Chapter 2

Low Breakdown Regression Procedures

Introduction

The very real possibility of a complete breakdown in OLS regression leads to searching for alternative solutions. A regression procedure is said to be *robust* if the presence of contaminated data does not impede the procedure's ability to capture the general trend of the data. This idea is quantified into a measure referred to as the *breakdown point* of a particular regression method. Simplistically, a large breakdown point means that the method has the ability to withstand large amounts of contaminated data without a ruining of the analysis. Technically, using the definition given in Simpson, Ruppert and Carroll (1992), who cite Donoho and Huber (1983), the breakdown point $\mathrm{BP}\left(T, \mathbf{X}_y\right)$ of an estimator *T*, given the data matrix $\mathbf{X}_y$, is

$$\mathrm{BP}\left(T, \mathbf{X}_y\right) = \min\left\{\frac{m}{n} : \sup_{\mathbf{X}_y^*} \left\| T(\mathbf{X}_y) - T(\mathbf{X}_y^*) \right\| = \infty\right\},$$

where the supremum is over all possible choices of $\mathbf{X}_y^*$ that consist of $n-m$ data points and *m* arbitrary points. With an upper bound of 50%, it is desirable to have a regression method that achieves a large breakdown point while still retaining high efficiency versus OLS under the regression model with normal errors.

There are several competing methods in the literature that are possible alternatives to OLS. Some fall into the category of nonparametric regression, where the prescribed model is set aside in favor of fitting the scatterplot by some data-driven method. These methods are generally not robust. Some of the major nonparametric regression methods are "nearest neighbor" methods. These include kernel regression, loess (Cleveland (1979)), and local polynomial regression. The basic philosophy is that the fit at a given location should be based more on local points and less on distant points. Of these three methods, loess is the only robust regression

technique (although not a high breakdown technique), having the ability to keep an outlier from pulling the fit in its neighborhood away from the general trend. A fourth method is spline smoothing. This is the optimal scatterplot fitting procedure when the objective function is based on a penalty function with certain smoothing constraints placed on the possible functions used to fit the data. This procedure is sensitive to outliers, and falls into the non-robust category as well.

Parametric methods still retain the prescribed model but replace the optimization function with something that is more robust to unusual data points. These types of methods are the focus of this chapter as well as Chapter 4. This chapter introduces some of the earlier work in this field, which falls into the category of low breakdown regression procedures. Each method is described in detail and illustrated via the stackloss case study.

Computational considerations are also addressed. While all of the robust regression procedures are computationally intensive in varying degrees, difficulties and drawbacks can arise in certain methods. This is especially true regarding the high breakdown regression procedures found in Chapter 4.

§2.1 M and Bounded Influence Regression

Consider again the objective function for OLS, but written as

$$\min_{\forall \mathbf{b}} \sum_{i=1}^{n} \rho(r_i),$$

with $\rho(t) = t^2$ representing the OLS argument inside the summation. Having $\rho(t) = t^2$ is optimal under the classical regression assumptions of iid normal errors, as it leads to parameter estimates that are BLUE (best linear unbiased estimators). In addition, under these conditions the OLS parameter estimates are also the Maximum Likelihood Estimates (MLE's). These are, of course, academic results. Violations to the assumptions are not uncommon in real data, as unusual observations may be present, for example.

In robust regression, the function ρ can be selected to either downweight or bound any arguments rising from unusual observations. This becomes the basis for *M regression* (Huber (1981)), which has the objective function

$$\min_{\forall \mathbf{b}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i'\mathbf{b}}{\hat{\sigma}}\right),$$

where $\hat{\sigma}$ is some appropriately chosen estimate of σ. The choice is generally limited to robust measures of scale. One possible estimate that is frequently used is the *median absolute deviation* (MAD), defined as

$$\hat{\sigma} = 1.4826 \cdot \underset{\forall i}{med} \mid r_i - \underset{\forall i}{med}\, r_i \mid ,$$

where the constant 1.4826 makes MAD a consistent estimator of σ under normal errors.

The ρ-function downweights observations with scaled residuals that are deemed too large in magnitude. Taking derivatives with respect to **b** leads to solving *p* "*altered" normal equations*,

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\mathbf{x}_i = \mathbf{0},$$

where $\psi = \rho^{(1)}$ and $\hat{\boldsymbol{\beta}}$ is the solution for **b**. These altered normal equations form a system of nonlinear equations. There are several methods available to solve them; two examples are Newton-Raphson and Iterated Reweighted Least Squares (IRLS). During this research, IRLS is the method used. A more detailed discussion of IRLS is found in Section 2.3. At convergence, IRLS produces the M regression parameter estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y},$$

where **W** is an $n \times n$ diagonal matrix. The diagonal elements of **W** consist of observation weights $w_1, w_2, \ldots, w_n$, where $0 \le w_i \le 1, \forall i$. Thus, **W** is commonly referred to as a weight matrix. Each weight, w_i, determines how much emphasis the regression will put on a particular observation. A large weight (near 1) *should* indicate a good observation. An outlier or high influence point, on the other hand, *should* get a reduced weight or perhaps even a zero weight. Keeping to these weighting guidelines, especially in the presence of multiple outliers and/or high influence points,

reflects on the quality of the regression procedure and, specifically, on its breakdown point. In M regression the weights are calculated as

$$w_i = \frac{\psi(r_i / \hat{\sigma})}{r_i / \hat{\sigma}}.$$

In order to illustrate this procedure and its application to robust regression, further mention of the ψ-function is warranted beforehand. This discussion is found in the next section; but first consider an alternative to M regression. A second class of estimators that incorporates both an observation's scaled residual and its location in the regressor space is the bounded influence (BI) estimator (Huber (1981)). Here, the objective function is

$$\min_{\forall \mathbf{b}} \sum_{i=1}^{n} \eta\left(\mathbf{x}_i', \frac{y_i - \mathbf{x}_i'\mathbf{b}}{\hat{\sigma}} \right).$$

Taking derivatives with respect to **b** yields another set of altered normal equations. The most common form for these equations is to incorporate the location of each observation into a pre-defined quantity, denoted as π_i. In other words, π_i is a function of **X** alone. One form of BI regression, using what is commonly referred to as the Schweppe-style weighting scheme, solves the altered normal equations

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{\pi_i \hat{\sigma}} \right) \mathbf{x}_i = \mathbf{0}.$$

The choice of π_i, referred to as the BI weight, affects which data structures the BI-estimator can and cannot handle. For example, if all π_i are 1, then this reduces to the M-estimator. Here, outliers at high leverage locations cannot be handled well. This is a direct result of the argument being a scaled residual, with high influence points generally having small residuals due to their "pull." Thus, the high influence point is typically not downweighted much, if at all. On the other hand, suppose that we define $\pi_i = (1 - h_{ii}) / \sqrt{h_{ii}}$, as recommended by Welsch (1982). This BI-estimator can handle an outlier with high leverage. This is because the observation weights are now $w_i = \psi(r_i / \pi_i \hat{\sigma}) / (r_i / \pi_i \hat{\sigma})$, which becomes $w_i = \psi("DFFITS_i") / "DFFITS_i"$. Here $"DFFITS_i"$ refers to the fact that the argument is nearly $DFFITS_i$, except that the standard

deviation used here is $\hat{\sigma}$ instead of s_{-i}. Since *DFFITS* is a high influence diagnostic, a single high influence point is downweighted due to a large $DFFITS_i$. In this setting, the π_i values are also referred to as *leverage weights*.

The bounded influence estimator introduced above is actually a member of the family of estimators that solve altered normal equations having the form

$$\sum_{i=1}^{n} \mathbf{x}_i \, u(\mathbf{x}_i) \, \psi\left(\mathbf{x}_i, \frac{r_i(\hat{\boldsymbol{\beta}})}{v(\mathbf{x}_i) \;\; \hat{\sigma}} \right) = \mathbf{0},$$

with $u(\mathbf{x}_i) = v(\mathbf{x}_i) = \pi_i$. The M estimator also takes this form, having $u(\mathbf{x}_i) = v(\mathbf{x}_i) = 1$. This class of estimators itself belongs to the broader family of *generalized M* (or GM, for short) estimators, which solve

$$\sum_{i=1}^{n} \mathbf{x}_i \varphi\left(\mathbf{x}_i, \frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}} \right) = \mathbf{0}.$$

Examples to illustrate the characteristics of M and BI regression are forthcoming after some technical issues are addressed in the next two sections.

