Robust and Rank-Based Regression

The historical approach to fitting linear models of the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$ proceeds by finding coefficient estimates $\widehat{\beta}_j$ that minimize the sum of squared errors: $\sum (Y_i - \widehat{Y}_i)^2$. Although these estimators are computationally simple and possess optimality properties, they can be dramatically affected by even a single extreme outlier data point. We have previously seen this problem with the sample mean, which is itself the least-squares solution to the model $Y_i = \beta_0 + \varepsilon_i$. We will briefly survey a few approaches that have been taken to develop estimators of the β_j coefficients that are not as easily affected as the least-squares estimators. The two methods that we will examine are i) a generalization of least-squares estimators called M-estimators, and ii) a generalization of our earlier ranking methods that have been called R-estimators, but are more recently referred to as rank-based regression methods. For each of these two approaches a first method was developed to downweight outlier data points, but was later shown to be susceptible to high leverage points (outliers in the X space) in regression problems, and newer methods have emerged to address both outlier and leverage problems.

M estimation and MM estimation

One approach that has been used to lessen the impact of outliers in linear models is to use absolute deviation as the fitting criterion, finding coefficient estimates $\widehat{\beta}_j$ that minimize $\sum |Y_i - \widehat{Y}_i|$, otherwise known as L_1 regression. A further generalization made by Huber, is to instead minimize:

$$\sum_{i=1}^{n} \rho \left(\frac{Y_i - \widehat{Y}_i}{\widehat{\sigma}_i} \right),$$

where $\rho()$ is a symmetric function and $\widehat{\sigma}_i$ is an estimate of the standard deviation of the errors ε_i. One such $\rho()$ function, the Tukey bisquare, is shown in the text. These M estimators (so-called because they are based on maximizing or minimizing some criterion) have the advantage of being able to downweight outliers while retaining efficiency when compared to least-squares methods. However, the original M estimators can be affected by leverage points (outliers in the X space) in regression problems. Another approach that was developed to protect against outliers and leverage points is least trimmed squares methods, where the criterion being minimized is

$$\sum_{i=1}^{q} |Y_i - \widehat{Y}_i|^2_{(i)},$$

where the q smallest residuals are being minimized. Since these methods do not use all of the data, some efficiency is lost. More recently, Yohai and others have developed extensions to these methods that are called MM estimators that protect against both outliers and leverage points, and retain good efficiency.

R estimation and rank-based regression

At the time when Huber and others were developing the theory of M estimators, methods based on ranking were known as R estimators and were not considered to be as generalizable as M estimators. Later Hettmansperger and others showed that rank-based estimators could also be cast as estimators that minimize the following criterion:

$$\sum_{i=1}^{n} a(R(Y_i - x_i'\widehat{\beta}))(Y_i - x_i'\widehat{\beta}),$$

where $R(Y_i - x_i'\widehat{\beta})$ is the rank of $Y_i - x_i'\widehat{\beta}$ and $a()$ is a suitably standardized score function. One important difference in these rank-based models from the M-estimator-based models above, is that the intercept is estimated separately after first finding the slope estimates, commonly the intercept estimate is taken as the median of the residuals from the model fitted above. These estimators, called Wilcoxon estimators, can be used for any general linear model, and have high efficiency compared to least squares estimators. However, these estimators can be affected by leverage points in regression problems. An extension to these estimators, called weighted Wilcoxon (WW) estimators, were later developed that could protect against outliers and leverage points and are efficient.

References

Hampel, F. R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. 1986. Robust Statistics, The Approach Based on Influence Functions, New York: John Wiley & Sons, Inc.

Huber, P.J. 1981. Robust Statistics, New York: John Wiley & Sons, Inc.

McKean, J.W., and Vidmar, T.J. 1994. A Comparison of Two Rank-Based Methods for the Analysis of Linear Models, The American Statistician 48: 220-229.

Terpstra, J.T., and McKean, J.W. 2005. Rank-Based Analyses of Linear. Models using R, Journal of Statistical Software 14: 1-26.

Yohai V.J. 1987. High Breakdown Point and High Efficiency Robust Estimates for Regression, Annals of Statistics 15: 642-656.