Chapter 1
Classical Regression Analysis

Introduction

Within the professional statistical community there commonly circulates the joke that the world is essentially linear and normal. The humor lies in the realization that these two assumptions form the cornerstone of many statistical procedures over the broad range of fields where statistics is applied. Under these conditions there exist a plethora of elegant statistical theory that is nicely packaged into a broad arsenal of weapons that can be utilized to analyze data. It may even suggest to the researcher the design points of an experiment needed to optimize some specific criteria. This is excellent news to a researcher, but may also lead to a false sense of security, or confidence, in the obtained results.

Real data is generally messy data. As stated in Simpson, Ruppert, and Carroll (1992), citing Hampel, Ronchetti, Rousseeuw, and Stahel (1986), "routine data are thought to contain 1% to 10% gross errors." Consider the field of regression that studies the functional relationship between two sets of variables. Even if the data exhibit a linear trend, contaminated, or unusual, observations may be present. There may be seemingly wild observations, spurious freaks of nature. Yet such observations may also indicate that the regression model is really misspecified, that the assumed linear model is not correct. Or, perhaps human error occurred either in performing the experiment or in recording the data. In the end, a statistician is really at the mercy of the validity of the assumptions that he or she makes when solving a problem or performing an analysis. Assuming a linear model and independent normally distributed random errors is very convenient, because then the subsequent analysis to be performed is specified. However, *assumptions of convenience* are not necessarily the correct assumptions. Furthermore, selecting one particular method of analysis solely because it is "the one which everybody is familiar with and always uses" places an arbitrary limit on the quality of work being performed.

Computer science has been responsible for a revolution of sorts in the field of applied statistics. Since the introduction of the personal computer in the mid 1970's, computer technology has been spiraling upwards at an incredulous pace. Access to powerful machines has become almost commonplace. Methodologies that were once considered too difficult to compute are now viable options. The door is now open to aggressively pursue new methods, new algorithms, new analyses. Classical statistics is no longer the only real option. In some sense, convenience has been surpassed by technology.

This research focuses on the study of robust, high breakdown linear regression modeling. This discipline is extremely computationally intensive. Thus, much of the published work in this area has been in the past decade, which was generally based on work done in the 1980's (of course, some ideas were proposed earlier, but generally not practiced). Currently, the major high breakdown techniques involve a combinatorics-based analysis that becomes overwhelming even with modest sample sizes. For practical purposes, random subsampling is employed to reduce the computations required to obtain the estimators. This does, of course, inject some random sampling variability into the analysis. As it will be shown, two researchers analyzing the same data by the same procedure can obtain vastly different results. Reproducibility, or the lack thereof, becomes a real issue.

The goals of this research are to introduce a new regression methodology that

- obtains robust, high breakdown regression parameter estimates,
- provides an informative summary regarding possible multiple outlier structure and
- has no practical issues concerning the reproducibility of the analysis.

It is assumed that the specified linear model is an adequate description of the general trend of the data, but that a normality assumption regarding the error distribution is no longer appropriate due to the presence of extreme errors. To ease the reading of this research, observations that follow the general trend may be referred to as good or clean, while those

observations not following the general trend may be described to as bad, outlying, contaminated or spurious. These colloquial terms may be supplemented by more technical terms as needed.

Several regression techniques exist that require an initial estimator that is subsequently updated. This research follows a similar path, but requires the development of an adequate initial estimator, one not plagued by subsampling variability and the ill effects of non-reproducibility. An acceptable algorithm will possess both a way to detect multiple outliers as well as rules for handling such observations. Many current methodologies incorporate a one-step analysis, which means that the initial estimator is updated with just one improvement calculation. Within their respective frameworks, iteration until convergence reduces the estimator's breakdown point (Ruppert and Simpson (1990)). A new algorithm is designed to allow for full iteration of an estimator without relinquishing the breakdown point of the initial estimator. The internal stability of these current one-step estimators will be scrutinized closely. Advantages gained by using the proposed method will be demonstrated.

