Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)
March 20–22, Vienna, Austria ISSN 1609-395X
Kurt Hornik, Friedrich Leisch & Achim Zeileis (eds.)
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Trellis Extensions

Richard M. Heiberger and Burt Holland

Abstract

We have designed several multi-panel graphical displays for which the panels are defined by the Cartesian product of one set of variables or parameters with another set. The traditional scatterplot matrix `splom` is the special case where both the row and column sets of variables are the same and the panels are ordinary scatterplots. Our displays extend the interpretation of the model formula to allow the panels to be functions of parametric transformations of the variables in the formula. We show examples of scatterplot matrices with different sets of variables along the rows and columns, of a ladder of powers display, of interaction plots, of ANCOVA plots, of Least Squares plots, and of Time Series plots.

The Trellis system of graphics in the S language, including both R and S-PLUS, is based on the paradigm of repeating the same graphical specifications for a structured set of variables. The majority of the methods supplied in the `trellis` library are based on a typical formula having the structure

```
y ~ x | a * b
```

where

`y`	is either continuous or factor
`x`	is either continuous or factor
`a`	is factor or shingle
`b`	is factor or shingle

and each panel is a plot of `y ~ x` for the subset of the observations defined by the levels of `a` and `b`.

The scatterplot matrix `splom` differs from the majority of the methods in two ways. First, each of the panels is a plot of a different set of variables. Second, each of the panels is based on the entire set of observations.

Several of our examples extend the concept of a structured presentation of plots of different sets of variables, or of different parametric transformations of the same set of variables.

Several of our examples extend the interpretation of the model formula, that is the semantics of the formula syntax, to allow easier exposition of standard statistical techniques.

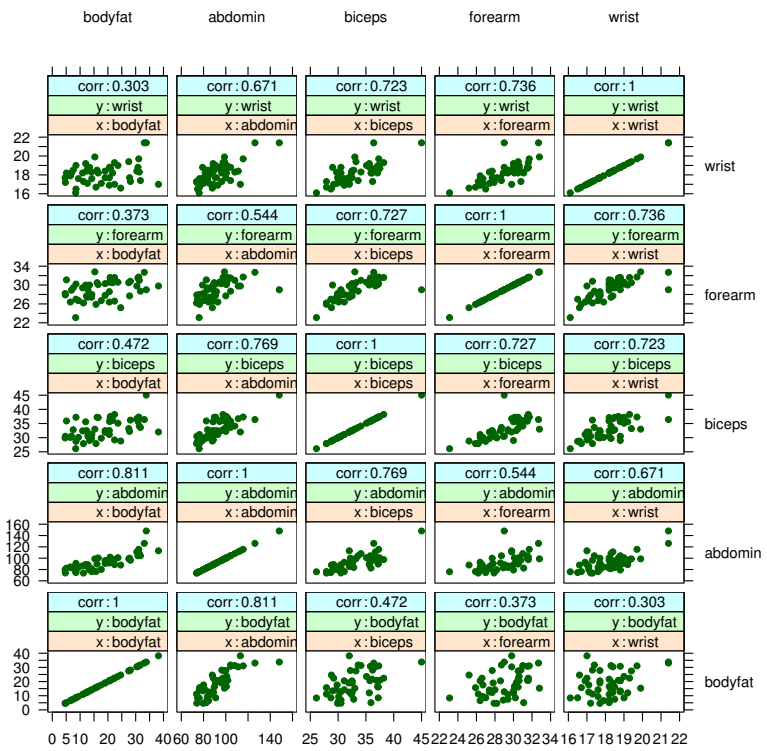

Figure 1: R: `xysplom( ~ fat, corr=T)`

This paper is a report on a work in progress. All the plots have been designed for our forthcoming text (Heiberger and Holland, 2003). The time series plots have already appeared in Heiberger and Teles (2002a and 2002b).

1 Scatterplot matrices

A scatterplot matrix is a Trellis display in which the panels are defined by a Cartesian product of variables. In the standard scatterplot matrix constructed by `splom`, the same set of variables define both the rows and columns of the matrix.

Figure 1 shows a scatterplot matrix constructed with our function `xysplom`. We indicate the differences in formating compared to the standard Trellis function `splom`.

1. We place a dotplot on the main diagonal of the matrix.
2. We optionally display the correlation of the variables in each panel of the scatterplot matrix.
3. We display all tick labels on the left and bottom, and suppress tick labels in the interior.
4. We label the variables on the top and right.

