Chapter 3

Robust Parametric Regression

Introduction

Problems that arise when analyzing real data led to the exploration of methods that attempt to robustify the most often used and misused model in statistical regression analysis, the linear model. These methods give an alternative procedure to least squares for *estimation* of the regression coefficients.

The main focus of this chapter is M-estimation in the regression setting, which is explained in detail, along with some other recent advancements in robust estimation, in Section 3.A. Section 3.B offers an example using M-estimation. The use of ψ functions, necessary for M-estimation, will be presented in Section 3.C. Since there is no closed form solution for regression parameters in M-estimation, an algorithm for solution will be discussed in Section 3.D, along with a few computational considerations. Section 3.E will provide a brief discussion of the topics covered in the chapter.

3.A M, Generalized-M (GM), Bounded-Influence (BI), Mallow's 1-Step (M1S) and Schweppe's 1-Step (S1S) Estimators

In M-estimation, the goal is to choose the regression coefficients that minimize some function of the residuals. The method of ordinary least squares, as described in Chapter 1, has as its solution the coefficients that minimize the sum of squared residuals. This solution is undesirable when the data contain outliers, since an observation with a large error term will have a much larger effect (relative to the other observations) on the estimated coefficients. M-estimation involves the minimization of the sum of a general function of the rescaled residuals in the objective function. Thus the M-estimate is the set of coefficients, $\hat{\beta}^{M}$, that minimize

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i'\beta}{\hat{\sigma}}\right), \qquad (3.A.1)$$

where ρ is an appropriately chosen, smooth loss function and $\hat{\sigma}$ is a robust estimate of scale. The parameters of the ρ function are chosen to regulate the amount of robustness of the final estimates. Notation for the rescaled residuals will be $r_i^* = \frac{y_i - \mathbf{x}_i'\hat{\beta}^M}{\hat{\sigma}}$. The only type of scale estimates that will be considered in this paper are regression-based estimates (estimates based on the residuals from the regression model). The most commonly used scale estimate, and the one that will be used here, is the rescaled median absolute deviation (mad), which is defined as

$$\hat{\sigma} = s = \frac{\underset{i}{\text{med}}\left|r_i - \underset{j}{\text{med}}(r_j)\right|}{0.6745}. \qquad (3.A.2)$$

The mad is robust in that it makes use of the median function, and the rescaling by 0.6745 makes it a consistent estimate of σ for normal errors.

Note that if $\rho(u) = u^2$, then $\hat{\beta}^M = \hat{\beta}^{OLS}$. In OLS estimation, the observations with the largest residuals have the most influence on the parameter estimates. The ρ function in M-estimation, and its corresponding parameters, are chosen so that large rescaled residuals, which are assumed to result from outliers, have a smaller contribution to the objective function and, therefore, have less influence on the estimated coefficients, $\hat{\beta}^M$. Note the dependency on the correctness of the model and the assumption that a large residual must be the result of random error, and not due to lack-of-fit.

The robust normal equations that result from the minimization of the objective function in (3.A.1) are

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i'\hat{\beta}}{\hat{\sigma}}\right)\mathbf{x}_i = \sum_{i=1}^{n} \psi(r_i^*)\mathbf{x}_i \equiv \mathbf{0} \qquad (3.A.3)$$

where $\psi(r_i^*) = \frac{d}{dr_i^*}\rho(r_i^*)$. Note that ψ is generally a nonlinear function of its argument, and thus (3.A.3) has no closed form solution for β.

M-estimates were developed as a method of, and are primarily used for, identifying and downweighting the effects of outliers in the y-direction. Krasker and Welsch (1982) warn against the possibility of arbitrarily large influence of an observation because of the multiplier $\mathbf{x}_i$ in (3.A.3), which occurs when high leverage observations are present in the data.

Bounded Influence (BI) estimators, also termed Mallows' estimators (Mallows (1975)), differ from M-estimators in that they were developed in an attempt to account for both an observation's residual and its leverage value. The robust normal equations of (3.A.3) are altered by use of π-weights, which are based purely on the value of the independent variables for a given observation and are typically chosen to vary indirectly with leverage. The normal equations for BI estimators are given by

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{y_i - \mathbf{x}_i'\hat{\beta}}{\hat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}. \qquad (3.A.4)$$

As mentioned earlier, since ψ is a nonlinear function of its argument, the only solutions that exist to (3.A.3) and (3.A.4) are iterative numerical methods. One method of solution will be discussed in Section 3.C.

