NBER WORKING PAPER SERIES

THE GOODNESS OF MATCH

Edward N. Wolff*

Working Paper No. 72

NATIONAL BUREAU OF ECONOMIC RESEARCH, INC.
261 Madison Avenue
New York, N.Y. 10016

December, 1974

Preliminary; Not for Quotation

*New York University and NBER

The Goodness of Match

I. Introduction

Statistical matching has received increasing popularity in the last five years as a method of creating synthetic microdata sets. Benjamin Okner merged the 1966 Internal Revenue Service (IRS) Tax File with the 1967 Survey of Economic Opportunity (SEO) by first dividing the two files into broad "equivalence classes" and employing a distance function to choose the best record to match (Okner, 1972). Edward Budd and Daniel Radner matched the Current Population Survey File with the IRS Tax File by ranking records in each file by income level and linking similarly ranked observations (Budd, 1971). Richard Rockwell combined the 1970 Public Use Sample (PUS) with the SEO file by dividing each sample into equivalence classes on the basis of five common variables. Horst Alter linked the 1970 Canadian Survey of Consumer Finances with the 1970 Family Expenditure Survey by using multiple regression analysis to minimize the "distance" between matched observations (Alter, 1974). Currently, Nancy and Richard Ruggles are undertaking a match of the 1970 PUS with the 1960 PUS, the 1969 IRS Tax File, and the Social Security's Longitudinal Employer-Employee Data File by creating matching classes on the basis of interval analysis ar.d combining records on a stochastic basis within these intervals (Ruggles, 1974).

Though the statistical techniques vary, the matching problem is essentially the same in each case and can be stated formally, as Christopher Sims does, as follows: Given "observations on X,Y from one sample and on X,Z from another sample, when will it be true that by matching observations according to X, an artificial Y,Z sample will result whose distribution is the true joint Y,Z distribution?" (Sims, 1972, p. 355). Though the imputed Y,Z distribution will, in gerneral, be different from the true Y,Z distribution,[1] the closeness of the two yields a natural criterion of the goodness of match.

By making certain simplifying assumptions, we can make this criterion operational. First, it can be assumed that the closeness of the correlation coefficient between Y and its imputed Z value to the true correlation coefficient between Y and Z reflects the closeness of the two joint distributions.[2] Though the true correlation between Y and Z is, in general, unknown, we can determine a lower and upper bound on the true correlation as a function of the matching variables and use the range as a measure of the goodness of match. In order to do this, we must, secondly, assume that for each observation in the first file there exists an observation in the second file with exactly the same X values, and conversely.[3]

[1] The two distributions will be exactly the same in the special case, which Sims mentions, when X,Y and Z are mutually independent, and in the special case when Y is a linear transformation of X and Z a linear transformation of X.

[2] The covariance between Y and Z, it should be recognized, is only one moment of their joint distribution. Moreover, it is implicitly assumed that Y, Z, and X are continuous variables. Furthermore, we shall ignore the problem of sample estimation of the correlation coefficients and the discrepancies between sample estimates and population values. In a sense, we shall treat the sample as the full population.

[3] In practice, the major problem arising from matching two files is that there rarely is an observation in the second file with the same X values as a given observation in the first file, and conversely.

The goodness of match depends on how much of the relation between Y and Z is transmitted through X--that is, on how X "mediates" between Y and Z-- and we will therefore call X the mediating variables and Y and Z the mediated variables.[4] Since the functional form the lower and upper bounds on the true correlation between Y and Z takes depends on the number of X variables, we shall treat the problem in three stages: (a) The case of one mediating variable. (b) The case of two mediating variables. (c) The case of n mediating variables.

[4] In principle there may be more than one Y or Z variable. Without loss of generality we can assume that there is only one of each, since the correlation of each pair Y_i, Z_j can be treated independently of the other pairs.

II. The Case of One Mediating Variable

Let X, Y, and Z be random variables with zero mean and unit variance.[5] Let

$$\begin{aligned} \rho &= \mathrm{cor}(Y,Z) \\ q &= \mathrm{cor}(X,Y) \\ r &= \mathrm{cor}(X,Z) \end{aligned}$$

Then,

$$\rho = \frac{E(YZ)-E(Y)E(Z)}{\sigma_Y \sigma_Z} = E(YZ)$$

Likewise,

$$\begin{aligned} q &= E(XY) \\ r &= E(XZ) \end{aligned}$$

Moreover,

$$\begin{aligned} E(X(Y-qX)) &= E(XY) - qE(X^2) \\ &= q - q \\ &= 0 \end{aligned}$$

