ANALYZING PAIR DATA AND POINT DATA ON SOCIAL RELATIONSHIPS, ATTITUDES AND BACKGROUND CHARACTERISTICS OF COSTA RICAN CENSUS BUREAU EMPLOYEES

C. H. Proctor, North Carolina State University

Introduction

Linear statistical models have been widely used in sociology and even their more refined versions are being properly applied and wisely interpreted.[1] So perhaps it is a fitting time to create a little confusion with the expectation that the problems eventually will be ironed out. The source of the present difficulty is sociometric or interpersonal data. In particular, it is often found possible to characterize, with a variable value, the state of the relationship between two persons. The problem is how to investigate the association between such pair variables and the more customary point data on the same persons.

The issue has been quite a bit discussed under the title of "ecological correlations" for the case that groups of persons or households are analyzed as well as the persons or households. The notion of "nested models" in analysis of variance parlance with components of variance or intraclass correlation formulation can aid in interpreting such data but this viewpoint seems rather specialized. It views the aggregate as a sum of group components and individual components, and the variables measured on the individual parts are the same as those measured on the aggregate. In the setting we wish to work, the variables are different.

To make the problem hopefully easier to discuss we will take a concrete example. The data of this example were collected by self-administered questionnaires from the employees of the Costa Rican Census Office in 1952. There are three classes of variables that will be analyzed. These are:

(1) Three attitude-toward-work items,

(2) Interpersonal structure measured by two sets of sociometric choice data using the criteria: want-to-work-with and have-coffee-with, and

(3) Seven background variables (five sections of the office as represented by four indicator variables, schooling, age and sex).

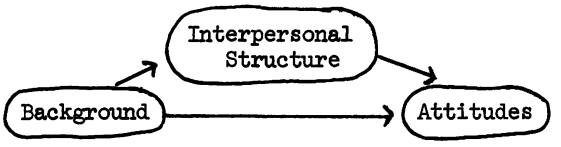
The interpersonal or sociometric variables were scored as mutual mention, one-way, or indifference for every pair of persons. From the 63 persons responding, there are 1,953 unordered pairs and 3,906 ordered pairs that can be formed. The other attitude and background variables were scored for each person and can be called "point" variables to contrast with the two sociometric or "pair" variables.

Preliminary Calculations

The initial approach to the data was pre-statistical, in that there was no thought of a probability mechanism, no parameters and thus no estimation. A regression analysis was done using the 1,953 pairs as the units of analysis. The three interpersonal structure pair variables were called: strength, direction, and unbalance and will be denoted z_1, z_2, and z_3, respectively. Their values were as follows: $z_1 = 0$ for indifference, $z_1 = 1$ for one-way and $z_1 = 2$ for mutual mention; $z_2 = +1$ when the first person mentions the first, $z_2 = -1$ for the reverse and $z_2 = 0$ if indifference or mutual; while $z_3 = -1$ if one-way and zero otherwise.

The point variables were used to form signed differences (first person's score minus second person's) and also to form absolute differences. These transformed variables then become pair variables derived from point variables. Thus there were 6 attitude plus 3 interpersonal plus 14 background equals 23 pair variables.

The causal mechanism that was posited had background variables giving rise to the attitude scores and also to the interpersonal structure with the structural variables further affecting the attitude scores. The path diagram is as follows:



Thus, two sets of regression equations for the pair variables were examined:

(1) Both attitude and structure on background;

(2) Attitude on structure with both adjusted for background.

When the coefficients were examined and when interpretations were attempted the results made good sense but there was also a good bit of nonsense introduced by the naive approach. Several points emerged:

(a) The signed differences and direction (z_2) variables are both asymmetric, in that they change

sign when the order of the pair is reversed, but the absolute differences along with strength (z_1) and unbalance (z_3) are symmetric. That is, they do not depend on the order of the pair. These two kinds of variables ought not be mixed in the same regression analysis.