§2.2 The ψ-function

As seen in the previous section, both the M and BI regression procedures have altered normal equations that involve a ψ-function. This ψ-function dictates some robustness properties of the estimator. It determines whether to downweight or bound a large value of its argument, which generally involves a scaled residual. One possible choice is the Huber ψ-function, defined as

$$\psi(t) = \begin{cases} -c_H, & if \; t < -c_H, \\ t, & if \; |t| < c_H, \\ c_H, & if \; t > c_H. \end{cases}$$

where c_H is a constant (often referred to as the *tuning parameter*). To attain 95% efficiency under normal errors for the location model, let $c_H = 1.345$.

Also frequently used is the bisquare ψ-function, which is defined as

$$\psi(t)=\begin{cases}0, & if\ t<-c_B,\\ t\left(1-\left(t/c_B\right)^2\right)^2, & if\ |t|<c_B,\\ 0, & if\ t>c_B.\end{cases}$$

The tuning parameter, c_B, is given the value of 4.685 to achieve 95% efficiency under normal errors for the location model.

The above critical values were derived for the location model, and are appropriate for use in M regression. However, for BI regression it is generally preferred to incorporate the correction factor

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

to account for the presence of π_i in the ψ-function argument that is used to produce the BI estimator (Walker (1984), Birch and Agard (1993)). In this case, the Huber and bisquare ψ-functions use

$$c_H=1.345\sqrt{2pn}/(n-2p)$$

and

$$c_B=4.685\sqrt{2pn}/(n-2p)$$

respectively.

In order to perform IRLS, each observation must be assigned a weight. Iteration means that these weights need to be recalculated at each step. The ψ-functions are used to calculate the weights given to each observation. General weight functions are of the form $w(t)=\psi(t)/t$. The Huber weight function is expressed as

$$w(t)=\begin{cases}-c_H/t, & if\ t<-c_H,\\ 1, & if\ |t|<c_H,\\ c_H/t, & if\ t>c_H,\end{cases}$$

with the bisquare weight function becoming

$$w(t)=\begin{cases}0, & if\ t<-c_B,\\ \left(1-\left(t/c_B\right)^2\right)^2, & if\ |t|<c_B,\\ 0, & if\ t>c_B.\end{cases}$$

A direct comparison between these two widely used weight functions is made graphically in Figure 2.1. In general, the Huber gives full weight (equal to 1) to arguments in some neighborhood of zero, and gradually reduces the weight as the argument becomes larger in magnitude. The arguments are scaled residuals and, hence, zero is the target value. The bisquare, on the other hand, is a smoother weight function. Starting with full weight at arguments of zero, it gradually assigns smaller weights until the tuning parameter is reached.

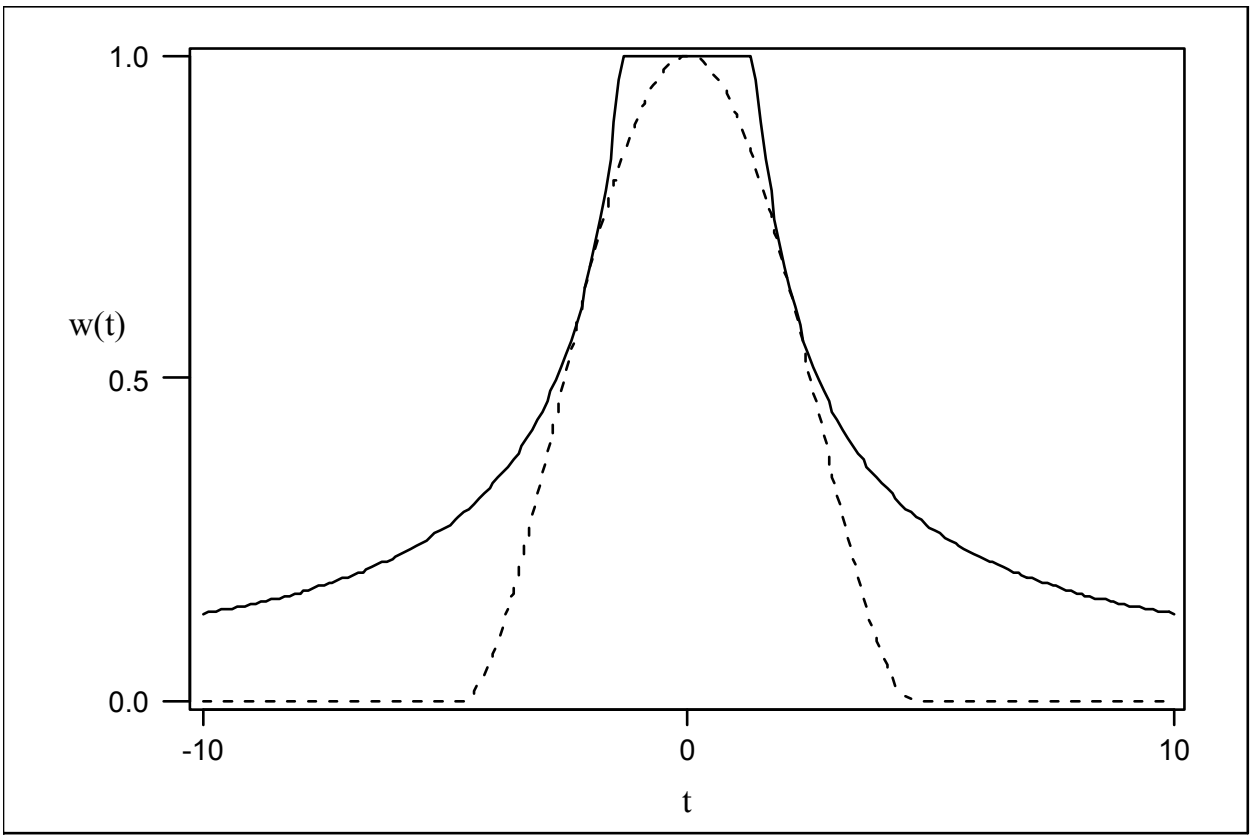

Figure 2.1: Huber(solid line) and bisquare (dashed line) weight functions.

Each argument beyond this value (in magnitude) receives a zero weight. There are many other ψ-functions that could be used. Andrews et. al. (1972) studied many competing ψ-functions in the context of location estimation. The Huber and bisquare functions were highly rated in overall performance. Together, they offer the option of using either a nondescending function (conservative weighting scheme) or a redescending function (liberal weighting scheme) respectively.

§2.3 Iterated Reweighted Least Squares (IRLS)

An estimator may be defined through a set of altered normal equations that cannot be solved explicitly in closed form. M and BI estimators are of this type. Iterated Reweighted Least Squares (IRLS) is one alternative method to obtain this estimator. Beginning with an initial estimator, the IRLS procedure recursively updates the estimate until convergence is reached. The following describes the IRLS procedure in more detail.

Step 1: Select an initial estimator. One such possible estimator is to use OLS as the starting value, and then proceed to try and improve upon it.

Step 2: Calculate the residuals, $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. From this vector the scale estimate $\hat{\sigma}$ is calculated. MAD could be used here. Also, this scale estimate may be iterated, i.e. recalculated at each iteration step, if desired.

Step 3: Weights are now calculated, their form determined by the normal equations. In M regression, the weights are calculated as $w_i = \psi(r_i / \hat{\sigma})/(r_i / \hat{\sigma})$. The weights in BI regression involve π_i's that need to be determined before the weights are calculated. According to Walker (1984), obtain weighted hat diagonals, $h_{ii,w}$, using the weighted least squares (WLS) hat matrix $\mathbf{H}_{\mathbf{W}} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$. Then, $\pi_i = (1 - h_{ii,w})\Big/\sqrt{h_{ii,w}}$ and each observations weight is calculated by $w_i = \psi(r_i / \pi_i\hat{\sigma})/(r_i / \pi_i\hat{\sigma})$.

Step 4: Use WLS to produce a new interim estimator from $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$. If this current estimate is *different* from the previous estimate then go back to *Step 2*. Here, different could be measured as the maximum relative change of the estimate vector. For example, if $\max_{\forall j} \left| \left(\hat{\beta}_{j,new} - \hat{\beta}_{j,previous} \right) \Big/ \hat{\beta}_{j,previous} \right| \leq \delta$, where $\delta \to 0^+$, then convergence has been achieved.

§2.4 Case Study: Stackloss Data

Since there is an absence of high leverage points in the stackloss data, it is better to use M regression rather than bounded influence regression. Using the Huber ψ-function (with $c_H = 1.345$) yielded, at convergence, the fitted equation

$$\hat{y}_i = -41.0512 + 0.8267\, x_{1i} + 0.9385\, x_{2i} - 0.1286\, x_{3i},$$

with the robust ANOVA table, obtained using the method described in Birch (1992), and parameter estimate summary given in Table 2.1.

