.50
7	278.80	6.54	288.40	98.82	4.60	5.54	0	1.41	4.86	6.34	13.43
8	260.40	3.12	268.70	80.85	1.77	5.57	1.41	0	5.41	3.37	8.84
9	268.50	10.72	272.80	113.30	7.01	5.39	4.86	5.41	0	12.26	11.29
10	224.10	8.69	230.90	61.82	8.13	16.25	6.34	3.37	12.26	0	8.00
11	264.00	8.08	262.80	86.73	9.04	12.50	13.43	8.84	11.29	8.00	0
12	264.60	11.53	262.60	96.22	12.09	14.12	13.47	10.30	10.57	8.84	0.72
13	258.60	3.34	265.90	66.87	2.44	9.42	11.92	5.80	13.35	11.15	12.75
14	238.90	3.23	246.50	61.36	2.33	9.95	6.41	1.85	10.16	3.33	8.61
15	338.80	3.68	342.10	97.61	7.23	9.27	17.00	12.22	21.55	17.45	9.01
16	262.40	1.17	268.70	68.54	1.55	7.72	6.91	2.31	10.95	4.42	5.97
17	240.60	11.20	241.00	81.55	8.82	13.68	19.71	12.58	10.94	17.54	10.23
18	228.00	5.22	233.10	64.55	3.41	10.48	9.04	3.67	8.10	5.67	7.52
19	224.60	6.18	229.10	67.04	4.00	10.62	8.92	3.85	6.96	5.78	7.27
20	240.30	5.34	248.50	54.98	6.00	15.39	8.29	3.77	16.63	1.71	9.40
21	786.80	165.10	820.60	317.40	174.50	170.90	193.10	193.90	231.60	226.00	236.70

s_{ij}	12	13	14	15	16	17	18	19	20	21
1	264.60	258.60	238.90	338.80	262.40	240.60	228.00	224.60	240.30	786.80
2	11.53	3.34	3.23	3.68	1.17	11.20	5.22	6.18	5.34	165.10
3	262.60	265.90	246.50	342.10	268.70	241.00	233.10	229.10	248.50	820.60
4	96.22	66.87	61.36	97.61	68.54	81.55	64.55	67.04	54.98	317.40
5	12.09	2.44	2.33	7.23	1.55	8.82	3.41	4.00	6.00	174.50
6	14.12	9.42	9.95	9.27	7.72	13.68	10.48	10.62	15.39	170.90
7	13.47	11.92	6.41	17.00	6.91	19.71	9.04	8.92	8.29	193.10
8	10.30	5.80	1.85	12.22	2.31	12.58	3.67	3.85	3.77	193.90
9	10.57	13.35	10.16	21.55	10.95	10.94	8.10	6.96	16.63	231.60
10	8.84	11.15	3.33	17.45	4.42	17.54	5.67	5.78	1.71	226.00
11	0.72	12.75	8.61	9.01	5.97	10.23	7.52	7.27	9.40	236.70
12	0	17.87	11.81	13.03	9.25	13.79	10.54	9.85	12.27	252.30
13	17.87	0	2.34	10.63	2.64	5.94	2.33	3.28	7.05	179.80
14	11.81	2.34	0	11.68	0.88	8.88	1.18	1.75	1.88	197.20
15	13.03	10.63	11.68	0	6.22	18.16	14.15	15.59	11.99	157.10
16	9.25	2.64	0.88	6.22	0	9.33	2.45	3.23	2.07	186.10
17	13.79	5.94	8.88	18.16	9.33	0	3.94	3.63	17.08	237.10
18	10.54	2.33	1.18	14.15	2.45	3.94	0	0.11	5.27	217.40
19	9.85	3.28	1.75	15.59	3.23	3.63	0.11	0	6.19	225.20
20	12.27	7.05	1.88	11.99	2.07	17.08	5.27	6.19	0	196.20
21	252.30	179.80	197.20	157.10	186.10	237.10	217.40	225.20	196.20	0

Table 5.11: Cluster history for example data.

Step	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	18	20	21
2	1	2	3	4	2	6	7	8	9	10	11	12	13	14	15	16	17	18	18	20	21
3	1	2	3	4	2	6	7	8	9	10	11	11	13	14	15	16	17	18	18	20	21
4	1	2	3	4	2	6	7	8	9	10	11	11	13	14	15	14	17	18	18	20	21
5	1	2	3	4	2	6	7	7	9	10	11	11	13	14	15	14	17	18	18	20	21
6	1	2	3	4	2	6	7	7	9	10	11	11	13	14	15	14	17	18	18	10	21
7	1	2	3	4	2	6	7	7	9	10	11	11	13	13	15	13	17	18	18	10	21
8	1	2	1	4	2	6	7	7	9	10	11	11	13	13	15	13	17	18	18	10	21
9	1	2	1	4	2	6	7	7	9	10	11	11	13	13	15	13	17	13	13	10	21
10	1	2	1	4	2	2	7	7	9	10	11	11	13	13	15	13	17	13	13	10	21
11	1	2	1	4	2	2	7	7	7	10	11	11	13	13	15	13	17	13	13	10	21
12	1	2	1	4	2	2	7	7	7	10	11	11	13	13	2	13	17	13	13	10	21
13	1	2	1	4	2	2	7	7	7	10	11	11	13	13	2	13	13	13	13	10	21
14	1	2	1	4	2	2	7	7	7	10	10	10	13	13	2	13	13	13	13	10	21
15	1	2	1	4	2	2	7	7	7	7	7	7	13	13	2	13	13	13	13	7	21
16	1	2	1	4	2	2	7	7	7	7	7	7	2	2	2	2	2	2	2	7	21
17	**1**	**2**	**1**	**4**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**21**
18	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	21
19	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Using $\mathbf{MVE}_2(\mathbf{B})$ as a distance metric, Table 5.10 provides the similarity matrix, **S**. The main cluster is required to contain at least $h = 13$ observations. The cluster history for a complete-linkage clustering of **S** is displayed in Table 5.11, dendrogram as Figure 5.10. After a total of 17 steps the initial main cluster is defined as $C_0 = \{2, 5:20\}$.