The remainder of Chapter 1 deals with the classical regression framework and a review of some of the available diagnostic tools used in detecting violations of the assumptions. A case study is introduced to serve as a means for both illustration and comparative purposes. Next, the literature search is provided in three parts. Chapter 2 introduces robust regression, with an emphasis on low breakdown regression methods. Various multivariate location and dispersion estimators are the topic of Chapter 3, the material presented here being vitally important to the Chapter 4 discussion of high breakdown regression methods. Upon completion of this trilogy of background material, the proposed method is introduced in Chapter 5. Some theoretical results concerning the proposed method are the focus of Chapter 6. To offer more insight into the performance of the proposed method in relation to the currently available methods, Chapter 7 provides two additional case studies, each one being cited often in the literature. Monte Carlo simulation becomes the topic for Chapter 8, where competing regression methods are compared and contrasted regarding performance under a wide variety of dataset scenarios. Finally, Chapter 9 summarizes the findings of the research, as well as providing additional comments about the proposed method and areas of future research.

§1.1 Background and Notation

Regression is basically the study of the relationship between two sets of variables of interest. The first set contains the *k* regressor variables, $x_1, x_2, \ldots, x_k$, which are assumed to be fixed and measured without error. The regressor variables are often referred to as the independent variables. Their numerical values can either be designed in advance or measured as observational data. In univariate regression, the setting assumed throughout this research, the second set contains a single response variable, *y*. This is also referred to as the dependent variable. As mentioned in Myers (1990), potential uses for regression include model building (including variable screening), parameter estimation, and prediction of the response variable.

The statistical model explains the response variable as a function of the regressor variables, adding a random error term to account for individual differences, as in

$$y_i = f(x_{1i}, x_{2i}, ..., x_{ki}) + \varepsilon_i, \quad i = 1, 2, ..., n,$$

where *n* is the number of observations in the data. The family of candidate functions will be restricted to linear functions of the type

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i .$$

Linear is in reference to the *p* unknown parameters, $\boldsymbol{\beta}$, and denotes that each of the parameters enters the linear model as "$\beta_j x_{ji}$".

In keeping with the usual linear regression notation, the data are organized into an $n \times 1$ response vector, $\mathbf{y}$, and an $n \times p$ regressor matrix, $\mathbf{X}$. By defining $\boldsymbol{\beta}$ as the $p \times 1$ parameter vector, and $\boldsymbol{\varepsilon}$ as the $n \times 1$ random error vector, the linear model becomes $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Elementwise, this is expressed as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

It is convenient to define a few other matrices that play a role in the computation process of various regression techniques. Form the $n\times(p+1)$ matrix $\mathbf{X}_y$ by augmenting the vector **y** to the matrix **X**, yielding

$$\mathbf{X}_y = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} & y_1 \\ 1 & x_{12} & x_{22} & \dots & x_{k2} & y_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} & y_n \end{bmatrix}.$$

When the focus shifts to multivariate location and scale estimation, such as in Chapter 3, it becomes necessary to eliminate the column of ones from both **X** and $\mathbf{X}_y$. Define **Z** as the $n\times k$ matrix containing only the *k* regressor variables,

$$\mathbf{Z} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix},$$

with $\mathbf{Z}_y$ representing the $n\times p$ matrix formed by augmenting the vector **y** to **Z**, given as

$$\mathbf{Z}_y = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} & y_1 \\ x_{12} & x_{22} & \dots & x_{k2} & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & \dots & x_{kn} & y_n \end{bmatrix}.$$

There will be many instances where individual observations are referenced. To this end, the ith row of **X** will be referred to as the $1\times p$ row vector $\mathbf{x}_i'$, and the $1\times k$ row vector $\mathbf{z}_i'$ will represent the ith row of **Z**. When the response variable is included, the notation becomes $\mathbf{x}_{y,i}'$ and $\mathbf{z}_{y,i}'$ for the ith row of $\mathbf{X}_y$ and $\mathbf{Z}_y$, respectively.

Standard "hat" notation is used to depict estimates. For example, $\hat{\boldsymbol{\beta}}$ represents the $p\times 1$ vector of parameter estimates. Then, the $n\times 1$ vector of predicted responses can be calculated

via $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, with $\hat{y}_i$ denoting the predicted response at the i[th] regressor location. Furthermore, the $n \times 1$ vector of residuals can be found as $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, with r_i representing the residual for the i[th] observation. Later in Chapter 1, more regression diagnostics will be introduced.

