Computer-Aided Illustration of Regression Diagnostics

T. Nummi, M. Nurhonen and S. Puntanen, Tampere

Abstract: In this paper we describe a microcomputer system developed by the authors for illustration of various concepts related to regression analysis. Particular attention is paid to the diagnostics for analysing influential observations and outliers. Naturally graphical methods play a central role in this illustration. The system is mainly planned for a student taking a course in regression analysis or a person who is applying regression analysis and wants to know the meaning of diagnostics in practice. The programming language is the APL and thus the user familiar with the APL can easily extend the system by his own functions. The system is implemented on Apple Macintosh personal microcomputer and it is a part of a larger system, called KONSTA 88, which is planned for illustrating statistical concepts.

Key words: Influential observation, leverage, microcomputer software, outlier, perturbation, residual, teaching of statistics.

1. INTRODUCTION

The recent growth of the computational capacity has activated lots of matrix formulas, related to regression diagnostics, which earlier were merely of theoretical interest. The availability of possible diagnostics, such as various residuals, leverages, Cook's distance etc., causes no problems for most users. As a matter of fact, the number of diagnostics suggested in the literature is almost confusingly high. However, their interpretation and practical meaning may not be that easy or obvious.

In this paper we describe a microcomputer system developed by the authors for illustration of various concepts related to regression analysis. Particular attention is paid to the diagnostics for analysing influential observations and outliers. Naturally graphical methods play a central role in this illustration. The system is mainly planned for a student taking a course in regression analysis or a person who is applying regression analysis and wants to know the meaning of diagnostics in practice. The system is implemented on Apple Macintosh personal microcomputer and the programming language is the APL. The user familiar with the APL can easily extend the system by his own functions.

The main motivation in developing the system was that the pure numerical values of the diagnostics do not tell the whole story. Namely the situation may often be so complicated that an interactive graphical manipulation of the data set is needed. On the other hand, one has to learn that the quantitative measures may tell something which is hidden from the eye.

Let us imagine what happens in a student's mind when he sees a scatterplot of the type of Figure1a. One natural idea to come to a student's mind could very well be

Compstat 1988
© Physica-Verlag Heidelberg for IASC
(International Association for Statistical
Computing), 1988

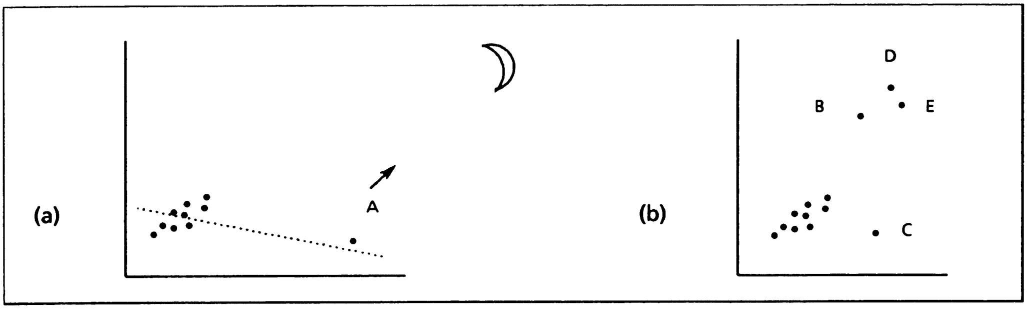


FIGURE 1. Illustration of moving and deleting data points.

something like: "I really would like to move the data point A and see what happens then to the regression line." The student might indeed be curious to know how far and to which direction he has to move the data point to obtain certain regression line or correlation coefficient. Similarly, one might wonder about the consequences of deleting some of the points B, C, D or E in Figure 1b. Furthermore, we might imagine two students arguing with each other about the place where point A should be moved so that a statistic would attain a given value. But what are we involved with now? Isn't it quite near to a kind of a "computer game"? Yes, we think so, and in fact this is one of our main ideas: this program could also be a *toy* for the user of regression analysis. We feel that this toy (like other "good" toys) might be educational.

Our program includes for example a game, where the player sees a scatterplot on the screen and he has to guess some statistics like the correlation coefficient or he has to draw the regression line. Then he is asked to add one data point such that the statistics would change to certain values, or the program adds one point and asks for the changes. On each step the player will be fined or rewarded according to his answers. For a novice user of regression analysis the practice of eyes is an essential matter; cf. Anscombe (1973) and Mosteller *et al.* (1981).

As Cook and Weisberg (1982, p. 101) state, the basic idea in influence analysis is quite simple: we introduce small perturbations in the problem formulation, and then monitor how the perturbations change the outcome of the analysis. The leading idea behind our system is just to demonstrate this principle so that the user could concretely and interactively participate in this process.

