

Applied Regression Including Computing and Graphics Workshop

Workshop Report: 3 May 2001, Glasgow

Ian Wilson
University of Aberdeen

i.wilson@maths.abdn.ac.uk

The late John Tukey in his book *Exploratory Data Analysis* stated “Exploratory data analysis is detective work – numerical detective work — or counting detective work — or graphical detective work”. The approach of Dennis Cook and Sanford Weisberg is to use graphical methods, including E.D.A techniques, to study how response variables depend on predictors – inferential detective work. This workshop demonstrated how dynamic graphical methods and inferential statistics can be taught together to form a new view of regression.

This article reports a workshop by Professor Sanford Weisberg — *Applied Regression Including Computing and Graphics (ARICG)* –run by the LTSN Maths, Stats & OR Network that was held on the 3 May 2001 at the Department of Statistics, Glasgow University. The workshop aimed to give an insight into the graphical and regression techniques that are described in the textbook: *Applied Regression Including Computing and Graphics* by R. Dennis Cook and Sanford Weisberg (Wiley) – denoted ARICG in this article. This event introduced material from the book (unfortunately in short supply in the UK at the time), and introduced the free software *Arc*. The functionality of *Arc* is so intrinsic to the book – and the methods – that it is fair to describe the book as the manual to *Arc*. On Wiley’s web site it states that the book “provides a bona fide user manual for the *Arc* software.” Further support for *Arc* comes from an excellent web site, <http://www.stat.umn.edu/arc/>, where software for the Windows, Macintosh and Linux operating systems; additional documentation; and add-ins to provide greater usability and extensions can be downloaded.

This course, for those involved in teaching statistics to undergraduates, was a distillation of ARICG, concentrating on the ideas that differentiate the authors’ approach to regression. Large parts of the book that cover standard regression and graphical techniques (with a unique slant) were passed over. Prof Weisberg concentrated on part III of ARICG, regression graphics, with a small departure into logistic regression in the last session of the day. The workshop comprised of 4 sessions of about an hour each and short computer laboratory session that introduced the *Arc* computer package – hands on experience of this vital weapon in the *Applied regression* armoury was essential to gain an understanding of the mind-set.

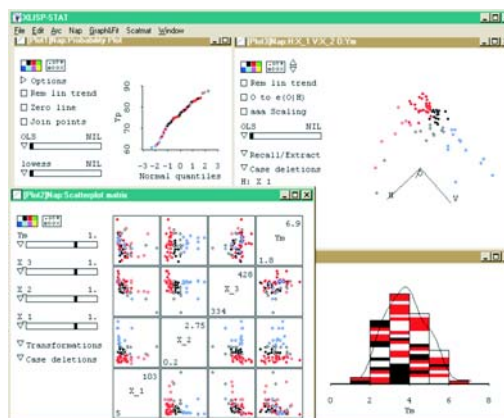


Fig 1 Brushing points across different views of the data using the Arc naphthalene dataset distributed with arc.

The Arc Software Package

The program *Arc* is a free software package designed to work as a teaching tool with ARICG, and also function as an analysis package in its own right. The package is written in XLISP-STAT (which is written by Luke Tierney). Effective use of the software requires only techniques covered in detail in ARICG, and no knowledge of XLISP-STAT is required. However, simulation experiments that are a vital part of the course require some knowledge of Lisp. For those more used to C or S type computer languages, the syntax and feel of Lisp is very different. This program is unlikely to be anyone's only statistics package, and students will have to learn other packages – for presentation quality graphics and other statistical techniques. Add-ins are available that increase the usability – one offering is a plug-in for Excel that allows data to be loaded directly from Excel into *Arc*. Parts of *Arc*'s functionality are available in other environments, for example, dynamic graphics and brushing across linked views will be familiar to users of XGOBI (now enhanced as GGOBI at <http://www.ggobi.org/> and accessible directly from R). The seamless integration within *Arc* of graphics and regression is not matched to my knowledge (for free software anyway). Fig 1 illustrates some of the basic graphical procedures, with dynamic brushing linked across all four views of the data. This intelligent and interactive style of graphics is very impressive. Students more familiar with the graphics included in the ubiquitous spreadsheet packages and their “business graphics” – with such horrors as 3-dimensional exploding piecharts — should find inspiration here.

Seeing Results Through Graphs

The *Arc* software is designed for more than informative graphics. This software integrates exploratory graphics, regression techniques and residual analysis seamlessly. The main thrust of the workshop was that access to increased computer power has enabled iterative fitting and criticism of complex models, yet we have not gone further and used this computation power to help to guide us in the statistical model formulation. The point that Prof Weisberg drove was that his approach to regression provides a context to dynamic graphics — visualisation and dimension reduction — rather than these graphics being just pretty pictures.