§1.2 Ordinary Least Squares

In addition to the model specification are the assumptions on the random error term. Classical regression assumes that the errors are independent, identically distributed (iid) random variables from a normal distribution having mean 0 and (constant) variance σ^2. The parameter estimates, $\hat{\boldsymbol{\beta}}$, are found as the solution to the following optimality criterion, referred to as *Ordinary Least Squares* (OLS):

$$\min_{\forall \mathbf{b}} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik})^2 .$$

The argument is essentially the sum of squared differences between the actual response values, denoted y_i, and their corresponding predicted response values, denoted $\hat{y}_i$, which are based on $\mathbf{x}_i'$. Thus, OLS minimizes the sum of squared residuals and hence the name least squares.

When the random errors are truly iid normal, the OLS estimator is the "best" linear unbiased estimator, or BLUE. This means that the OLS estimator has the smallest variance among all possible linear unbiased estimators. Furthermore, the maximum likelihood estimator (MLE) equals the OLS estimator under these conditions.

§1.3 Terminology

There is cause for concern when the data contain observations that are extreme in either the response variable or the regressor space. The term *outlier* will refer to an observation that is extreme in the response variable, relative to the general trend of the data. *Leverage* is the term used to describe the position of an observation in the regressor space. A low leverage point is positioned near the central tendency of the regressor space, while a high leverage point resides in

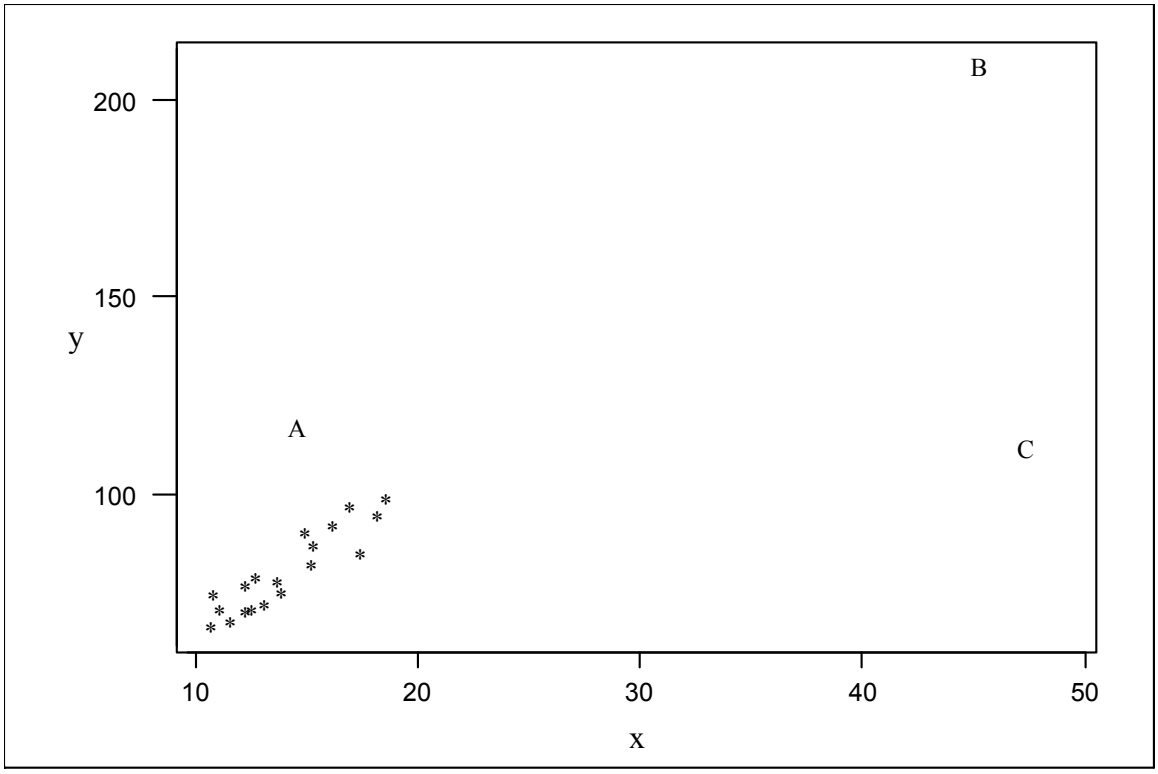

Figure 1.1: Scatterplot illustrating leverage and influence. Point A is a low leverage outlier, point B is a good leverage point, and point C is a bad leverage point.

some extreme location. Following the terminology introduced by Rousseeuw and van Zomeren (1990), a "good leverage point" is an observation, $\mathbf{x}_i'$, possessing large leverage while its response, y_i, fits the general trend of the data well. A "bad leverage point" is an observation, $\mathbf{x}_i'$, with large leverage and the response, y_i, does not fit the general trend of the data well. The later is also referred to as a *high influence point*. Figure 1.1 illustrates the distinction between these three situations. Point A is a low leverage outlier; point B, a high leverage-low influence point, is a good leverage point; and point C, a high influence point, is a bad leverage point.