One typical application of our system is that the user moves an observation on the screen and simultaneously sees the (graphically presented) effect on the diagnostics he is interested in. Similarly the user can see e.g. how the estimated regression line is moving or the residual variance is changing when one observation or a group of observations is moved. The points which stand out from the group on their diagnostic measures can be "flagged" if the user wants; the user can then perturb the data and ask for a new flagging. One possibility is to have a scatterplot on one window and the standard regression output summary of these data on another window and then to perturb the plot and to observe the consequences to the output. It is also possible to generate data sets with desired values of diagnostics. The user can also violate the standard assumptions of the linear model in various ways (e.g. non-normal, heteroscedastic, correlating errors), generate observations from these violated models and then observe the consequences.

One important aspect is to clarify the differences and similarities between an outlier and an influential case. Other central concepts to be illustrated are e.g. multiple outliers, jointly influential observations, masking, leverage values, transformation to normality, collinearity etc.

2. EXAMPLES

In this section we consider some examples of the use of KONSTA 88 in illustrating concepts related to regression analysis. It is clear that it is not at all an easy task to show on paper how the program works because the leading idea behind the program is that it should work dynamically, interactively and visually. In the following examples we consider the simple regression model with one explanatory variable, i.e., $y = \beta_0 + \beta_1 x + \varepsilon$, where $var(\varepsilon) = \sigma^2$.

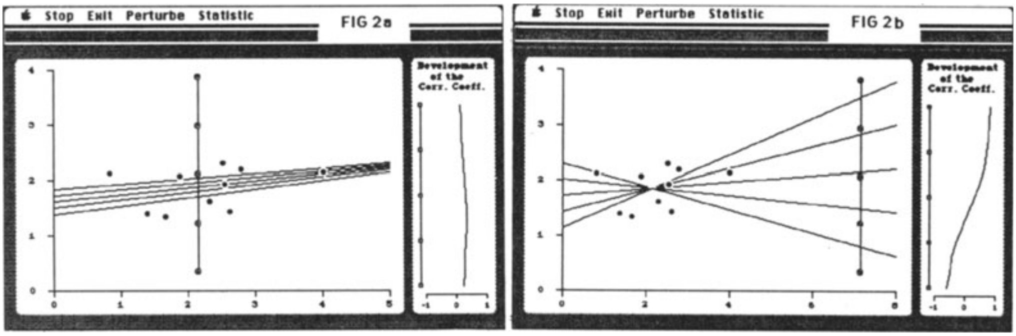

Stop Exit Perturbe Statistic
FIG 2a
Development of the Corr. Coeff.
4
3
2
1
0
0
1
2
3
4
5
-1
0
1
Stop Exit Perturbe Statistic
FIG 2b
Development of the Corr. Coeff.
4
3
2
1
0
0
2
4
6
8
-1
0
1

FIGURE 2. ILLUSTRATION OF THE MOVEMENT OF THE POINT ALONG A LINE. This figure shows (or at least gives some idea on) how the user can move a data point and observe the consequences on the screen. In Figures 2a and 2b the user has defined a line along which the point will move. The user can also define the statistics whose development, while moving a data point, he wants to see plotted. The path of the data point can be marked also completely free by hand or the user can define a circle along which the point will move.

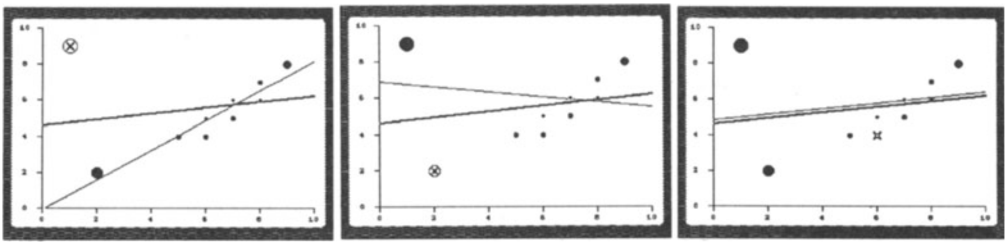


FIGURE 3. ILLUSTRATION OF THE CONSEQUENCES OF DELETING A DATA POINT. Each point is marked with a circle whose area is proportional to the corresponding Cook's distance. The points will be deleted one by one and the resulting regression line is drawn. The deleted point is marked by a cross.

