Tools and Concepts in Data Analysis[1]

Günther Sawitzki
Institut für Angewandte Mathematik
Im Neuenheimer Feld 294
D 6900 Heidelberg

[1] Presented at Softstat '89, the 5th Conference on Scientific Use of Statistical Software, Heidelberg April 2nd-April 6th, 1989. Part of the material has already been presented at the Biometrical Colloquium Freiburg, 8. 12. 1988 and the Biometrical Colloquium, Celle 1989.

To appear in F. Faulbaum et al. (eds.): Softstat '89. Proceedings of the 5th Conference on Scientific Use of Statistical Software, Heidelberg April 2nd-April 6th, 1989.

Summary

Data analysis has two aims: finding informative features in data, and bringing them to human perception. And, in doing this, data analysis has to avoid artifacts coming from random fluctuation, and from perception. Some of the tools, and some of the current concepts in data analysis are reviewed. Diagnostic indices are discussed as special statistics highlighting informative features of data. After an excursion, a look at interactive graphics and some of the statistics involved is given.

1. Introduction

As a spin off of classical statistics, isolated (unidimensional) numbers are often used to express informativity, for example when reporting achievable significance levels. Mere numbers give very restrictive ways to express informativity. The current progress in graphical facilities has made (two dimensional) displays a main utility for data analysis. As a consequence, displays is what data analysis is offering at first sight today.

It is not the point of data analysis to produce a display. The point is to know which displays to look at, and how to look at them.The first part of this talk will start with a classical display, the residual plot, and from here will lead to several diagnostic displays as used in data analysis. The current means to study these displays, namely interactive graphics and linked windows, are the second topic of this talk. In a final part, we will ask for the statistical meaning of these analyses and link back to "classical" aims like testing hypotheses, finding best fits etc.

Let us approach data analysis from well-known grounds and start with a linear regression situation. For illustration, we will use the Scottish Hill Runners' data (Atkinson 1986). This data set allows a substantial analysis by classical tools, while - even for those not trained in data analysis - clearly it is pushing classical approaches to their limits (Figure 1).

Our starting point is the classical linear model

$$\mathbf{Y} = \mathbf{X}\beta + e \qquad e \sim N(0,\sigma^2) \quad \text{iid.} \qquad (*)$$

Given a matrix of covariates $\mathbf{X}=(X_1,\ldots,X_k)'$, we would estimate the vector of unknown parameters by the least square method, so $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$. This defines a fitting function $\hat{Y}(X) = X'\hat{\beta}$ and we can use the residuals $r_i=Y_i-\hat{Y}(X_i)$ to judge the quality of the fit by looking at the coefficient of determination $R^2=(SSY-SSr) / SSY$, where $SSY=\sum (Y_i-\overline{Y})^2$, $SSr = \sum (r_i-\bar{r})^2$. To check for the validity of the underlying assumptions (linearity, homoskedasticity, error distribution), the residual

scatterplot of r_i against $\hat{Y}(X_i)$ would be consulted. Studying the plot of the residuals (Figure 2) is something any "classical" statistician would do to judge the appropriateness of assumptions of the model.

If the model (*) holds, there is no purpose to do any further analysis. Due to the random fluctuation, we would find structures and systematic features - but they will only be outcomes of random fluctuation. But to find features not covered by the model (*), to look for the unexpected, is a task for data analysis.

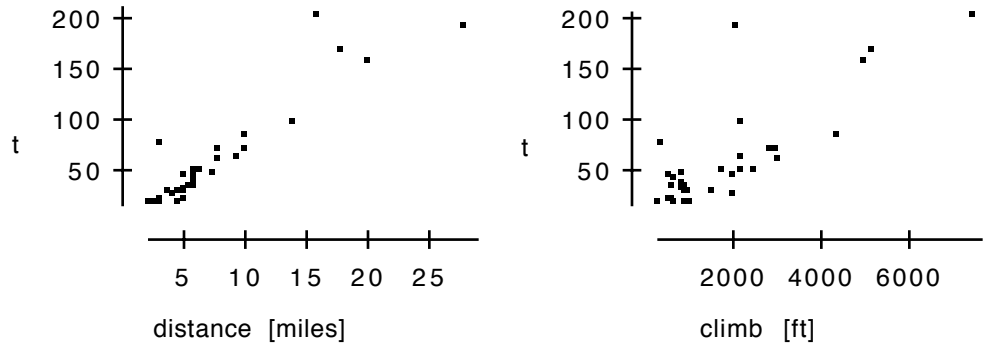

Figure 1: Scottish Hill Runners' data: dcatterplots of the target variable time t [min] against potential regressors distance and climb. Note the clear linear relation in the main regressor's scatterplot (time against distance) with three obvious outliers. What are the effects of these outliers ?

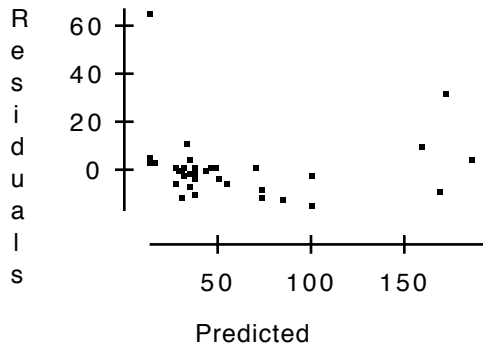

Figure 2: Residual plot for Scottish Hill Runners' data, with obviously non-homogeneous residuals and trend.