(b) Although it was interesting to see how background affects structure, it does not seem advisable to "adjust" structure for background when examining how structure affects attitudes. This is a substantive question of outlook to the causal mechanism.

(c) The F-ratios and regression coefficient t-test statistics under the independent observation model between the signed differences (attitudes on background) were stupendously large and do require to be corrected.[2] The regression coefficients themselves are, however, numerically equal to those of the point variable analysis.

(d) The regression of structure on background may be calculated and interpreted under the usual model, namely fixed independent variable with independent homogeneous error, but counting 1,953 observations.

(e) Using absolute differences makes the model equation assumptions problematic and this will be treated below.

The computations were then redone by first regressing the attitude point variates on the background point variables and then calculating residual attitude point scores. These residuals were used to compute absolute difference pair data which were then regressed on strength and direction, while the signed difference of residuals pair data was regressed on direction. The regression coefficients of these three analyses appear in Table 1.

The regression coefficients, and in this instance even the accompanying t-values, that were produced by the computer can be used to screen relationships. Under this criterion attitude differences over whether coworkers should be friends showed dependence on both the coffee and the work interpersonal structures. When the strength of the relationship increased, the attitude differences decreased. The distribution of absolute differences themselves are shown by coffee relationship strength class in Figure 1. Also shown there are the log (Absolute difference + .1) transforms. This log transform was suggested by an argument to be presented shortly and appears to have stabilized the within-class distributions.

The analysis of variance of the log transformed absolute differences shows that $\bar{y}_0 = -.81$, $s_0^2 = .53$ for $n_0 = 1917$ pairs at zero strength; $\bar{y}_1 = -.89$, $s_1^2 = 78$ for $n_1 = 19$ pairs (one-way coffee choices); and $\bar{y}_2 = -1.50$, $s_2^2 = .62$ for $n_2 = 17$ pairs of strength 2 (mutual coffee choices). The pooled estimate of error variance was $s_p^2 = .5316$. Thus the differences themselves gradually decreased with an increase in the strength of the relationship. The distributions are reasonably well behaved as Figure 1 shows and the variance homogeneity is encouraging. Incidentially, the analysis of variance for the absolute differences led to F = 3.91 while for the log transformed differences F = 7.69 for the same case of three levels of strength. This difference in F values could be due to the large population fourth moment of the untransformed data, and an unexpectedly large overestimate of the error variance.

Statistical analysis

To interpret these means and variances some stochastic model needs to be suggested. There seemed to be two main approaches to explaining the data. Either census bureau employees begin by taking a variety of stands on the attitude question and then move together if their relationship is close or they begin by being initially identical and move apart if their relationship is not close. The model we will use supposes a process of random separation or drifting apart of two people that is modified by a gradual passage to uniformity or agreement if a relationship exists.

Although the following derivation is historically faithful to its conception and is given to bolster its use, the model should not be judged exclusively by this line of reasoning. Let $d_{ij}(t)$ be the difference in attitude scores between two persons i and j at time t. This has a sign. $d_{ij}^2(t)$ is a more appropriate measure of distance as it has no sign, and will be used in the following. When dealing with quantities that have no sign the subscripts i,j will be taken with $i < j$ (only the upper triangle of the matrix is used). Now $d_{ij}^2(t+1)$ should be related to $d_{ij}^2(t)$ in some way. If this relationship is represented multiplicatively as:

$$d_{ij}^2(t+1) = \beta d_{ij}^2(t)\epsilon_{ij}(t+1)$$

$$= \left(e^{2\alpha X_{ij}}\right) d_{ij}^2(t) \left[e^{2\delta_{ij}(t+1)}\right]$$

