Chapter 10

Summary and Future Research

10.A Summary

The objective of this research is to form a regression technique that is robust to both outliers and to misspecification of the underlying model by the user. Chapter 1 through Chapter 5 reviewed both outlier resistant and model robust procedures. The proposed methodology (ORMRR) was presented in Chapter 6, a method that utilized outlier robust developments in both parametric and nonparametric regression, in addition to the model robust technique of Mays and Birch (1996).

Chapter 7 presented the asymptotic bias and variance calculations of ORMRR and also of methods considered competitors of ORMRR, including M-Regression and Loess. Simulations were utilized to verify the accuracy of these procedures, thus validating them as a legitimate criteria for comparison purposes.

Chapter 8 presented a comparison of each of these procedures using the theoretically optimal mean squared error values across a range of sample sizes, degrees of model misspecification, and proportions of contamination of the error distribution for a particular underlying model. ORMRR was shown to either have superior results or very close to superior results in a large majority of the situations. Loess, for example, was competitive only for larger sample sizes, larger amounts of model misspecification, and smaller amounts of contamination of the error distribution. M-Regression was competitive only for very small amounts of model misspecification.

In addition, a criterion ($d_\rho(\theta)$) for data-driven selection of the bandwidth and mixing parameter involved in calculating the predicted values for ORMRR was presented. This criterion was developed to guard against a large amount of influence of outlying observations on the selection of the parameters.

The criterion was studied by comparing the data-driven parameters selected by the technique with the theoretically optimal parameters calculated from the asymptotic bias and variance formulas. In addition, the mse of ORMRR when fit using $d_\rho(\theta)$ was compared to the mse of Loess when fit using $d_\rho(\theta)$ to select the parameter f, and that of Loess when using the default value of f = 0.5. Once again, ORMRR was either superior or near superior an overwhelming majority of the time, particularly over Loess when using the default parameter value of f = 0.5. This is significant because of the fact that many analysts will tend to use the default specifications of a statistical software package either because they do not understand enough about the procedure to change it, or they are unwilling to invest the time necessary to select a more appropriate parameter value.

Chapter 9 described the extension of ORMRR to multiple regressor variables, and offered an example of the ORMRR, Loess, and M-Regression fits in multiple regressor space. Once again, ORMRR showed great improvement over its competitors, though only for this individual data set. However, this does give an indication that ORMRR may prove to be even more superior in the multiple regressor case than in the univariate case, motivating future research into its behavior across a variety of situations.

In general, the estimator performed extremely well in comparison to both parametric and nonparametric procedures in a variety of sample sizes, degrees model misspecification, and proportions of contamination. It is a fully defined technique that can be used on any univariate data set, with a data driven parameter selection criterion in place that is based on cross-validation.

10.B Future Research

10.B.1 Automatic Selection Criteria

The selection criterion that has been developed works well in the selection of both the bandwidth b_n and the mixing parameter λ, but the theoretically optimal mse values presented in Chapter 8 indicate that there is room for improvement. As there is never ending debate about bandwidth selection criteria in the literature, we are doubtful that we can ever be fully satisfied

with whatever criterion has been selected. This leads us to hope to improve upon $d_\rho(\theta)$ as much as possible, in order to gain mse values for ORMRR using the data-driven technique closer to the theoretically optimal mse values.

10.B.2 Local Bandwidth/Mixing Parameter

Another consideration that could be made is to use local bandwidths in the nonparametric fit. It is widely accepted that a feasible way of dealing with outliers in nonparametric regression is to use a larger bandwidth in a neighborhood of the outlier in order for more observations to have an influence on the fit. The general consensus in the literature, however, is that a global bandwidth is more appropriate when dealing with small sample sizes.

A local mixing parameter would add varying amounts (according to location in the x-space) of the nonparametric fit of the residuals to the parametric fit. This would almost certainly prove useful since the linear model may fit reasonably well in some areas and very poorly on others, as seen in the example in Chapter 7.

10.B.3 Outlier Diagnostics

The outlier diagnostics proposed in Chapter 7 offer a starting point for research into how well these may work in simulations. They appear to have a good theoretical basis, since ORMRR appears to be superior in removing most of the structure from the data, resulting in residuals that contain information almost exclusively about variance (as opposed to bias).

10.B.4 Confidence Intervals

The confidence intervals introduced in Chapter 7 also need to be investigated in order to determine their validity. The development of confidence bands for the entire fit could lead to an increased use of ORMRR because of the wide use of such bands by practical analysts. These

should be researched across a range of sample sizes, degrees of model misspecification, and proportions of contamination of the error distribution.

10.B.5 Multiple Regression Extension

As mentioned previously, Chapter 9 was merely an introduction to the multiple regression extension of ORMRR with an example on one data set. Since the results are very promising, we hope to continue research in this area immediately. This is particularly true because of the appeal of multiple regressor techniques to real world data. We hope to study this form of ORMRR as extensively as the univariate form in order to prove its feasibility. We will probably also consider data that are not equally spaced in the regressor space, as this becomes a much more critical assumption in higher dimensions.