2. Diagnostic Indices and Diagnostic Plots

We will start again with the classical residual plot. By defining **diagnostic indices,** the information contained in this plot will be unfolded into several diagnostic plots for data analysis. These plots will exhibit different features inherent in the data.

To study the residual plot, we must define what we are looking for. We have to say what we mean by interesting, or informative events. A first idea would be to look at observations which are **"outliers"**, in the sense of observations which are special with respect to the regression. For this purpose the usual residual plot does not help: since all points affect the regression, any outlier would have influenced the regression, and hence its own residual. In this way, an outlier might

produce unduly low residuals and thus hide itself. For outlier detection better residuals $r_{(i)}$ are constructed for data element i by calculating the regression omitting observation i, and then taking scale estimator and residuals from this regression. The "externally studentized residuals" $r_{(i)}/s_{(i)}$ are distributed under the hypothesis as Student's t on (n-p-1) degrees of freedom, and they are not inflated by gross errors in the i^{th} observation. Unfortunately, like the residuals, these are not mutually independent - but there is no help against this. So instead of looking at the raw residual plot, one would look at the externally studentized residual plot when searching for outliers.

We can ask a different question: which observations are of special importance for the regression ? The importance can be potential or factual: again the simple problem splits into two. We can ask which observation are of **potential influence on the regression** - so we would look at the leverage $\partial\widehat{Y}(X_i)/\partial Y_i$. For the linear regression situation, since $\widehat{Y}(X) = X'\widehat{\beta} = X'(\mathbf{X'X})^{-}\mathbf{X'Y}$ this reduces to the leverages $\partial\widehat{Y}(X_i)/\partial Y_i = h_i = (\mathbf{X}(\mathbf{X'X})^{-}\mathbf{X'})_{ii}$, that is studying the diagonal elements of the "hat matrix" $H := \mathbf{X}(\mathbf{X'X})^{-}\mathbf{X'}$, the leverages.

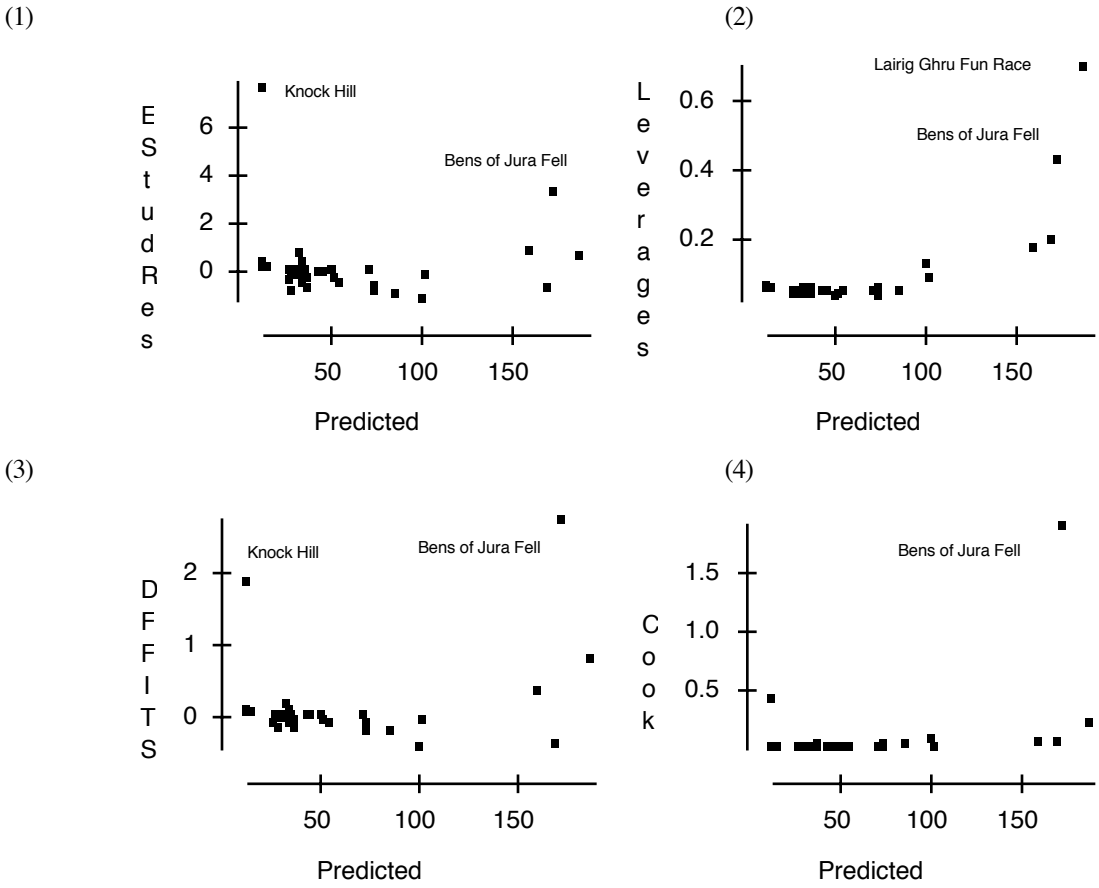

Figure 3: diagnostic plots for studying (1) outliers, observation of potential (2) and factual (3) influence on the fit, and influential observations for parameter estimation (4). The Knock Hill race is the race which is most off with respect to the other data. The highest influence on regression (3) or parameter estimation however comes from the Bens of Jura Fell Race. The most critical race as far as the design (i.e. covaritate values) is concerned is the Lairig Ghru Fun Race.