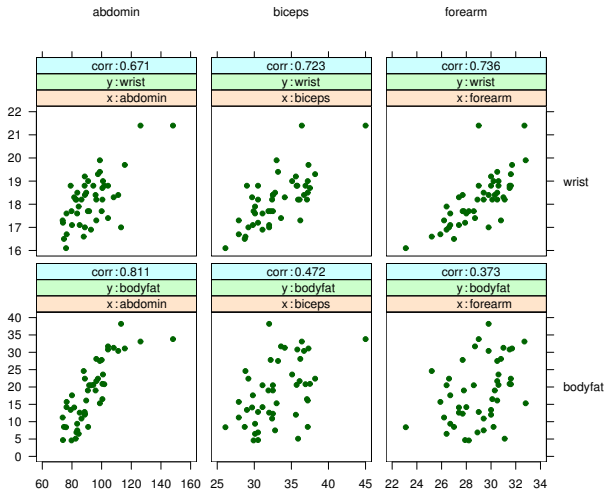

Figure 2: R:
`xysplom(bodyfat + wrist ~ abdomin + biceps + forearm, corr=T)`

Figure 2 shows a scatterplot matrix with different sets of variables on the x and y axes.

1. We can think of this display as a block off-diagonal of the standard scatterplot matrix.

2. We label the individual variables on the top and right.

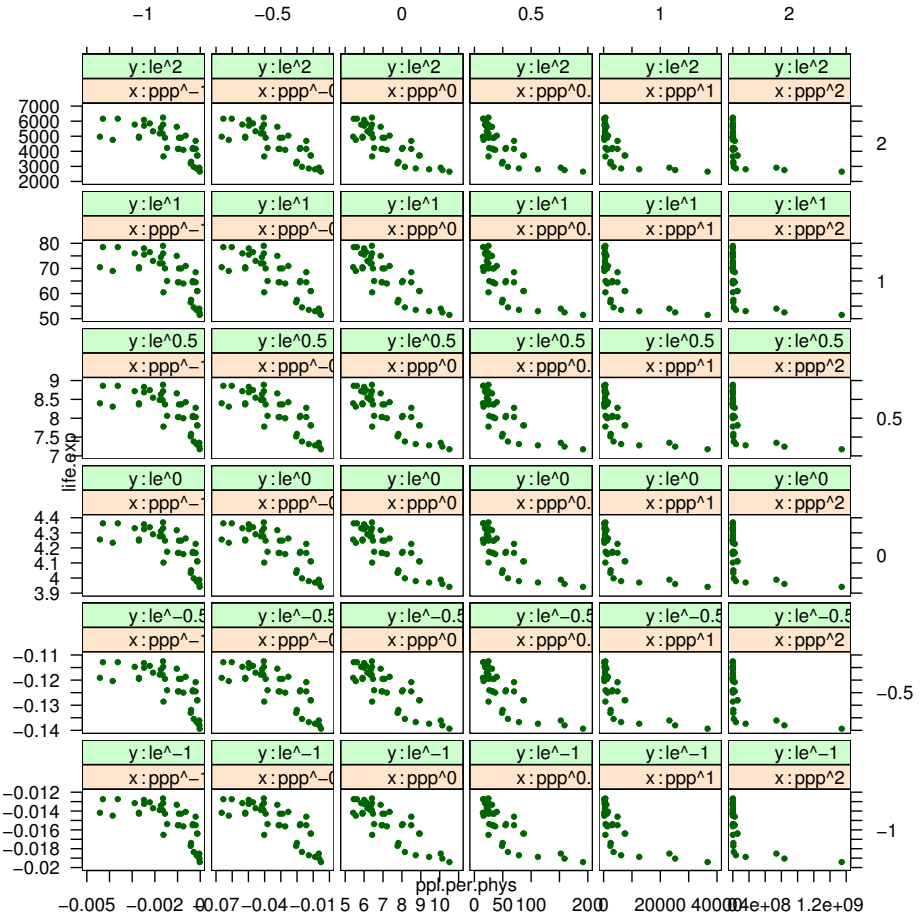

Figure 3: R: `ladder(life.exp ~ ppl.per.phys, data=tv)`

2 Ladder of powers: Cartesian product of powers

In the ladder of powers plot in Figure 3, the rows and columns of the matrix are defined by the Cartesian product of a series of power transformations of a row and column variable. The power transformations are the set: $-1, -.5, 0, .5, 1, 2$ with the 0 power interpreted as the logarithm.

1. We label the individual panels with the power of y and x.
2. We label the rows and columns of the matrix by the power.
3. We label the individual variables in the outer margin.