Step 5: The OLS estimate based solely on the 17 main cluster observations is

$$\hat{\boldsymbol{\beta}}_0' = [-37.6525 \quad 0.7977 \quad 0.5773 \quad -0.0671].$$

The scale estimate from this regression fit is computed across all 21 residuals and becomes $\hat{\sigma} = 1.4604$.

Step 6: $H = \{1:3, 5:20\}$ and the weight vector $\boldsymbol{\omega}$ becomes

$$\boldsymbol{\omega}' = [1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0].$$

Step 7: Being essentially an update to *Step 1*, the data is characterized by

$$\mathbf{m}_H\left(\mathbf{Z}_{\mathrm{y}}\right)' = \begin{bmatrix} 59.8421 & 21.0000 & 86.0000 & 17.1053 \end{bmatrix}$$

and

$$\mathbf{C}_H\left(\mathbf{Z}_{\mathrm{y}}\right) = \begin{bmatrix} 87.8070 & 25.4444 & 24.5556 & 95.4620 \\ 25.4444 & 10.5556 & 7.5000 & 29.3889 \\ 24.5556 & 7.5000 & 30.5556 & 24.3333 \\ 95.4620 & 29.3889 & 24.3333 & 108.3216 \end{bmatrix}.$$

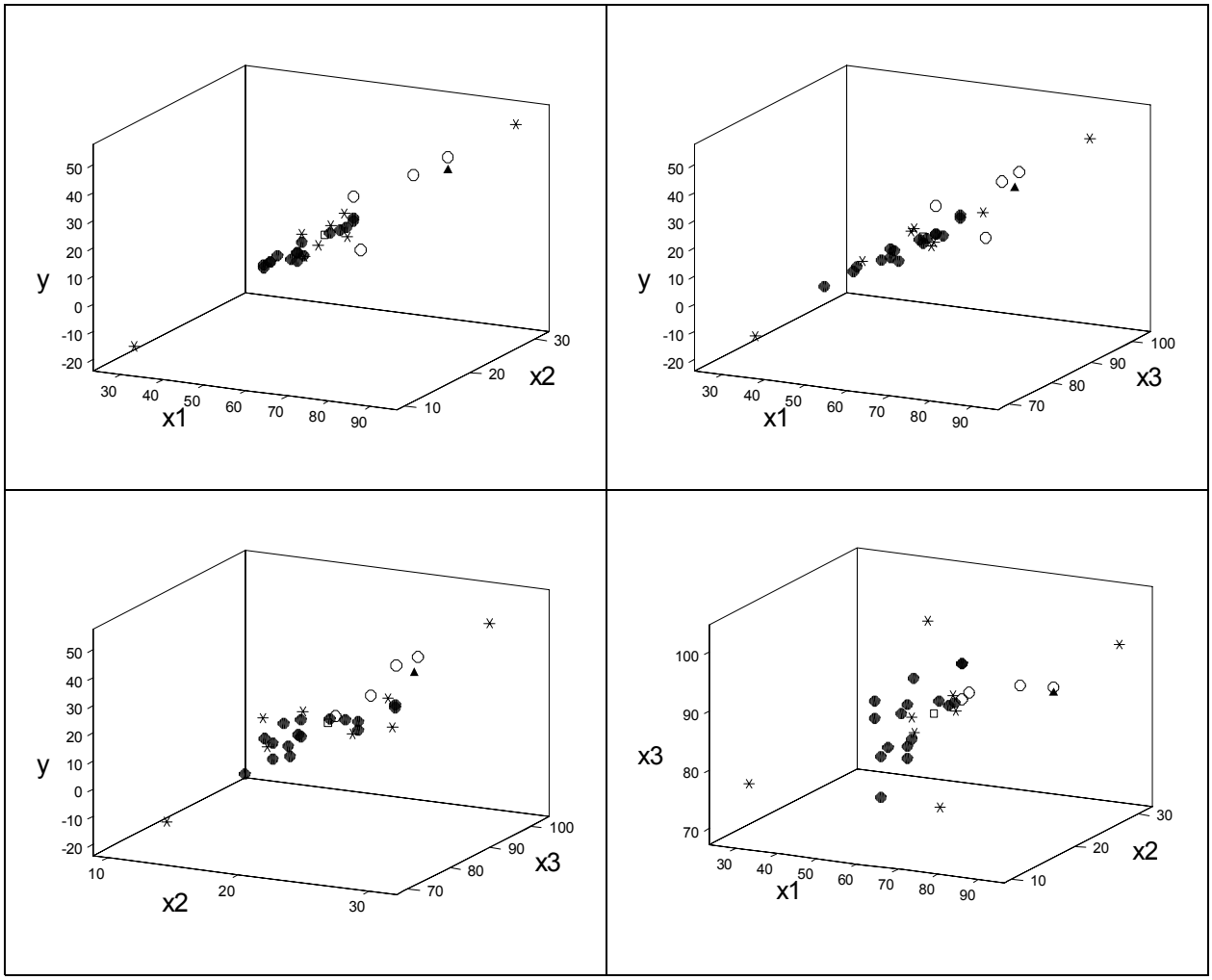

Figure 5.11: Illustration of the revised anchor point layout for the stackloss data, as well as the visual representation of the high influence points. Center of the anchor points a square, other anchor points are asterisks, observations 1, 3, 4 and 21 are circles, observation 2 a solid triangle, remaining data are solid circles.