The OLS procedure is highly sensitive to extreme data. An outlier may have such a large residual that, upon squaring, it overwhelms the OLS objective function. In order to reduce this enormous squared residual, the entire regression may be pulled toward this one point. The

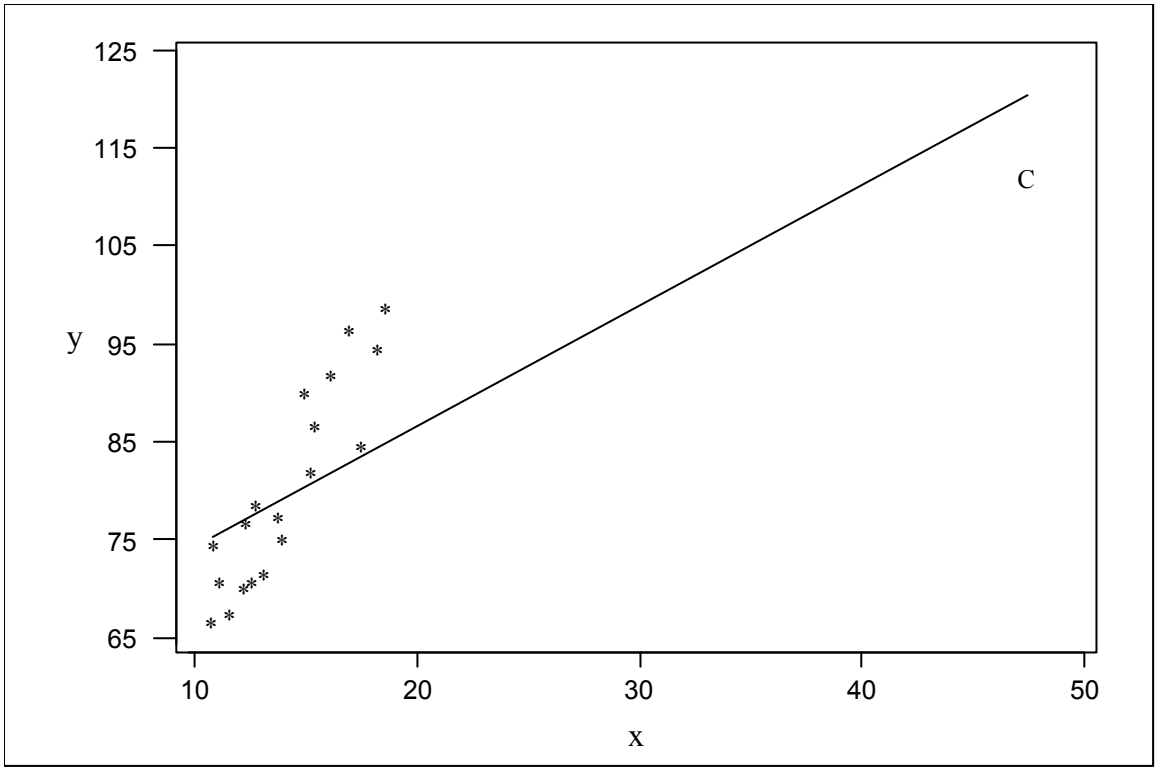

Figure 1.2: The "pulling" effect of a high influence point in OLS regression.

reduction of this particular squared residual more than offsets the increase in the other squared residuals. If the extreme observation is a high influence point, then the OLS estimator can change dramatically away from the general trend and may even reverse the signs for various coefficients. Figure 1.2, illustrates the "pulling" effect that a high influence point has on the OLS fitted regression line (using the data in Figure 1.1, but omitting points A and B to eliminate any interaction). As an aside, low leverage outliers tend to affect the intercept, producing a fit that not aligned with the general trend.