But we could as well look at the **factual influence on the regression** - of course scaled by the estimated variation. This would lead to considering the DFFITS $\hat{Y}_i - \hat{Y}_{(i)} / s(\hat{Y}_{(i)})$ where $s(\hat{Y}_{(i)})$ is the estimated standard error of $\hat{Y}_{(i)}$

$$\text{DFFITS}_i = \sqrt{\frac{h_i}{1-h_i}} \frac{r_i}{s(i)\sqrt{1-h_i}}$$

Finally, we can interpret the regression problem as an estimation problem. So we ask for the influence of observation i on the **estimation of the parameter** β. A corresponding statistic is Cook's distance D_i which is to be compared with the F(p+1,n-p-1) distribution,

$$D_i = \frac{(\hat{\beta}-\hat{\beta}_{(i)})^t \mathbf{X}\mathbf{X}^t (\hat{\beta}-\hat{\beta}_{(i)})}{(p+1)\hat{\sigma}^2}$$

We started with one classical display, the residual plot. By making more precise what we mean by "informative" we split up this single display into four different displays - each showing a different aspect of the information. All was contained in the original display. These displays do not add information. But they make information of interest stand out more clearly. For more information on these diagnostic indices see Cook and Weisberg (1982) or Chatterjee and Hadi (1988). The definition of diagnostic indices, adapted to various purposes, is one of the tasks in data analysis. We are not going to study these plots in details as presented by now, because today's dynamical graphics provide better possibilities.

3. Excursion: Validation

So far, the critical issue was outliers. If this were the only problem, robust methods would be appropriate. Data analysis however tries to be prepared for the unexpected. So assumptions about models and distributions as used in robust statistics cannot be the basis of data analysis.

Like any approach in statistics, data analysis has to take care not to be mislead by random fluctuation. Other than the classical statistical approaches, data analysis can rarely draw on predefined model assumptions to check the validity of its findings. In the definition of the externally studentized residuals above we already met a typical approach in data analysis: to use the data at hand to crossvalidate findings. We discarded one observation at a time and calculated the statistic of interest based on the remaining observations.

This approach can be generalized: judge a statistic $T=T(X_1,\ldots,X_n)$ by comparing it to the empirical distribution of corresponding statistics $T'=T(X_{i_1},\ldots,X_{i_m})$, where $i_1,\ldots,i_m$ runs through all subsets of size $m<n$. This **crossvalidation** idea has been extensively applied and studied in data analysis. If we have a model at hand in which we can estimate the error distribution, we can go a step beyond crossvalidation and **bootstrap** our experiment. Crossvalidation, bootstrap and other resampling plans are a common approaches in data analysis.

4. Excursion: Perception of Graphic Displays

Data analysis tries to find informative features in data, and to bring them to human perception. Appropriate statistics can help finding (or identifying) informative events. Bringing them to human perception is another point. As an illustration, consider the residual plot (figure 4a).

