Psychological Reports, 1997, 81, 1387-1391.

COMPUTING SUM OF PRODUCTS ABOUT THE MEAN WITH PAIRWISE ALGORITHMS[1]

J. GABRIEL MOLINA, PEDRO M. VALERO, AND JAIME SANMARTÍN

Department of Behavioural Sciences Methodology
University of Valencia

Summary.—We discuss pairwise algorithms, a kind of computational algorithm which can be useful in dynamically updating statistics as new samples of data are collected. Since test data are usually collected through time as individual data sets, these algorithms would be profitably used in computer programs to treat this situation. Pairwise algorithms are presented for calculating the sum of products of deviations about the mean for adding a sample of data (or removing one) to the whole data set.

In this paper, we discuss algorithms for computing the sum of squares and sum of products of the deviations about the mean. *Sum of Squares* (*SS*) and *Sum of Products* (*SP*) algorithms are basic since they are a common calculation used in computing several basic statistics such as

$$\text{Variance } (X) = \frac{SS(X)}{N} \qquad \text{Covariance } (X,Y) = \frac{SP(X,Y)}{N}$$

$$\text{Correlation } (X,Y) = \frac{SP(X,Y)}{SS(X) \cdot SS(Y)}$$

Among the several algorithms proposed for obtaining the sum of squares and sum of products of the deviations about the mean, definition algorithms (also called two-pass algorithms) have been most widely used in teaching, since they reflect mathematically the meaning of the algorithm. This is also true for desk-calculator algorithms (also called textbook algorithms) because they are the simplest for desk calculation. In Table 1, these algorithms are shown.

Lesser known algorithms are the updating algorithms, also called recurrence, recursive, or on-line algorithms. The basis for these algorithms is an iteration formula for deriving the statistic for n values, from the statistic for the first $n - 1$ of these. In this way, the value of the statistic is obtained as each observation is introduced into the calculation. Some advantages of these algorithms have been shown by Searle (1983). In Table 2 some updating sum-of-squares and sum-of-products algorithms suggested by different authors are shown.

[1]Address enquiries to J. G. Molina, Facultad de Psicologia, Universidad de Valencia, An. Blasco Ibanez 21, 46010 Valencia, España. The algebraic derivation of the algorithms is on file with the National Auxiliary Publications Service, c/o Microfiche Publications, POB 3513, Grand Central Station, New York, NY 10168. Request Document NAPS-05432 and remit $5.00 for fiche or $7.75 for photocopy.

TABLE 1
CLASSICAL SUM OF SQUARES AND SUM OF PRODUCTS ALGORITHMS

Algorithms	Sum of Squares (SS)	Sum of Products (SP)
Definition	$SS = \sum_{i=1}^{n} (x_i - M)^2$	$SP_{XY} = \sum_{i=1}^{n} (x_i - M_X)(y_i - M_Y)$
Desk-calculator	$SS = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$	$SP_{XY} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$

Chan, Golub, and LeVeque (1983) presented an algorithm for calculating the sum of squared deviations which they called a *pairwise algorithm.* This algorithm allows the combination of the sum-of-squares values of two different samples of arbitrary size to obtain the value for the total sample. According to this algorithm, if we have two samples of observations ($\{x_i\}_{i=1}^{m}$ and $\{x_i\}_{i=m+1}^{m+n}$) of which we know:

$$S_{1,m} = \sum_{i=1}^{m} x_i$$

$$S_{m+1,m+n} = \sum_{i=m+1}^{m+n} x_i$$

$$SS_{1,m} = \sum_{i=1}^{m} \left(x_i - \frac{1}{m} S_{1,m}\right)^2$$

$$SS_{m+1,m+n} = \sum_{i=m+1}^{m+n} \left(x_i - \frac{1}{n} S_{m+1,m+n}\right)^2$$

Then, if we combine all the observations into a sample of size $m+n$ [$\{x_i\}_{i=1}^{m+n} = \{x_i\}_{i=1}^{m} + \{x_i\}_{i=m+1}^{m+n}$], we can calculate the sum of squares of deviations about the mean with Formula [1].

$$SS_{1,m+n} = SS_{1,m} + SS_{m+1,m+n} + \frac{m}{n(m+n)} \left(\frac{n}{m} S_{1,m} - S_{m+1,m+n}\right)^2 \qquad [1]$$

If, on joining two samples, one of them had only one observation, then the pairwise algorithm would adopt the characteristics of an algorithm as those presented in Table 2. We can see this in the pairwise algorithm of Chan, *et al.* (1983), taking into account that the second sample ($\{x_i\}_{i=m+1}^{m+n}$) involves only one observation (x_i). In this way, it is evident that an updating algorithm is a particular case of a pairwise algorithm.

