ESTIMATING DESIGN EFFECT FROM A CORRELOGRAM

C. H. Proctor, North Carolina State University

A survey of student attitudes was done in 1978 by Dr. Cates (then a doctoral candidate in education at North Carolina State University) who collected answers from 1,974 students in a random sample of 129 classrooms. In calculating sampling standard errors one would ordinarily make use of variation among classrooms. In this particular case classroom identification was not available. It had been deleted in the interests of privacy and confidentiality. However, the questionnaire responses were coded and punched as they arrived - classroom by classroom - and this ordering of students was preserved on the dataset.

We made use of the shape of the correlogram in order to estimate the intra-cluster or intra-classroom correlation coefficient and were thus able to compute a design effect quantity or deff as written by Leslie Kish [1]. Our procedure was to express the large sample expectations of the circular serial correlation coefficients as a function, a linear function as it turns out, of the distance or gap and then furnish a line fitted to the observed correlogram. The slope of this line estimates $-\rho_c/m$ where ρ_c is the intra-cluster correlation while $m = 15.3$ or the average cluster size.

The actual correlogram showed some decline (see Figure 1) continuing beyond the gap of 15 whereas preliminary theoretical results would require the correlogram there to level out at a little below zero. In part such a departure could be due to the variation in cluster sizes around 15 but there may also be reason to expect some autocorrelation from classroom to classroom due to having a school system in common. Consequently, the derivation of the following results will be done for a model having both the intra-cluster correlation parameter ρ_c and also supposing a Markoff process from cluster to cluster with parameter γ.

We will suppose provisionally that all clusters have the same size, namely m, and that there are n of them. If such had actually been the case we could have assigned each m students to their classroom and used the variation among the n cluster means to estimate sampling error. The purpose of the assumption is, of course, to gain an idea of the general relationship between the shape of the correlogram and the intra-cluster correlation coefficient without becoming bogged down in all the possible ways the n classrooms with their unequal sizes could be ordered.

Let y_{ij} be the score for the $j^{\underline{th}}$ student in the $i^{\underline{th}}$ classroom. The following model equation refers to a hypothetically infinite super-population that is taken to underly the measurements:

$$y_{ij} = \mu + a_i + \gamma a_{i+1} + b_{ij} \quad ,$$

$$i = 1, 2, \ldots, n, \text{ and } j = 1, 2, \ldots, m \quad . \qquad (1)$$

The n a's and nm b's are independent with zero means, random effects, with $E(a_i^2) = \sigma_a^2$ and $E(b_{ij}^2) = \sigma_b^2$. For convenience we set $a_{n+1} = a_1$ for a circular ordering and the autocorrelations will be handled similarly.

Consider first the following autocovariances:

$$nmc_d = \sum_{i=1}^{n} \left[\sum_{j=1}^{m-d} (y_{ij} - \bar{y})(y_{i,j+d} - \bar{y}) + \sum_{k=1}^{d} (y_{i,m-d+k} - \bar{y})(y_{i+1,k} - \bar{y}) \right] \qquad (2)$$

$$nmc_{d'} = \sum_{i=1}^{n} \left[\sum_{j=1}^{2m-d'} (y_{ij} - \bar{y})(y_{i+1,d'-m+j} - \bar{y}) + \sum_{i=1}^{n} \sum_{k=1}^{d'-m} (y_{i,2m-d'+k} - \bar{y})(y_{i+2,k} - \bar{y}) \right] \qquad (3)$$

for $d = 1, 2, \ldots, m$ and $d' = m+1, m+2, \ldots, 2m-1$. The expectations of these autocovariances can be found by substituting terms of the model equation (1) into expressions (2) and (3) while making use of the following relations:

$$E(a_i - \bar{a})(a_j - \bar{a}) = -\sigma_a^2/n \quad , \qquad (4a)$$

$$E(a_i - \bar{a})^2 = (n-1)\sigma_a^2/n \quad , \qquad (4b)$$

$$E(b_{ij} - \bar{b})(b_{k\ell} - \bar{b}) = -\sigma_b^2/nm \quad . \qquad (4c)$$

It may be verified that:

$$E(c_d) = -d(1-\gamma+\gamma^2)\sigma_a^2/m + [1+\gamma^2-(1+\gamma)^2/n]\sigma_a^2 - \sigma_b^2/nm \quad , \qquad (5)$$

$$E(c_{d'}) = -d'\gamma\,\sigma_a^2/m + [2\gamma-(1+\gamma)^2/n]\sigma_a^2 - \sigma_b^2/nm \quad , \qquad (6)$$

$$E(c_{d''}) = -(1+\gamma)^2\,\sigma_a^2/n - \sigma_b^2/nm \quad , \qquad (7)$$

for $d = 1, 2, \ldots, m$; $d' = m+1, m+2, \ldots, 2m-1$ and $d'' = 2m+1, 2m+2$, etc.