Table 2.1: Robust analysis of variance table and parameter estimate summary for the M-Huber regression of the stackloss data, using a tuning parameter of 1.345.

Source	df	SS	MS	F	p-value
Regression	3	1823.19	607.73	87.63	0.000
Error	17	118.27	6.96		
Total	20	1941.46			

Parameter	Estimate	s.e.	t	p-value
Intercept	-41.0512	9.7153	-4.2254	0.0006
x_1	0.8247	0.1180	7.0062	0.0000
x_2	0.9385	0.3240	2.8964	0.0100
x_3	-0.1286	0.1273	-1.0103	0.3265

Table 2.2: Robust analysis of variance table and parameter estimate summary for the M-bisquare regression of the stackloss data, using M-Huber as the initial estimate and a tuning parameter of 4.685.

Source	df	SS	MS	F	p-value
Regression	3	1753.99	584.66	56.81	0.000
Error	17	104.68	6.16		
Total	20	1858.67			

Parameter	Estimate	s.e.	t	p-value
Intercept	-41.6703	11.9329	-3.4921	0.0028
x_1	0.8528	0.1486	5.7386	0.0000
x_2	0.8730	0.4061	2.1498	0.0463
x_3	-0.1224	0.1568	-0.7807	0.4457

Next, the M-bisquare regression analysis (with $c_B = 4.685$) is performed using the above M-Huber fit as the initial fit. This yielded the fitted equation

$$\hat{y}_i = -41.6703 + 0.8528x_{1i} + 0.8730x_{2i} - 0.1224x_{3i},$$

with Table 2.2 providing the robust ANOVA table and parameter estimate summary.

Table 2.3: Robust analysis of variance table and parameter estimate summary for the M-bisquare regression of the stackloss data, using a tuning parameter of 3.5 for the bisquare function and initial M-Huber estimates with a 1.01 tuning parameter.

Source	df	SS	MS	F	p-value
Regression	3	736.35	245.45	177.88	0.000
Error	17	14.14	0.83		
Total	20	750.49			

Parameter	Estimate	s.e.	t	p-value
Intercept	-37.0132	4.5890	-8.0656	0.0000
x_1	0.8229	0.0667	12.3434	0.0000
x_2	0.5068	0.1639	3.0926	0.0066
x_3	-0.0739	0.0603	-1.2254	0.2371

Another M-bisquare regression was run on the stackloss data, but with different tuning parameters. Here, $c_H = 1.01$ and $c_B = 3.5$, the settings used by Birch (1992). This resulted in the fitted equation

$$\hat{y}_i = -37.0132 + 0.8229\,x_{1i} + 0.5068\,x_{2i} - 0.0739\,x_{3i},$$

with the robust ANOVA table and parameter estimate summary as given in Table 2.3.

Neither of the first two M regression fits are performing that well, based on the much larger MSE values that were obtained by these two analyses in comparison to the MSE of the third analysis. This is due to the fact that, in general, the four outliers are not being downweighted much. However, all four outliers are zeroed out in the third case. The final observation weights for all three fitted equations can be viewed in Table 2.4. It is clear that the choice in tuning parameters is of some importance. Furthermore, it appears that using those values that attain 95% efficiency under normal errors for the location model can hinder the robustness of the method. The incorporation of a data-driven selection process for the tuning parameter would appear to be beneficial.

Table 2.4: Comparison of final observation weights for the three M regression analyses of the stackloss data.

Ob.	M-Huber $c_H = 1.345$	M-Bisquare $c_H = 1.345$ $c_B = 4.685$	M-Bisquare $c_H = 1.01$ $c_B = 3.50$	Ob.	M-Huber $c_H = 1.345$	M-Bisquare $c_H = 1.345$ $c_B = 4.685$	M-Bisquare $c_H = 1.01$ $c_B = 3.50$
1	1	0.911	0	**12**	1	0.985	0.997
2	1	0.936	0.911	**13**	1	0.929	0.414
3	0.817	0.828	0	**14**	1	0.988	0.809
4	0.526	0.563	0	**15**	1	0.943	0.847
5	1	0.964	0.951	**16**	1	0.992	0.999
6	1	0.919	0.864	**17**	1	0.994	0.982
7	1	0.963	0.992	**18**	1	1	1
8	1	0.993	0.956	**19**	1	0.998	0.963
9	1	0.942	0.919	**20**	1	0.974	0.710
10	1	0.999	1	**21**	0.383	0.234	0
11	1	0.977	0.951				

§2.5 Illustrating the Robustness Properties of OLS, M and BI Regression

In order to compare various regression methods, it is desirable to know in advance the true parameters. This knowledge provides the yardstick for analyzing how well the estimates are performing. Robustness properties can then be checked for various situations. For illustration and comparison purposes, the following six regression procedures are compared:

(1) OLS.

(2) M regression, Huber ψ, $c_H = 1.345$, iterated MAD.

(3) M regression, bisquare ψ, $c_B = 4.685$, iterated MAD.

(4) BI regression, Huber ψ, $c_H = 1.345 \cdot c$, iterated MAD.

(5) BI regression, bisquare ψ, $c_B = 4.685 \cdot c$, non-iterated MAD.

(6) WLS: $w_i = 1$, if good point; $w_i = 0$, otherwise.

OLS is used as the initial estimate for the Huber ψ-function based regressions. The results of these Huber ψ-function based regressions (at convergence) are used as the initial estimates for their corresponding bisquare ψ-function based methods.

The last regression method, WLS, is just OLS performed on only the good observations. This provides some insight into the general trend of the data, but cannot be considered a viable option since data construction information was used to determine the good observations. The main focus is on the first five regression methods.

Three different data situations will be examined. In Example 2.1 a low leverage outlier is present in a set of otherwise well-behaved data. Then, this outlier is moved to a high leverage position in Example 2.2. Finally, Example 2.3 contains an additional high leverage point to form a small cluster outside the general trend of the data.

§2.5.1 Example 2.1: A Single Low Leverage Outlier

To construct the data, eight observations were generated using the linear model

$$y_i = 100 - 4x_i + \varepsilon_i$$

under the assumption of iid normal errors, where $\varepsilon_i \sim N\left(\mu = 0, \sigma^2 = 25\right)$, with the regressor variable generated via $x_i \sim U[10, 20]$. Next, one observation is added. It is a low leverage outlier amidst data that is otherwise well behaved. The data is shown graphically in Figure 2.2, labeled by observation number. The data itself, along with their ordinary hat diagonals, are presented in Table 2.5. Clearly, observation 9 is a low leverage outlier, having the second smallest hat diagonal while missing the general trend completely.