The first workshop session concentrated on trying to give a new insight into our own attitudes to regression problems, outlining an approach that de-emphasised models, the emphasis being instead on viewing data as conditional distributions, so that regression is the study of the distribution of the response Y conditional on the

predictors, X . This was done by testing our intuition about multidimensional distributions and attempting to disconcert us about what graphics actually say, while developing theory to provide a context for dynamic graphics in regression.

Regression Graphics

Graphical investigation of a regression with one predictor is widely taught at an elementary level in schools and universities. A scatterplot of the response versus the predictor can tell us whether a simple linear regression is sufficient. However, this approach does not generalise well to multiple predictors. The main novelty of the course, and the textbook, is that of visualising regression using the *structural dimension* of data. Structural dimension is the number of *linear* combinations of predictors needed to explain the structure of Y , the response variable. The structural dimension can be identified by its *sufficient summary plot*. Such plots have the minimal complexity needed to explain the structure of the model and to make predictions.

Informally, if there is no structure in the data, then scatterplots of Y against all the predictors show no pattern, and you have a zero-dimensional structure, and a sufficient plot is a histogram of the response. For a one-dimensional structure then Y depends on X through a linear combination of the (possibly transformed) predictors, and a scatterplot of Y against this linear combination is a sufficient plot. When we only have two predictors then this is intuitively easy. The 3-dimensional plot of Y on the two predictors will have a linear projection – see Fig 2 for an example of 1-dimensional structure. The plot on the left is rotated such that the projection of the predictors onto the scatterplot gives a simple relationship, the plot is sufficient to explain the data; it is a sufficient summary plot.

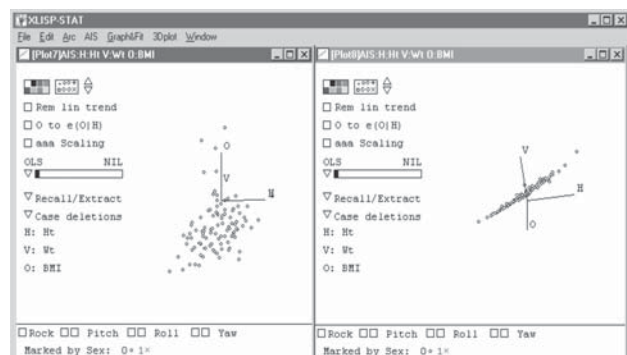


Fig 2 Two views of a 1-dimensional structure, plot from the Australian Athlete data set distributed with *Arc*.

A model is two-dimensional, if you cannot find a combination of rotation or change of variables such that the structure in data is only visible with a 3-d plot, i.e. only by rotating 2 dimensional scatterplots. This intuitively simple construction provides a basis for the extension of this to problems with more than two constructors.

This approach can also incorporate binomial and logistic regression; generalised linear models, and associated statistical techniques, however the course was too short to cover these adequately. Fig 3 shows a plot from the collisions dataset that illustrates how we may use a matrix scatterplot to investigate the structure of responses to a binary variable. The methods described in this workshop would be good for teaching a stand-alone course in regression, supported by the textbook. However as it stands somewhat out of the main current of statistical practice and notation it may be difficult to integrate within a statistics undergraduate degree. The Arc software is excellent, and brings together functionality

on graphics, regression, and diagnostics and I would certainly recommend it for practical data analysis.

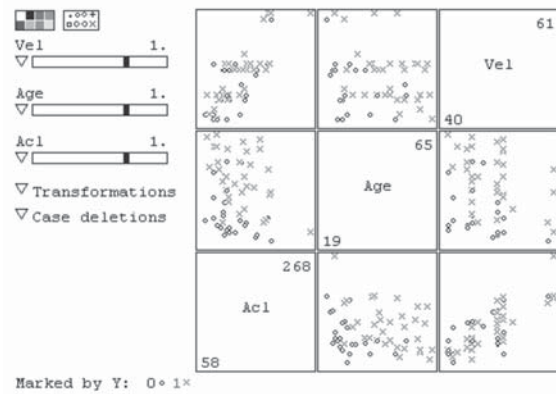


Fig 3 Viewing binary responses. Data on “deaths” of crash test dummies, from Arc dataset collisions distributed with Arc. Crosses represent dummies that “died”.

A Quick Tour of...Internet Mathematician

<http://www.evl.ac.uk/vts/maths/>