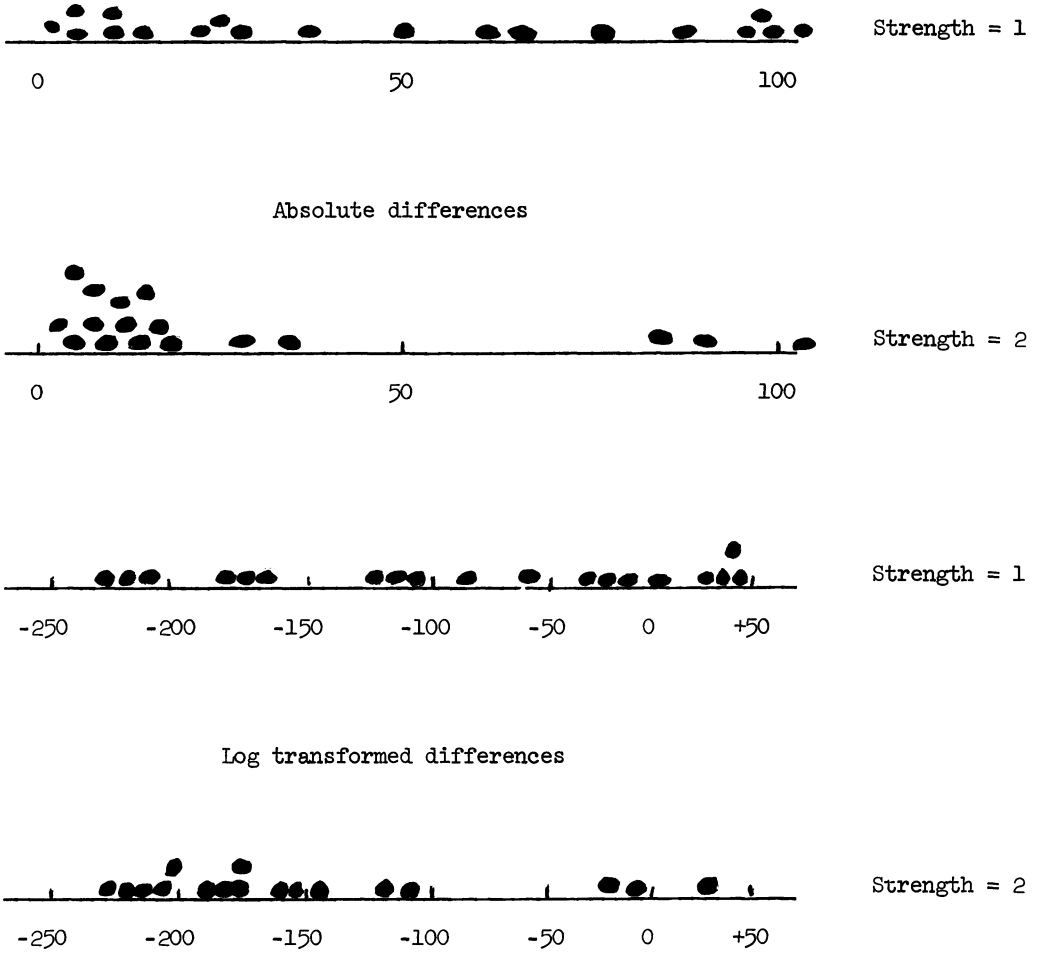
where X_{ij} is the strength of the relationship between i and j and $e^{2\alpha X_{ij}}$ is the systematic multiplier while $e^{2\delta_{ij}(t+1)}$ is the random

Table 1. Regression coefficients for pair analyses, boxed coefficients exceed 2 standard errors.