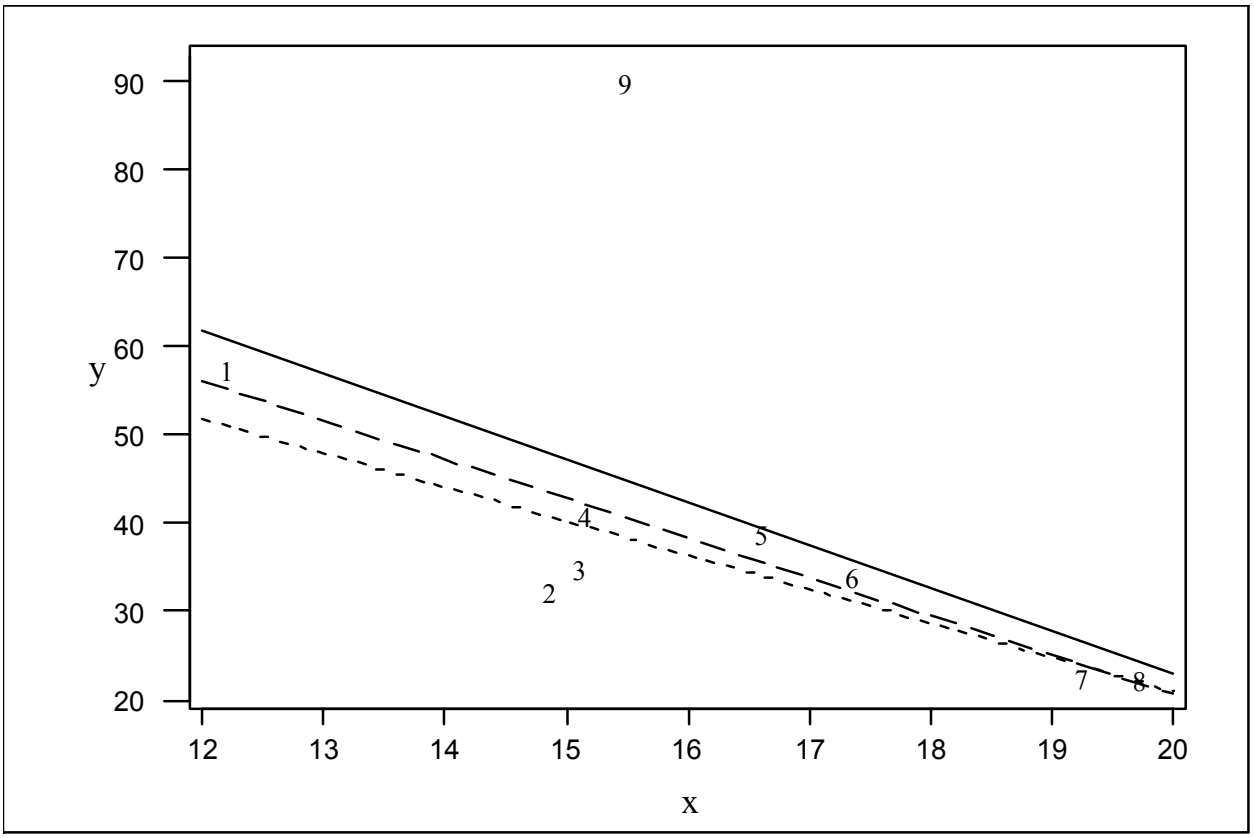

Figure 2.2: Various regression fits for Example 2.1. OLS is the solid line, M-bisquare is the dashed line, and BI-bisquare is the dotted line.

Table 2.5: Raw data and hat diagonals for Example 2.1.

Ob.	y_i	x_i	h_{ii}
1	57.340414	12.211352	0.476
2	32.252039	14.876992	0.151
3	34.772027	15.114503	0.138
4	40.789918	15.157843	0.136
5	38.930530	16.615033	0.115
6	34.106102	17.369173	0.142
7	22.631159	19.247127	0.323
8	22.403825	19.726283	0.396
9	90	15.5	0.122

Analysis of this data by each of the six competing methods yields the parameter estimates found in Table 2.6. Notice that the slope estimates are not dramatically different, although the BI-bisquare slope is a bit smaller (and better) than the first four slopes. Larger discrepancies occur among the intercept estimates. With OLS, the outlier pulls the estimate upwards to an inflated value of 119.895, when the true value is 100. Improvement over the OLS solution is reached as the outlier becomes downweighted. Three M/BI analyses produce very similar estimates, with the fourth, BI-bisquare, being the best. For reference, the WLS estimates are quite close to the true parameter values.

Again referring to Figure 2.2, it is seen that the OLS fit (the solid line) is consistently above the general trend, due to the "pull" of observation 9. An improved fit is found with the M-bisquare (dashed line) procedure, while an even better fit results from the BI-bisquare procedure. Notice that observation 1, possessing the largest hat diagonal, "pulls" the M-bisquare fit upwards a bit, while the BI-bisquare fit is unaffected by the leverage of this observation.

Table 2.6: Regression coefficients for Example 2.1.

	OLS	**M Huber**	**M bisquare**	**BI Huber**	**BI bisquare**	**WLS**
Intercept	119.895	109.625	109.394	109.842	97.976	100.788
x	-4.840	-4.439	-4.440	-4.413	-3.849	-4.014

Table 2.7: Final weights given to the nine observations of Example 1 for each of the six regression methods.

	OLS	M Huber	M bisquare	BI Huber	BI bisquare	WLS
1	1	1	0.9557	1	0.6425	1
2	1	0.3341	0.1084	0.8571	0.8523	1
3	1	0.4880	0.4762	1	0.9461	1
4	1	1	0.9795	1	0.9970	1
5	1	1	0.8898	1	0.9666	1
6	1	1	0.9661	1	0.9848	1
7	1	1	0.9812	1	0.9900	1
8	1	1	0.9964	1	0.9988	1
9	1	0.0770	0	0.3857	0	0

By viewing Table 2.7, it is seen that the outlier (observation 9) is downweighted to 0 by using a bisquare ψ-function for both M and BI regression. Using a Huber ψ-function still reduces the weight given to the outlier, although in BI regression it is 0.3857.

In conclusion, it is witnessed that both M and BI regression can handle a low (to moderate) leverage outlier, with the BI-bisquare method being better than the other three methods in this particular example. OLS, on the other hand, has difficulties obtaining accurate estimates. The final regression equation, based on the BI-bisquare, is

$$\hat{y}_i = 97.976 - 3.849\, x_i .$$

An interesting development occurs in the M-bisquare analysis of Example 2.1. It turns out that the IRLS algorithm does not converge to the estimator presented in Table 2.6. Rather, the algorithm alternates between the two potential solutions (local minima) shown in Table 2.8. Notice that there are only modest differences between weight vectors and estimated standard deviations (which themselves are noted to be poor estimates of scale). There are two potential problems in obtaining convergence. First, the scale estimate is being updated. The existence (and consistency, for that matter) proofs of Maronna and Yohai (1981) regarding the family of GM estimators require that the simultaneous estimate of scale (that is, the update of $\hat{\sigma}$ during IRLS) be found by solving

$$\sum_{i=1}^{n} \chi\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right) = \mathbf{0},$$

where χ is typically taken to be an even function (Huber (1981, pg. 109)). Huber (1981, pg. 176) uses this form as well, and mentions the lack of a convergence proof when MAD is employed as the scale estimate.

A second difficulty regarding IRLS convergence is that in BI regression the leverage weights are also being updated (this is, of course, not required). This can produce a situation where the minimization process has, basically, two different reference frames. In other words, the objective function is being modified by these updates, and each local minima that is obtained is really a representation of a minimization with respect to its own set of constants. The IRLS convergence proof for solving M and BI objective functions supplied by Huber (1981) assumes that the leverage weights remain at their initial settings. Furthermore, convergence has not been proved for the "modified leverage weight" iteration procedure that was outlined in Section 2.3. Guaranteed convergence is shown for the "modified residual" procedure, which is the Gauss-Newton procedure using metrically Winsorized residuals in place of ordinary residuals (see Huber (1981, pg. 182) for more details).

Table 2.8: Two local minima for M-bisquare analysis of Example 2.1.

	(i) $\hat{\boldsymbol{\beta}}' = [109.394 \quad -4.440]$ $\hat{\sigma} = 2.911$ obj.fct. = 0.329	**(ii)** $\hat{\boldsymbol{\beta}}' = [109.694 \quad -4.454]$ $\hat{\sigma} = 2.817$ obj.fct. = 0.340
1	0.9557	0.9471
2	0.1084	0.0861
3	0.4762	0.4562
4	0.9795	0.9804
5	0.8898	0.8787
6	0.9661	0.9621
7	0.9812	0.9803
8	0.9964	0.9960
9	0	0

Due to potential convergence problems, the proposed method, introduced later in Chapter 5, will not update either the scale estimate or the leverage weights during a given IRLS procedure. While this example led to only minor differences between the two resulting estimators, more dramatic differences in local minima are indeed possible.

It is further noted that if y_9 equals 60 instead of 90, then neither the M-bisquare procedure nor the BI-Huber procedure converge. So, this problem is not limited to either M or BI regression, nor to either bounded or redecending ψ-functions.

Finally, the selection between the two competing solutions for Example 1 was made on the basis of evaluating the M-bisquare objective function for both. From the values in Table 2.8, it is seen that case (i) is optimal, and that it is also slightly closer to the true parameters (even though its scale estimate is slightly higher than that of case (ii)).

§2.5.2 Example 2.2: A Single High Influence Point

Observation 9 from Example 2.1 (the low leverage outlier) is removed and replaced by a high influence point, thus forming Example 2.2. This new dataset and its ordinary hat diagonals are found in Table 2.9. Notice that observation 9 now has a huge hat diagonal of 0.818.

The comparisons are much different here than they were for Example 2.1. Recall, the true parameter vector is $\boldsymbol{\beta}' = [100 \quad -4]$. Figure 2.3 provides the scatterplot for Example 2.2, again with the OLS fit (solid line) displayed. It is obvious that OLS is performing poorly. Since all nine observations receive the same weight, the high influence point pulls the regression towards itself. Here, both parameter estimates are extremely poor, with the slope estimate having the wrong sign.

The resulting estimates for each of the six methods can be viewed in Table 2.10, with the associated final weights shown in Table 2.11. Figure 2.3 illustrates the obvious improvement in

Table 2.9: Raw data and hat diagonals for Example 2.2.