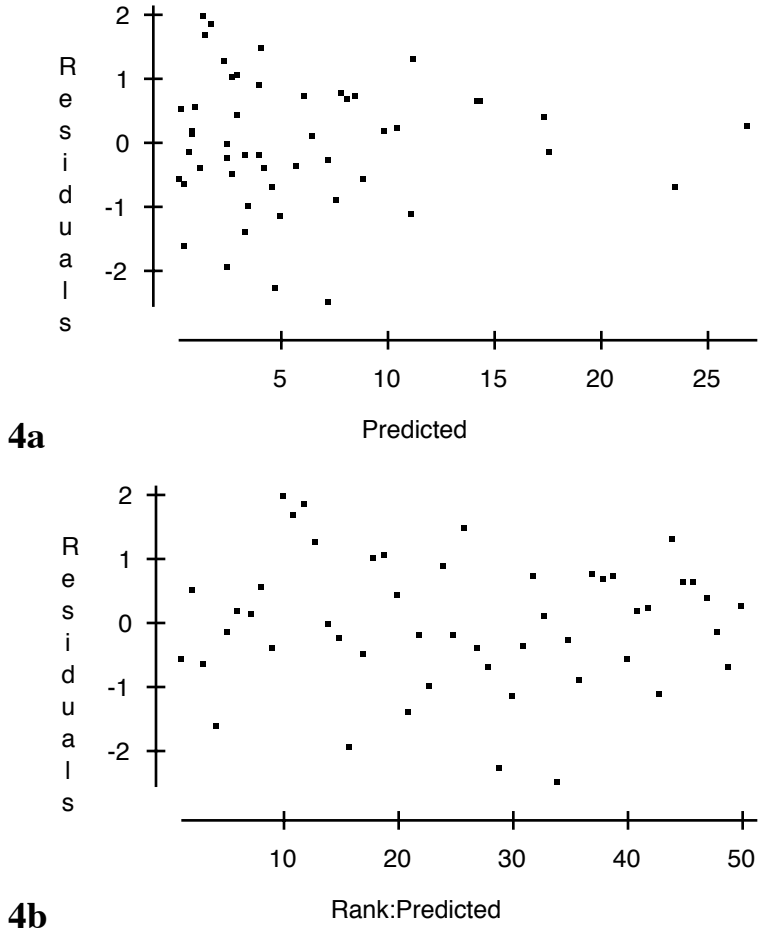

Figure 4: Two views of a residual plot. Note the wedge shaped outline in plot 4a which is not present in plot 4b. Both plots show the same residuals, with the x-axis rank transformed on plot b.

Nearly everybody will immediately see a wedge-shaped structure in this plot. So if you are looking for heterogeneity in variance, you will be alert that variance might not be constant. Now look at figure 4b. There is nothing particularly alerting about this plot. The sad news is: both plots are representing the same residuals. The only difference is that now the residuals are plotted against the rank of the prediction. So the statistical information in both plots is the same. But the concentration of the design points lets the residual plot come out wedge shaped in the first plot. As a matter of fact, these residuals were generated by a simulated simple linear model with homoskedastic normal errors. The skew distribution of the design points makes us perceive plot 4a as different from plot 4b.

The core is: we perceive a shape, or more precisely a "Gestalt". We do not perceive fluctuation (which would be the same in both plots). This is what makes graphical displays a delicate instrument. Data analysis is forced to think carefully of how to communicate informative events to

human perception (Tufte, 1983). An example of a particularly successful display construction is John Tukey's Box&Whisker plot. Let us us look at this gem in data analysis displays (figure 5).

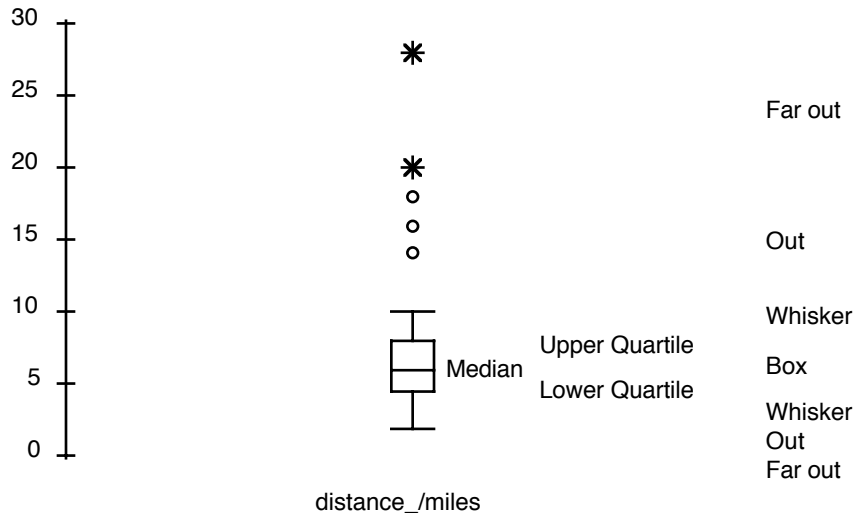

Figure 5: Box&Whisker plot. A *box* showing the quartiles, *whiskers* unto highest data not more than 1.5* interquartile range from upper/lower quartile. *Out* upto 3.0* interquartile range from upper/lower quartile. *Farout* beyond.

The Box&Whisker plots achieve to present general information about the core of the data, with information hiding in this area. On the other hand, they highlight the exceptional. The exceptional data might be just tail effects, or it might be genuine outliers - they are worth a second look anyway. As an application of the Box&Whisker-Plots, figure 6 shows the activity cycle of ants during the day, as observed for a sample of ant colonies. The general characteristic stands out clearly: the "patrollers" are out first, then the nest maintenance ants have their highest activity, and then the foragers. The characteristic time pattern , the "daily round", can be easily perceived from these plots. What is obvious here is very hard to illustrate or find by classical statistical models.

4. Interactive Graphics

After this excursion to dangers and gems of data displays, let us return to the Scottish Hill Runners' data as our leading example. We have split our residual plot into four informative quantities. We now want to make use of their information. What comes to help here is graphical interaction. A common tool in graphical interaction is an identification tool: you move a pointer (linked to a mouse or other input device) across the plot to select a point, press a button and the software replies by showing the identification of the point selected (figure 7).