In summary, the OLS procedure lacks resistance to as little as one unusual observation. Coefficients and their standard errors, predictions, diagnostics, hypothesis tests, and other numerical measures can all become very misleading. If one blindly follows the OLS procedure without any exploratory data analysis, it may not become apparent when the analysis has severe deficiencies. Conclusions may be reached that are not really supported by the data. It becomes

imperative to make a judicious choice when selecting the regression method to be used in the analysis.

In order to protect the analysis against the shortcomings of the OLS procedure, diagnostic tools have been developed to aid in the detection of outliers, leverage points, and high influence points (Myers (1990)). These concepts carry over to the robust regression framework as well.

§1.4 Leverage and the Hat Matrix

As mentioned in the previous section, the potential that an observation has to dominate or overwhelm an estimator based solely on its location in the regressor space, ignoring totally the response variable, defines an observation's *leverage*. A key distinction between classical regression and robust regression lies in the quantification of this concept.

First, consider the *k*-dimensional regressor space, where the intercept variable is not incorporated. The *Mahalanobis distance measure*, d_i, for the ith observation is defined as

$$d_i = \sqrt{(\mathbf{z}_i - \mathbf{m})'\mathbf{C}^{-1}(\mathbf{z}_i - \mathbf{m})} ,$$

where $\mathbf{m}$ is the $k \times 1$ mean vector and $\mathbf{C}$ is the usual $k \times k$ covariance matrix.

Next, consider the case where the intercept variable is included. The *hat matrix*, defined as the $p \times p$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, gets its name from the fact that it takes the original response values and transforms them into the predicted values, as $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$. This matrix, however, has more to offer to the analysis. The hat matrix is symmetric, idempotent, and has several interesting properties (Myers (1990)):

- The trace of $\mathbf{H}$ equals p, its dimensional size: $\sum_{i=1}^{n} h_{ii} = p$.
- The ith hat diagonal, h_{ii}, is bounded:

  Model with an intercept: $1/n \le h_{ii} \le 1, \quad \forall i$,

  Model without an intercept: $0 \le h_{ii} \le 1, \quad \forall i$.

- If $h_{ii} = 1,$ then $h_{ij} = h_{ji} = 0$ for $\forall i \neq j$.
- Every row of **H** (and, by symmetry, every column) sums to one: $\sum_{j=1}^{n} h_{ij} = 1, \forall i$.

Additionally, a monotone relationship exists between the i[th] hat diagonal, h_{ii}, and the i[th] Mahalanobis distance, d_i, namely

$$h_{ii} = \frac{d_i^2}{n-1} + \frac{1}{n}.$$

Each hat diagonal measures how far its observation is, in a "standardized sense", from the center (defined as the mean vector for all *n* observations) of the regressor space. A contour plot of specific values for h_{ii} on the regressor space would result in a series of ellipsoids about the center, gradually encompassing more and more design points as h_{ii} increases. Thus, in classical regression analysis it is the hat diagonal of an observation that determines its leverage. Using the fact that the sum of all *n* hat diagonals must equal *p*, and thus $2p/n$ is twice the average hat diagonal, one rule of thumb (Myers (1990)) is to consider any $h_{ii} > 2p/n$ as representing a high leverage point.

The major problem arising from this discussion is that the hat diagonals are mean-based results. As witnessed by their relationship to the Mahalanobis distances, these hat diagonals are based on the sample mean location estimator, **m**, and the corresponding sample covariance shape estimator, **C**. Neither estimator resists unusual observations, so both can become very misleading. In the presence of unusual observations, the hat diagonals may not represent the true notion of leverage very well, if at all. This result leads to the discussion of more resistant leverage estimators. Specifically, a *robust Mahalanobis distance* measure will be formulated by replacing **m** and **C** by more resistant estimators of multivariate location and dispersion. More is said on this subject later, especially in Chapters 3 and 4.