Ob.	y_i	x_i	h_{ii}
1	57.340414	12.211352	0.260
2	32.252039	14.876992	0.152
3	34.772027	15.114503	0.146
4	40.789918	15.157843	0.145
5	38.930530	16.615033	0.118
6	34.106102	17.369173	0.112
7	22.631159	19.247127	0.121
8	22.403825	19.726283	0.129
9	60	30	0.818

fitting the general trend of the data when using BI-bisquare regression (dotted line). The high influence point is no longer posing any difficulty.

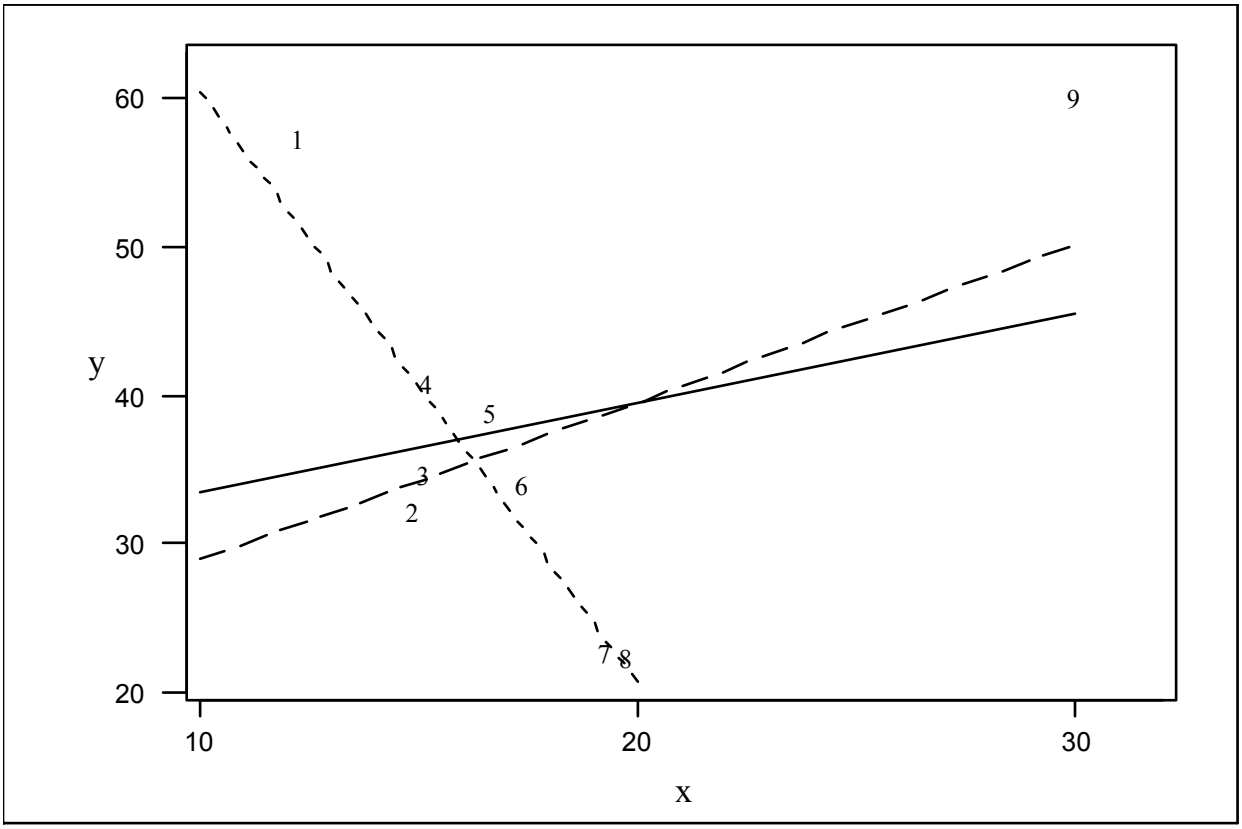

Figure 2.3: Various fits to the Example 2.2 data. OLS is the solid line, M-bisquare is the dashed line, and BI-bisquare is the dotted line.

Table 2.10: Regression coefficients for Example 2.2.

	OLS	M Huber	M bisquare	BI Huber	BI bisquare	WLS
Intercept	27.400	19.154	18.297	54.682	100.194	100.788
x	0.603	1.045	1.062	-1.145	-3.979	-4.014

Table 2.11: Final observation weights for the six competing regression procedures in the analysis of Example 2.2.

	OLS	M Huber	M bisquare	BI Huber	BI bisquare	WLS
1	1	0.4684	0.3657	0.8169	0.9200	1
2	1	1	0.9961	1	0.9749	1
3	1	1	0.9998	1	0.9917	1
4	1	1	0.9529	1	0.9998	1
5	1	1	0.9896	1	0.9949	1
6	1	1	0.9920	1	0.9975	1
7	1	0.7161	0.7214	1	0.9991	1
8	1	0.6861	0.6977	1	0.9993	1
9	1	1	0.8903	0.2047	0	0

Both M regression estimators tend to follow OLS. A single high influence point dominates the analysis. As seen in Table 2.11, observation 9 was hardly downweighted with the bisquare ψ-function, and given full weight with the Huber ψ-function.

It is seen that bounded influence regression is superior to the other methods, particularly when incorporated with the bisquare ψ-function. It is on par with the WLS procedure, which uses data construction information to assign weights. The redescending bisquare function enables the method to fully eliminate the high influence point, witnessed by the zero weight in Table 2.11. The final regression equation is

$$\hat{y}_i = 100.194 - 3.979\, x_i ,$$

a very accurate result for this dataset.

For completeness, there were no convergence problems with any of the procedures regarding the analysis of this data.

§2.5.3 Example 2.3: A High Influence Cluster

Next, consider the data of Example 2.2 with one additional high influence point as observation 10, becoming Example 2.3. This data will serve to illustrate the low breakdown points associated with M and BI regression. In Table 2.12, which displays both this new dataset and the ordinary hat diagonals, it is seen that there are, indeed, two large leverage points ($h_{99} = 0.422$ and $h_{10\,10} = 0.484$). It is seen from Figure 2.4 that observations 9 and 10 actually form a high influence cluster.

The parameter estimates for the same six regression methods are displayed in Table 2.13, with no convergence problems witnessed. Keeping the WLS results aside, the five remaining

Table 2.12: Raw data and hat diagonals for Example 2.3.

Ob.	y_i	x_i	h_{ii}
1	57.340414	12.211352	0.231
2	32.252039	14.876992	0.149
3	34.772027	15.114503	0.144
4	40.789918	15.157843	0.143
5	38.930530	16.615033	0.117
6	34.106102	17.369173	0.108
7	22.631159	19.247127	0.100
8	22.403825	19.726283	0.101
9	60	30	0.422
10	60	31	0.484

Table 2.13: Regression coefficients for Example 2.3.

	OLS	M Huber	M bisquare	BI Huber	BI bisquare	WLS
Intercept	20.175	13.308	10.537	17.749	18.399	100.788
x	1.053	1.425	1.610	1.152	1.115	-4.014