Step 8: Being essentially an update to *Step 2*, the eigenvectors and eigenvalues for $\mathbf{C}_H\left(\mathbf{Z}_y\right)$ are

$$\mathbf{E}=\begin{bmatrix}\mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_4\end{bmatrix}=\begin{bmatrix}0.6433 & -0.0616 & -0.4999 & 0.5766\\ 0.1956 & -0.0195 & 0.8391 & 0.5072\\ 0.1942 & 0.9767 & 0.0297 & -0.0865\\ 0.7143 & -0.2047 & 0.2123 & -0.6347\end{bmatrix}$$

and

$$\boldsymbol{\lambda}=\begin{bmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\\ \lambda_4\end{bmatrix}=\begin{bmatrix}208.9584\\ 23.7567\\ 3.0979\\ 1.4268\end{bmatrix},$$

respectively. As shown graphically in Figure 5.11, the updated anchor set matrix becomes

$$\boldsymbol{\Omega}=\begin{bmatrix}59.8421 & 21.0000 & 86.0000 & 17.1053\\ 28.8012 & 11.5593 & 76.6294 & -17.3623\\ 90.8830 & 30.4407 & 95.3706 & 51.5728\\ 60.8450 & 21.3181 & 70.1088 & 20.4353\\ 58.8392 & 20.6819 & 101.8913 & 13.7752\\ 62.7795 & 16.0699 & 85.8253 & 15.8577\\ 56.9047 & 25.9301 & 86.1747 & 18.3528\\ 57.5431 & 18.9776 & 86.3447 & 19.6359\\ 62.1411 & 23.0224 & 85.6553 & 14.5746\end{bmatrix}.$$

Step 9: The updated **B** matrix is displayed in Table 5.12 (an update to *Step 3*).

Step 10: Being essentially an update to *Step 4*, the weighted mean vector and covariance matrix for **B** is

$$\mathbf{m}_H\left(\mathbf{B}\right)'=\begin{bmatrix}-42.4490 & 0.9566 & 0.5554 & -0.1088\end{bmatrix}$$

and

*Table 5.12: The updated 21×4 **B** matrix for the stackloss case study.*

Obsn.	Intercept	X_1	X_2	X_3
1	-42.779	0.978	0.561	-0.119
2	-42.386	0.936	0.550	-0.096
3	-43.353	0.984	0.540	-0.110
4	-43.012	0.904	0.760	-0.108
5	-42.424	0.958	0.548	-0.110
6	-42.347	0.965	0.519	-0.110
7	-41.709	0.966	0.528	-0.119
8	-42.401	0.957	0.554	-0.110
9	-42.318	0.970	0.516	-0.111
10	-42.709	0.954	0.564	-0.106
11	-42.561	0.961	0.537	-0.105
12	-42.456	0.958	0.552	-0.108
13	-44.269	0.930	0.643	-0.093
14	-41.561	0.955	0.582	-0.126
15	-43.081	0.943	0.549	-0.088
16	-42.405	0.952	0.553	-0.104
17	-42.586	0.957	0.555	-0.107
18	-41.941	0.953	0.563	-0.113
19	-41.695	0.946	0.582	-0.115
20	-41.550	0.955	0.559	-0.117
21	-39.840	0.802	0.987	-0.147

$$\mathbf{C}_H(\mathbf{B}) = \begin{bmatrix} 0.4304 & 0.0012 & -0.0068 & -0.0042 \\ 0.0012 & 0.00017 & 0.00023 & -0.00007 \\ -0.0068 & 0.00023 & 0.00075 & 0.00004 \\ -0.0042 & -0.00007 & 0.00004 & 0.00008 \end{bmatrix}.$$

Using $\mathbf{C}_H(\mathbf{B})$ as a distance metric, Table 5.13 provides the updated similarity matrix, **S**. The cluster history for a complete-linkage clustering of **S** is displayed in Table 5.14, dendrogram as Figure 5.12. After a total of 15 steps the main cluster is defined as $C_0 = \{2, 5 : 12, 14, 16 : 20\}$. There are $g = 5$ minor clusters, $C_1 = \{1, 3\}$, $C_2 = \{4\}$, $C_3 = \{13\}$, $C_4 = \{15\}$ and $C_5 = \{21\}$.

Table 5.13: Updated similarity matrix using $s_{ij} = \left(\mathbf{b}_i - \mathbf{b}_j\right)' \left(\mathbf{C}_H\left(\mathbf{B}\right)\right)^{-1} \left(\mathbf{b}_i - \mathbf{b}_j\right)$.

s_{ij}	1	2	3	4	5	6	7	8	9	10	11
1	0	20.82	2.63	86.03	8.25	12.64	8.69	5.35	9.82	5.54	6.69
2	20.82	0	23.51	109.40	4.87	6.75	7.79	5.24	7.70	5.15	5.08
3	2.63	23.51	0	99.81	12.19	14.71	13.67	8.73	11.36	8.69	7.19
4	86.03	109.40	99.81	0	109.90	136.10	115.40	93.34	131.20	89.97	101.40
5	8.25	4.87	12.19	109.90	0	1.50	1.50	1.12	1.51	1.34	2.41
6	12.64	6.75	14.71	136.10	1.50	0	2.26	4.19	0.31	4.96	4.21
7	8.69	7.79	13.67	115.40	1.50	2.26	0	2.22	1.69	3.72	3.59
8	5.35	5.24	8.73	93.34	1.12	4.19	2.22	0	3.29	0.29	0.89
9	9.82	7.70	11.36	131.20	1.51	0.31	1.69	3.29	0	4.25	3.02
10	5.54	5.15	8.69	89.97	1.34	4.96	3.72	0.29	4.25	0	1.30
11	6.69	5.08	7.19	101.40	2.41	4.21	3.59	0.89	3.02	1.30	0
12	5.39	5.10	8.23	93.73	1.30	4.25	2.51	0.03	3.28	0.29	0.61
13	19.76	18.79	25.15	81.38	14.62	22.24	24.20	15.16	23.24	11.49	19.09
14	9.91	14.63	20.91	85.52	6.18	11.78	5.23	5.88	11.25	6.67	11.10
15	18.08	7.90	14.96	88.67	13.50	17.27	17.10	8.89	15.75	8.39	5.69
16	7.88	4.54	9.88	85.47	3.96	7.79	5.20	1.15	6.51	1.43	0.96
17	5.36	5.07	8.02	94.17	1.15	4.10	2.86	0.09	3.24	0.14	0.68
18	6.84	6.86	12.26	79.61	3.86	8.74	3.69	1.23	7.44	1.93	2.93
19	10.38	10.74	18.00	64.02	8.89	16.21	8.39	4.77	14.65	5.45	7.59
20	8.61	10.50	14.58	74.19	7.78	13.41	6.03	3.71	11.43	5.03	5.39
21	380.50	370.50	437.40	202.60	372.40	413.50	378.90	366.30	417.90	361.20	398.60