Regression Run	Independent Variable	Dependent Variables are indices of extent of work-rational orientation to: Salary Allocation	Explaining Mistakes	Friendship among Co-workers
I. Point Regression:	Jefes	-.076	-.120	.451
	Secretaries	.328	.117	.219
	IBM	.323	-.339	.170
	Coding	.251	.379	-.052
	Age	-.071	.106	.007
	Sex	-.109	-.023	-.195
	School	.038	-.079	-.033
II. Pair Regression:				
Work Pairs	Strength	-.043	.152*	[-.105]
	Unbalance	.069	-.199	-.015
Coffee Pairs	Strength	-.073	.046	[-.098]
	Unbalance	-.078	-.164	.131
III. Pair Regression:				
Work Pairs	Direction	.034	.010	-.118
Coffee Pairs	Direction	-.137	.101	.113
Separate Pair Regressions:				
Work	Strength	-.060	.164*	[-.129]
Coffee	Strength	-.079	.070	[-.121]

*These coefficients were close to twice their standard errors.

Fig. 1. Distributions of absolute differences and of log transformed differences by strength of coffee pair relationship.



one. Taking logs and cancelling 2's leads to:

$$\log|\, d_{ij}(t+1)\,| = \alpha X_{ij} + \log|\, d_{ij}(t)\,| + \delta_{ij}(t+1),$$

or by rewriting the terms in an obvious way ($\log|\, d_{ij}\,| = D_{ij}$):

$$D_{ij}(t+1) = \alpha X_{ij} + D_{ij}(t) + \delta_{ij}(t+1).$$

In order to further normalize the distribution of the δ_{ij}'s the values $y_{ij} = \log(|d_{ij}| + .1)$ were analyzed but the discussion will proceed with D_{ij}. Moving toward a slight bit more of generality we can dispense with α as the single linear regression coefficient and use an effects model as:

$$D_{ij}(t+1) = \alpha(X_{ij}) + D_{ij}(t) + \delta_{ij}(t+1)$$

in which α is a function; namely:

$$\alpha(0) = 0$$
$$\alpha(1) = \alpha_1$$
$$\alpha(2) = \alpha_2.$$

If it be supposed that $V(\delta_{ij}(t)) = \sigma^2_{ij}$ for all t and the δ_{ij}'s are independent from one time to the next then we can follow the course of $D_{ij}(t)$ through time as a function of X_{ij} and deduce some convenient distributional properties. If $X_{ij} = 0$ and we take to begin the process $D_{ij}(0) = D_0$ say, then

$$D_{ij}(1) = 0 + D_{ij}(0) + \delta_{ij}(1) = D_0 + \delta_{ij}(1)$$
$$D_{ij}(2) = 0 + D_{ij}(1) + \delta_{ij}(2) = D_0 + \delta_{ij}(1) + \delta_{ij}(2)$$
$$\vdots$$
$$D_{ij}(t) = D_0 + \sum_{k-1}^{t} \delta_{ij}(k).$$

However if $X_{ij} = 1$ or 2 then (taking $X_{ij} = 1$ for example)

$$D_{ij}(1) = \alpha_1 + D_0 + \delta_{ij}(1)$$
$$D_{ij}(2) = \alpha_1 + \alpha_1 + D_0 + \delta_{ij}(1) + \delta_{ij}(2)$$
$$\vdots$$
$$D_{ij}(t) = t\alpha_1 + d_0 + \sum_{k=1}^{t} \delta_{ij}(k).$$

Similarily:

$$D_{ij}(t) = t\alpha_2 + D_0 + \sum_{k=1}^{t} \delta_{ij}(k), \text{ for } X_{ij} = 2.$$

For data that are from a one-point-in-time survey, such as are those of the present example, one cannot know t. It could be supposed that attitude reassessment occurs very frequently so that t (the number of reassessments) is very large and then we would suppose that α_1 and α_2 are very small but as t increases $t\alpha_1 \to \lambda_2$ and the variances of each of the $\delta_{ij}(t)$ also are very small so that again as t increases $V(\sum_{k=1}^{t} \delta_{ij}(k)) \to \sigma^2_\delta$.

the model equations then become:

$$D_{ij}(\text{survey time}) = D_0 + \delta_{ij}, \quad \text{if } X_i = 0$$

$$(1) \quad D_{ij}(\text{survey time}) = D_0 + \lambda_1 + \delta_{ij}, \quad \text{if } X_{ij} = 1$$

$$D_{ij}(\text{survey time}) = D_0 + \lambda_2 + \delta_{ij}, \quad \text{if } X_{ij} = 2$$

with $V(\delta_{ij}) = \sigma^2_\delta$.