The similarities of BI estimators and M-estimators are obvious from the normal equations. The addition of the π-weights in the normal equations for BI estimators address an observation's leverage. Observation i is downweighted if it has a large leverage value, causing its contribution to the objective function to be abated.

Simpson, Ruppert, and Carroll (1992) introduced a one-step Mallows' estimator (M1S) that makes use of the Newton-Raphson method of determining the root of a nonlinear equation. The initial estimates are obtained, which are suggested to be high breakdown estimates (estimates that are robust to the situation where a large fraction of the dataset is worthless), such as the least median of squares (LMS) or least trimmed squares (LTS) estimators of Rousseeuw (1984) and Rousseeuw and Leroy (1987). The high breakdown estimates are then updated by one iteration

of the Newton-Raphson method. These one-step estimators are shown to retain the high breakdown point of the initial estimators, while improving upon the efficiency of the LTS estimators. Krasker and Welsch (1982), however, make reference to the separation of leverage and outlierness in (3.A.4). They state that downweighting an observation based on its leverage (which is the effect of π_i in (3.A.4)), without consideration of whether or not that data point follows the trend of the majority of the data, cannot be efficient.

Schweppe (Handschin, Kohlas, Fiechter, and Schweppe (1975)) introduced an estimator that *simultaneously* considers the leverage and outlierness of an observation that is defined as

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{y_i - \mathbf{x}_i' \beta}{\hat{\sigma} \cdot \pi_i} \right) \mathbf{x}_i = \mathbf{0} . \qquad (3.A.5)$$

The argument for ψ in (3.A.4) is a function of both the leverage of an observation, and its corresponding residual. Thus, a *high leverage* point is downweighted only if its *residual* is relatively large.

Coakley and Hettmansperger (1993) introduced a one-step estimator based on Schweppe's estimator in (3.A.5). It is similar to the M1S of Simpson *et al* in that it has as its initial estimates high breakdown estimators, such as LTS, which are updated by one iteration of a Newton-Raphson type numerical algorithm. It is demonstrated that the π-weights in (3.A.5) result in a bounded influence, efficient estimator.

Welsch (1977) suggested using $\pi_i = \frac{\sqrt{h_{ii}}}{1 - h_{ii}}$ as the π-weights. The motivation for using this form of the π-weights is simple and is based on a re-expression of $DFFITS_i$, the diagnostic in OLS regression mentioned in Chapter 2 that can be rewritten as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}} = \frac{y_i - \hat{y}_i}{s_{-i}} \cdot \frac{\sqrt{h_{ii}}}{1 - h_{ii}} .$$

Substituting Welsch's weights into (3.A.4) results in an argument for ψ that is identical to $DFFITS_i$ except for the use of $\hat{\sigma}$ instead of s_{-i} (these similar diagnostics will be referred to as $rDFFITS_i$, in reference to the robust application). This difference should be negligible since s_{-i} is

meant to be a scale estimate that is not inflated by the i^{th} residual, as should the mad be since it is a *robust* estimate of scale. A DFFITS type of argument is reasonable for the ψ function, since observations with large $DFFITS_i$ values have some combination of high leverage and outlyingness.

As mentioned earlier, it is common for the π-weights in (3.A.5) to be a function of the leverage value h_{ii}. However, a danger that occurs in basing the π-weights on these by-products of OLS estimation is their inherent non-robustness. While h_{ii} is a good indicator of leverage when only one leverage point exists, it is deflated in the presence of other leverage points due to the combined effect of these points on the center of the x-space. Simpson *et al* (1992) suggest robust alternatives for leverage value approximations based on the minimum volume ellipsoid (mve), which is the smallest ellipsoid (by volume) that contains 50% of the observations in x-space. The mve is commonly used because of its 50% breakdown point.

3.B Example using M-Regression

Consider a simulated data set generated from the true underlying model $y = (x-5)^2$ with errors generated from a normal distribution with standard deviation $\sigma = 3$. An extreme observation in the response was injected at x = 5, a low leverage point. Figure 3.1 is a plot of the simulated data, the true mean function, the fitted curve using ordinary least squares, and the fitted curve using M-regression. Note the influence of the extreme observation on the fit using least squares. The fit using M-regression downweights that observation to the point that it has no influence, and the result is a fit that accurately characterizes the true mean function.