s_{ij}	12	13	14	15	16	17	18	19	20	21
1	5.39	19.76	9.91	18.08	7.88	5.36	6.84	10.38	8.61	380.50
2	5.10	18.79	14.63	7.90	4.54	5.07	6.86	10.74	10.50	370.50
3	8.23	25.15	20.91	14.96	9.88	8.02	12.26	18.00	14.58	437.40
4	93.73	81.38	85.52	88.67	85.47	94.17	79.61	64.02	74.19	202.60
5	1.30	14.62	6.18	13.50	3.96	1.15	3.86	8.89	7.78	372.40
6	4.25	22.24	11.78	17.27	7.79	4.10	8.74	16.21	13.41	413.50
7	2.51	24.20	5.23	17.10	5.20	2.86	3.69	8.39	6.03	378.90
8	0.03	15.16	5.88	8.89	1.15	0.09	1.23	4.77	3.71	366.30
9	3.28	23.24	11.25	15.75	6.51	3.24	7.44	14.65	11.43	417.90
10	0.29	11.49	6.67	8.39	1.43	0.14	1.93	5.45	5.03	361.20
11	0.61	19.09	11.10	5.69	0.96	0.68	2.93	7.59	5.39	398.60
12	0	15.42	6.74	8.01	0.93	0.05	1.39	5.06	3.87	371.10
13	15.42	0	19.02	24.94	18.19	14.00	18.75	21.36	25.84	311.40
14	6.74	19.02	0	26.25	9.77	7.07	4.39	5.08	5.99	300.00
15	8.01	24.94	26.25	0	4.19	8.06	9.74	12.72	11.35	409.80
16	0.93	18.19	9.77	4.19	0	1.16	1.31	3.87	2.80	370.30
17	0.05	14.00	7.07	8.06	1.16	0	1.80	5.69	4.65	371.50
18	1.39	18.75	4.39	9.74	1.31	1.80	0	1.25	0.80	340.30
19	5.06	21.36	5.08	12.72	3.87	5.69	1.25	0	0.78	308.70
20	3.87	25.84	5.99	11.35	2.80	4.65	0.80	0.78	0	337.30
21	371.10	311.40	300.00	409.80	370.30	371.50	340.30	308.70	337.30	0

Table 5.14: Updated cluster history for example data.

Step	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	1	2	3	4	5	6	7	8	9	10	11	8	13	14	15	16	17	18	19	20	21
2	1	2	3	4	5	6	7	8	9	10	11	8	13	14	15	16	8	18	19	20	21
3	1	2	3	4	5	6	7	8	9	8	11	8	13	14	15	16	8	18	19	20	21
4	1	2	3	4	5	6	7	8	6	8	11	8	13	14	15	16	8	18	19	20	21
5	1	2	3	4	5	6	7	8	6	8	11	8	13	14	15	16	8	18	19	19	21
6	1	2	3	4	5	6	7	8	6	8	11	8	13	14	15	11	8	18	19	19	21
7	1	2	3	4	5	6	7	8	6	8	11	8	13	14	15	11	8	18	18	18	21
8	1	2	3	4	5	6	7	5	6	5	11	5	13	14	15	11	5	18	18	18	21
9	1	2	3	4	5	6	6	5	6	5	11	5	13	14	15	11	5	18	18	18	21
10	1	2	1	4	5	6	6	5	6	5	11	5	13	14	15	11	5	18	18	18	21
11	1	2	1	4	5	6	6	5	6	5	5	5	13	14	15	5	5	18	18	18	21
12	1	2	1	4	2	6	6	2	6	2	2	2	13	14	15	2	2	18	18	18	21
13	1	2	1	4	2	6	6	2	6	2	2	2	13	14	15	2	2	14	14	14	21
14	1	2	1	4	2	2	2	2	2	2	2	2	13	14	15	2	2	14	14	14	21
15	**1**	**2**	**1**	**4**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**13**	**2**	**15**	**2**	**2**	**2**	**2**	**2**	**21**
16	1	2	1	4	2	2	2	2	2	2	2	2	13	2	1	2	2	2	2	2	21
17	1	2	1	4	2	2	2	2	2	2	2	2	1	2	1	2	2	2	2	2	21
18	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	21
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	21
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