These equations have the flavor of an ANOVA model. Before using the F-tests suggested by this formulation, a problem would seem still to be the dependence among the δ_{ij}. Here the patterns are numerous but still finite. The major departure could be expected as a correlation between a certain δ_{ij} and the associated ones of the forms $\delta_{ij'}$ or $\delta_{i'j}$ where the primed subscript is unequal to the corresponding unprimed one.

Since there are 1953 pairs there become 1,906,128 pairs of pairs. Of these, 119,133 or about 6.3% have a subscript in common. Table 2 shows twelve types of pairs of pairs (the upper number is of mutually exclusive pairs and the lower gives the overlapped pairs) as counted from the coffee structure.

If such seems reasonable, it may be supposed that the δ_{ij}'s of overlapped pairs of pairs were correlated while the mutually exclusive pairs were not. To investigate such a possibility, a sample of 10 pairs from Table 2 was drawn at random from pairs with strengths 1 and 2 (See Table 3). The 45 pairs of pairs were found to include 39 mutually exclusive and 6 overlapped. The ten residuals ($\overline{y_{ij}}$ minus fitted y_{ij}) were recorded and from them the differences were computed and then the variances of differences of the two types were calculated separately. These data are given in Table 4, along with the original data for strengths 1 and 2 in Table 3. This gave:

Table 2. Counts of Non-overlapping and Overlapping Pairs of Pairs by Combination of Strengths of the Pair (number before plus sign is of non-overlapping pairs of pairs)

Strength of One Pair	Strength of Other Pair			No. of Pairs
	0	1	2	
0	1,723,122 +113,364			$n_0 = 1917$
1	32,788 +3,635	163 +8		$n_1 = 19$
2	30,428 +2,161	306 +17	118 +18	$n_2 = 17$

$s^2_{ex.} = .8373$, and $s^2_{lap.} = .0815$, respectively. If δ_{ij} and $\delta_{i'j'}$ are mutually exclusive it is supposed that $V(\delta_{ij} - \delta_{i'j'}) = 2\sigma^2_\delta$ but if not, then $V(\delta_{ij} - \delta_{i'j'}) = V(\delta_{ij} - \delta_{ij'}) = 2\sigma^2_\delta(1-\rho)$. One could, therefore use $.0815/.8373 = .0973$ as an estimate of $1 - \rho$. Thus $\hat{\rho} = .90$, quite a sizeable correlation and worthy of special attention.

To obtain unbiased estimates of the parameters D_0, λ_1 and λ_2 is straight forward using the mean values from the three classes (recalling $\bar{y}_0 = -.81$, $\bar{y}_2 = -.89$ and $\bar{y}_2 = -1.50$). The results were: $\hat{D}_0 = -.81$, $\hat{\lambda}_1 = -.89 + .81 = -.08$, $\hat{\lambda}_2 = -1.50 + .81 = -.69$.

The crucial parameter in representing the effect of social structure on attitudes would seem to be λ_2 whose estimate is $\hat{\lambda}_2 = -.69$. To test if $\lambda_2 = 0$ is tantamount to deciding if there is an effect. The suggested test procedure is as follows.