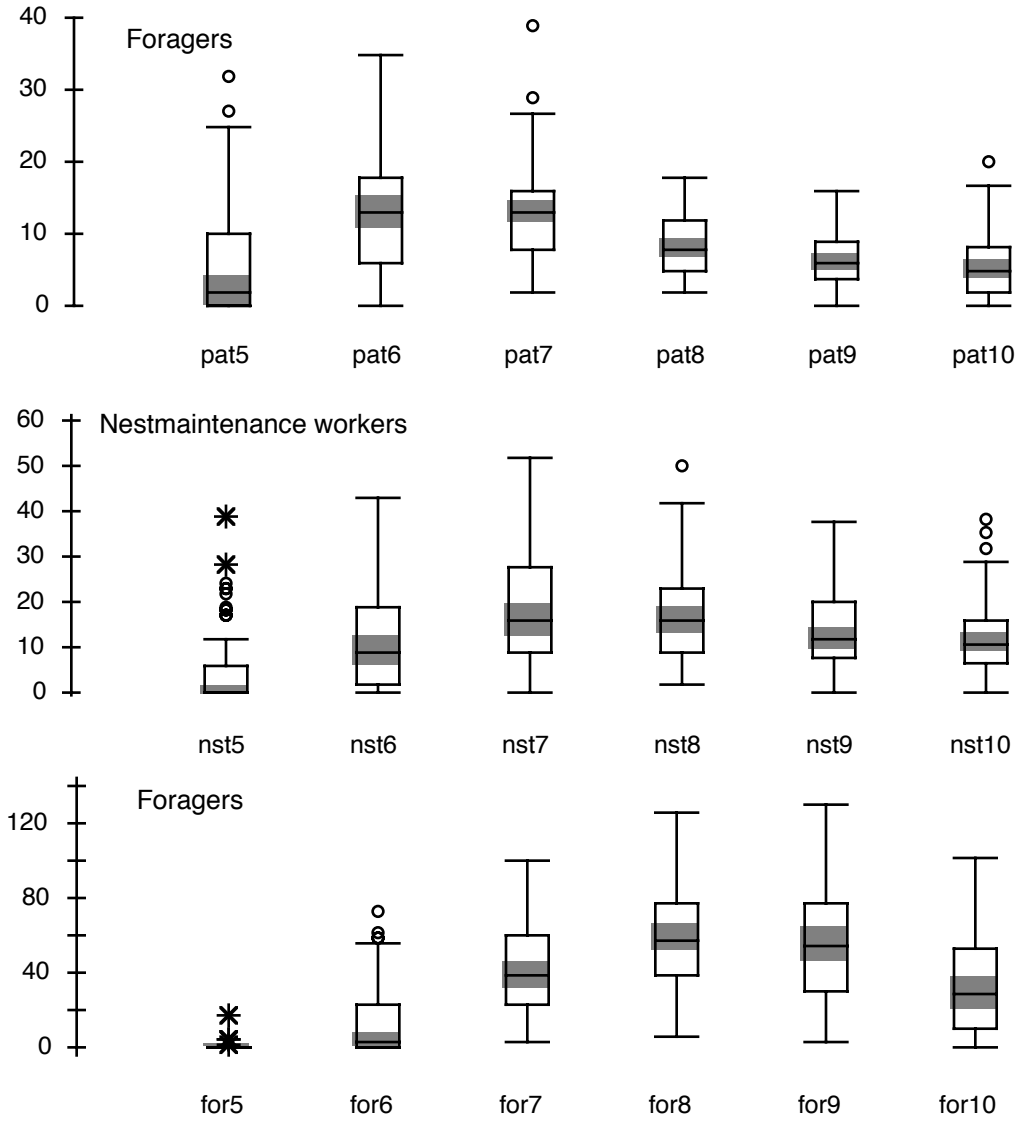

Figure 6: The daily round of ants: activity count of ants, drawn as box&whisker plots, against time of day. Each box&whisker plot represents data from several colonies, by activity group. The different activity patterns of the groups is evident: first activity of patrollers, with maximum at 6:00 (pat6), then main nest maintenance activity with peak at 7:00 (nst7), then foraging with peak at 8:00 (for8). Still to be improved: the shaded bands provide a quick test for pairwise comparison of medians. What would be adequate bands for a profile curve ?

The challenge still is in higher dimensions, and the availability of high computing power made a step in higher dimensions feasible. With a two-dimensional display, one needs a means to reduce dimensionality. There are essentially four ways of reduction:

- put additional dimensions in time and produce a movie representation
- put additional dimensions in space and produce "draftsman's plots"
- put additional dimensions in quality (colour, sound, vibration,…)
- put additional dimensions in interaction.

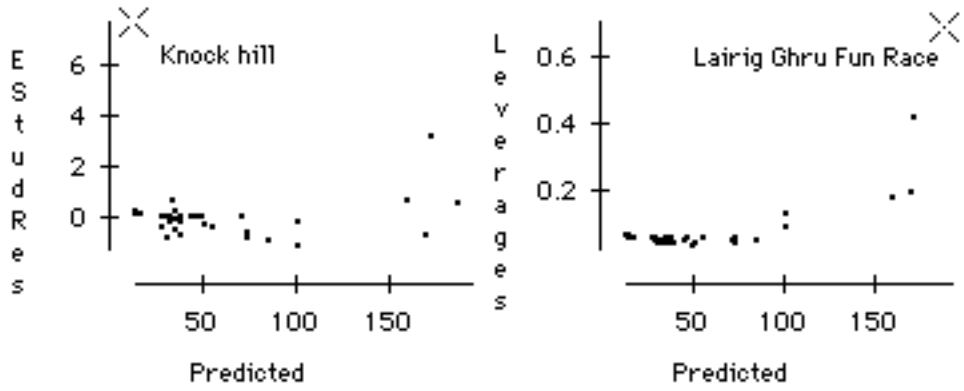

Figure 7: Identification: tag of selected item is shown when selecting with a mouth or an equivalent input device. The main outlier in the in the plot of the externally studentized residuals corresponds to the Knock Hill race; Lairig Ghru Fun Race has exceptional leveerage.