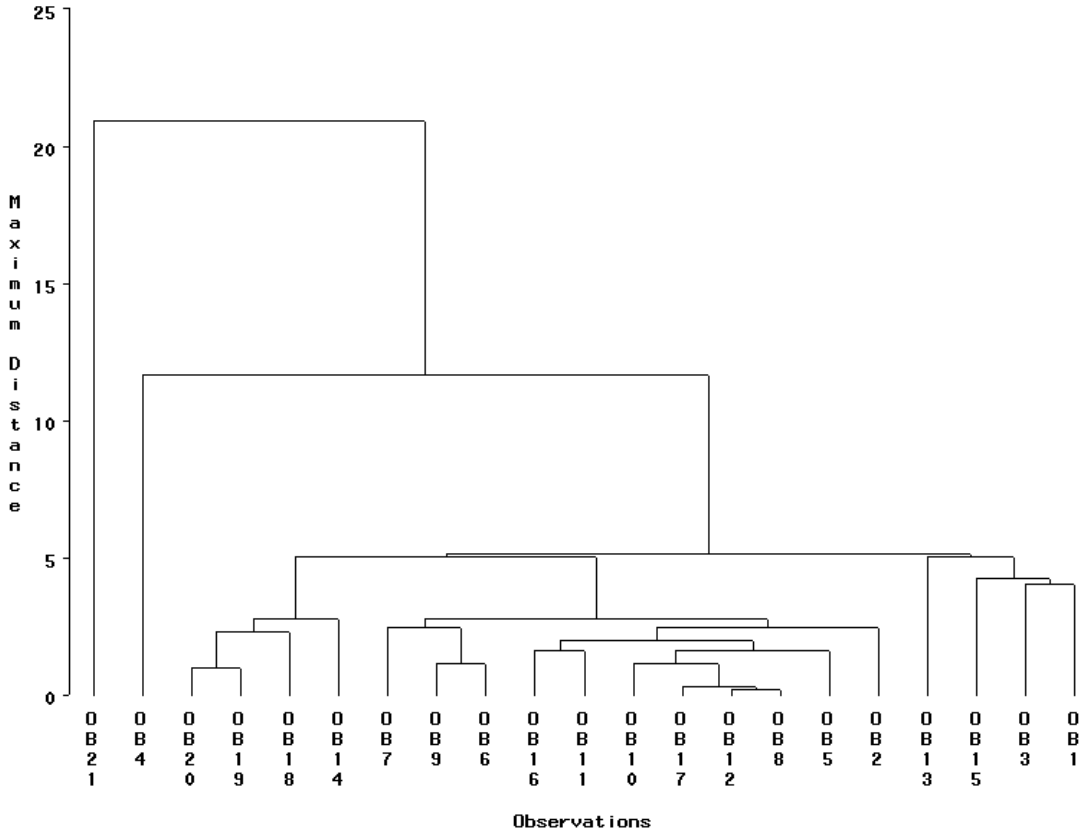

Figure 5.12: Dendrogram for final clustering of the stackloss data.

Step 11: The OLS estimate based solely on the 15 main cluster observations is

$$\hat{\boldsymbol{\beta}}_1' = \begin{bmatrix} -34.7830 & 0.8675 & 0.4533 & -0.1171 \end{bmatrix}.$$

The scale estimate from this regression fit is computed across all 21 residuals and becomes $\hat{\sigma}_1 = 1.2654$.

Step 12: The regressor space is characterized by

$$\mathbf{m}_H(\mathbf{Z})' = \begin{bmatrix} 59.8421 & 21.0000 & 86.0000 \end{bmatrix}$$

and

$$\mathbf{C}_H(\mathbf{Z}) = \begin{bmatrix} 87.8070 & 25.4444 & 24.5556 \\ 25.4444 & 10.5556 & 7.5000 \\ 24.5556 & 7.5000 & 30.5556 \end{bmatrix},$$

which are used to determine the BI leverage weight vector $\boldsymbol{\pi}$, where in this case the elements $\pi_i = 1, \forall i$. Additionally, the BI regression tuning constant is $c = 4.685\sqrt{2pn}/(n-2p) = 4.671$.

Step 13: At convergence, the IRLS solution (using a non-iterated scale estimate) of a BI regression estimator for C_0 is

$$\hat{\boldsymbol{\beta}}_2' = \begin{bmatrix} -35.0358 & 0.8676 & 0.4519 & -0.1140 \end{bmatrix}.$$

Step 14: The minor clusters are evaluated regarding activation candidacy via the statistics $DFFITS^2_{+1} = 0.0026$ (with $w_1 = 0.018$ and $w_3 = 0.00018$), $DFFITS^2_{+2} = 0$ (with $w_4 = 0$), $DFFITS^2_{+3} = 0.5604$ (with $w_{13} = 0.643$), $DFFITS^2_{+4} = 0.3685$ (with $w_{15} = 0.906$), and $DFFITS^2_{+5} = 0$ (with $w_{21} = 0$).

Table 5.15: Final observation weights for the CBI regression estimator.

Obsn.	w_i	Obsn.	w_i	Obsn.	w_i
1	0	**8**	0.9760	**15**	0.9012
2	0.9359	**9**	0.9445	**16**	0.9992
3	0	**10**	0.9978	**17**	0.9903
4	0	**11**	0.9605	**18**	0.9996
5	0.9719	**12**	0.9948	**19**	0.9782
6	0.9140	**13**	0.6355	**20**	0.8062
7	0.9930	**14**	0.8881	**21**	0

Step 15: Given $\delta = 4$, then the activation set becomes $J = C_1 \cup C_3 \cup C_4 = \{1, 3, 13, 15\}$. Subsequently, $DFFITS^2_{+J} = 0.2111 < 4$, with $w_1 = 0$, $w_3 = 0$, $w_{13} = 0.636$ and $w_{15} = 0.901$ (the IRLS weights at convergence, upon the activation of *J*).

Step 16: As at least one minor cluster observation obtained a nonzero weight, the final cluster-based bounded influence (CBI) regression estimator becomes

$$\hat{\boldsymbol{\beta}}'_{CBI} = \hat{\boldsymbol{\beta}}'_{+J} = [-37.2844 \quad 0.8121 \quad 0.5368 \quad -0.0709].$$

Based on $\hat{\boldsymbol{\beta}}_{CBI}$, the final CBI scale estimate is computed using all 21 residuals as

$$\hat{\sigma}_{CBI} = 1.2675,$$

with final CBI weights for the individual observations as shown in Table 5.15.

From this analysis it is interesting to point out the use of observation 2. This is considered to be a good leverage point in the literature and the CBI clustering algorithm placed this observation in the main cluster. Meanwhile, observations 1, 3, 4 and 21 each receive a zero weight.

The inferential side of the CBI regression analysis for the stackloss data will be provided in the next section, along with the inferential analysis for the simple linear regression example of Section 5.3.

§5.6 Inferential analysis with the CBI regression estimator

Finally, this chapter concludes with a brief discussion of the methodology that is used to produce standard errors for the coefficients of the CBI regression estimator. Already available is the BI-bisquare robust analysis of variance procedure of Birch (1992). The CBI procedure provides the mechanism that allows a BI-regression to improve upon a low breakdown point. Yet the CBI estimator still consists of a bounded influence IRLS solution, with final CBI-based observation weights and MVE-based leverage weights. Therefore, this robust analysis of variance framework is applicable to the CBI estimator.