It can be fairly judged that $\hat{\lambda}_2$ will be covered by the central limit theorem. To estimate its variance one must take into account the covariances of overlapped pairs of pairs. These arise in three places in each of $V(\bar{y}_0)$ and $V(\bar{y}_2)$ as well as in $\text{Cov}(\bar{y}_0, \bar{y}_2)$. The following computation is based on Table 2.

$$V(\bar{y}_0 - \bar{y}_2) = V(\bar{y}_0) + V(\bar{y}_2) - 2\,\text{Cov}(\bar{y}_0, \bar{y}_2)$$

$$= \left[\left(\frac{1}{1917} + \frac{2(.9)113,364}{1917^2}\right) + \left(\frac{1}{17} + \frac{2(.9)18}{17^2}\right) - \frac{2(.9)2161}{1917 \cdot 17}\right][.532]$$

$$= \left[\left(\frac{1}{1917} + \frac{1}{17}\right) + 2(.9)\left(\frac{113,364}{1917^2} + \frac{18}{17^2} - \frac{2161}{1917 \cdot 17}\right)\right][.532]$$

$$= [.10762344][.532] = .05725567$$

As an example, consider the expression for $V(\bar{y}_0)$. The estimate $\bar{y}_0$ itself is the mean of 1917 quantities, but not independent quantities. From Table 2 one finds that 113,364 pairs are correlated.

Table 3. Values of $y_{ij} = \log (|d_{ij}| + .1)$ for Strengths 1 and 2.

Strength 1		Strength 2	
(i,j)	y_{ij}	(j,j)	y_{ij}
3,6	-2.18	4,27	-2.18
5,9	-2.01	7,16	-1.32*
5,54	- .01	10,40	-2.24
8,34	+ .09*	10,50	-2.18*
10,27	-2.07	11,13	-1.69*
15,26	- .94	11,14	-2.02*
16,43	-1.71	11,54	-1.49*
18,25	+ .10*	13,14	-1.88*
20,41	-1.75	13,54	-1.97
23,56	-1.15*	14,54	-1.64
28,43	-1.30	15,18	+ .19
29,61	- .19	15,56	-1.87
30,55	- .57	30,42	-1.83
35,41	- .53	30,58	- .11
37,43	+ .13	32,51	-1.12
42,55	- .29	40,50	-2.12*
46,52	-1.79	42,58	- .05
47,62	+ .20		
55,58	-1.19		

$n_1 = 19$ $\quad$ $n_2 = 17$

$\bar{y}_1 = -.8096$ $\quad$ $\bar{y}_2 = -.8926$

*Pair sampled to estimate ρ.

Table 4. Differences (in absolute values) of residuals from Table 3 for overlapping pairs of pairs and for mutually exclusive pairs of pairs.

Mutually Exclusive					Overlapping
.01	1.60	1.61	.86	.69	.06
1.24	1.25	.44	.37	.30	.33
.80	.81	.42	.70	.43	.20
1.66	1.67	.07	.17	.10	.19
1.17	1.18	.26	.56	.39	.53
1.50	1.51	.27	.80	.63	.14
.97	.98	.12	.49	.24	
1.36	1.37	.36	.16		

Sum of Sqs. = 32.6538 SS = .4891

Average squared difference = .8373 Av. Sq. Diff. = .08152

****

Thus

$$V(\bar{y}_0) = V\left[\frac{1}{1917}(y_1 + y_2 + \cdots + y_{1917})\right]$$

$$= \left(\frac{1}{1917}\right)^2 \left[\sum_i V(y_i) + 2 \sum_{i<j} \sum Cov(y_i, y_j)\right]$$

$$= \left(\frac{1}{1917}\right)^2 \quad [1917\ \sigma_\delta^2 + 2(113,364)\rho\ \ \sigma_\delta^2]$$

The standard error of $\hat{\lambda}_2$ then is S.E. $(\hat{\lambda}_2)$ = .2393 and a test of $\lambda_1 = 0$ may be based on the critical ratio of .69/.2393 = 2.88. There is evidence that $\lambda_1 \neq 0$ from the fact that the probability of a standard normal deviate exceeding 2.88 is about .002.

