Statistical Science
2007, Vol. 22, No. 1, 40–43
DOI: 10.1214/088342307000000078
Main article DOI: 10.1214/088342306000000682

Rejoinder: Fisher Lecture: Dimension Reduction in Regression

R. Dennis Cook

1. INTRODUCTION

I am grateful to all of the discussants for their comments which raise a number of important and insightful issues, and add significantly to the breadth of ideas. Following a few introductory comments on the need for a new regression genre that centers on dimension reduction, I turn to the discussants' remarks.

The development in the 1960s and early 1970s of diagnostic methods for regression produced a major shift in regression methodology. When a diagnostic produces compelling evidence of a deficiency in the current model or data it is natural to pursue remedial action, leading to a new model and a new round of diagnostics, proceeding in this way until the required diagnostic checks are passed. By the late 1970s this type of iterative model development paradigm was widely represented in the applied sciences and was formalized in the statistical literature by Box (1979, 1980) and Cook and Weisberg (1982). With the availability of desktop computing starting in the mid-1980s, it is now possible to apply in reasonable time batteries of graphical and numerical diagnostics to many regressions.

Advances in computing and other technologies now allow scientists to routinely formulate regressions in which the number p of predictors is considerably larger than that normally considered in the past. Such large-p regressions necessitate a new type of analysis for at least two reasons. First, the standard iterative paradigm for model development can become untenable when p is large. Recognizing the variety of graphical diagnostics that could be used and the possibility of iteration, a thorough analysis might require assessment of many plots in addition to various numerical diagnostics. Experience has shown that the paradigm can often become imponderable when applied with too many predictors. Second, in some regressions, particularly those associated with high-throughput technologies, the sample size n may be smaller than p, leading to operational problems in addition to ponderability difficulties. These issues have caused a shift in the applied sciences toward a different regression genre with the goal of reducing the dimensionality of the vector $\mathbf{X} \in \mathbb{R}^p$ of predictors as a first step in the analysis, effectively raising an old idea to a position of prominence.

Today, dimension reduction is ubiquitous in the applied sciences, represented primarily by principal component methodology. Fifteen years ago I rarely encountered intra-university scientists seeking help with principal component reductions in regression. Such settings no longer seem unusual. The reasons for this are as indicated previously: While I occasionally see problems with $n < p$, more frequently n is several times p, while p itself is too large for a full commitment to iterative model development guided by diagnostics. This, in addition to the reasons stated in Section 2 of the article, leads me to conclude that the case for dimension reduction methodology has been made, methodology based on firm parametric foundations with subsequent robust and nonparametric counterparts. Whether the ideas and methodological directions I proposed will meet this goal is less clear, but I am still convinced that they hold promise when $\mathbf{X}$ and Y are jointly distributed.

In contrast to a comment by Christensen, I think parametric dimension reduction is currently as important, if not more important, than other forms, partly because dimension reduction methodology has existed mostly in a world apart from core Fisherian theory, making it difficult to appreciate what could be achieved. For this reason I welcome Christensen's development of connections with multivariate linear model theory.

2. APPLICABILITY

According to Christensen, a key issue in the development of models (2), (5), (10) and (13) is whether they are "broadly reasonable." I agree. Moreover, the emerging picture does seem to be one of broad reasonableness for the reasons indicated in the following sections.

R. Dennis Cook is Professor, School of Statistics, University of Minnesota, 224 Church Street S. E., Minneapolis, Minnesota 55455, USA (e-mail: dennis@stat.umn.edu).