Even a non-statistical engineer/physicist being benefited by SAS - the reliable source of statistical arts.

Tadao SHIBAYAMA
(Retired: Nagoya Municipal Industrial Research Institute)
Present Address (Home): Sunshine Kanayama 502, 14-13 Furuwataricho, Naka-ku, 460-0025, Nagoya, JAPAN
64111256@people.or.jp or td.shbym@mediawave.or.jp

1. The first experience of mine with SAS software references.

In the 1988 annual conference of the Japanese Society of Quality Control, I got fortunately acquainted with a staff of SAS Institute Japan Ltd. who afterwards let me know SAS/QC® software, version 6, the FACTEX procedure, particularly, just to introduce it in Japan widely. Looking into the Software Reference [1] that was handed me kindly, I was impressed by the concise, readable and really convincing description with the literatures location. By the guides so dependable and convenient, I could get into archives of analysis and design of experiments, and, furthermore, into archives of general linear models, and of more, all described in SAS/STAT® User's Guide [2].

2. Learning the GLM procedure and some others of SAS/STAT® software.

A chapter of SAS/STAT® User's Guide (version 6) [2] vol.1, Ch.9, The Four Types of Estimable Functions, begins its explanation from an equation of model as rewritten* here as follows

(1a) y = X.cc + vv (= yy + vv provided that yy = X.cc)

- *A doublet cc stands for a greek small letter (beta) in the original whereas a doublet vv for a greek small letter (epsilon) in the original, both to be convenient in popular word processing systems.

It means that a response column vector y as realised is equal to a sum of an error column vector vv and a true response column vector yy (:= X.cc) as built from a true effect elements column vector cc by the design matrix X. A response column vector y as actually measured is employed for the response column vector y as realised in the lefthand side of eqn (1a). Whereas the true effect elements column vector cc in the righthand side is replaced by a fitted effect elements column vector cv. Furthermore, the error column vector vv is replaced by a fitted residues column vector vy. Then, an observation equation follows for the fitted effect elements so that

(1b) y = X.cv + vy (= yv + vy provided that yv = X.cv)

with a doublet cv for a roman small letter (b) in the original - the fitted effect elements column vector whereas a doublet vy for a roman small letter (e) in the original - the fitted residues column vector.

With the given response column vector y as measured, if the sum of squares of the fitted residues (= vy'.vy) is minimized, then, follows the normal equation such that

(1c) X'X.cv = X'.y with a solution cv = (X'X)⁻.X'.y

to be indeterminate so inducing problems of generalized inverses or of arbitrary constraints.

3. The contraction and restoration matrices.

If rows of the design matrix X are linearly dependent, one of the rows is a linear combination of the other rows. Furthermore, if some of the other rows are linearly dependent, one of the other rows is a linear combination of the rest. And so on. Finally, the design matrix X is contracted to an estimable full matrix Lu (say) as composed of qu (say) linearly independent rows of the design matrix X. A contraction matrix K can give the estimable full matrix Lu if that is operated on the design matrix X. Whereas a restoration full matrix J can restore the design matrix X because any row of it is a linear combination of the rows of the estimable full matrix Lu. Consequently so that

(2a) Lu = K.X and X = J.Lu

An estimable part matrix L is defined with some rows of the estimable full matrix Lu replaced by zero rows. Still operated by the restoration full matrix J, it gives a design part matrix Xp just equal to J.L. It replaces the design matrix X in eqn (1b) giving an observation equation such that

(2b) y = Xp.cvP + vyP (= yvP + vyP provided that yvP = Xp.cvP)

With the given response column vector y as measured, the sum of squares of the fitted residues (= vyP'.vyP) is minimized, then, follows the normal equation such that

(2c) Xp'Xp.cvP = Xp'.y with a solution cvP = (Xp'Xp)⁻.Xp'.y

to be indeterminate, again, so inducing problems of generalized inverses or of arbitrary constraints.

4. General form of estimable functions and various sums of squares.

Rows Lu⁻1, Lu⁻2, ⋯ , Lu⁻qu of an estimable full matrix Lu, if combined linearly and arbitrarily, give an estimable row vector Lu⁻o in a general form such as below -- it is multiplied to an effect elements column vector cc (or cv) to give an estimable function Lu⁻o.cc (or Lu⁻o.cv).

(3a) Lu⁻o = Lu_1.Lu⁻1+ Lu_2.Lu⁻2+ ⋯ + Lu_qu.Lu⁻qu

The combination coefficients Lu_1, Lu_2, ⋯ , Lu_qu are selected arbitrarily giving many estimable row vectors, many estimable full matrices Lu ('s), and many estimable part matrices L ('s).

A sum of squares SS_yv or SS_yvP to be employed in hypothesis testing follows so that

(3b) SS_yv = yv'.yv = cv'.X'X.cv = (Lu.cv)'.J'J.(Lu.cv)

(3c) SS_yvP = yvP'.yvP = cvP'.Xp'Xp.cvP = (L.cvP)'.J'J.(L.cvP)

and the sums of squares of SAS Type I-IV, too. (Instead of the matrix product J'J, an inverse matrix $[L'.(X'X)^{-}.L]^{-1}$ is employed in SAS/STAT® User's Guide (version 6) [2] p.110.)

Many estimable functions are detailed in the User's Guide [2] p.115-124 and in the important references [2a][2b]. A proceeding [2c] of SAS is seemingly especially important if available.

The works of Goodnight [2a][2b][2c] are important in very clear explanation of theory and practice of normal equations to be solved by forward elimination and backward substitution, too.

5. Estimability problem and others.

Results of GLM procedure above are not affected by indeterminacies or specific constraints as proved by estimability. It is remarked that indeterminacy in an effect elements vector cc or cv is traceable in terms of effect elements null base vectors [3] so that the proof is to be visualized.

Meanwhile, canonical (usual) constraints are equivalent to isolability (i.e. orthogonality) of effect components, as established by Lagrange's art of indeterminate coefficients [4].

These discussions help a beginner to be benefited by SAS whereas the immense but readable compilation is obviously the most reliable source of statistical arts for professionals ever.

REFERENCES

[1] SAS Institute Inc. (1989): SAS/QC® Software Reference, Version 6, 1st edn.

[2] SAS Institute Inc. (1989): SAS/STAT® User's Guide, Version 6, 4th edn, vol. 1 and 2.

[2a] J.H.Goodnight (1978): SAS Technical Report R-101.

[2b] J.H.Goodnight (1978): SAS Technical Report R-106.

[2c] SAS Institute Inc. (1976): Proceedings of the First International User's Conference.

[3] Tadao Shibayama (2003): Bulletin ISI 54th Session LX, Contrib. Papers Book 2, p.424-425.

[4] Tadao Shibayama (2003): Paper to be presented in 17th Asia Quality Symposium.

RÉSUMÉ

Un ingénieur/physician non statistique aussi bien bénéficiant par SAS - la source digne de confiance d'arts statistiques. *Tadao SHIBAYAMA (en retraite de l'Institute de la Municipalité de Nagoya de Recherché de l'Industrie), Nagoya, JAPON. La immense compilation de courants arts statistiques aide ingénieurs/physicians non statistiques aussi bien à savoir les exacts princips par la description dans les manuels des instructions lesquels repérent précises locations des originals.*