The previous analysis (suggested estimates and tests) appeals to an analogy with the analysis of variance. The model equations are the same except for the dependence among the δ_{ij}'s. When that dependence is characterized by a non-zero covariance ($=\rho\sigma_\delta^2$ say) between δ_{ij} and δ_{kh} when the sets {i,j} and {k,h} overlap, but zero otherwise then some interesting problems arise concerning the covariance matrix of the δ_{ij}'s. Let the single (generic subscripts r and s replace the pair subscripts (i,j) and (k,h) in which the range of r and s is 1, 2, ..., n(n-1)/2(= m say), and let the ordering on r and s be such that r < s if i < k or when i = k if j < h (recall that i < j and k < h by convention). Now the m by m covariance matrix of the $\delta_{ij} = \delta_r$'s will be denoted σ_δ^2 V where V contains 1's along the main diagonal and a scattering of ρ's and 0's off the main diagonal.

The model equations (1) can be rewritten in matrix notation as:

(2) $\underset{\sim}{y} = X \underset{\sim}{\theta} + \underset{\sim}{d}$ with $E(\underset{\sim}{d}) = \underset{\sim}{0}$ and $E(\underset{\sim}{d}\ \underset{\sim}{d}') = \sigma_d^2 V$.

Here y and d are m x 1 and X is m x p where p is the number of parameters.

In the particular example given above $\underset{\sim}{\theta}' = (D_0, \lambda_1, \lambda_2)$ and the 1,953 by 3 incidence matrix X consisted of 0's and 1's. The estimate of $\underset{\sim}{\theta}$ given by least squares theory is:

$$\hat{\theta} = (X'V^{-1}X)^{-1} X' V^{-1} Y$$

The naive estimators suggested above are these with V = I.

For the present data such estimates (with $V \neq I$) were not calculated, and this for two reasons. The size of V is such as to make its inversion a doubtful computing operation and also the value of ρ was not known. Consequently, it is not known how closely the naive estimates $\hat{\underset{\sim}{\theta}} = (-.81, -.08, -.69)$ are to the least squares ones or, more to the point, how the variances of the naive estimates compare to those of the least squares ones.

Although the naive procedures may soon become replaced by the least squares or even better ones, let us try to cast them in a slightly more general notation. The naive

estimates are $\underset{\sim}{\bar{\theta}} = (X'X)^{-1} X'\underset{\sim}{y}$ as mentioned above. The p by p covariance matrix of $\bar{\theta}$ is given by:

$$E(\underset{\sim}{\bar{\theta}}\,\underset{\sim}{\bar{\theta}}') = \sigma_d^2 (X'X)^{-1} X' V X(X'X)^{-1} = \underset{\sim}{\Sigma}_{\bar{\theta}} \text{ say.}$$

For the usual ANOVA arrangements of X the entries of this matrix can be calculated using counts of overlapped pairs as in Table 1 above. The efficiency of the method of sub-sampling the pairs as described above and the calculation form for estimating ρ is not known. It was dictated by my limited computing resources. For example, the estimate of ρ cannot be expected to be unbiased even though the estimates of σ_δ^2 and $(1 - \rho)\,\sigma_\delta^2$ are. The estimates for these based on $s^2_{ex.}$ and $s^2_{lap.}$ are themselves not unbiased due to some slight negative covariances among the residuals.[4]

FOOTNOTES

1. This comment was strongly prompted by seeing the paper by W.H. Sewell and V.P. Shah, "Parent's education and children's educational aspirations and achievement," Am. Soc. Rev., 33:191-209, April 1968.

2. The factor involved appears to be the multiplier $(n-p-1)/[\binom{n}{2} - p - 1]$ which is to be applied to the F-ratios and its square root to the t-values of the paired variables analysis to obtain those of the single variable analysis. Here p is the number of independent variables in the regression.

3. Kendall, M.G. and Stuart, Alan. The Advanced Theory of Statistics, Vol. 2, New York, Hafner, 1961, p. 87.

4. Anscombe, F.J., "Examination of Residuals," in Proceedings of the Fourth Berkeley Symposium, University of California Press, 1961.