INSTANT MATRIX DESIGN: STRAIGHT LINE WITH ERRORS IN BOTH VARIABLES

Irving H. Siegel, The W. E. Upjohn Institute for Employment Research*

My "supermatrix" and "normal-identity" approaches to the adjustment of linear data are applicable not only to the familiar cases in which the values of the dependent variable are subject to error but also to instances in which all the data may be inexact. At the 1967 ASA meeting, I illustrated the normal-identity approach to straight-line-fitting when both the observed y's and x's are uncertain (i.e., $y_i = Y_i - s_i$ and $x_i = X_i - t_i$). [1] I now present the supermatrix equivalent.

In the two adjustment procedures, residuals are introduced explicitly as unknown constants into the observation equations. These equations thus become observational identities of the form $y_i + s_i \equiv a + bx_i + bt_i$, or $y_i \equiv a + bx_i + \Delta_i$, where $\Delta_i = bt_i - s_i$ incorporates paired x and y residuals.

In the normal-identity approach, summary statements are first derived from the observational identities and then reduced to conventional "normal equations" by the suppression of certain aggregates. The eliminated aggregates correspond to plausible assumptions concerning the residuals. In the case considered here, there are three summary statements and four unknowns -- a, b, and two unsuppressed aggregates. A relation between the aggregates is accordingly postulated.

While the normal-identity approach features the "compression" of information, the supermatrix approach leaves the data intact. The observational identities and aggregate residual conditions are organized in unprocessed form (or nearly so) into a very large matrix system that is sufficient to determine a, b, and all the residuals. Although this supermatrix system is condensable to the usual normal equations, solution may be effected in any manner deemed expedient.

To derive the supermatrix system, we may start with what is left of the normal identities after the appropriate residual aggregates are equated to zero. As already noted, we have three equations in four unknowns, two of which are residual aggregates:

$$\left\{\begin{array}{l} \sum y = na + b\sum x \\ \sum xy = a\sum x + b\sum x^2 + b\sum tx \\ \sum y^2 + \sum sy = a\sum y + b\sum xy \end{array}\right\}$$

Next, we rewrite one of the unknown sums in terms of the other: $\sum sy = k\sum tx$, where k is a parameter. [2] This relationship is actually a disguised form of $\sum s^2 = k\sum t^2$, which states that the sum of squared deviations in the y direction is k times the sum in the x direction.

The first normal equation may also be written as $\sum \Delta = 0$, and the second and third yield $\sum (kx + by)\Delta = 0$. These two sums are aggregate residual conditions needed for completion of the supermatrix system, the design matrix of which is a square of the order $\overline{n+2} \times \overline{n+2}$.

The supermatrix system is this:

$$\left[\begin{array}{c|cc} 0 & 1 & 1 \quad \cdots \\ & kx_1+by_1 & kx_2+by_2 \quad \cdots \\ \hline 1 \; x_1 & 1 & \\ 1 \; x_2 & & 1 \\ \vdots & & \ddots \end{array}\right] \cdot \left[\begin{array}{c} a \\ b \\ \hline \Delta_1 \\ \Delta_2 \\ \vdots \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ \hline y_1 \\ y_2 \\ \vdots \end{array}\right]$$

The square design matrix is partitioned so that the packages of unprocessed information may easily be identified. If a numerical value is specified for k, a trivial amount of prior multiplication is required.

Since the design supermatrix contains b, solution for b yields a quadratic equation expressed in b! From this equation, shown in my paper of May 1967 and in some of the literature cited there, the value of b is readily ascertained.

---

*The author's views should not be ascribed to the Upjohn Institute.

[1] See I. H. Siegel, "From Identities to Normal Equations: An Easy Approach to Least Squares", 1967 Social Statistics Section Proceedings of the American Statistical Association, pp. 354-356.

[2] In the 1967 paper (p. 355), k was inadvertently omitted.