Original form of v^2, the robust ANOVA scale estimate (Birch (1992))

The construction of the BI-bisquare regression robust analysis of variance table begins with a weighted sums of squares (SS) breakdown utilizing the final CBI observation weight vector, where

$$SS_{TOTAL} = \sum_{i=1}^{n} w_i y_i^2 - \frac{\left(\sum_{i=1}^{n} w_i y_i \right)^2}{\sum_{i=1}^{n} w_i},$$

$$SS_{ERROR} = \sum_{i=1}^{n} w_i r_i^2 \left(\hat{\boldsymbol{\beta}}_{CBI} \right)$$

and

$$SS_{MODEL} = SS_{TOTAL} - SS_{ERROR}.$$

The regression model and the random error component have

$$df_{MODEL} = p - 1$$

and

$$df_{ERROR} = n - p$$

degrees of freedom, respectively. Mean squares are reported in the usual fashion, with

$$MS_{MODEL} = \frac{SS_{MODEL}}{df_{MODEL}}$$

and

$$MS_{ERROR} = \frac{SS_{ERROR}}{df_{ERROR}}.$$

However, the mean square error is not used during hypothesis testing. Given the CBI scale estimate, $\hat{\sigma}_{CBI}$, the BI leverage weight vector $\boldsymbol{\pi}$, and $\hat{\boldsymbol{\beta}}_{CBI}$, a robust mean square error is found by

$$v^2 = \frac{\frac{n^2}{n-p}\hat{\sigma}_{CBI}^2 \sum_{i=1}^{n} \psi^2\left(\frac{r_i\left(\hat{\boldsymbol{\beta}}_{CBI}\right)}{\pi_i \hat{\sigma}_{CBI}}\right)}{\sum_{i=1}^{n} \psi'\left(\frac{r_i\left(\hat{\boldsymbol{\beta}}_{CBI}\right)}{\pi_i \hat{\sigma}_{CBI}}\right)},$$

where $\psi'(r) = \frac{\partial \psi(r)}{\partial r}$ and with the CBI estimator utilizing a bisquare ψ-function. This statistic, in turn, is employed in the model F-test as

$$F = \frac{MS_{MODEL}}{v^2} \dot{\sim} F_{p-1,\, n-p}.$$

Furthermore, the parameter variance-covariance matrix is given by

$$\mathbf{V} = v^2 \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1},$$

where $\mathbf{W}$ is a diagonal matrix containing the final CBI observation weights. The significance test for the ith parameter then becomes

$$t = \frac{\hat{\beta}_{CBI,\,i}}{v_{ii}} \dot{\sim} t_{n-p},$$

where $\hat{\beta}_{CBI,\,i}$ is the ith element of $\hat{\boldsymbol{\beta}}_{CBI}$ and v_{ii} is the ith diagonal element of $\mathbf{V}$.

Modified form, v_w^2, of the robust ANOVA scale estimate

As will be demonstrated later in Chapter 8 via Monte Carlo simulation studies, v^2 has a tendency to become biased, overestimating the level of variability exhibited by an underlying trend, as the level of contamination increases. The expression $n^2/(n-p)$ in the aforementioned formula for v^2 is the focal point for a recommended modification. Specifically, under sparse

contamination where ordinary BI regression performs well, perhaps only a few observations receive a zero weight. There likely may be nearly n observations following the general trend. Yet under a high contamination scenario, the number of observations following the general trend could approach a mere half the dataset. In order to combat the effect that the level of contamination has on the bias associated with v^2, the *effective sample size*, n_w, defined as

$$n_w = \sum_{i=1}^{n} w_i ,$$

is used to replace n. The modified version of the robust analysis of variance scale estimate then becomes

$$v_w^2 = \frac{\dfrac{n_w^2}{n_w - p}\hat{\sigma}_{CBI}^2 \sum_{i=1}^{n} \psi^2\left(\dfrac{r_i\left(\hat{\boldsymbol{\beta}}_{CBI}\right)}{\pi_i \hat{\sigma}_{CBI}}\right)}{\sum_{i=1}^{n} \psi'\left(\dfrac{r_i\left(\hat{\boldsymbol{\beta}}_{CBI}\right)}{\pi_i \hat{\sigma}_{CBI}}\right)},$$

with the parameter variance-covariance matrix becoming

$$\mathbf{V} = v_w^2 \left(\mathbf{X'WX}\right)^{-1}.$$

Furthermore, the error degrees of freedom are now computed as

$$df_{ERROR} = n_w - p ,$$

which is utilized in the model F-test as well as in the individual parameter t-tests.

This inferential statistical analysis is now applied to each of the two illustrative examples used within Chapter 5. First, the Section 5.3 example dataset has a robust analysis of variance table that is given as Table 5.16. The effective sample size was 9.639 (from the total degrees of freedom, plus one) and the error variance used for testing purposes was $v_w^2 = 27.793$ (a vast improvement over $\hat{\sigma}_{CBI}^2 = 8.2277^2 = 67.695$ since $\sigma^2 = 25$). The overall model was significant at the 5% level, with a p-value of 0.000+. The coefficient of determination was $R^2 = 0.962$. The final CBI coefficients were nearly identical to those from the OLS fit to the clean data alone (i.e. removing observations 10 and 11). The dendrogram (Figure 5.13) illustrates the high

influence pair as being dramatically different from the remaining ten observations; their complete downweighting is also indicated from the observation weight plot of Figure 5.13.

Table 5.16: Robust analysis of variance table and parameter estimates summary for the CBI analysis of the Section 5.3 example dataset.

Cluster History

Step	Clusters	
Initial	$C_0 = \{2:9\}$	$n = 12$
Final	$C_0 = \{2:9, 12\}$ $C_1 = \{1\}$, $C_2 = \{10\}$, $C_3 = \{11\}$	$h = 7$

Sequential CBI Regression Summary

Parameter	Initial OLS: $\hat{\beta}_0$	OLS C_0: $\hat{\beta}_1$	BI C_0: $\hat{\beta}_2$	Scale
Intercept	196.5825	195.8304	195.7777	$\hat{\sigma}_0 = 8.0168$
X	-3.7664	-3.7094	-3.7099	$\hat{\sigma}_1 = 8.2277$
				$\hat{\sigma}_{CBI} = 8.2277$

Minor Cluster, I	$DFFITS^2_{+I}$	Activate
C_1	0.1329	*Yes*
C_2	0	*No*
C_3	0	*No*

Candidates, J	$DFFITS^2_{+J}$	Activate
C_J	0.1329	*Yes*

Robust Analysis of Variance

Source	df	SS	MSR/MSE/v_w^2	F/Pvalue/R^2
Regression	1	4848.707	4848.707	174.460
Error	7.639	192.657	25.219	0.000+
Total	8.639	5041.364	27.793	0.962