§1.4.1 Altered Hat Matrix

The hat matrix, defined in Section 1.4, is a function of the regressor variables only. To address an observation's possible outlier nature in the response variable, the $n \times n$ *altered hat matrix*, $\mathbf{H}_y$, is defined as $\mathbf{H}_y = \mathbf{X}_y(\mathbf{X}_y'\mathbf{X}_y)^{-1}\mathbf{X}_y'$. The altered hat matrix is also symmetric and idempotent (Myers (1990)). The trace of the altered hat matrix is now $p+1$, as opposed to p for the regular hat matrix. The other three properties bulleted for $\mathbf{H}$ remain valid for $\mathbf{H}_y$. This matrix has a more useful role in outlier analysis. It can be shown that

$$\mathbf{H}_y = \mathbf{H} + \frac{\mathbf{r}\,\mathbf{r}'}{SSE},$$

where $\mathbf{r}$ and SSE are, respectively, the residual vector and sum of squared residuals (errors) obtained from the OLS analysis of the data. As $\mathbf{H}_y$ is a function of $\mathbf{H}$, OLS residuals and *SSE*, its elements are not robust to outliers or high influence points. Further discussion is left for Chapter 5, where the altered hat matrix played an integral role in the genesis of the proposed method.

§1.5 Outlier Diagnostics

For diagnosing outliers in the response variable perhaps it is most obvious to begin with an analysis of the residuals. However, as seen previously in Figure 1.2, high leverage points generally have small residuals due to their "attraction" of the fitted curve. This suggests that a cautious approach be taken when drawing inferences from the set of residuals.

Since residuals retain the units of the response variable, the quantification of what constitutes a large residual is scale dependent. Therefore, an estimate for σ, denoted by *s*, is needed in order to rescale the residuals. Usually, this estimate is taken to be the root mean square error (*RMSE*) of the classical regression analysis of variance (ANOVA) table. Then, the *internally studentized residual* is defined as

$$r_i' = \frac{r_i}{s\sqrt{1-h_{ii}}}.$$

The phrase "internally studentized" refers to the fact that the i[th] observation is used in the calculation of s. Because the numerator and denominator are not independent, r_i' does not follow a true t distribution. However, considering it t-like, one common rule (Myers (1990)) is to deem as outliers those observations with $|r_i'| > 2$.

It can be beneficial to view the regression analysis both with and without a particular observation and gauge the change in certain aspects of the regression analysis. This type of cross validation is often referred to as "single point deletion analysis" or the "PRESS procedure." For notation, a "-i" subscript indicates that the analysis is performed without the i[th] observation. For example, $\hat{y}_{i,-i}$ is the fitted value at $\mathbf{x}_i'$ when the regression is performed without $\mathbf{x}_i'$ and y_i. Initially, this appears to involve performing n separate regressions, each having $n-1$ observations. Due to algebraic simplifications, however, only the full data regression needs to be performed. For example, the residual $r_{i,-i} = y_i - \hat{y}_{i,-i}$, commonly referred to as the "PRESS residual," can be obtained as $r_{i,-i} = r_i/(1-h_{ii})$.

An outlier in the response tends to inflate the variance estimate s^2. It is then desirable to develop an alternative to the internally studentized residual for use as an outlier diagnostic. Obtaining a variance estimate, s_{-i}^2, in the PRESS setting, where

$$s_{-i} = \sqrt{\frac{(n-p)s^2 - r_i^2/(1-h_{ii})}{n-p-1}},$$

leads to the *externally studentized residual*, which is given as

$$Rstudent_i = \frac{y_i - \hat{y}_i}{s_{-i}\sqrt{1-h_{ii}}}.$$

The diagnostic statistic Rstudent follows a t distribution with $n-p-1$ degrees of freedom, denoted by t_{n-p-1}, under the classical assumptions, as the numerator and denominator are now independent (Myers (1990)). An outlier is recognized by a large Rstudent value, where large is often defined by the expression $|Rstudent_i| > t_{1-\alpha/2,n-p-1}$. As the influence of an observation

increases, generally this observation will possess a larger Rstudent value than it does an internally studentized residual. This is because the standard deviation estimate used in the denominator without this observation will generally be smaller than the standard deviation estimate obtained when using all of the data.

§1.5.1 Influence Diagnostics

In Section 1.3 it was mentioned that a high leverage outlier is also referred to as a high influence point. The concept of *influence* can be defined as the relative importance of a particular observation's presence on the resulting estimator. Because of the "pulling" effect that a high leverage outlier has on the OLS estimator, its removal may lead to a dramatically different estimator. Thus, this particular observation is said to be highly influential. Single point deletion analysis becomes the basis for creating a standard set of influence diagnostics.