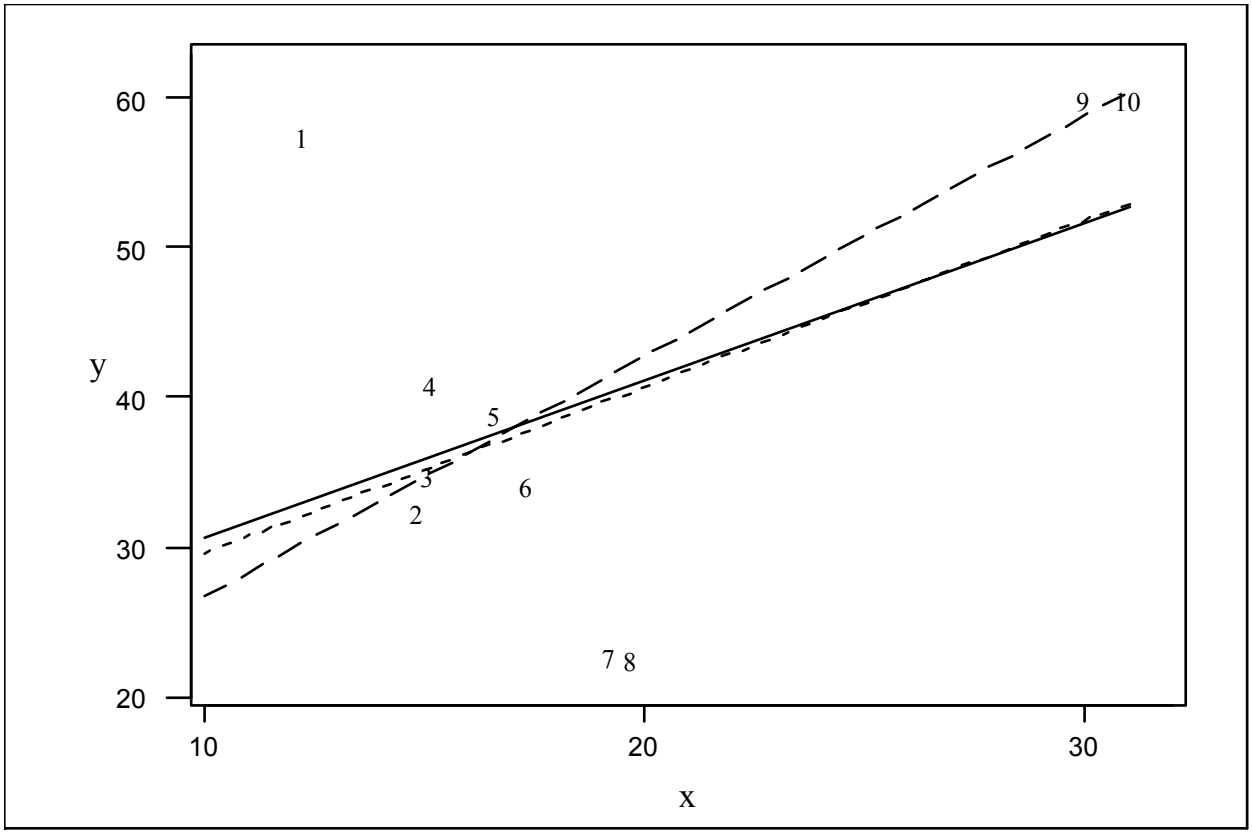

Figure 2.4: Various regression fits for Example 2.3. OLS is the solid line, M-bisquare is the dashed line, and BI-bisquare is the dotted line.

methods produce very similar estimates. Figure 2.4 displays the OLS, M-bisquare and BI-bisquare fits (bisquare chosen over Huber due to the presence of high influence points), none of which follow the general trend. The final weights are found in Table 2.14, which shows very little to no downweighting of the high influence pair across the board. In fact, an occasional good observation is downweighted instead. Essentially, the five regression procedures break down due to the presence of a cluster of high influence points. The general trend that is described by the WLS method cannot be captured with any of these other methods.

In conclusion, M and BI regression both run into trouble when clusters of high influence points are present in data. As is the case with OLS, these methods can break down with a small

Table 2.14: Final observation weights for the six competing regression procedures in the analysis of Example 2.3.

	OLS	M Huber	M bisquare	BI Huber	BI bisquare	WLS
1	1	0.3025	0.0000	0.7938	0.7626	1
2	1	1	0.9778	1	0.9977	1
3	1	1	1.000	1	0.9999	1
4	1	1	0.8527	1	0.9914	1
5	1	1	0.9879	1	0.9991	1
6	1	1	0.9156	1	0.9973	1
7	1	0.4452	0.0408	1	0.9500	1
8	1	0.4239	0.0133	1	0.9451	1
9	1	1	0.9939	1	0.8881	0
10	1	1	0.9991	1	0.8820	0

percentage of contaminated data. M regression utilizes scaled residuals, but, since residuals are generally small at high influence points, poor estimation can result. This can even occur when using a more resistant initial estimator (rather than OLS) because the intermittent IRLS estimates may migrate towards the high influence point during the iteration process. BI regression, using Welsch leverage weights, incorporates a single-point deletion *DFFITS* diagnostic to combat high influence. Since this diagnostic cannot handle clusters of high influence points (or, more generally, jointly influential observations), success is limited to data with few bad observations. This is because if one high influence point is removed, the other high influence point(s) will still control the regression and no significant difference in fits will be observed. Therefore, we have no useful regression equation to report in Example 2.3; a better regression method is required.

§2.5.4 Concluding Remarks on the Previous Three Examples

It has been illustrated that OLS is a very non-robust regression procedure. In fact, OLS has a $1/n$ breakdown point (Rousseeuw and Leroy (1987), pg. 10). It has also been shown that it is not difficult to obtain data with few bad points such that M and BI regression perform poorly. Any generalized M estimator has an upper bound of $1/(p+1)$ (Simpson, Ruppert, and Carroll (1992)) for its breakdown point. More importantly, however, is that there is no guarantee that a particular GM estimator even attains this upper bound as its breakdown point; it may be much

smaller. For example, the breakdown point for M regression is only $1/n$ because of its sensitivity to one high leverage point (Rousseeuw and Leroy (1987, pg. 149)). While BI regression can handle at least one bad observation, this procedure may not reach the upper bound either; it depends on the dataset being analyzed.

Maronna, Bustos, and Yohai (1979) consider the "gross error" breakdown point for GM estimators when the regressors are themselves random variables, and show that the upper bound of this quantity tends to 0% as p increases.

To conclude this discussion, the following are a few guidelines (Birch (1993)) for selecting among the OLS, M, and BI regression procedures:

1. If $\hat{\boldsymbol{\beta}}_{OLS} \approx \hat{\boldsymbol{\beta}}_{M} \approx \hat{\boldsymbol{\beta}}_{BI}$, use OLS as there are no unusual observations.
2. If $\hat{\boldsymbol{\beta}}_{OLS} \approx \hat{\boldsymbol{\beta}}_{BI} \neq \hat{\boldsymbol{\beta}}_{M}$, use M as there is (are) a low to moderate leverage outlier(s).
3. If $\hat{\boldsymbol{\beta}}_{OLS} \approx \hat{\boldsymbol{\beta}}_{M} \neq \hat{\boldsymbol{\beta}}_{BI}$, use BI as there is (are) a high influence point(s).

Due to the low breakdown properties involved these are, of course, only suggestions. Setting 1 may be observed yet all three estimators are misleading, as in Example 3. Therefore, the study of multiple outliers and joint influence has definite merit. In conjunction, further expansion into regression techniques that are more robust (i.e. have larger breakdown points) is required.

§2.6 Joint Influence Diagnostics

Single-point deletion statistics were previously discussed in Sections 1.4.1 and 1.4.2. For example, $DFFITS_i$ represents the (scaled) difference in the fitted value at $\mathbf{x}_i'$ by regressing both with and without the i^{th} observation. As seen in the earlier sections of Chapter 2, M or BI regression may remedy the breakdown of OLS due to the presence of one bad point. However, it was also witnessed that M and BI can both fail when a cluster of bad points (as few as two

