Chapter 9

Future Directions and Summary

Introduction

This research represents an investigation of the feasibility of a radically different approach to solving the linear model estimation problem given the potential presence of high levels of unusual observations. The proposed methodology, as detailed in Chapter 5, was the culmination of research spanning several evolutions of the algorithm. Many nuances have been learned about both the fundamentals that propel the CBI regression procedure into a viable regression methodology as well as previously unreported or downplayed issues regarding the current high breakdown procedures. This final chapter serves to summarize what was learned as well as to discuss possible avenues of future research that may be beneficial. While the philosophical foundation for the CBI methodology has been laid out and shown to have definite merit, further enhancements may very well lead to even better performance characteristics across the spectrum of contamination levels.

§9.1 Current Methodologies

§9.1.1 Ordinary Least Squares

This optimal solution for a linear model with iid normal errors has a 0% breakdown point, making it highly susceptible to a single outlying point. Thus was born the field of robust regression.

§9.1.2 M-regression

This method does well to handle a single low leverage outlier, but is limited in handling of multiple outliers to an upper bound breakdown point of $1/_{p+1}$. M regression does not do well, in general, when dealing with high influence points. M regression is very competitive (efficient) with OLS regression when the data is well behaved with no outliers.

§9.1.3 BI-regression

This method does well to handle a single outlier, particularly if it is a high influence point. M regression does better without high influence points, however. Bounded influence regression is also limited in the handling of multiple outliers due to an upper bound breakdown point of $1/_{p+1}$. BI regression is also very competitive (efficient) with OLS regression when the data is well behaved with no outliers.

§9.1.4 LMS regression

This early 50% breakdown point regression method suffers from slow convergence and is not efficient when compared to OLS if the data is well behaved with no outliers. This method is extremely computationally expensive to solve explicitly, and random subset search algorithms do not guarantee that the correct estimate will be produced. The LTS objective function may have local minima of similar magnitudes in various regions of the parameter space. Hence, dramatically different estimates may be produced from the analysis of the same dataset. Additionally, internal instability can arise and lead to a condition whereas the LMS fit follows a trend that is dramatically different from the larger general trend of the data.

§9.1.5 LTS regression

The slow convergence problem of LMS was circumvented with the introduction of LTS, another 50% breakdown point regression method. LTS is also is not efficient when compared to OLS if the data is well behaved with no outliers. This method is also extremely computationally expensive to solve explicitly, and random subset search algorithms do not guarantee that the correct estimate will be produced. As with LMS, local minima residing throughout the parameter space may result in analysis reproducibility issues, as witnessed frequently throughout this research. Again, analysis reproducibility is not guaranteed and, in practice, often not witnessed. Internal instability is still an inherent drawback, as it is with LMS. LTS is primarily employed as an initial estimate, with efficiency improved by making a one-step improvement; two such methods being M1S regression and S1S regression.

§9.1.6 M1S regression

Having a one-step improvement estimator form, M1S requires an initial estimator with the LTS estimator the recommended choice. During the course of this research it was noted that the M1S estimator could become dramatically different even when two LTS estimates are virtually identical. This fact makes the use of M1S with an LTS start a rather ominous method. Furthermore, examples were presented such that the LTS initial fit was quite accurate in depicting the general trend, yet the M1S estimate strayed rather dramatically from the general trend. Section 5.3 (and Figure 5.7) offered one such puzzling, even troubling, result.

§9.1.7 S1S regression

A second one-step improvement estimator, S1S was an improvement over M1S regarding efficiency. The S1S method utilizes an extra bounded influence weight that is not present in the M1S objective function, but still requires an initial estimator with LTS again being the recommended choice. As with M1S, the S1S estimator can become dramatically different even when two LTS estimates are virtually identical. Furthermore, as with M1S, examples were presented such that the LTS initial fit was quite accurate in depicting the general trend, yet the S1S estimate strayed rather dramatically from the general trend (again refer to Section 5.3 and Figure 5.7). S1S estimation is the current state-of-the-art method in the literature.