More important in practice, since collection of data is usually done

TABLE 2
UPDATING ALGORITHMS FOR SUM OF SQUARES AND SUM OF PRODUCTS

Algorithms	Sum of Squares (SS)	Sum of Products (SP)
Welford (1962)	$SS_i = SS_{i-1} + \left(\frac{i-1}{i}\right)(x_i - M_{i-1})^2$	$SP_i = SP_{i-1} + \left(\frac{i-1}{i}\right)(x_i - MX_{i-1})\,(y_i - MY_{i-1})$
Van Reeken (1968)	$SS_i = SS_{i-1} + (x_i - M_{i-1})^2 - \frac{(x_i - M_{i-1})^2}{i}$	
Youngs-Cramer (1971)	$SS_i = SS_{i-1} + \frac{(ix_i - S_i)^2}{i(i-1)}$	$SP_i = SP_{i-1} + \frac{(ix_i - S_{x_i})(iy_i - S_{Y_i})}{i(i-1)}$
Cotton (1975)	$SS_i = x_i^2 + (i-1)(s_{i-1}^2 + M_{i-1}^2) - iM_i^2$	
Searle (1983)	$SS_i = SS_{i-1} + i(i-1)(M_i - M_{i-1})^2$	

sequentially (for example, different test data sets are obtained through time as new subjects take a test), pairwise algorithms represent a computational solution that fits this circumstance. Statistics for the combination of these data-sets can be obtained without the time-consuming task of recalculating them with all of the raw data.

Anyway, other pairwise algorithms should be available for program developers to support other calculations than the sum of squares about the mean. Moreover, it would be desirable to think over pairwise algorithms that can be applied when a definite data set is withdrawn from the whole data set. These goals are reflected in section below on the sum of the products for pairwise algorithms.

Sum of Products for Pairwise Algorithms

Suppose we have two samples of observations (sizes m and n, respectively) from two variables X and Y [$\{x_i,y_i\}_{i=1}^{m}$ and $\{x_i,y_i\}_{i=m+1}^{m+n}$] of which we know their respective means [$M_{X1,m}$, $M_{Y1,m}$ and $M_{Xm+1,m+n}$, $M_{Ym+1,m+n}$] and sum of products ($SP_{XY1,m}$ and $SP_{XYm+1,m+n}$). So, the sum of products of the whole sample $[\{x_i,y_i\}_{i=1}^{m+n} = \{x_i,y_i\}_{i=1}^{m} + \{x_i,y_i\}_{i=m+1}^{m+n}]$ can be calculated through the following pairwise algorithm [2]:

$$SP_{1,m+n} = SP_{1,m} + SP_{m+1,m+n} + \frac{mn}{m+n}(M_{Xm+1,m+n} - M_{X1,m})(M_{Ym+1,m+n} - M_{Y1,m}) \qquad [2]$$

To compute the sum of products for the sample resulting from extracting a data set from the whole data set, the next variant of the pairwise algorithm [2] would be useful.

Let X and Y be two variables of which we have a sample of observations $[\{x_i,y_i\}_{i=1}^{m+n}]$ from which we take a subsample $[\{x_i,y_i\}_{i=r}^{s}]$ of size n. Known for the whole sample and the subsample are their sizes ($m+n$ and n, respectively), their means ($M_{X1,m+n}$, $M_{Y1,m+n}$ and $M_{Xr,s}$, $M_{Yr,s}$), and their sums of products ($SP_{X1,m+n}$, $SP_{Y1,m+n}$ and $SP_{Xr,s}$, $SP_{Yr,s}$), so the sum of squares for the data-set resulting from extracting the subsample from the whole data-set ($\{x_i,y_i\}_{i=1}^{m-n} = \{x_i,y_i\}_{i=1}^{m+n} - \{x_i,y_i\}_{i=r}^{s}$) could be calculated using the computation algorithm [3]:

$$SP_{1,m-n} = SP_{1,m+n} - SP_{r,s} - \frac{mn}{m-n}(M_{X1,m+n} - M_{Xr,s})(M_{Y1,m+n} - M_{Yr,s}) \qquad [3]$$

The increasing use of updating methods (Grant, 1987) and the generality and usefulness of the pairwise algorithms suggest that they are worthy of being implemented in programs focused on bank, test, or item analysis which involve a dynamic manipulation of the data. And, more generally, applications wherein is involved dynamic updating of statistics as a result of data management.

REFERENCES

CHAN, T. F., GOLUB, G. H., & LEVEQUE, R. J. (1983) Algorithms for computing the sample variance: analysis and recommendations. *The American Statistician*, 37, 242-247.

COTTON, E. W. (1975) Remark on Stably updating mean and standard deviation of data. *Communications of the ACM*, 18, 458.

GRANT, I. H. W. M. (1987) Recursive least square. *Teaching Statistics*, 9, 15-18.

SEARLE, S. R. (1983) The recurrence formulae for means and variances. *Teaching Statistics*, 5, 7-10.

VAN REEKEN, A. J. (1968) Dealing with Neely's algorithms. *Communications of the ACM*, 3, 149-150.

WELFORD, B. P. (1962) Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4, 419-420.

YOUNGS, E. A., & CRAMER, E. M. (1971) Some results relevant to choice of sum and sum-of-product algorithms. *Technometrics*, 13, 657-665.

Accepted November 3, 1997.