The zero-gap autocovariance c_o is the average sum of squared deviations,

$$nmc_o = \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y})^2 \quad ,$$

and has expectation $\sigma_b^2 + (1+\gamma^2)\sigma_a^2$.

The intra-cluster correlation may be defined as:

$$\rho_c = \sigma_a^2/(\sigma_b^2 + \sigma_a^2) \qquad (8)$$

so that the variance of the sample mean of n randomly selected clusters becomes:

$$V(\bar{y}) = \sigma_t^2[\rho_c/n + (1-\rho_c)/nm] = \sigma_t^2[1 + (m-1)\rho_c]/nm \quad , \qquad (9)$$

where $\sigma_t^2 = \sigma_a^2 + \sigma_b^2$. The variance for the mean of nm randomly selected individuals is σ_t^2/nm so that design effect becomes

$$\text{deff} = [1 + (m-1)\rho_c] \quad . \qquad (10)$$

Notice that the population of interest for inference has been taken as that shown in (1) with $\gamma = 0$.

The serial correlations are the ratios $r_d = c_d/c_o$ which, as sample size (n) increases, tend to:

$$r_d \to \rho - d\rho(1-\gamma+\gamma^2)/(1+\gamma^2)m$$

$$r_{d'} \to 2\gamma\rho/(1+\gamma^2) - d'\rho\gamma/(1+\gamma^2)m \qquad (11)$$
$$r_{d''} \to 0 \quad ,$$

where $\rho = (1+\gamma^2)\sigma_a^2/[\sigma_b^2 + (1-\gamma^2)\sigma_a^2]$.

These results suggest examining the correlogram first for evidence of a difference of slope at m . If the serial correlations of gap greater than m seem to be constant and near zero then declare $\gamma = 0$. Notice that in this case $\rho_c = \rho$ and

$$r_d \to \rho - d\rho/m \quad , \quad d = 1,2, \ldots , m \quad , \qquad (12)$$

so that the slope between gap 1 and gap m can be multiplied by -m to estimate ρ . If the correlogram is still falling for gaps beyond m then one uses the two slopes, one from d = 1 to d = m and the other from d = m to d = 2m - 1 , to solve for γ and ρ from equation (11).

In practice we first drew the eyefitted line shown in Figure 1, but later we used unweighted regression on splines, see Smith [2], with the observed correlations as dependent variables and the variables shown in Table 1 as X, XF and XL as independent. That is, the variable XF, being constant beyond m , will furnish the fit for the case $\gamma = 0$ while fitting to both XD and XL will cover the case $\gamma \neq 0$. The results of the fittings are shown in Table 2.

Being correlation coefficients on 1,974 cases, one might expect an error mean square of the serial correlations of $1{,}974^{-1} = .000507$ which is reasonably close to the values .000740 and .000631 of Table 2. The improvements in fit lead to t-values of t = 2.62 for variables @5 and of t = 2.03 for @8 so that both slopes will be used. Dividing the slope for r_d by that for $r_{d'}$ in (11) one finds

$$b_1/b_2 \hat{=} (1-\gamma+\gamma^2)/\gamma \quad , \text{ or}$$
$$2\hat{\gamma} = 1 + b_1/b_2 - [(1+b_1/b_2)^2 - 4]^{\frac{1}{2}} \quad . \qquad (13)$$