Parameter Estimates

Parameter	Estimate	Std. Error	t	P-value
Intercept	198.0978	4.607	43.004	0.000+
X	-3.8170	0.289	-13.208	0.000+

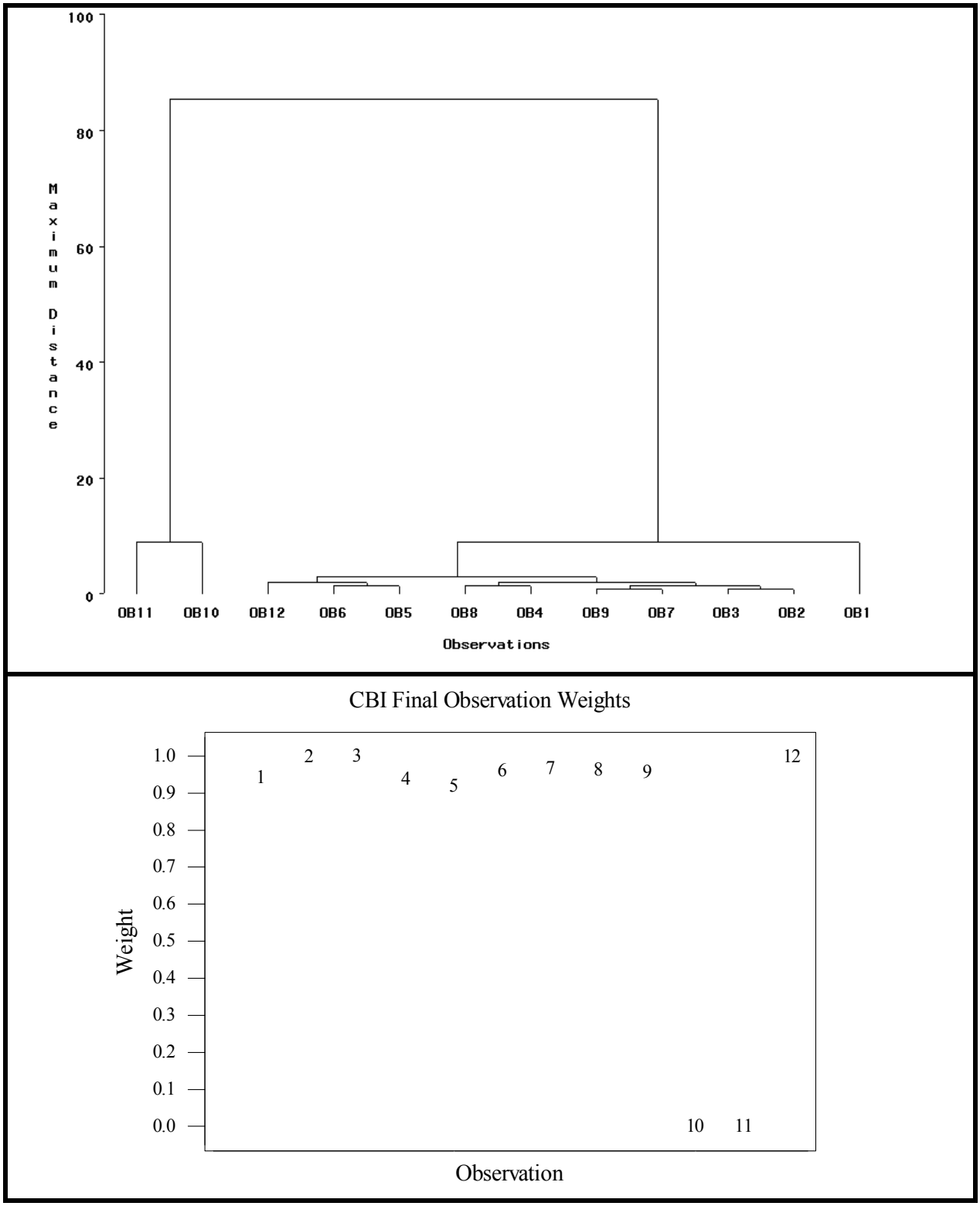

Figure 5.13: Final clustering dendrogram and final CBI observation weights for the Section 5.3 example dataset.

Table 5.17: Robust analysis of variance table and parameter estimates summary for the CBI analysis of the stackloss dataset.

Cluster History

Step	Clusters	
Initial	$C_0 = \{2, 5:20\}$	$n = 21$
Final	$C_0 = \{2, 5:12, 14, 16:20\}$ $C_1 = \{1, 3\}$, $C_2 = \{4\}$, $C_3 = \{13\}$, $C_4 = \{15\}$, $C_5 = \{21\}$	$h = 13$

Sequential CBI Regression Summary

Parameter	Initial OLS: $\hat{\beta}_0$	OLS C_0: $\hat{\beta}_1$	BI C_0: $\hat{\beta}_2$	Scale
Intercept	-37.6525	-34.7830	-35.0358	$\hat{\sigma}_n = 1.4604$
X1	0.7977	0.8675	0.8676	$\hat{\sigma}_1 = 1.2654$
X2	0.5773	0.4533	0.4519	$\hat{\sigma}_{CBI} = 1.3327$
X3	-0.0671	-0.1171	-0.1140	

Minor Cluster, I	$DFFITS^2_{+I}$	Activate
C_1	0.0026	*Yes*
C_2	0	*No*
C_3	0.5604	*Yes*
C_4	0.3685	*Yes*
C_5	0	*No*
Candidates, J	$DFFITS^2_{+J}$	**Activate**
C_J	0.2111	*Yes*

Robust Analysis of Variance

Source	df	SS	MSR/MSE/v^2	F/Pvalue/R^2
Regression	3	754.421	251.474	162.603
Error	11.887	16.535	1.391	0.000+
Total	14.887	770.956	1.547	0.979

Parameter Estimates

Parameter	Estimate	Std. Error	t	P-value
Intercept	-37.2844	4.794	-7.777	0.000+
X1	0.8121	0.069	11.724	0.000+
X2	0.5368	0.170	3.154	0.004
X3	-0.0709	0.063	-1.130	0.140