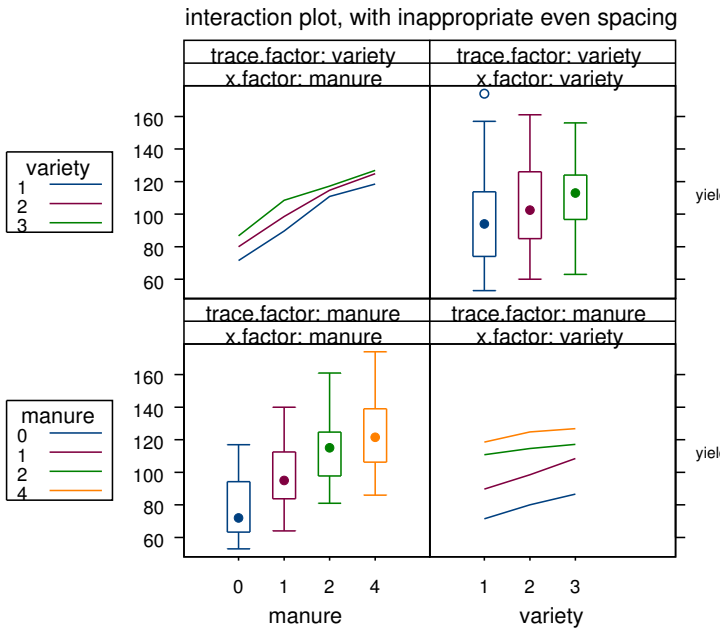

Figure 4: S-PLUS: interaction plot, with inappropriate even spacing.

```
barley2$manure <- factor(barley2$manure)
interaction2wt(yield ~ manure + variety, data=barley2)
```

3 Interaction plots

The interaction plot in Figure 4, for a design with several factors, defines the rows and columns of the display by the Cartesian product of the factors.

1. Each off-diagonal panel is a standard interaction plot.
2. Panels in mirror-image positions interchange the trace- and x-factors.
3. We use the main diagonal for the boxplot of the factor.
4. The rows are labeled with a key that shows the line type and color for the trace factor by which the row is defined.
5. The boxes in the boxplot are colored to match the traces in the same row.
6. The columns are labeled by the x-factor.

In this example, one of the factors was spaced at 0,1,2,4 with the anticipation that a quadratic effect would be visible. We display in Figure 5 the correct display in which we detect that the factor is both ordered and numeric and adjust the placement of the bwplot and the interaction plot accordingly. We needed to modify the R `panel.bwplot` function.

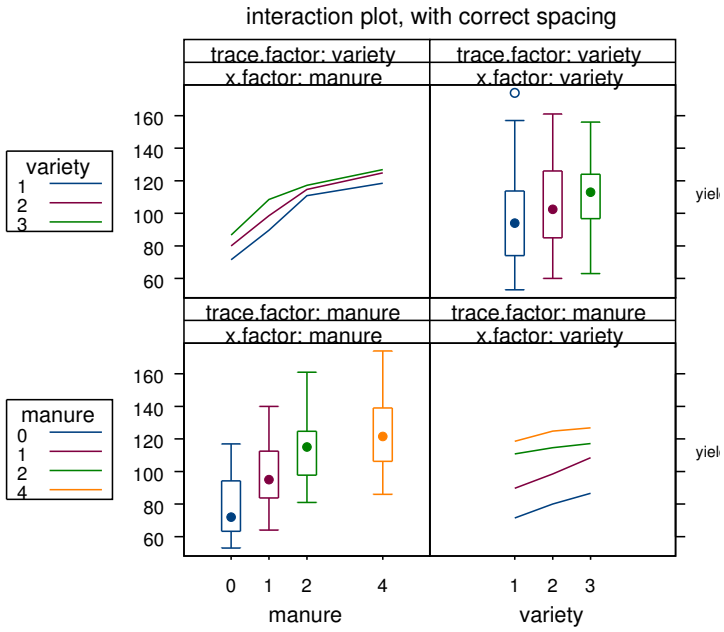

Figure 5: S-PLUS: interaction plot, with correct spacing.

```
barley2$manure <- ordered(barley2$manure)
contrasts(barley2$manure) <- contr.poly(unique(barley2$manure))
interaction2wt(yield ~ manure + variety, data=barley2)
```