Using the values of b_1 and b_2 from Table 2 we find $\hat{\gamma} = .5026$ for variable @5 and $\hat{\gamma} = .6923$ for @8 . In the presence of nonzero γ the estimate of $\hat{\rho}$ based on b_1 becomes:

$$\hat{\rho} = mb_1(1+\hat{\gamma}^2)/(1-\hat{\gamma}+\hat{\gamma}^2) = mb_2(\hat{\gamma}^{-1}+\hat{\gamma}) \quad , \qquad (14)$$

and that for ρ_c is

$$\hat{\rho}_c = [1 + (1+\hat{\gamma}^2)(\hat{\rho}^{-1} - 1)]^{-1} \qquad (15)$$

and one finds that $\hat{\rho} = .1548$ and $\hat{\rho}_c = .1276$ for @5 , while $\hat{\rho} = .0882$ and $\hat{\rho}_c = .0614$ for @8 . Referring to equation (10) the corresponding estimates of deff become 2.82 and 1.88 . I recommend that the researcher use a rather conservative design effect of 3 in correcting sampling standard errors. That is, standard errors that had been calculated on the basis of the 1,974 individual scores were multiplied by $\sqrt{3}$.

This overly cautious recommendation was based in part on my uncertainty as to whether the formula (9) for variance of a sample mean should have ρ in place of ρ_c . I recalled having advised on the sample design but I was not sure whether the sample had turned out as one drawn of classrooms or of schools. If schools had been used as sampling units then ρ rather than ρ_c should appear in the formula for $V(\bar{y})$. The reader may verify that design effects in this case become 3.37 and 2.35 so that taking 3 for design effect is a compromise here.

It was finally verified that the selection was done as a simple random sample of classrooms. The evidence for nonzero γ was apparently brought about by a grouping of the schools before coding by geographic, size and population density criteria. That is, the correlation represented by γ was, in this case, created mechanically and thus deserved to be discounted as is done by using ρ_c rather than ρ in (9) .

The choice of estimators was made without extensive support from knowledge of the joint distribution of the serial correlations. Thus, it is likely that improvements to equations (13), (14) and (15) can be found. Perhaps simulation studies could be done to verify the reasonableness of the procedure illustrated for testing that $\gamma = 0$. However, these refinements would only be justified if the method were to see wide usage, while one might rather hope that classroom identification be kept available and sampling variances could then be calculated more directly.

Table 1. Serial Correlation Coefficients and Independent Variables Used to Fit Lines to the Correlograms for Two Questionnaire Items, Objective Five (@5) and Objective Eight (@8) .

Gap	Serial Correlation @5	Serial Correlation @8	Independent Variables XF	Independent Variables XL
1	.209	.147	1	.0
2	.145	.078	2	.0
3	.155	.058	3	.0
4	.093	.049	4	.0
5	.119	.076	5	.0
6	.074	.054	6	.0
7	.082	.016	7	.0
8	.078	.010	8	.0
9	.101	.030	9	.0
10	.101	.044	10	.0
11	.115	.068	11	.0
12	.087	.076	12	.0
13	.101	.060	13	.0
14	.098	.018	14	.0
15	.075	.041	15	.0
16	.041	.064	16	.7
17	.029	.047	16	1.7
18	.036	.026	16	2.7
19	.051	-.009	16	3.7
20	.067	.015	16	4.7
21	.026	.010	16	5.7
22	.012	.013	16	6.7
23	.060	.011	16	7.7
24	.019	.045	16	8.7
25	.055	.005	16	9.7
26	.047	.016	16	10.7
27	.013	-.009	16	11.7
28	-.039	-.003	16	12.7
29	.014	-.005	16	13.7

Source: Correlations were calculated from data described in Cates [3].

Table 2. Results of Fitting Lines to Correlogram*

	Q5	Q8
Total corrected sum of squares	.07236014	.03335531
Error sum of squares after fitting XF	.02431875	.01901611
Error sum of squares after fitting X and XL	.01923830	.01651628
Difference in Error Sums of Squares	.00508045	.00259983
Error mean square	.00073993	.00063140
Coefficient of X (b_1)	-.00605949	-.00306589
Coefficient of XL	.00199875	.00036884
Sum of coefficients (b_2)	-.00406074	-.00269705

*Entries in this table appear as output of the GLM procedure in SAS [4], namely:

```
PROC GLM; MODEL Q5 Q8 = XF ;
PROC GLM; MODEL Q5 Q8 = XD XL ;
```

applied to the data of Table 1.

References

[1] Kish, Leslie (1965). Survey Sampling, Wiley, N. Y.

[2] Smith, Patricia L. (May 1979). Splines as a useful and convenient statistical tool, The American Statistician, vol. 33, pp.57-62.