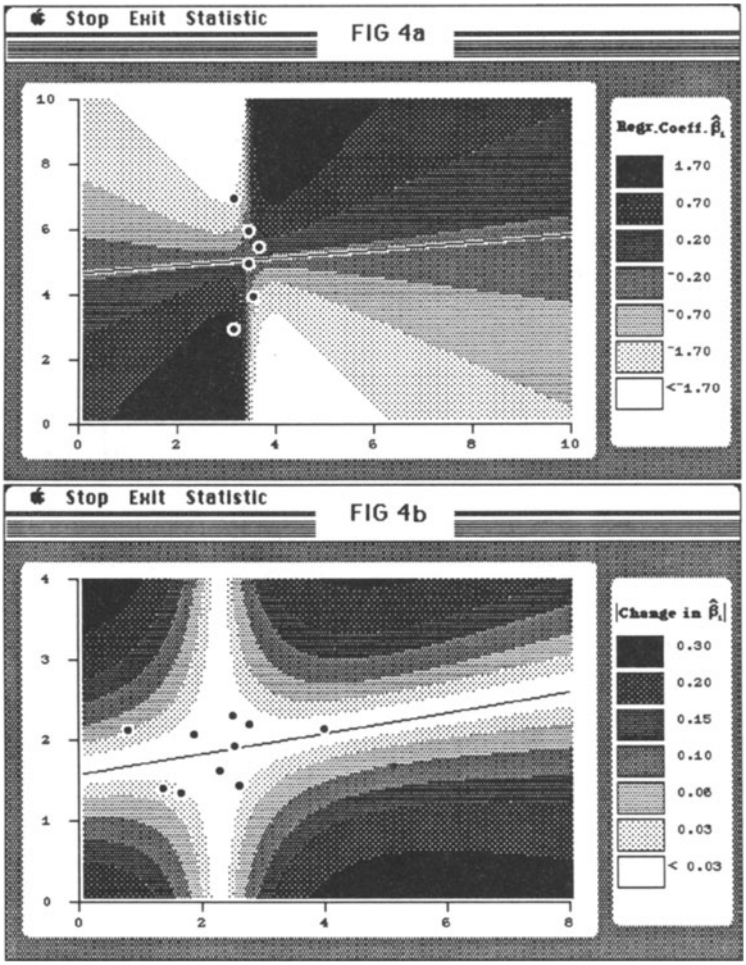


FIGURE 4. SHADED PLOT ILLUSTRATING THE EFFECT OF AN ADDITIONAL POINT ON $\hat{\beta}_1$ (SLOPE).

In Figure 4a dark shades correspond to large values of $\hat{\beta}_1$ indicating that if a point is added in that place then the corresponding new value of $\hat{\beta}_1$ is high. Light shades correspond to small new values of $\hat{\beta}_1$.

In Figure 4b we have a different scatterplot and here the absolute value of the change in $\hat{\beta}_1$, due to the additional point, is calculated.

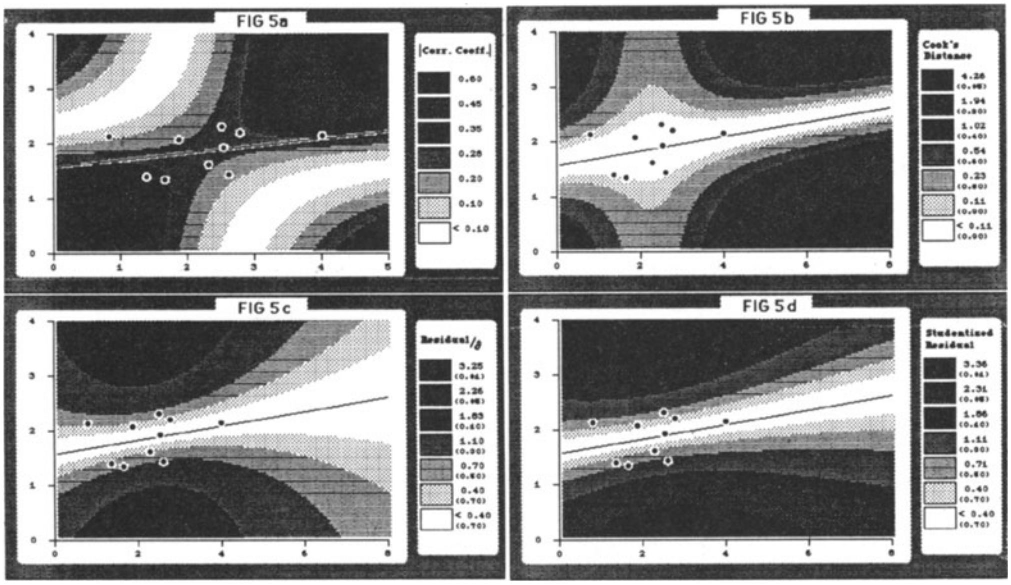
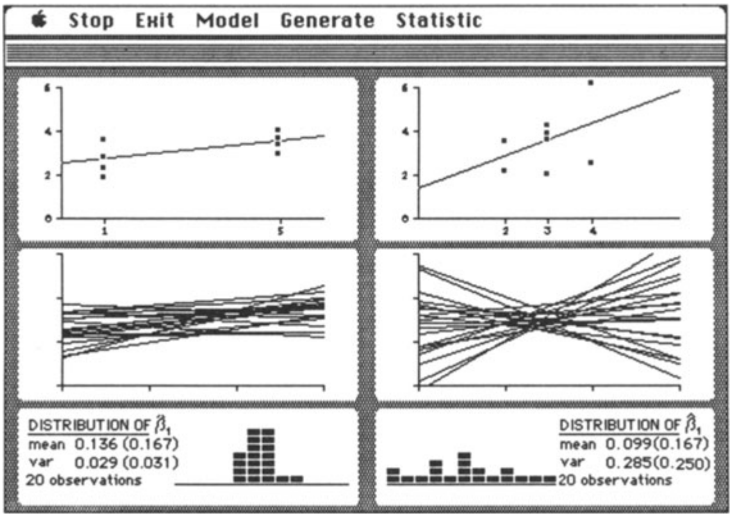