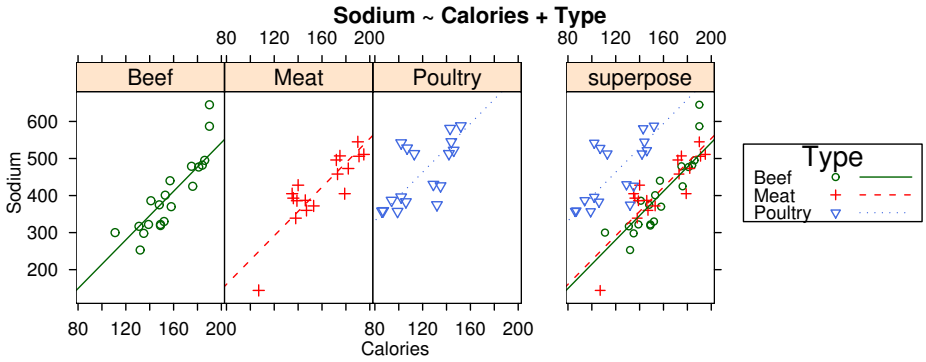

Figure 6: R: `ancova(Sodium ~ Calories + Type, data=hotdog)`

4 ANCOVA plots

ANCOVA plots are defined by the Cartesian product of display format. In Figure 6, an example with one factor and one covariate, we show separate panels for each level of the factor on the left side and superposed panels on the right side. The display extends naturally to ANCOVA with two factors and one covariate.

1. The `ancova` function constructs both the ANOVA table and the ANCOVA plot from a single specification of a model formula.

2. Depending on the level of overlap of the x- and y-ranges and the collinearity of the groups, it may be more advantageous to look at the set of separate panels or at the single superposed panel. Therefore we display both.

3. The `ancova` function can display any one of horizontal slope, common slope, or distinct lines. The specifications are:

   horizontal lines: `ancova(Sodium ~ Type, x=Calories, data=hotdog)`

   common slope: `ancova(Sodium ~ Calories + Type, data=hotdog)`

   distinct lines: `ancova(Sodium ~ Calories * Type, data=hotdog)`

   Figure 6 shows common slope.

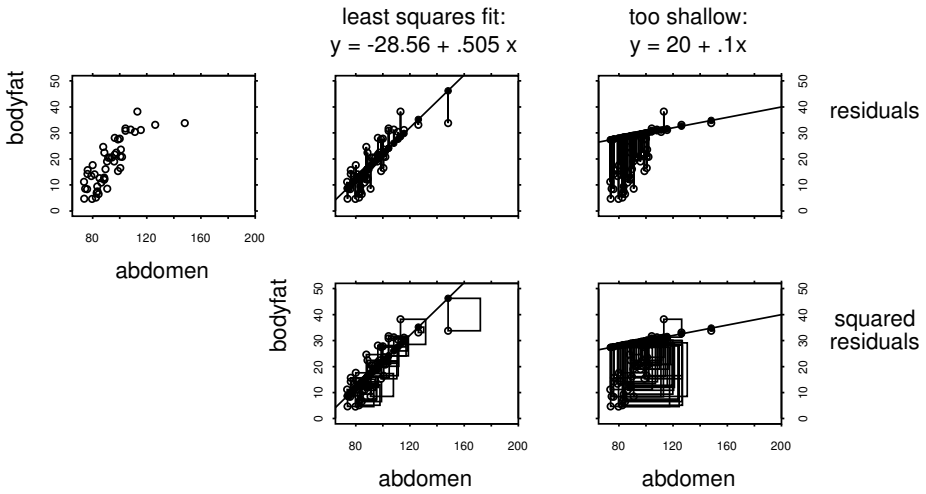

Figure 7: This plot was manually constructed using traditional graphics. The next round will construct the panel function as an outer product of panel function statements.

5 Explanation of least squares

Least squares is the most frequently used technique for estimating the parameters of a straight line fit of a response variable y to an explanatory variable x. Explaining it to students, particularly where it gets its name, is not always easy. Figure 7, the Cartesian product of three straight lines by three display techniques, helps clarify the intent of the method.

1. We show two different straight lines, one per column of the display. One is the least squares line, the other is too shallow.

2. We show two rows, one with the fitted lines and the residuals from the lines, and one with the squared residuals.

3. The "Least Squares" method minimizes the sum of the squared residuals. The second row of our display shows the individual squared residuals. The sum of their areas is the "Sum of Squares". This form of the display is from Smith and Gonick (1993). It is clear from the picture that the least squares line has a smaller summed area of the squares than the other line.

4. The squares in the display are real squares in the measurement units of the piece of paper (or display screen) on which they are drawn. The vertical dimension of the squares is measured in y units. The horizontal dimension is the same number of inches on the paper as the vertical dimension. We had to do some background work, back-computing from the aspect ratios of the plots, to make this come out right.