The influence of the i[th] observation can be measured with a scaled "difference in fits" (DFFITS) statistic, defined as

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

Furthermore, it can be shown (see Myers (1990)) that this statistic can also be written as

$$DFFITS_i = Rstudent_i \left[\frac{h_{ii}}{1-h_{ii}} \right]^{1/2}.$$

This diagnostic is comprised of an outlier component, $Rstudent_i$, and a leverage component, $\left[h_{ii}/(1-h_{ii}) \right]^{1/2}$. Thus, the *DFFITS* statistic considers both aspects in its evaluation of an observation for influence. Belsley, Kuh and Welsch (1980) suggest using $\left| DFFITS_i \right| \geq 2\sqrt{p/n}$ as a guideline for determining which observations have large influence.

Another measure of influence for the i[th] observation looks at the scaled difference between the parameter estimates in the vectors $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{-i}$. This "difference in the betas" is commonly abbreviated as *DFBETAS*. There are actually $n \cdot p$ diagnostics here, one for each

parameter-observation combination. Each $DFBETAS_{j,i}$ measures the influence of the ith observation on the jth parameter estimate and is computed as

$$DFBETAS_{j,i} = \frac{\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_{j,-i}}{s_{-i}\sqrt{c_{jj}}},$$

where c_{jj} is the jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. As with *DFFITS*, larger magnitude *DFBETAS* correspond to higher influence. In another parallel with *DFFITS*, *DFBETAS* can also be written in a form that combines both outlier and leverage terms. Additionally, only one regression is required to calculate this set of diagnostics. See Myers (1990) for more details. Belsley, Kuh and Welsch (1980) suggest using $\left| DFBETAS_{j,i} \right| \geq 2/\sqrt{n}$ as a guideline for determining which observations have large influence.

To view the influence of an observation on the entire set of parameter estimates at once, a composite of the p $DFBETAS_{j,i}$ (per observation) form a diagnostic called Cook's Distance, which is defined as

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

This diagnostic represents a standardized distance between the two estimate vectors. Large values of *Cook's D* (say $Cook's\ D_i > 1$) correspond to observations that have high influence over the set of parameter estimates. The simplified formula

$$Cook's\ D_i = \left[\frac{r_i'^2}{p}\right]\left[\frac{h_{ii}}{1-h_{ii}}\right]$$

again shows that outlier and leverage components determine the diagnostic, and that only one regression needs to be performed.

As not all influential observations are a detriment to the regression, another useful diagnostic would measure the benefit of an observation's presence in gaining precision of the parameter estimates. Precision can be viewed in terms of the generalized variance (*GV*) of the parameter estimates, defined as

$$GV = \left|(\mathbf{X}'\mathbf{X})^{-1}\sigma^2\right|.$$

As the quality of the estimation is reflected by small values of this determinant, evaluate this expression both with and without the i[th] observation. The ratio of these two generalized variances becomes the diagnostic *CovRatio*, calculated as

$$CovRatio_i = \frac{\left|(\mathbf{X}'_{-i}\mathbf{X}_{-i})^{-1}s^2_{-i}\right|}{\left|(\mathbf{X}'\mathbf{X})^{-1}s^2\right|}.$$

Here, a value greater than 1 indicates that the i[th] observation decreased the generalized variance, and resulted in an improvement in the regression. Guidelines for this diagnostic, again provided by Belsley, Kuh and Welsch (1980), are that either $CovRatio_i < 1-3p/n$ or $CovRatio_i > 1+3p/n$ indicate an observation that had a large influence on the precision of the regression.

All of the aforementioned diagnostics are intended for use as signals or warnings to the researcher concerning some aspect of the analysis. Some commonly used yardsticks, i.e. critical values that help guide the analyst, were provided. Using past experience with the diagnostics is generally the best avenue towards a proficient exploratory analysis.

§1.6 Case Study: Stackloss Data

Brownlee (1965) introduced data that have become a benchmark for checking the performance of new regression techniques, referred to as the stackloss data. The data itself is tabulated in Appendix A.1.

A plant is oxidizing ammonia to nitric acid. The data has 21 observations, which correspond to 21 consecutive days of operation. The response variable, *y* (stackloss), is ten times the percentage of ammonia that is lost as unabsorbed nitric oxides. There are three regressor variables: Air flow is x_1, temperature of cooling water is x_2, and acid concentration (coded) is x_3. It is generally accepted in the literature that observations 1, 3, 4, and 21 are outliers.