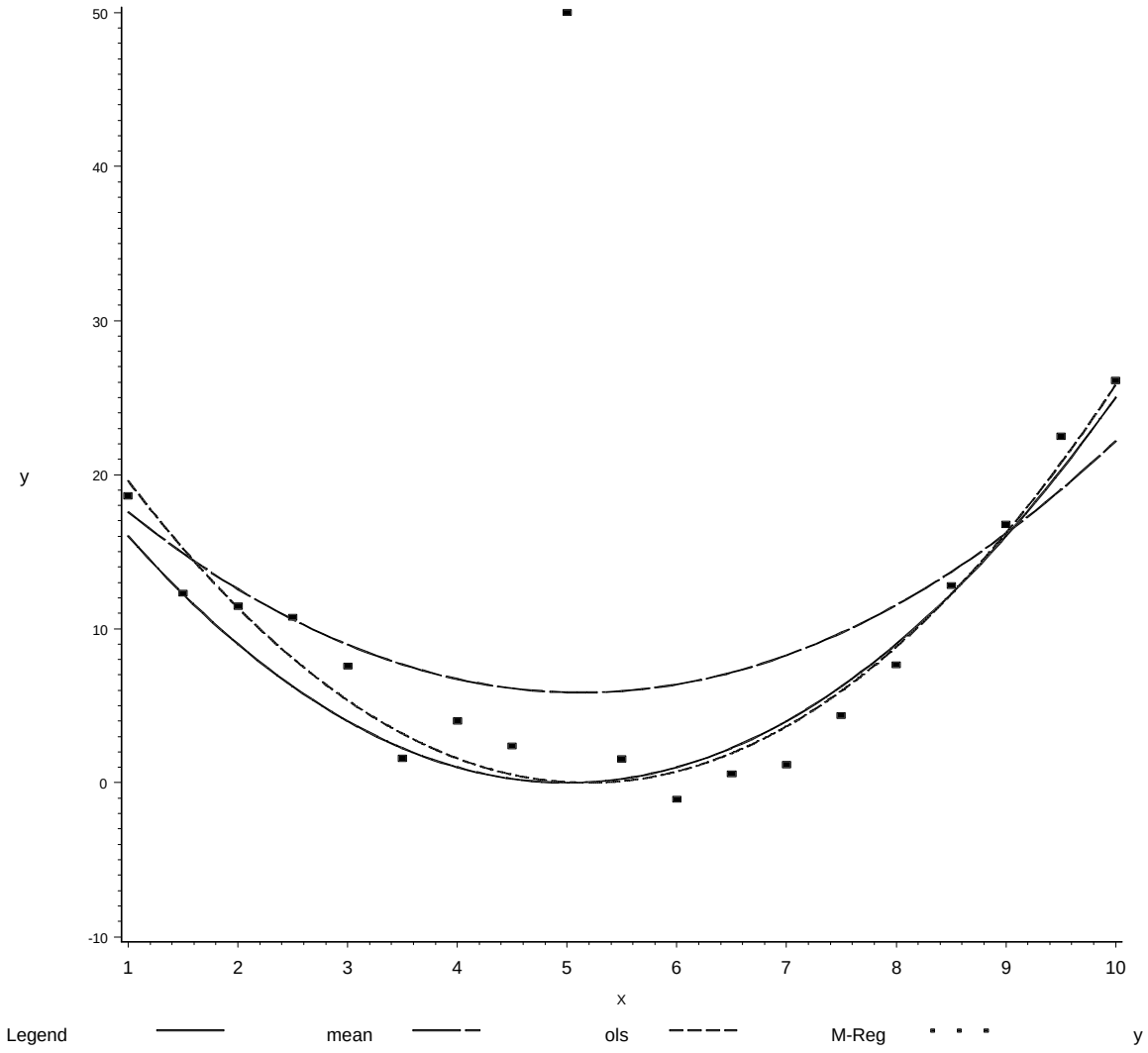


Figure 3.1 Example using M-Regression to fit quadratic model to data generated from the model $y = (x-5)^2$. Note the effect of the extreme point (x = 5) on the OLS fit and the absence of an effect on the M-regression line.