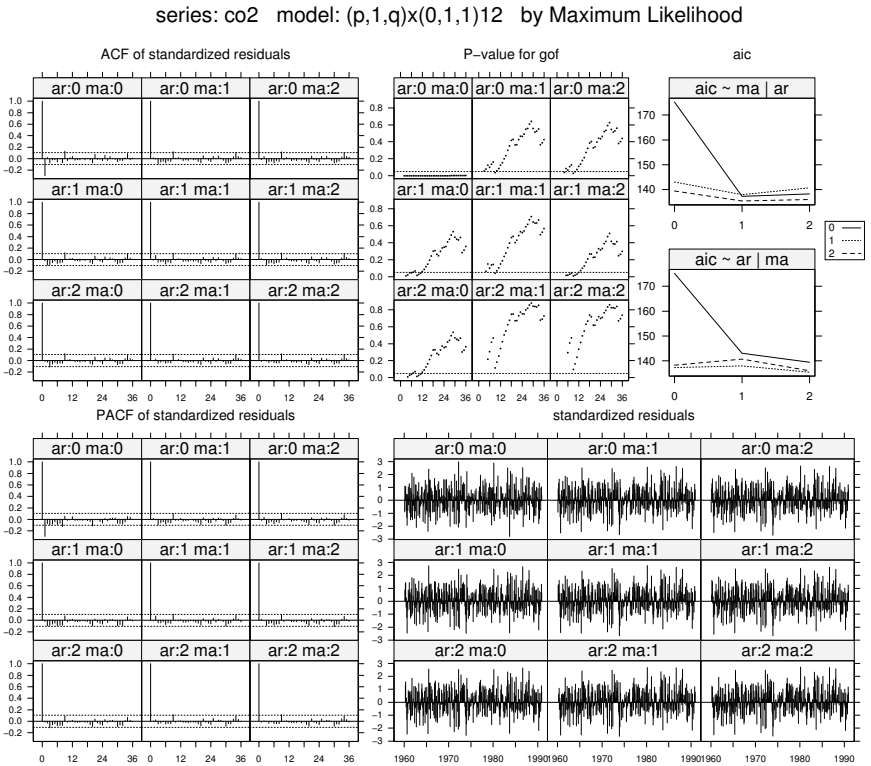

Figure 8: S-PLUS: Displays for Direct Comparison of ARIMA Models

6 Time series plots

The ARIMA method of time series analysis depends critically on identifying the best autoregressive (AR) and moving average (MA) parameters prior to estimating the regression coefficients. Heiberger and Teles (2002a and 2002b) construct a series of displays of the usual diagnostic measures for ARIMA models. Figure 8 is one of the sets of plots from their paper.

1. Four of the sets of graphs on the display are Cartesian products of candidate AR and MA parameters. Placing all of them on one sheet of paper in a coordinated way makes it relatively easy to read the graph and complete the identification phase of the analysis.

2. The fifth set of graphs is also a Cartesian product, similar to the interaction plot, in which the AR and MA parameters alternate being the trace- and the x-factors and the Akaike Information Criterion (AIC) is the y-variable.

7 Programming in R

7.1 R and S-Plus

The language and the user interface are intended to be the same. The internals are quite different. Much of what we have presented here was written as new panel functions. As is noted in the R documentation, the details of panel function construction are quite different in the two implementations.

Our sense is that the grid and lattice packages in R, as the newer implementation of the Trellis concepts, are more coherent than the ground-breaking programming of S-PLUS.

7.2 Parsing trellis formulas

The lattice function `parseLatticeFormula` could not be used in our panel functions. It conflates two ideas: parsing the formula and selecting columns from the data.frame.

References

Heiberger, Richard M. and Holland, Burt (2003). *Statistical Analysis and Data Display: An Intermediate Course*. Springer-Verlag, New York, preliminary edition.

Heiberger, Richard M. and Paulo Teles, (2002a). "Displays for Direct Comparison of ARIMA Models," *The American Statistician*, **56**, 131–138, 258–260.

Heiberger, Richard M. and Paulo Teles, (2002b). Software for "Displays for direct comparison of ARIMA models". http://lib.stat.cmu.edu/S/ARIMA-trellis.

Smith, Woollcott and Larry Gonick. (1993). *The Cartoon Guide to Statistics*. HarperCollins.

Affiliation

Richard M. Heiberger, Burt Holland
Department of Statistics
Temple University
Philadelphia
PA 19122-6083, USA