FIGURE 5. SHADED PLOTS ILLUSTRATING THE EFFECT OF AN ADDITIONAL POINT ON SOME STATISTICS. Dark shades in Figure 5a correspond to large absolute values of correlation coefficient r indicating that if a point is added in that place then the corresponding new value of $|r|$ is high. In Figures 5b, 5c and 5d the dark area indicates that if a point is added in that place then that point's Cook's distance, residual (divided by $\hat{\sigma}$), or externally Studentized residual, respectively, is large. We can clearly observe e.g. that the leverage value affects less the externally Studentized residual than the ordinary residual.



```
DISTRIBUTION OF β1
mean 0.136 (0.167)
var  0.029 (0.031)
20 observations
```

```
DISTRIBUTION OF β1
mean 0.099(0.167)
var  0.285(0.250)
20 observations
```

FIGURE 6. Here we generate observations from two identical models, but under different values of x.

The program repeatedly generates observations from both models. After each generation the estimated regression line will be drawn (upper window), it will be "added" to the set of earlier lines (middle window) and also a histogram of observed values of $\hat{\beta}_1$ (or some other statistics) will be drawn.

3. CONCLUDING REMARKS

In this paper we have described some aspects of the illustration of the concepts related to regression analysis. We emphasize simple possibilities for illustrations; at least simple methods should naturally precede the more complicated ones.

It should be emphasized that our computer system is not supposed to be a pure calculator: its main goal is to offer possibilities for an interactive and dynamic illustration of the concepts related to regression diagnostics. The leading principle is to allow the user himself perturb the data (e.g. by deleting, moving or adding points) and then immediately recognize the consequences.

We also agree with Denby and Pregibon (1987) when they argue that exploratory data analysis, using rather simple techniques, should precede formal regression modelling and diagnostic checking. Our system, which is part of a larger system called KONSTA 88, offers flexible possibilities for applying low-tech graphical techniques for looking at the data. KONSTA 88 is an interactive microcomputer system whose main aim is to illustrate theoretical statistical concepts particularly on the first statistics courses.

Furthermore, we would like to emphasize the possibilities to use the program as a toy: certainly the regression fans have earned their own toys. The program has several features which might be rather useless, in the serious classical sense, for the ordinary user of regression analysis. However, we think that they may give new ideas for the user's imagination, ideas which otherwise might not occur to him so easily.

Anyway, though life is short (cf. Cook and Weisberg 1982, p. 8) it seems fair to devote some part of it to regression diagnostics.

ACKNOWLEDGEMENT

This paper was prepared when the first author was an acting research assistant in the Academy of Finland.

REFERENCES

Anscombe, F.J. (1973). Graphs in statistical analysis. *The American Statistician*, 27, 17-21.

Cook, R. Dennis and Weisberg, Sanford (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

Denby, Lorraine and Pregibon, Daryl (1987). An example of the use of graphics in regression. *The American Statistician*, 41, 33-38.

Mosteller, F.; Siegel, A.F.; Trapido, E. and Youtz, C. (1981). Eye fitting straight lines. *The American Statistician*, 35, 150-152.