PARALLEL LINES IN RESIDUAL PLOTS

Shayle R. Searle

Biometrics Unit, Cornell University, Ithaca, NY 14853

BU-945-M* October, 1987

ABSTRACT

Attention is drawn to a little-known feature of residuals plotted against predicted values: when observed values of y are repeated in the data, then the plot exhibits a series of parallel lines having slope -1.

It is common practice after fitting a model to data to make a plot of residuals, of $y - \hat{y}$, say, against predicted values $\hat{y}$. The utility of doing so is promoted and described in many texts (e.g., Draper and Smith, 1981, p. 147). Figure 1 shows an example of such a plot. It is for data that consist of 80 records of an observer's visual assessment of the percentage of a plot of potato plants infected with a leaf disease. The 80 plots consist of four plots on each of 20 varieties of potatoes.

At first sight the startling feature of Figure 1 is the occurrence of parallel lines. But, in fact, this is not startling. Pathologists eyeballing the percentage of a plot of potato plants that has leaf disease record their assessment as a code, A,B,C,$\cdots$; and later convert that code to a pre-assigned percentage of disease, in these data to just the six percentages 10, 25, 50, 75, 90 and 97%. Furthermore, these data had 6, 9, 8, 14, 18 and 25 plots recorded with those percentages, respectively. And the six parallel lines of Figure 1 correspond precisely to these six groups of data. (Points in the figure labeled x, 2 and 3 represent 1, 2 and 3 observations, respectively, and those with a thick-looking symbol represent two or more sets of closely adjacent points.)

The reason for this characteristic is very simple. The plot is of $y - \hat{y}$ against $\hat{y}$. Consider that plot just for y of some particular value, c say. Then it is a plot of conditional variables, of $[(y - \hat{y})|y = c]$ against $(\hat{y}|y=c)$, i.e., of $c - (\hat{y}|y=c)$ against $(\hat{y}|y=c)$. Clearly, this is a straight line, one for each c, every such line having a slope of -1.

---

* Technical Report number BU-945-M in the Biometrics Unit.

Although this result is so simple, it seems to have largely evaded the literature, both books and journals, e.g., Draper and Smith (1981) do not mention it. One place where it *has* been seen is McCullagh and Nelder (1983, p. 216), as kindly pointed out by H.V. Henderson.

Several general properties of this feature of plots of residuals against predicted values are worthy of note. (i) When analyzing data with a new observation $y = c$ added, that additional observation will give rise to a point on the line already established by the $y = c$ data without the additional observation. (ii) Even with data having many different observed values, these parallel lines do exist, although they may not be readily apparent. A vivid illustration of this would be with truncated data, as suggested by M.P. Meredith: if several observations were truncated to the same value and most other observed values each occurred only once, then the straight line corresponding to the truncated value would be patently apparent in the residual-predicted plot. In general, though, it is the presence of only a few and clearly different values that makes the lines very apparent; e.g., Figure 1 with only six different values. A special case of this would be for y being a dichotomous variable, e.g., the data of Neter, Wasserman and Kutner (1984, p. 358): there would be just two distinct lines in the residual-predicted plot. (iii) Even when a data value occurs only once, it implicitly gives rise to a line that would be evident were there more than one datum with that value. (iv) Parallel lines of this nature occur no matter what model is fitted to y, and correspondingly no matter how $\hat{y}$ is calculated, be it based on linear or non-linear estimation. So long as $y - \hat{y}$ is plotted against $\hat{y}$ the parallel lines will exist either explicitly, for repeated y-values, or implicitly, for single y-values. (v) The slope of all such lines is always -1, and the line corresponding to $y = c$ crosses the line $y - \hat{y} = 0$ at $\hat{y} = c$.

References

Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*, (2nd ed.), John Wiley & Sons, New York.

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models*, Chapman & Hall, London.

Neter, J., Wasserman, W. and Kutner, M. (1984). *Applied Linear Regression Models*, Irwin, Homewood, IL.

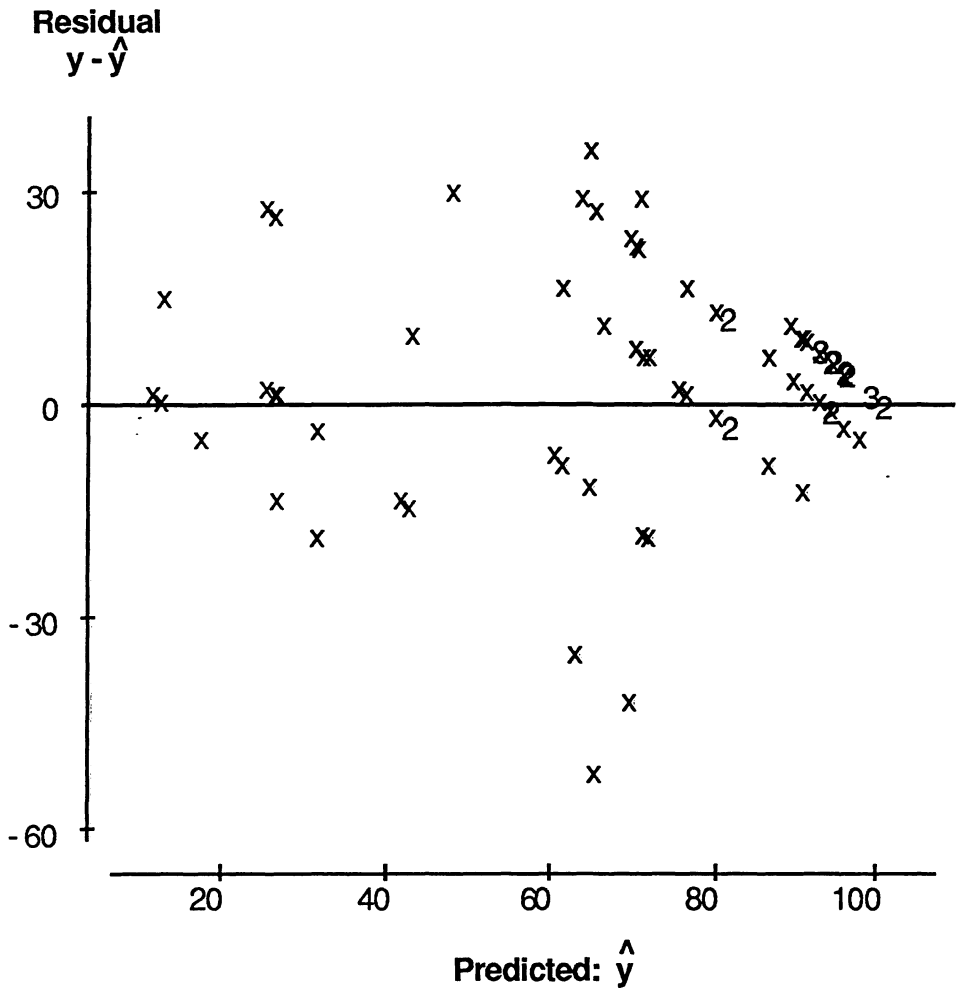

Figure 1: Plot of $y - \hat{y}$ against $\hat{y}$.