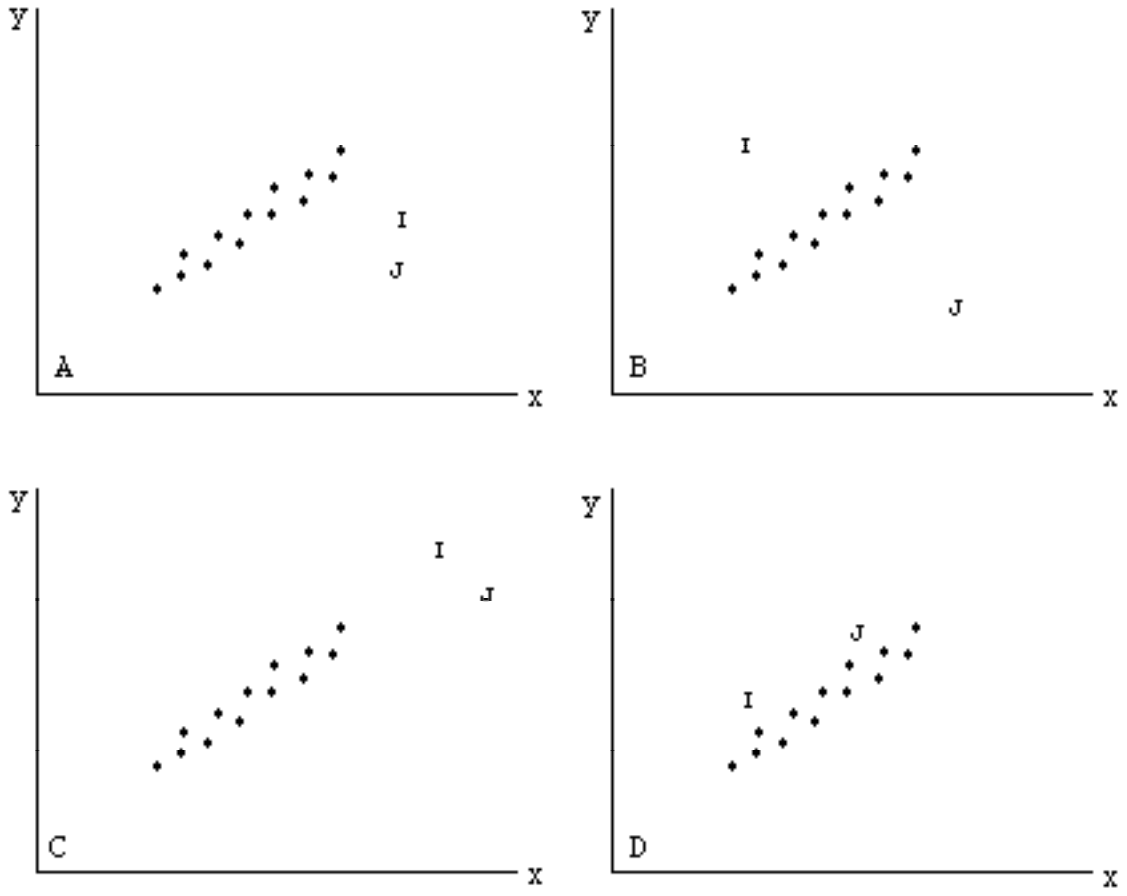

Figure 2.5: Illustration of joint influence in SLR. (A) and (B) show a jointly influential pair; (C) and (D) show a non-influential pair.

in some cases for BI, or just one high influence point for M) are present in the data. Thus, it becomes important to have a means for identifying these multiple outliers. Using a scatterplot is great for simple linear regression (one regressor), as illustrated in Figure 2.5 (similar to Figure 1 in Hadi (1985)), but not practical beyond two regressors.

A group of observations are *jointly influential* if their removal substantially changes the regression. Joint influence can then be viewed in terms of multi-point deletion diagnostics. Let I represent a set of indices used to define a subsample of the dataset. For notation purposes, a "-I"

subscript indicates that all observations in I are omitted from either the definition of, or the calculation of, the specific argument.

The first joint influence diagnostic is the extension of $Cook's\ D$ to subsets, indexed by I, given as

$$Cook's\ D_I = \frac{(\hat{\boldsymbol{\beta}}_{-I} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{-I} - \hat{\boldsymbol{\beta}})}{p\ s^2}.$$

This diagnostic uses the mean square error (MSE), denoted by s^2, obtained via the full regression (i.e. using all n observations) analysis in the denominator. For the single point deletion case, this reduces to

$$Cook's\ D_i = \frac{1}{p}\frac{r_i^2}{s^2(1-h_{ii})}\frac{h_{ii}}{(1-h_{ii})}.$$

For both the single and cluster cases, Cook and Weisberg (1980) suggest viewing $Cook's\ D$ as F-like to obtain a critical value such as $F_{0.95,p,n-p}$, for example, with influential subsets corresponding to those values of $Cook's\ D$ that exceed this critical value. In this section, $Cook's\ D_I > 1$ is used instead.

Recall that in the single point deletion case, $DFFITS_i$ is focused on the scaled difference in prediction at the location of the ith observation, given by

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}}.$$

When a cluster of m observations is considered, there are m of these fitted values. Thus, to quantify the diagnostic into one number, a different interpretation is required. It then becomes convenient to view $DFFITS_i$ instead by the equivalent expression

$$DFFITS_i^2 = \frac{\sum_{j=1}^{n}(\hat{y}_{j,-i} - \hat{y}_j)^2}{s_{-i}^2}.$$

To extend this notion to a multiple deletion diagnostic, divide this quantity by *m* to represent an average effect coming from each of the *m* observations being deleted. The new difference in fits diagnostic becomes

$$DFFITS_I^2 = \frac{\sum_{i=1}^{n} (\hat{y}_{i,-I} - \hat{y}_i)^2}{m\, s_{-I}^2}.$$

A direct comparison is available between the *DFFITS* diagnostic and the *Cook's D* diagnostic. Aside from a constant, the two diagnostics are only different in the scale estimator that is used, whether the full data or the reduced data is used in its calculation. It can be shown that the relationship between the two diagnostics is

$$DFFITS_I^2 = \frac{p\, s^2}{m\, s_{-I}^2} Cook's\ D_I .$$

Other influence diagnostics have direct extensions to the multiple point deletion scenario as well. For instance, the diagnostic *CovRatio* becomes

$$CovRatio_I = \frac{\left|(\mathbf{X}'_{-\mathrm{I}}\mathbf{X}_{-\mathrm{I}})^{-1} s_{-I}^2\right|}{\left|(\mathbf{X}'\mathbf{X})^{-1} s^2\right|}.$$

Another diagnostic, introduced by Andrews and Pregibon (1978), looks like *CovRatio* but is very different. It is based on the determinantal ratio

$$R_I = \frac{\left|\, \mathbf{X}'_{y,-\mathrm{I}}\mathbf{X}_{y,-\mathrm{I}} \,\right|}{\left|\, \mathbf{X}'_y\mathbf{X}_y \,\right|},$$

and measures the extremeness of cluster I in a geometric sense. Note that $\mathbf{X}_y$ (not $\mathbf{X}$) is the matrix involved, that no inversion is required and that no variance estimates appear in this diagnostic, in contrast to the *CovRatio* diagnostic. Small values of R_I indicate joint influence. Andrews and Pregibon (1978, pg. 89) show that R_I is proportional to a $Beta\big((n-p-m)/2,\ m/2\big)$ variate, and thus the mean of this distribution, $(n-p-m)/(n-p)$,

could be used as a reference point. Specifically, any R_I that is less than half of this mean would correspond to a subset with high joint influence. For the single-point deletion case, this diagnostic reduces to

$$R_i = 1 - h_{ii} - \frac{r_i^2}{(n-p)\, s^2} .$$

Another measure of joint influence for clusters is provided by Welsch (1982). He recommends using

$$Q_I^2 = \frac{\mathbf{B}_I'(\mathbf{X}_{-I}'\mathbf{X}_{-I})\mathbf{B}_I}{(n-m)\, s_{-I}^2},$$

where

$$\mathbf{B}_I = (1/m)\sum_{i \in I}(n-m)(\mathbf{X}_{-I}'\mathbf{X}_{-I})^{-1}\mathbf{x}_i'(\mathrm{y}_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{-I}) .$$

As a conservative yardstick, Welsch (1982, pg. 159) suggests that any $Q_I^2 \geq \frac{9p}{n-1}$ be reflective of an influential subset. The Q_I^2 statistic simplifies in the single-point deletion case to

$$Q_i^2 = (n-1)\frac{r_i^2}{s_{-i}^2 (1-h_{ii})^2}\frac{h_{ii}}{(1-h_{ii})} .$$

Furthermore, this is a function of the single-point deletion diagnostic *DFFITS*, written as

$$Q_i^2 = \frac{n-1}{1-h_{ii}} DFFITS_i^2 .$$

To illustrate these measures of joint influence, consider the data of Example 2.3, in Section 2.5.3. Recall that observations 9 and 10 form a high influence cluster. First, consider the single point deletion influence diagnostics for observation 10. With components obtained using the REG procedure in SAS®, three of the diagnostics are calculated as follows:

$$R_{10} = 1 - 0.4841 - \frac{7.1793^2}{(10-2)\,178.5191} = 0.4798$$

$$Q_{10}^2 = (10-1)\frac{7.1793^2}{189.74914(1-0.4841)^2}\frac{0.4841}{(1-0.4841)} = 8.6194$$

$$Cook's\ D_{10} = \frac{1}{2}\frac{7.1793^2}{178.51905(1-0.4841)}\frac{0.4841}{(1-0.4841)} = 0.2626$$

A more general SAS/IML® program is used to allow for clusters of size one or more. The above results were verified and are displayed in Table 2.15. In four of the five diagnostics, observation 10 does not appear to be a high influence point (only $Q_I^2 > 2$ signals the high influence of this observation).

Table 2.15: Influence diagnostics for observation 10 of Example 2.3, Section 2.5.3.

I = {10} s^2 = 178.51905 s^2_{-I} = 189.74914		
Diagnostic	**Value**	**Yardstick**
Cook's D	0.2626	> 1
R	0.480	< 0.4375
Q^2	8.619	> 2
DFFITS	0.703	> 0.8944
CovRatio	2.190	< 0.40

Table 2.16: Joint influence diagnostics for observations 9 and 10 of Example 2.3.