Table 1.1: Analysis of variance table and parameter estimates summary for the OLS analysis of the stackloss data.

Source	df	SS	MS	F	p-value
Regression	3	1890.41	630.14	59.90	0.000
Error	17	178.83	10.52		
Total	20	2069.24			

Parameter	Estimate	s.e.	t	p-value
Intercept	-39.92	11.90	-3.36	0.004
x_1	0.7156	0.1349	5.31	0.000
x_2	1.2953	0.3680	3.52	0.003
x_3	-0.1521	0.1563	-0.97	0.344

A multiple linear regression (MLR) model using all three regressors was fit with OLS. As seen from Table 1.1, which contains the analysis of variance (ANOVA) table and parameter estimates summary, acid concentration (x_3) is not significant (p-value equals 0.344) in the presence of air flow (x_1) and temperature (x_2).

The standard outlier and influence diagnostics for the OLS regression analysis are given in Table 1.2. Based on $r'_{21} = -2.639 < -2$, observation 21 is judged to be an outlier. However, observations 1, 3, and 4 are not flagged at all. Compared to the other observations, these three observations have relatively large residuals yet have standardized residuals under 2 in magnitude. The difficulty is that in order to accommodate these points, the variance estimate (mean square error, MSE) becomes inflated. This, in turn, causes the standard errors for the coefficients to increase as well. It is seen, however, that the *Rstudent* values for these four outliers are somewhat larger in magnitude, especially for observation 21.

By viewing the hat diagonals it is seen that there are no high leverage points in the stackloss data. Only observation 17 has a hat diagonal above 0.381 (i.e. $2p/n$), but it barely exceeds this value.

Table 1.2: Diagnostics for the OLS analysis of the stackloss data.

Ob.	hat diagonal	residual	int. stud. residual	Rstudent	CooksD	DFFITS
1	0.301	3.234	1.193	1.209	0.153	0.794
2	0.317	-1.918	-0.716	-0.706	0.059	-0.482
3	0.174	4.555	1.546	1.617	0.126	0.744
4	0.128	5.697	1.881	2.051	0.130	0.787
5	0.052	-1.712	-0.543	-0.531	0.004	-0.125
6	0.077	-3.007	-0.966	-0.964	0.019	-0.280
7	0.219	-2.390	-0.834	-0.826	0.048	-0.438
8	0.219	-1.390	-0.485	-0.474	0.016	-0.251
9	0.140	-3.145	-1.046	-1.049	0.044	-0.424
10	0.200	1.267	0.436	0.426	0.011	0.213
11	0.155	2.636	0.884	0.878	0.035	0.376
12	0.217	2.779	0.968	0.966	0.065	0.509
13	0.157	-1.429	-0.480	-0.469	0.010	-0.203
14	0.205	-0.051	-0.018	-0.017	0.000	-0.009
15	0.190	2.361	0.809	0.800	0.038	0.388
16	0.131	0.905	0.299	0.291	0.003	0.113
17	0.412	-1.520	-0.612	-0.600	0.065	-0.503
18	0.160	-0.456	-0.154	-0.149	0.001	-0.066
19	0.174	-0.599	-0.204	-0.198	0.002	-0.091
20	0.080	1.412	0.453	0.443	0.004	0.130
21	0.284	-7.238	-2.639	-3.331	0.691	-2.101

The four outliers possess the four largest *Cook's D* values: observation 21 stands out at 0.691, while observations 1, 3, and 4 are 0.153, 0.126, and 0.130 respectively. None of these diagnostic statistics exceeds 1, the yardstick for significant influence, so no influential observations are detected. However, $DFFITS_{21}$ is large in magnitude, so while observation 21 has moderate leverage, it has more influence than any other observation.

In conclusion, the fitted OLS equation, when retaining all three regressors, is

$$\hat{y}_i = -39.92 + 0.7156\,x_{1i} + 1.2953\,x_{2i} - 0.1521\,x_{3i}\,.$$

Three outliers, observations 1, 3, and 4, are not really detected. Any misleading results due to this fact would then be unknown and hence not reported.

Other, more robust and outlier-resistant, regression techniques are now introduced that may improve the analysis of datasets such as the stackloss data where outliers are indeed present, but go undetected by the classical regression procedure, OLS.