All of them have been tried. The most spectacular one has been the last, as implemented in the PRIM system (PRIM: ADA for Projection, Rotation, Isolation of subgroups, Masking controlled by additional parameters). PRIM was one of the first systems to allow real interaction with data, in particular to allow interactive analysis of (projections of) higher dimensional data sets. The most impressive feature of PRIM was the possibility to rotate a data set in three dimensions. The ISP software package, originally implemented on workstations, provided a link to PRIM-H, thus making it useable for a wider audience. MacSpin was the first program to bring the PRIM facilities down to an affordable budget. By now, the PRIM facilities have become fairly standard in data analysis systems.

Although spectacular at first sight, the PRIM facilities have turned out to be limited in use. PRIM is useful in low dimensional data sets (3-4 dimensions) where you have a large scale structure which can not be reduced to planar models. The diabetes data set of Reaven and Miller (1979) is one of the rare success stories of the PRIM approach.The most fruitful solution so far is the most humble one: just to use a collection of scatterplots (for example to lay out different dimensions in space to produce a draftsman's plot: a matrix of the marginal scatterplots, possibly showing the corresponding histograms in the diagonal) but to **link** the plots. Interactivity is left free to explore the linked windows (figure 8). Once it is realized, this is trivial. But it took someone to have this idea. The earliest reports on interactive graphics for linked windows seems to be Newton (1978, quoted from Becker et al. 1987). A very impressive implementation by John McDonald in the Orion project controlled a "brush" with a mouse. While the brush was moved across one plot, the points under the brush got coloured depending on their distance to the center of the brush. The corresponding representations of the data points in other windows (showing other variables) were coloured in the same colours.

The techniques discussed in this section (Identification, PRIM, linked windows, brushing) are fairly common in graphical data analysis today. For a survey of this area, as seen from Murray Hill, see (Becker et al., 1987).