I = {9, 10} s^2 = 178.51905 s^2_{-I} = 189.74914		
Diagnostic	**Value**	**Yardstick**
Cook's D	33.835	> 1
R	0.011	< 0.375
Q^2	46983.58	> 2
DFFITS	22.163	> 0.8944
CovRatio	0.202	< 0.40

Table 2.17: Joint influence diagnostics for observations 4 and 21 of stackloss data.

I = {4, 21} **s^2 = 10.5194** **s^2_{-I} = 3.9855**		
Diagnostic	**Value**	**Yardstick**
Cook's D	1.160	> 1
R	0.207	< 0.44
Q^2	83.211	> 1.80
DFFITS	2.475	> 0.87
CovRatio	0.033	< 0.43

Next, the cluster $I = \{9,10\}$ is analyzed for joint influence, with Table 2.16 providing the diagnostics. It is obvious that this cluster has very large joint influence, as witnessed by the large *Cook's D*, small *R*, huge Q^2, large *DFFITS*, and small *CovRatio*. This shows that, aside from Q^2_{10}, observation 9 was masking the influence of observation 10 in the single-point deletion diagnostics (and vice versa). Removing one high influence point is not enough because the other high influence point still dominates the OLS fit.

Next, the stackloss data is used to further illustrate these diagnostics. Consider the cluster $I = \{4,21\}$, which represents two of the four outliers present in the data. Table 2.17 displays the joint influence diagnostics for this pair. All of the diagnostics indicate the presence of joint influence here. This is an interesting finding, as these two observations are not clustered. The proposed method will offer another viewpoint (later in Chapter 5) regarding joint influence for the stackloss data. Their (observations 4 and 21) relation to the general trend is not terribly similar.

§2.7 Multiple Outlier Detection Procedures

There are several procedures available in the literature that addresses the problem of identifying multiple outliers in the regression setting. Two such methods are now discussed. The first utilizes the joint influence diagnostics of the previous section, while the second is a simulation-based graphical procedure. For a more comparative discussion on multiple outlier

detection, refer to Kempthorne and Mendel (1990), Penny and Jolliffe (2001), Becker and Gather (2001) or Wisnowski, Montgomery and Simpson (2001).

§2.7.1 K-Clustering

The altered hat matrix, defined in Section 1.3.3 as $\mathbf{H}_y = \mathbf{X}_y(\mathbf{X}_y'\mathbf{X}_y)^{-1}\mathbf{X}_y'$, is utilized as the basis for three matrices whose elements are defined as (1) $|h_{y,ij}|$, (2) $h_{y,ij}$, and (3) $-h_{y,ij}$. These represent the absolute value of $\mathbf{H}_y$, $\mathbf{H}_y$ itself, and the negative of $\mathbf{H}_y$. Because of the relationship of the hat matrix (and altered hat matrix) with Mahalanobis distances, jointly influential subsets tend to display similar rows (and columns) within one of these three matrices. A pair of jointly influential observations in the same general location of the regressor/response space, such as in Figure 2.5-A, will exhibit similar diagonal elements of $\mathbf{H}_y$. The corresponding off-diagonal elements will be of comparable magnitude and positive. Other jointly influential pair configurations, such as Figure 2.5-B, may result in negative off-diagonal elements. The later two matrices allow for a determination of these configurations as well.

Gray and Ling (1984) introduced K-clustering as a means of identifying influential subsets (Allen (1984) offers a discussion on this method). The main premise was that, as stated above, influential subsets would have large (in magnitude) off-diagonal elements in the augmented hat matrix. Given a reordering of the data so that observations determining a jointly influential subset appear consecutively, and other non-influential observations are masked, one would obtain a block diagonal augmented hat matrix. K-clustering is a cluster method whose merging criterion is based on ensuring that a specified number, *K*, of observations exceed some threshold value, which corresponds to the completeness of a given block along the diagonal.

Each of the three matrices yields a set of candidate jointly influential subsets for each value of *K*. One could take *K* = 1, 2, and 3, which yields nine sets of candidate jointly influential subsets, for example. Each of these subsets is then evaluated by three subset influence measures mentioned earlier in Section 2.6; namely, $Cook's\ D_I$, R_I and Q_I^2. Results as to which subsets are

most influential are then tabled. This gives the researcher some indication as to the presence of any influential subsets. However, there may be many candidate sets, and furthermore, overlapping clusters may occur. This is an identification method only.

Selecting the appropriate threshold value is necessary in order to ensure that jointly influential observations are actually found. Because the elements of the altered hat matrix lack resistance to outliers and high leverage points (see Section 1.3.3), the classification process could become incomplete when larger amounts of unusual observations are present in the data.

The topic of influence is discussed further in Gray (1986, 1988), Gray and Ling (1985) and Hocking (1984).

§2.7.2 Stalactite Plot

A graphical technique called a stalactite plot was introduced by Atkinson (1994). This analysis of the data is basically a simulation of a serial subsampling algorithm. A forward search is performed on the data. Initially, a subsample of size $m = p$ is randomly selected, and OLS is used to fit this subsample. From this fitted equation, the "median" (actually the $\left[(n+p+1)/2\right]$ ordered) squared residual is retained for future use. The squared residuals for the observations not included in the subsample are scaled by dividing by their corresponding prediction variance, $1+h_{ii}$. To move forward in the search means to increase the subsample size by one. The subsample used in the next step will be the observations that correspond to the $m+1$ smallest squared residuals. This is repeated until eventually the entire data is fit. This completes one forward search. It should be noted that observations might enter and leave the analyzed subset. The variance estimate for this particular search is the smallest of all the observed "median" squared residuals.

This entire forward search procedure is repeated a large number of times, to provide for many random starts. Overall, the search with the smallest variance estimate will be deemed best. Using this criterion, the best stalactite plot will be based on this particular search. Also, it is

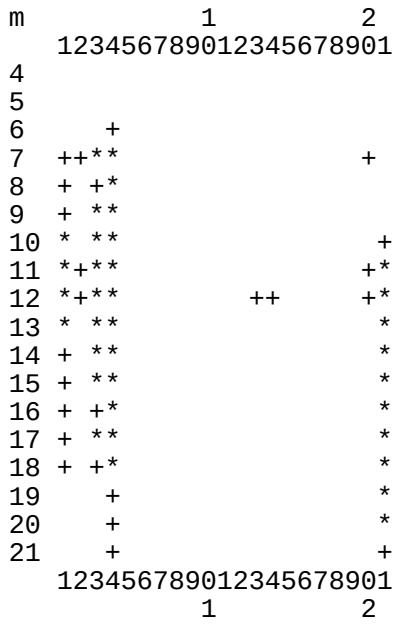

Figure 2.6: Best stalactite plot for stackloss data.

possible to construct a stalactite plot for each random search, and then sort them according to their respective variance estimates. However, note that the results of the many searches are not combined. Each search produces one variance estimate, and it is the job of the simulation to produce a search that minimizes this variance estimate.

The stalactite plot itself is similar to a contour plot. The plot consists of a grid, where each cell is based on some key. There is one column for each observation. There is one row foreach stage of the forward search, i.e. $m = p, p+1, \ldots, n$. Each scaled studentized residual, denoted as t_i and calculated at each stage of the search, is coded into the plot as to what degree this observation appears as an outlier. A t_i beyond 3 in magnitude gets labeled "*", while a t_i between 2 and 3 in magnitude gets labeled "+". The smaller magnitude t_i's are left blank to avoid unnecessary clutter.

Interpretation of a stalactite plot is to see bands of vertical "lines" that occur roughly over the same stages (rows) for several observations (columns). Multiple outliers are then found visually from analysis of the stalactite plot. As an illustration, Figure 2.6 is the best stalactite plot for the stackloss data based on 100 random searches, as presented in Atkinson (1994). It is seen that observations 1, 3, 4, and 21 are consistently displaying an outlying nature. Viewing the last row also shows the problem of masking with multiple outliers. Here, only two observations (4 and 21) are flagged as outliers, while 1 and 3 seem fine.

§2.8 Chapter Summary

It has been illustrated in this chapter that the OLS procedure requires very well behaved data. An outlier or a single high influence point can make the OLS analysis misguided. As a remedy, both M and BI regression have a limited application range. M regression works well in the presence of low leverage outliers, but can be dominated by a single high influence point. Bounded influence regression, on the other hand, can handle at least one high influence point effectively. More than one high influence point can become a problem, however, depending on the data structure. In any event, bounded influence regression is immune to only small amounts of high influence points.

Background information on what is currently available to address joint influence has been discussed in Sections 2.6 and 2.7. This includes a large diagnostic study (K-clustering) as well as a graphical analysis (stalactite plot) of the data.

The focus of the research now shifts to high breakdown regression procedures.