ASA Continuing Education Short Course

Regression Graphics: Ideas for Studying Regressions thru Graphics

by R. Dennis Cook

Sunday, August 8, 1999
at the Joint Statistical Meetings in Baltimore

Abstract

Dimension reduction a leitmotif of statistics. For instance, starting with a sample $z_1, \ldots, z_n$ from a univariate normal population with mean μ and variance 1, we know that the sample mean $\bar{z}$ is sufficient for μ. This means that we can replace the original n-dimensional sample with the one-dimensional mean $\bar{z}$ without loss of information on μ.

In the same spirit, *dimension reduction without loss of information* is a dominant theme of regression graphics. The goal of a regression study is to infer, as far as possible with the available data, about the conditional distribution of the response Y given the $p \times 1$ vector of predictors $\mathbf{X}$: How does the conditional distribution of $Y|\mathbf{X}$ change with the value assumed by $\mathbf{X}$? This goal is relevant in many scientific areas for forming prediction equations, judging the effectiveness of a treatment and for many other purposes.

A central goal of regression graphics is to reduce the dimension of $\mathbf{X}$ without loss of information on the conditional distribution of $Y|\mathbf{X}$ and without requiring a model. We call this *sufficient dimension reduction*, borrowing terminology from classical statistics. Sufficient dimension reduction leads naturally to *sufficient summary plots* which contain all of the information on the regression that is available from the sample. Such summary plots can be quite useful throughout a regression study, particularly for guiding the choice of a first model.

We will describe a new context for regression that centers on dimension reduction and sufficient summary plots, and that requires few scope-limiting constraints. The methodology to be described is very general and can be used in almost any regression. Most of the methodology will be illustrated using computer programs that will be made available to the participants without charge.

This course is intended for industry statisticians, graduate students and college instructors. Prerequisites are familiarity with standard regression methodology at the level of one of the major textbooks in the area, a first course in mathematical

statistics (Master's level preparation would be adequate), and basic knowledge of linear algebra.

Outline

The following outline is arranged in logical units that require different amounts of presentation time.

1. *Introduction.* In this first unit we will motivate the need for a well-defined context for graphics in regression, arguing that such a context is needed to achieve full benefit of visualization. It will also be shown that present graphical methodology can suffer from the lack of a context.

2. *Foundations.* Here we establish the population context that will serve as a focal point for the rest of the course. Central dimension-reduction subspaces and sufficient summary plots will be defined and discussed. Simple examples will be used for illustration.

   Two issues will arise at the end of this unit: How can central subspaces and sufficient summary plots be estimated? And, what might be gained in practice by doing so? The latter issue will be considered first, since it may be easier to understand the methodology when the end is in sight.

3. *Examples.* Several examples will be given in this unit showing that the population context established in the previous unit results in a better understanding of standard graphics (eg. residual plots), and how estimates of central subspaces and sufficient summary plots can benefit an analysis. Mild restrictions on the marginal distribution of the predictors (not on the conditional distribution of $Y|\mathbf{X}$) will be introduced.

   The graphical context under study allows for novel, comprehensive methods of detecting outliers and subpopulation structure (eg. a missing binary predictor like gender or location). These and other properties will be emphasized during the presentation.

   Equally important, the examples will serve to establish the role for graphics in regression under the umbrella of central subspaces.

4. *Standard Fitting Methods.* In this unit we will study the relationship between standard fitting methods (eg OLS, Huber's M estimate), and the central subspace. In particular, when the central subspace is one-dimensional, a 2D plot of the response versus fitted values can be a sufficient summary. The ability recognize such situations can be quite useful.

The rest of the units in the course focus on methodology at varying levels of detail.

5. *Inverse Response Plots.* In this unit we describe how to use inverse response plots to gain information on the dimension of the central subspace. Scatterplot matrices will play a key role.

6. *3D Plots.* In regressions with two predictors, central subspaces and sufficient summary plots can be estimated visually from 3D plots, as will be shown in this unit. Standard fitting methods can be used as well with mild constraints on the marginal distribution of the predictors. This methodology forms a basis for visual estimates of summary plots in regressions with more than 2 predictors.

7. *Graphical Regression.* In regressions with more than two predictors, the central subspace can often be estimated visually from a series of 3D plots using the methodology described in the previous unit. It is possible in some regressions to conduct a "complete" using graphics alone (no models).

8. *Numerical Methods.* We will discuss two numerical methods for estimating the central subspace: Sliced inverse regression (SIR) and sliced average variance estimation (SAVE). At the end of this unit participants will have all the information necessary to reconstruct the examples of unit 3.

9. *Net Effect Plots.* Time permitting, net-effect plots for studying the contributions of individual predictors will be discussed. These plots make use of many of the dimension-reduction ideas presented in earlier units.