Likewise,

$$E(X(Z-rX)) = 0$$

And:

$$\begin{aligned} E(Y-qX)^2 &= E(Y^2) - 2qE(XY) + q^2E(X^2) \\ &= 1 - q^2 \end{aligned}$$

$$E(Z-rX)^2 = 1 - r^2$$

[5] This is not a restrictive assumption. Suppose Y' has mean $\mu_{Y'}$ and variance $\sigma^2_{Y'}$ and Z' has mean $\mu_{Z'}$ and variance $\sigma^2_{Z'}$. Then $Y = \frac{Y'-\mu_{Y'}}{\sigma_{Y'}}$ and $Z = \frac{Z'-\mu_{Z'}}{\sigma_{Z'}}$, and each has zero mean and unit variance, and $\mathrm{cor}(Y',Z') = \frac{E(Y'-\mu_{Y'})(Z'-\mu_{Z'})}{\sigma_{Y'}\sigma_{Z'}} = E(YZ) = \mathrm{cor}\,(Y,Z)$

<u>Theorem 1</u>: $qr + \sqrt{(1-q^2)(1-r^2)} \geq \rho \geq qr - \sqrt{(1-q^2)(1-r^2)}$

Proof: Noting that $Y = qX + (Y-qX)$ and $Z = rX + (Z-rX)$,

$$\rho = E(YZ) = E[qX+(Y-qX)][rX+(Z-rX)]$$

$$\rho = E[rqX^2+rX(Y-qX) + qX(Z-rX) + (Y-qX)(Z-rX)]$$

$$\rho = rq +[E(Y-qX)(Z-rX)]$$

$$(\rho-rq)^2 = [E(Y-qX)(Z-rX)]^2$$

From Schwartz' inequality,

$$(\rho-rq)^2 \leq E(Y-qX)^2E(Z-rX)^2$$

Therefore,

$$(\rho-rq)^2 \leq (1-q^2)(1-r^2) \qquad \text{Q.E.D.}$$

Lower and upper bounds are shown for selected values of q and r in Table 1. The lower bound is symmetrical in q and r. Denoting the lower bound by L_1 ,

$$\frac{\partial L_1}{\partial q} = r + q\sqrt{\frac{1-r^2}{1-q^2}}$$

Therefore, when q and r have the same sign, the lower bound increases as either $|q|$ or $|r|$ increases. When q and r are non-negative, the lower bound is less than or equal to q and r, and equals q when r equals 1, and conversely. The upper bound behaves obversely to the lower bound. Denoting the upper bound by U_1 ,

$$\frac{\partial U_1}{\partial q} = r - q\sqrt{\frac{1-r^2}{1-q^2}}$$

Table 1

Lower and Upper Bounds for the Case of One Mediating Variable

A. Lower Bounds:

q \ r	.60	.70	.80	.90	.95	1.00
.60	-.280	-.151	.000	.191	.320	.600
.70	-.151	-.020	.132	.319	.442	.700
.80	.000	.132	.280	.458	.573	.800
.90	.191	.319	.458	.620	.719	.900
.95	.320	.442	.573	.719	.805	.950
1.00	.600	.700	.800	.900	.950	1.000

B. Upper Bounds:

q \ r	.60	.70	.80	.90	.95	1.00
.60	1.000	.991	.960	.889	.820	.600
.70	.991	1.000	.988	.941	.888	.700
.80	.960	.988	1.000	.982	.947	.800
.90	.889	.941	.982	1.000	.991	.900
.95	.820	.888	.947	.991	1.000	.950
1.00	.600	.700	.800	.900	.950	1.000

When q and r have different signs, the upper bound decreases as either $|q|$ or $|r|$ increases. When q equals 1, the upper bound is the same as the lower bound, r, and conversely when r equals 1. When q and r have the same sign, the upper bound reaches its maximum value of 1 when q equals r,[6] and decreases as q deviates from r, given r. Conversely, when q and r have different signs, the lower bound reaches its minimum when q equals -r, and increases as $|q|$ deviates from $|r|$, given r.

The upper and lower bounds give the range of values the true correlation coefficient between Y and Z may have. The size of the range is the difference between the two bounds and equals $2\sqrt{(1-q^2)(1-r^2)}$. Thus, the larger q and r, the smaller the range and the greater the certainty that the imputed correlation between Y and Z is close to the true one.

In the special case where $q = r = p$, the bounds take the following form:

<u>Lemma 1</u>: $1 \geq \rho \geq 2p^2 - 1$

When $|p|$ equals 1, ρ equals 1.[7] Since the lower bound is quadratic in p, the range quickly widens as $|p|$ deviates from 1 (See Figure 1).

[6] This can be shown by setting $\partial V_1/\partial q$ to zero and noting that the second derivative is negative.

[7] In this case, both Y and Z are linear functions of X, say $Y = aX+b$ and $Z = cX + d$. Therefore,

$$Z = (\frac{c}{a})Y + (d - \frac{cb}{a})$$

and Z is a linear function of Y.