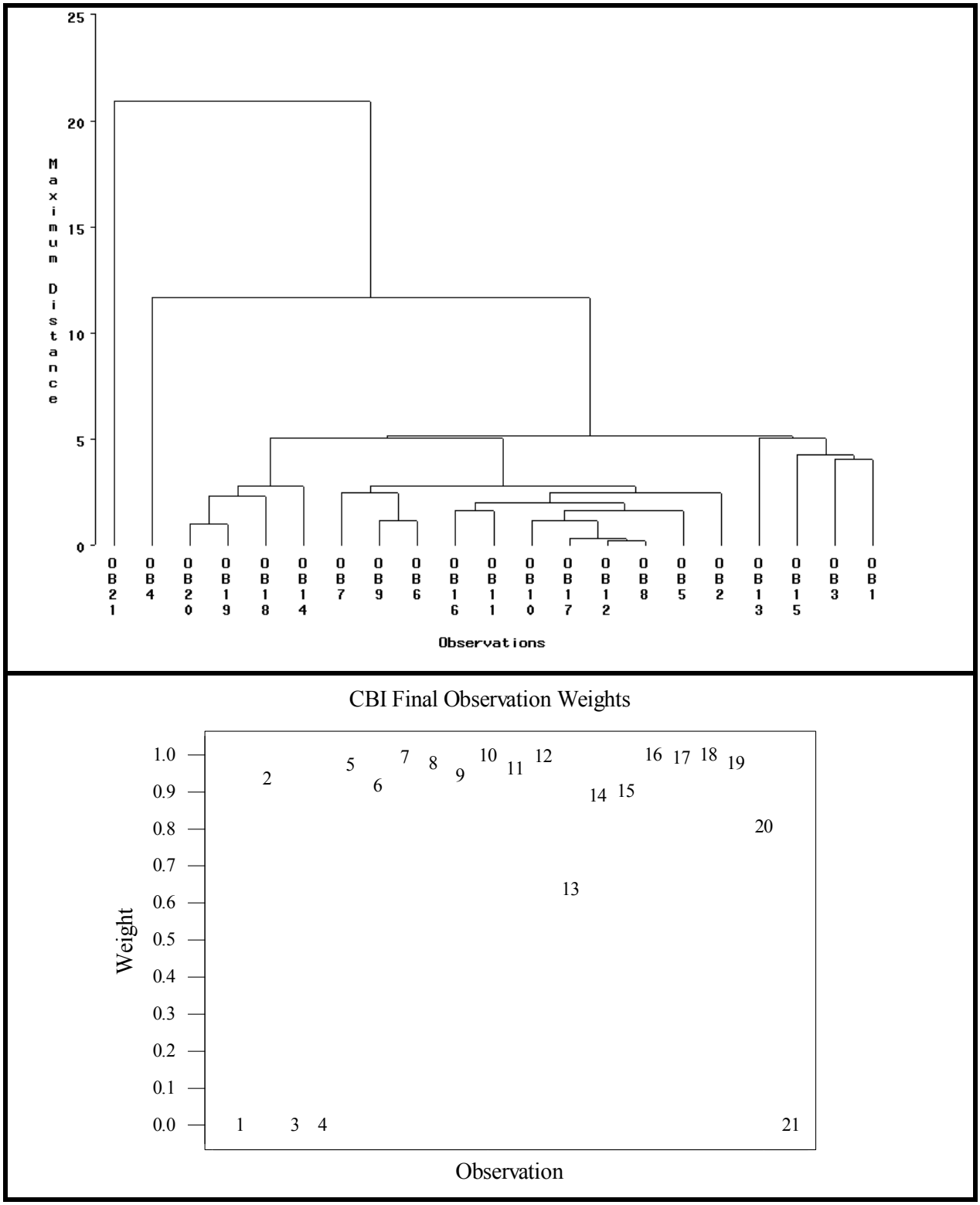

Figure 5.14: Final clustering dendrogram and final CBI observation weights for the Section 5.3 example dataset.

The stackloss dataset (Section 5.5) has a robust analysis of variance table that is given as Table 5.17. The effective sample size was 15.887 and the inferential error variance estimate was $v_w^2 = 1.547$. The overall model was significant at the 5% level, with a p-value of 0.000+. The coefficient of determination was $R^2 = 0.979$. It was seen that x_3 was not significant (at the 5% level) in the presence of x_1 and x_2 (p-value = 0.140). The observation weight plot of Figure 5.14 shows the complete downweighting of four observations, the same four that are typically addressed in the literature as being problematic. The dendrogram (Figure 5.14) illustrates how dramatically different observation 21 (and observation 4, for that matter) is with respect to the general trend observations. It is also interesting to note that observations 1 and 3 are most similar to observations 13 and 15, two observations that were downweighted upon activation (observation 13 more severely downweighted). In fact, observations 13 and 15 were removed from the main cluster between the first and second cluster phase, indicating that these two observations may be on the fringe of the general trend.

§5.7 Chapter Summary

The CBI methodology has been presented, the algorithm details provided and two examples offered for illustrative purposes. Theoretical properties of the CBI regression estimator will be presented next, in Chapter 6. Other case studies often mentioned in the literature along with Monte Carlo simulations will then offer insight into the performance capabilities of the CBI methodology.

The proposed CBI methodology is a comprehensive regression analysis procedure. The goal is to be competitive with methods such as LTS, M1S and S1S when the data is highly contaminated but also be able to compete with the efficient M and BI regression methods when the data has few or no problematic observations. That the user can rely on the CBI method to perform well across the spectrum of data contamination levels is an advantage, especially when the user may not be savvy with respect to the finer details of robust regression. Additionally, the CBI methodology provides valuable insight into the data structure, identifying multiple outliers or subgroups of similar observations. With a dendrogram illustrating the cluster history, a minor

cluster activation summary and a final CBI regression estimator, scale estimate and observation weights along with a robust ANOVA for inferential statistics, a CBI regression analysis provides an extensive amount of information in a compact tabular and graphical summary form.