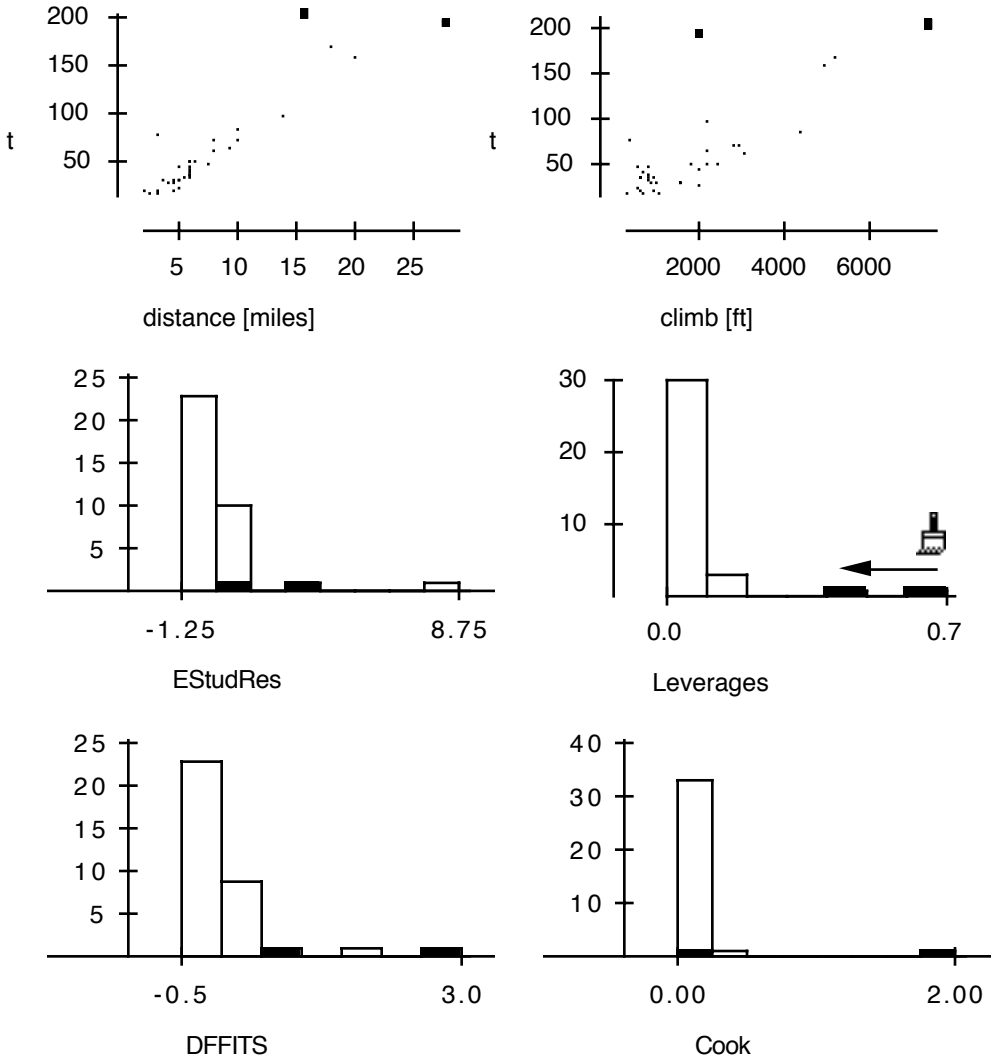

Figure 8: Linked windows with various diagnostic indices. Cases with high leverage are highlighted by a brushing technique in the leverage plot. Corresponding entries in the other plots are highlighted automatically.

5. Statistics of selection and brushing

A new quality in data analysis is gained with the ability to *interact* with the data, especially when there is an immediate graphical response. There is, however, a common overestimation of the importance of the technical tools. The new quality did not arrive with modern high power computer systems. It was already achieved with the first attempts to gain more insight by adding pen strokes and guidelines to pen and paper to help focus on interesting features. It was effectively buried with modern computers providing rapid numerical results, and it is only slowly regained with the new graphical facilities. Two examples for this claim will be given below.

First, let us study the situation of selecting in linked scatterplots. In the simplest situation, we select in one dimension, and look for the response in another dimension. So we can integrate both linked windows in one plot, a ususal scatterplot. Selection acts on one marginal, with response on the other. Again, to simplify in order to make the idea stand out more clearly, we think of a response

discretized into bins i=1…k. Any selection yields highlight counts $m_1,\ldots,m_k$ with $\sum m_i=m$. Now we can see: checking for a relation corresponds to a classical situation. Under the hypothesis of no relation, i.e. independence, the number of highlighted responses is distributed as multi-hypergeometrical distribution (figure 9).

For example, we can select the upper (or lower) half, and to split the response into two bins at the median (hence we have a hypergeometric distribution with parameters m, m_1, n where m_1=N=m div 2). In a test context, this leads to Quenouille's quadrant test (Quenouille 1952, 1959). The generalization to other selections and discretization of the response and to higher dimensions was already obvious in early times to Quenouille (Quenouille 1959).

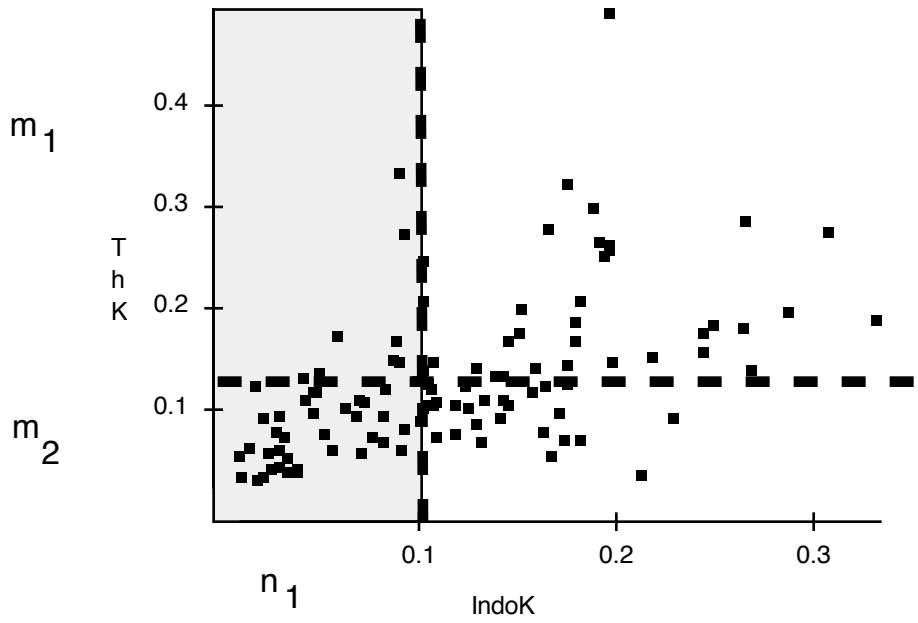

Figure 9: Selecting for an informative event. Here: selecting by medians in IndoK, with reponse split at medians. The informative event is the high occupation count in the lower quadrant. This corresponds to a variant of Quenouille's corner test.

For a second example, let us turn to the dynamic version, i.e. brushing. For simplicity, let us again focus on a one dimensional exploratory variable, with a one dimensional response. We brush in one variable until we see an "informative event" in the other. Again, as a prototype, let us split the response at the median. By brushing, we will get a response on one side of its median. A most obvious informative response is if we jump to the other side of the median (figure 10). In a test context, this again corresponds to a familiar situation: this is Olmstead and Tukey's corner test (Olmstead and Tukey, 1947)