§9.2 CBI Regression

This methodology blends the higher efficiency of low breakdown methods with the robustness of the high breakdown methods. It has been demonstrated via Monte Carlo study #1 (Section 8.1) that it more closely mimics BI regression performance when outliers are less frequent than does the current high breakdown methods. CBI regression does not have the random subsampling induced repeatability issues that LTS and, by its use as the initial estimator, M1S and S1S have demonstrated. As the basis for the CBI cluster phase, MVE estimation via the feasible solution algorithm has demonstrated no reproducibility problems. Results of Chapter 6 show that the CBI estimator is affine, regression and scale equivariant, properties achieved by each of the aforementioned regression procedures.

The value of the CBI methodology goes well beyond parameter estimation and extends into a presentation of the data structure and multiple outlier detection. Utilizing information regarding the cluster history, major and minor cluster classifications, the activation process and perhaps including even such intermediate computational details such as the anchor point regression set or the similarity matrix, a very comprehensive data analysis summary can be provided. Combined with the dendrogram summary of the cluster phase and the observation weight plot, the presentation of the CBI regression analysis becomes amenable to users that have limited statistical training, either in general or concerning the field of robust regression.

§9.3 Future CBI Research

The viability, usefulness and contribution of CBI as a very competitive high-breakdown linear regression methodology have been established with this research. Yet there are two topics of interest relating to CBI, areas of future study, which may lead to an improved, second generation methodology.

The first topic deals squarely with the results of the Monte Carlo simulation studies of Chapter 8. It was demonstrated that the CBI scale estimate v_w^2 was a much better performer than was v^2 and, perhaps more importantly, often outperformed $\hat{\sigma}_{LTS}^2$ and otherwise was quite competitive with this scale estimate. Yet it was demonstrated in the Monte Carlo studies involving 40% contamination (Section 8.5 to Section 8.8) that these scale estimates were generally much too large. While the CBI scale estimate was a vast improvement (over LTS), with minuscule bias, in one scenario (Section 8.5), however, and was a solid improvement in two other scenarios (Sections 8.6 and 8.8), the issue of bias could be investigated and perhaps improved.

The second topic focuses on internal instability and specifically, on the ability of an estimator to follow the general trend of a larger subset of observations instead of detecting a trend exhibited by a smaller subset of observations. This condition was witnessed by M1S and S1S during Example 2.3, where the general trend was abandoned due to the high influence cluster and certain centrally located good observations. LTS has witnessed this problem as well. The CBI algorithm demonstrated the ability to resist such departures from the general trend throughout the examples, case studies and simulation studies provided in this research. In some sense, one could view the scalar h as a tuning parameter that needs to be data-driven. Yet given the sub-sampling variability exhibited by the current methodologies as well as the sometimes erratic behavior of one-step estimation from nearly identical initial values, an algorithm that scans across several values of h appears to have many computational and reproducibility issues. Furthermore, the selection of h must be truly data-driven, requiring the user to select h by viewing summary output is at odds with one primary goal of this research: to have a fully automated methodology. However, the CBI algorithm has a starting mechanism that could be exploited in order to evaluate whether or not a particular dataset has a structure such that the general trend exhibited by a larger subset of the data is at odds with a trend exhibited by a smaller subset of observations. Consider the first main cluster, C_0, determined from the first stage cluster process. The focus is on the last merge that created C_0. Suppose that C_{0A} having n_{0A} observations merged with C_{0B} having n_{0B} observations to form C_0, having $n_0 \geq h$ observations. The remaining observations are elements of C_{0C}, having n_{0C} observations (essentially, one large minor cluster). There very well may be merit to an algorithm that compares (intermediate) regression results from (1) C_0, (2) $C_A = C_{0A} \cup C_{0C}$ and (3) $C_B = C_{0B} \cup C_{0C}$. The later two cases would focus on a general trend of $h+1$ observations to warrant an improvement. Taking advantage of the cluster history provides for a means to alleviate extensive computational searches and focus on just two possible alternatives.

Preliminary research suggested that a data-driven expansion of h utilizing such an approach might be a realistic achievement. Beyond the internal instability issue, of which the proposed CBI procedure did not exhibit during the research, such an evaluation process could improve CBI efficiency in low contamination situations by involving more observations earlier in the computation process, thereby further narrowing the performance disparity between BI and CBI under such conditions.