3.C ψ Functions

The robustness properties of the fit obtained by M-estimation rely heavily on the choice of the ψ function in Equation (3.A.3). In addition, the majority of the ψ functions that have been proposed rely on the specification of one or more parameters. This reliance of the ψ functions on these parameters forces M-estimation to also be dependent on them. In addition, the choice of the parameters is rather ambiguous and is usually done on the basis of efficiency versus normal errors.

In general, a ψ function must be continuous, odd, and bounded to achieve desirable robustness properties (such as a high breakdown point, high efficiency, etc.). Numerous candidate ψ functions have been proposed, and a few of the most popular will be presented here. For a more exhaustive list, see Andrews *et al* (1972).

In considering different forms for ψ, it is obvious that the most desirable ψ function would be one that maintains efficiency for normal errors, while protecting against heavy-tailed distributions. Such a ψ function was proposed by Huber, denoted ψ_H, given by

$$\psi_H(u) = \begin{cases} -c_h & u < -c_h \\ u & |u| < c_h \\ c_h & u > c_h \end{cases} . \qquad (3.B.1)$$

The tuning parameter c_H determines the robustness properties of estimates based on ψ_H. Figure 3.2 is a plot of Huber's ψ for c_H =1.345.

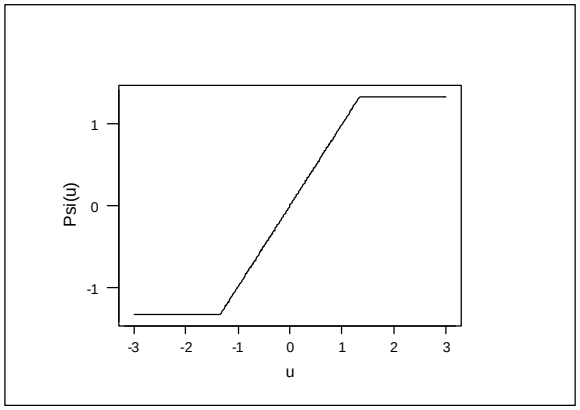


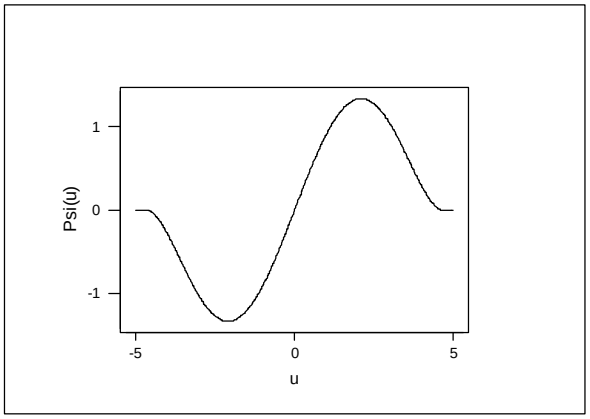
Figure 3.2 Huber's ψ function with tuning parameter c_H = 1.345.

Note that the function is bounded for large residuals, where large is quantified by the value of c_H.

The Bisquare ψ function offers a redescending form that gives zero influence to observations with large enough residuals. The tails of the Bisquare function are heavy, resulting in a slightly more efficient estimator than that produced by some of the other redescending functions. This ψ function is defined as

$$\psi_B(u) = \begin{cases} u\left(1-\left(\frac{u}{c_B}\right)^2\right)^2 & |u| \le c_B \\ 0 & |u| > c_B \end{cases} . \qquad (3.B.4)$$

The M-estimator based on a Bisquare ψ function with tuning parameter $c_B = 4.685$ can be shown to achieve 95% efficiency at the normal distribution.



.

Figure 3.3 Bisquare ψ function with tuning parameter $c_B = 4.685$.