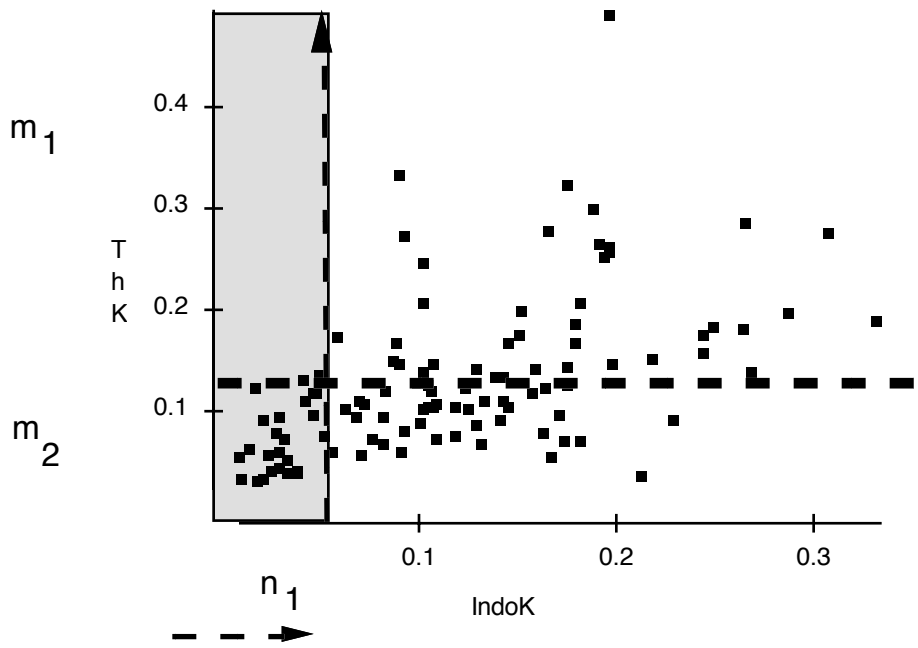

Figure 10: Brushing for an informative event. Here: until a jump over the median occurs in the response. This corresponds to a variant of Tukey's corner test.

6. Data driven modelling

An interesting field is that of data analytical modelling. One wants to go beyond an analysis of the data at hand, aiming for models. There are some approaches. For general regression, for example, there are the popular smoothing approaches, either by using splines or using kernel smoothers, both approaches being closely related (Silverman 1984). The status of the data driven selection of the smoothing parameter, the bandwidth selection, is discussed in (Härdle et al. 1988). CART is a quite distinct approach to tackle regression and classification problems with data analytical methods, essentially tiling the covariate space for homogeneous response. An extensive discussion of this approach can be found in Breiman et al. (1984).

These are data analyitical approaches to modelling. But in general, the relation between data analysis and model building is an open field.

Atkinson, A.C. (1986) Comment: Aspects of Diagnostic Regression Analysis. Statistical Sciences 1, 397 - 402

Becker, R.A., Cleveland, W.S., Wilks, A.R. Dynamic Graphics for Data Analysis (with discussion) Statistical Science 2 (1987) 355-395

Belsely, D., Kuh, E., and Welsch, R.E. (1980) Regression Diagnostics. John Wiley and Sons, NewYork

Breiman, L. and Friedman, J.H. Estimating optimal transformations for multiple regression and correlation (with discussion) J. Amer. Statist. Assoc. 80 (1985) 580-618

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J (1984) Classification and Regression Trees. Wadsworth, Belmont

Chatterjee, S., and Hadi, A. (1988) Sensitivity Analysis in Linear Regression. John Wiley and Sons, NewYork

Chambers, J. M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983) Graphical Methods for Data Analysis. Wadsworth Statistics/Probability Series. Wadsworth, Belmont.

Cook, R.D. Detection of Influential Observations in Linear Regression. Technometrics 19 (1977) 15-18.

Cook, R. D., and Weisberg, S. (1982) Residuals and Influence in Regression. Chapman and Hall, NewYork.

Gordon, D, M. (1986) The Dynamics of the Daily Round of the Harvester Ant Colony (pogonomyrmex barbatus). Animal Behaviour 34, 1402-1419

Härdle, W., Hall , P., and Marron, J.S. (1988) How far are Automatically Chosen Regression Smoothing Parameters from their Optimum ? (with Discussion). J. Amer. Statist. Assoc. 83, 86-101

Newton, C.M. (1978) Graphics: From alpha to omega in data analysis. In: Graphical Representation of Multivariate Data (P.C.C. Wang, ed.) 59-92. Academic Press, NewYork.

Olmstead, P.S., and Tukey, J.W. (1947) A Corner Test for Association. Ann. Math. Statist 18, 495 - 513

Quenouille, M.H. (1952) Associated Measurements. Butterworth's Scientific Publications.

Reaven, G.M., and Miller, R. (1979) An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis. Diabetologia 16, 17-24

Silverman, B.W. Spline Smoothing: the Equivalent Variable Kernel Method. Ann. Statist. 12 (1984) 898-916

Tufte, E. (1983) The Visual Display of Quantitative Information. Graphics Press, Cheshire.

Velleman, P.F. and Hoaglin, D.C. (1981): Applications, Basics and Computing of Exploratory Data Analysis. Duxbury Press, Boston.

The ants activity data (figure 6) are from observations reported in (Gordon 1986). All plots were produced using Paul Velleman's "Data Desk professional" on an Apple Macintosh Plus computer.