§9.4 Conclusion

A viable alternative in the field of robust, high-breakdown regression has been proposed here and introduced professionally previously (Lawrence, 1996). The current state-of-the-art robust, high-breakdown regression estimator is the S1S estimator, with several performance issues that lend themselves to improvement opportunities:

- *Inherent sub-sampling variability.* Local minima issues may lead to very different LTS initial estimates upon a reanalysis.
- *One-step variability.* A slight deviation in the initial estimator can dramatically alter the final S1S regression estimator. This type of variability can often be seen by viewing the intermediate IRLS computations for BI regression, where the estimator may jump around initially before a convergence path is observed.
- *Internal instability.* A general trend of a smaller subset of the data may be followed instead of the general trend of a larger subset of the data. This is often more apparent when there is a central cloud of good observations with no clear trend unto themselves (this reflects the classical regression variance problem where having good leverage points establishes the trend and thereby reduces estimator variances).
- *Interpretation of data structure.* The S1S weighting summary may be used to evaluate each observation, but the information is limited. Multiple outlier description and interrelationship information is not directly available.
- *Scale estimation.* The LTS scale estimate has been shown (Chapter 8) to produce gross over-estimates, on average, when the contamination level becomes large.

These issues are now related to the CBI regression algorithm.

- *No inherent sub-sampling variability.* The initial estimator is produced via clustering and, unlike the LTS sub-sampling algorithm, the feasible solution algorithm for MVE estimation is reproducible.
- *No one-step variability.* The CBI algorithm fully iterates BI regression intermediate estimates and the activation of minor cluster observations is controlled by a group influence diagnostic analysis.
- *Better internal stability.* Based upon the case studies, examples and simulations of this research, the CBI was better able to detect the general trend of a larger subset of the data instead of tracking the general trend of a smaller subset of the data.
- *Interpretation of data structure.* The CBI algorithm is quite forthcoming in the production of useful, informative descriptions of the data structure, multiple outliers, etc. via a compact two-page tabular and graphical summary.
- *Scale estimation.* From Chapter 8, the scale estimate v_w^2 displayed (sometimes dramatic) improvement over the LTS scale estimate regarding bias.

Thus, five specific drawbacks or weaknesses of the current state-of-the-art robust, high-breakdown regression procedure are successfully addressed by the proposed CBI procedure. Furthermore, from the Monte Carlo studies of Chapter 8, in addition to scale estimation other findings were as follows:

- *Unbiasedness of the regression coefficients.* The CBI algorithm always demonstrated unbiasedness across all eight studies, while S1S had a pronounced bias for Monte Carlo study #2 (Section 8.2).
- *Low Contamination Performance.* The CBI estimator outperformed each of the other high breakdown procedures during Monte Carlo studies #1, #2 and #3.
- *Standard Errors of the regression coefficients.* While standard errors are at the mercy of the scale estimate performance, the CBI standard errors were always competitive and often improved over the S1S standard errors.

- *Coefficient Stability*. The S1S coefficients generally had wider ranges, sometimes even dramatically wider ranges, than those witnessed by the CBI coefficients, which also had very competitive (if not better) IQR's.

Based upon these bulleted issues and findings, it has been shown that the CBI regression procedure is competitive with the current high breakdown methods under various high breakdown conditions and yet is more competitive with the current low breakdown methods under low breakdown conditions. This broad-based application range had served as one motivation for the development of the CBI regression procedure. Combining this achievement with the extra data summary information that is a by-product of the clustering phase, a statistically less-sophisticated user can obtain more valuable and understandable insight into the regression analysis than was available before.

In conclusion, cluster-based bounded influence regression provides a competitive high breakdown estimator, with features and performance characteristics that exceed the current state-of-the-art methods. While circumventing the perils of sub-sampling for an initial estimator, the CBI algorithm parlays clustering and regression computations into an insightful numerical and graphical presentation of the data structure and its general trend. Overall, the proposed methodology represents an advancement in the field of robust regression, offering a distinct philosophical viewpoint towards data analysis and the marriage of estimation with diagnostic summary.