

Rank-Based Robust Analysis of Linear Models. I. Exposition and Review

David Draper

Abstract. Linear models are widely used in many branches of empirical inquiry. The classical analysis of linear models, however, is based on a number of technical assumptions whose failure to apply to the data at hand can result in poor performance of the classical techniques. Two methods of dealing with this that have gained some acceptance are the *data-analytic* and *model expansion* approaches, in which graphical and numerical methods are employed to detect the ways in which the data do not meet the classical assumptions, and either the data are modified appropriately before the classical techniques are applied (data-analytic) or the model is broadened to take account of the departures discovered (model expansion). Another approach involves the use of *robust* methods, which are intended to be sufficiently insensitive to deviations from the classical assumptions that the data may be analyzed without modification or additional (explicit) modeling. In this article a comparison is made between the data-analytic, model expansion and robust approaches to linear models analysis, and the application of one type of robust methods, those based on *R-estimators* (which use the logic of rank tests to motivate inference on the raw data scale), to problems of estimation, testing and confidence and multiple comparison procedures in the general linear model is reviewed.

Key words and phrases: Robust estimation, general inferential strategies, rank-based linear model, *R-estimators*, Hodges-Lehmann, kernel-type density estimation, Bayesian robustness.

1. INTRODUCTION: THE CONTEXT OF ROBUSTNESS IN GENERAL INFERENCE STRATEGIES

The linear model is one of the most widely used tools yet devised by statisticians to aid in empirical inquiry. Applications of linear regression, analysis of variance (ANOVA) and analysis of covariance techniques abound in the biological, social, physical and behavioral sciences, as well as in industrial and other business settings. It is a basic truth in mathematical modeling, however, that powerful inferences are often arrived at only through powerful assumptions, and linear models provide no counterexample to this statement. It is worthwhile to consider these assumptions and to take up the question of what to do in practice

David Draper is a member of the Statistical Research and Consulting Group in the Department of Economics and Statistics at The RAND Corporation, 1700 Main Street, Santa Monica, California 90406.

when some or all of them are not reasonable for the data at hand.

The general fixed-effects linear model can be written in the form

$$(1.1) \quad Y_i = g(X_{i1}, \dots, X_{ip}) + e_i, \quad i = 1, \dots, N.$$

Here $(Y_i; X_{i1}, \dots, X_{ip})$ is the i th of N total observations on the quantitative dependent variable Y and the p quantitative or qualitative (nominal or ordinal) independent variables X_1, \dots, X_p , which are considered to be either under experimenter control or passively observed without random error; the e_i are thought of as stochastic errors or disturbance terms. The Y_i and e_i are taken to be random variables and the X_{ij} to be fixed known constants. $g(\cdot)$ is assumed to be of the known functional form

$$(1.2) \quad g(X_{i1}, \dots, X_{ip}) = \beta_0 + \sum_{j=1}^p X_{ij}\beta_j,$$

in which the β_j are unknown parameters. A number of assumptions are made in the classical analysis about