3.D Method of Solution for M-Estimators: Iterated Reweighted Least Squares

As mentioned previously, the function ψ generally is nonlinear, making a closed form solution for $\hat{\boldsymbol{\beta}}$ impossible. Equation (3.A.4) has $\mathbf{x}_i$ as a multiplier, resulting in p equations in p unknowns. The nonlinearity of ψ requires a numerical method of solution. The M-estimates from Iterated Reweighted Least Squares (IRLS) are given by

$$\hat{\boldsymbol{\beta}}^M = (\mathbf{X}^{P\prime}\mathbf{W}^M\mathbf{X}^P)^{-1}\mathbf{X}^{P\prime}\mathbf{W}^M\mathbf{y}$$

where $\mathbf{W}^M = \text{diagonal}\ (w_1, w_2, \ldots, w_n)$, $w_i = \psi\left(r_i^*\right)\big/ r_i^*$, and $\mathbf{X}^P$ is used to denote the X-matrix for parametric regression (to be distinguished from the X-matrix for nonparametric regression in subsequent chapters). Iteration is necessary since the weights in $\mathbf{W}$ depend on the unknown coefficients $\hat{\boldsymbol{\beta}}^M$. Figure 3.4 offers plots of the weight functions that result from using Huber's ψ function (with $c_H = 1.345$) and the Bisquare ψ function (with $c_B = 4.685$).

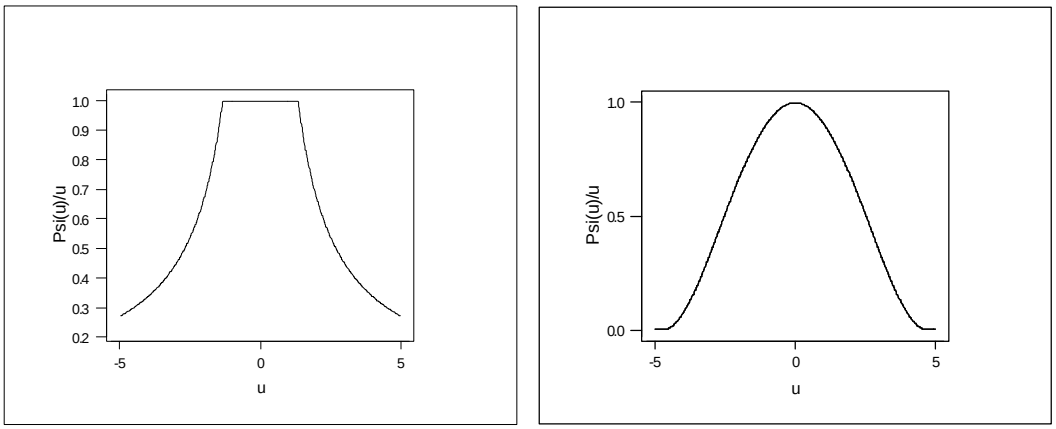


Figure 3.4 Huber (c_H=1.345) and Bisquare (c_B=4.685) weight functions used in the Iterated Reweighted Least Squares method of solution for M-estimates.

There are several reasons for using IRLS, but its primary purpose is its interpretability. The weights that are generated provide a means of identifying outliers in the data set. Observations with a large w-weight (near one) are identified as good points by the M-estimation procedure, while observations that receive a small w-weight (near zero) would be considered outlying observations. Note that there are other numerical methods that exist for obtaining the solution to Equation (3.A.3), including the Newton-Raphson method.

Using Welsch's π-weights causes the w-weights in the IRLS solution of (3.A.5) to have the form $w_i = \dfrac{\psi(\text{rDFFITS}_i)}{\text{rDFFITS}_i}$. Thus a point is downweighted if the value rDFFITS_i is large, indicating that it is a highly influential observation which is usually attributable to high leverage and outlierness.

Considerations

Birch (1980) demonstrated that if a ρ function that corresponds to a redescending ψ function is utilized in Equation (3.A.3) then the objective function may have several local minima, causing (3.A.5) to have several solutions. This creates a problem for procedures such as IRLS and NR, which rely heavily on the starting values to find the root - which usually ends up being the nearest root. Thus the starting values must be chosen carefully in order to find the estimates

that result in the *global* minimum of the objective function. This creates a dilemma for the user that wants extra protection against extreme outliers.

Birch also stated that the objective function has only one minimum value if a ρ function is chosen that corresponds to a monotone ψ function. This leads to the logical remedy of finding consistent M-estimates based on some monotone ψ function such as Huber's ψ, then use these estimates as initial values for calculating the M-estimates based on some redescending ψ function.

3.E Discussion

The specific setting that we are concerned with basically dictates the choice of the robust parametric method. Since we are assuming relatively uniformly distributed data in the x-space, problems with high leverage points are not anticipated to be a problem. Thus, parametric regression using M-estimation is the obvious choice - it does not complicate the methodology by accounting for leverage, and it is a proven method for handling outliers.