[3] Cates, Billy Reeves (1979). "Perceptions of Administrators, Teachers, and Students Concerning the Appropriateness of the Objectives of the North Carolina Middle Grades Occupational Exploration Program," Ph.D. Thesis, NCSU, Raleigh, NC.

[4] SAS User's Guide, 1979 Edition, SAS Institute Inc., Raleigh, NC.

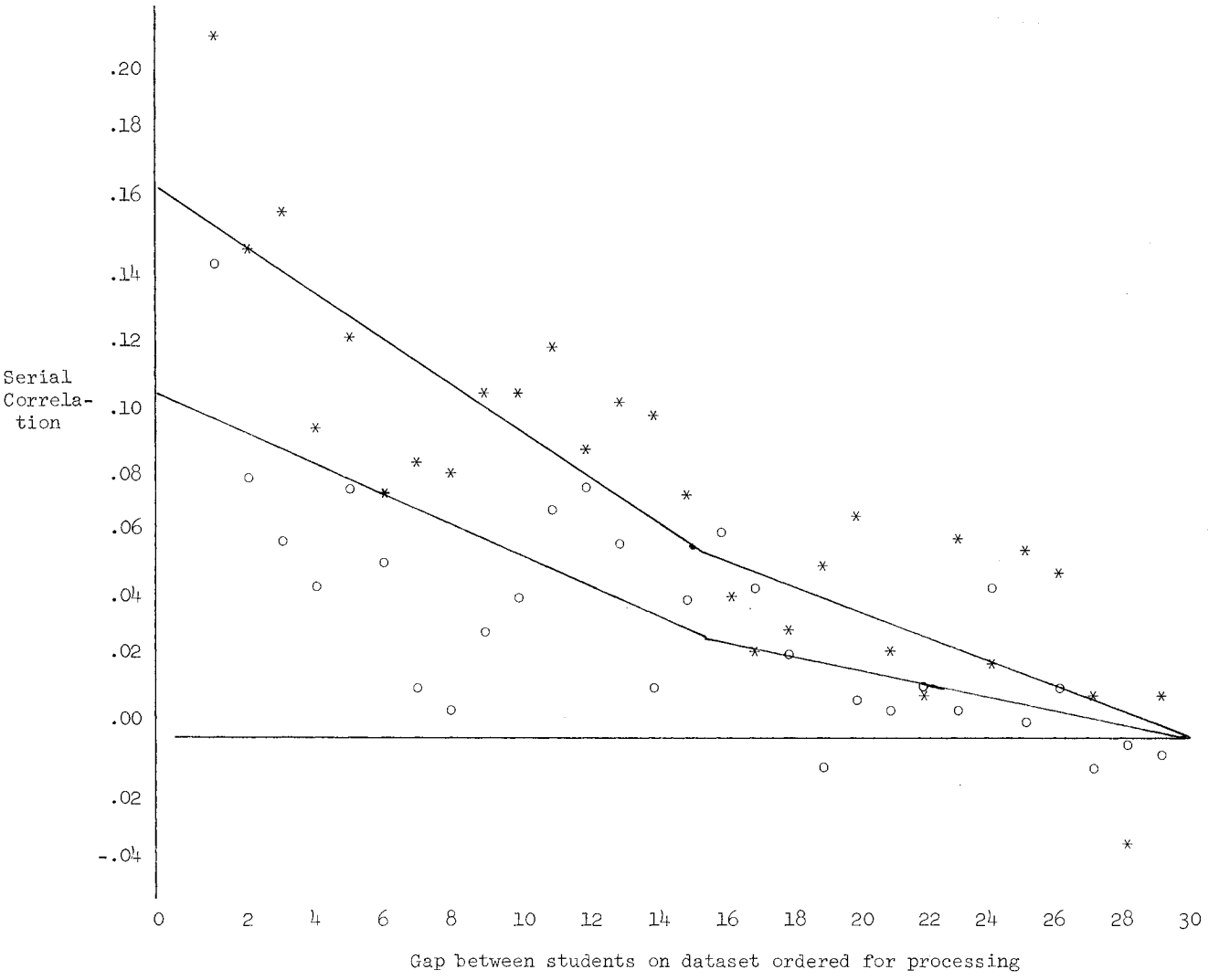

Figure 1. Correlograms for Two Questionnaire Items, Q5 (*'s) and Q8 (o's) based on data in Table 1 with Eye Fitted Lines.