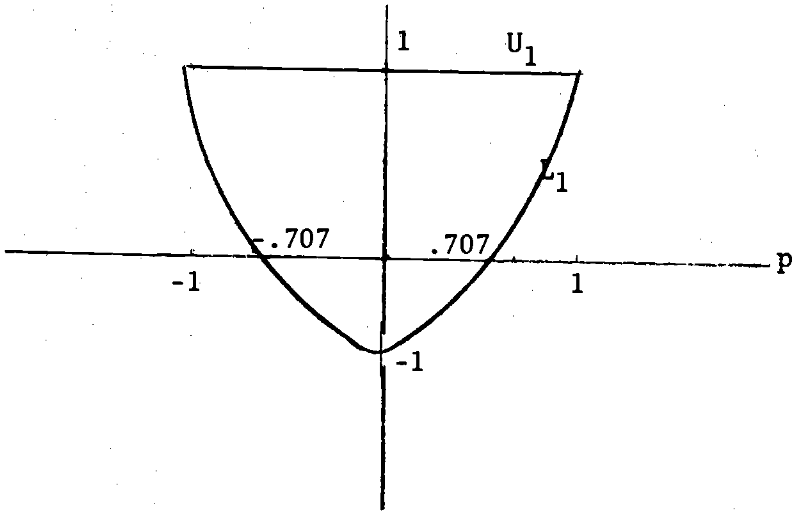

Figure 1: Lower and Upper Bounds when q = r

At p = .9, L_1 = .62; at p = .8, L_1 = .28; and at p = .7, L_1 = -.02. When |p| falls below .707, it cannot be ascertained whether ρ is positive or negative.

III. The Case of Two Mediating Variables

The same technique can be applied in the case of two intervening variables X_1 and X_2 as in the case of one to determine the upper and lower bounds on ρ. Let

$$
\begin{aligned}
q_1 &= \mathrm{cor}(X_1,Y)\\
q_2 &= \mathrm{cor}(X_2,Y)\\
r_1 &= \mathrm{cor}(X_1,Z)\\
r_2 &= \mathrm{cor}(X_2,Z)\\
s &= \mathrm{cor}(X_1,X_2)
\end{aligned}
$$

Then:

$$
\begin{aligned}
4\rho = E[q_1X_1 + (Y-q_1X_1) + q_2X_2 + (Y-q_2X_2)][r_1X_1 + (Z-r_1X_1)\\
+ r_2X_2 + (Z-r_2X_2)]
\end{aligned}
$$

$$
\begin{aligned}
4\rho - (3q_1r_1 + 3q_2r_2 - s(q_1r_2 + q_2r_1)) =\\
E(Y-q_1X_1)(Z-r_1X_1) + E(Y-q_1X_1)(Z-r_2X_2)\\
+ E(Y-q_2X_2)(Z-r_1X_1) + E(Y - q_2X_2)(Z - r_2X_2)
\end{aligned}
$$

Squaring both sides, collecting terms, using Schwarz[1] inequality, completing the square, and transposing terms yields:

$$
\begin{aligned}
[\rho + \tfrac{1}{4}(s(q_1r_2 + q_2r_1) - 3(q_1r_1 + q_2r_2)]^2 \leq\\
(1-q_1^2)(1-r_1^2) + (1-q_1^2)(1-r_2^2) + (1-q_2^2)(1-r_1^2) + (1-q_2^2)(1-r_2^2)\\
-(3q_1^2r_1^2 + 3q_2^2r_2^2 - 2q_1q_2r_1r_2)s^2/16\\
+(q_1^2r_1r_2 + q_2^2r_1r_2 + r_1^2q_1q_2 + r_2^2q_1q_2)s/8\\
-(3q_1^2r_1^2 + 3q_2^2r_2^2 - 2q_1q_2r_1r_2)/16
\end{aligned}
$$

To simplify for purposes of analysis, consider the case where $q_1=q_2=q$ and $r_1=r_2=r$. Then

Theorem 2: $$\frac{qr(3-s)}{2} + T_2 \geq \rho \geq \frac{qr(3-s)}{2} - T_2$$

$$\text{where } T_2 = \sqrt{(1-q^2)(1-r^2) - [\tfrac{qr}{2}(1-s)]^2}$$

Lower bounds are shown for selected values of q, r and s in Table II.[8] It is evident from the table that the lower bound increases as q and r increase, when q and r are positive. This can be shown formally as follows:

$$\frac{\partial L_2}{\partial q} = \frac{r(3-s)}{2} + \frac{q((1-r^2) + r^2(1-s)^2/4}{T_2}$$

where L_2 is the lower bound in the case of two X variables. Both terms are positive when q and r are positive and

negative when q and r are negative. Since the lower bound is symmetrical in q and r, the lower bound increases as $|q|$ or $|r|$ increases, when q and r have the same signs. When q and r are of opposite signs, the sign of $\partial L_2/\partial q$ and $\partial L_2/\partial r$ depends on the values of q, r, and s. The obverse holds for the upper bound, U_2:

[8] Impermissable combinations of q,r, and s are indicated by an asterisk in Table II. They occur when the term in the radical is negative, and this arises when Lemma 1 is violated. This is a necessary (though not sufficient) condition, as can be shown by proving the converse. Let:

$$p = \max(|q|,|r|)$$

$$(1-q^2)(1-r^2) \geq (1-p^2)^2$$

Since $$q^2 \leq 1 \text{ and } r^2 \leq 1,$$

$$(1-q^2)(1-r^2) \geq (qr)^2(1-p^2)^2$$

From Lemma 1,

$$s \geq 2p^2-1$$

$$p^2 \leq (s+1)/2$$

(Footnote 8 continued)

Table II

Lower Bounds for the Case of Two Mediating Variables

s	-1.00				-.50				0.0				.50				.70				.90			
r	.60	.70	.80	.90	.60	.70	.80	.90	.60	.70	.80	.90	.60	.70	.80	.90	.60	.70	.80	.90	.60	.70	.80	.90
.60	.191	.453	.960	*	.050	.258	.523	*	-.062	.113	.323	.620	-.184	-.037	.135	.353	-.224	-.085	.077	.282	-.262	-.130	.025	.219
.70	.453	.839	*	*	.258	.504	.895	*	.113	.307	.542	*	-.037	.117	.295	.519	-.085	.059	.224	.428	-.130	.005	.160	.352
.80	.960	*	*	*	.523	.895	*	*	.323	.542	.847	*	.135	.295	.478	.710	.077	.224	.389	.590	.025	.160	.313	.497
.90	*	*	*	*	*	*	*	*	.620	*	*	*	.353	.519	.710	*	.282	.428	.590	.785	.219	.352	.497	.665

$$\frac{\partial U_2}{\partial q} = \frac{r(3-s)}{2} - \frac{q((1-r^2) + r^2(1-s^2)/4)}{T_2}$$

If q and r have different signs, the upper bound decreases as $|q|$ or $|r|$ increases. When q and r have the same signs, the direction of movement depends on the values of q, r, and s.

As is also evident from Table II, the lower bound decreases as s increases, when q and r are positive.[9] Formally:

$$\frac{\partial L_2}{\partial s} = \frac{-qr}{2} - \frac{(1-s)q^2r^2}{4T_2}$$

$$\frac{\partial U_2}{\partial s} = \frac{-qr}{2} + \frac{(1-s)q^2r^2}{4T_2}$$

Since $s \leq 1$, $\partial L_2/\partial s$ is always negative when q and r have the same sign, and the lower bound increases as s decreases. When q and r have different signs, the sign of $\partial L_2/\partial s$ depends on the values of q, r and s. The obverse holds for the upper bound.

The upper bound can exceed 1 in certain cases when q and r have the same sign and the lower bound can fall short of -1 in certain cases when q and r have different signs. Therefore, the range $R \leq 2\sqrt{(1-q^2)(1-r^2)-[qr(1-s)]^2/4}$. The upper bound on the range therefore decreases as $|q|$ or $|r|$ increases and as s decreases.

[8] (continued)

$$(1-q^2)(1-r^2) \geq (qr)^2(\frac{1-s}{2})^2 = [\frac{qr}{2} (1-s)]^2$$

Therefore, if Lemma 1 is satisfied, the term in the radical will be non-negative.

[9] In the limiting case, when s equals 1:

$$qr + \sqrt{(1-q^2)(1-r^2)} \geq \rho \geq qr - \sqrt{(1-q^2)(1-r^2)}$$

When X_2 is a linear transformation of X_1 , the bounds degenerate into the limits in the case of one mediating variable.

IV. The Case of n Mediating Variables

Let:

$$\begin{aligned} q_i &= \mathrm{cor}(X_i,Y) \\ r_i &= \mathrm{cor}(X_i,Z) \\ s_{ij} &= \mathrm{cor}(X_i,X_j) \\ S &= \mathrm{cov}(XX') = \begin{bmatrix} 1 & s_{12} & \cdots & s_{1n} \\ s_{21} & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ s_{n1} & \cdots & & 1 \end{bmatrix} \end{aligned}$$

where $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$

$$A_i = Y - q_iX_i$$

$$B_i = Z - r_iX_i$$

To simplify the analysis of the results, we shall assume $q_1 = \ldots = q_n = q$ and $r_1 = \ldots = r_n = r$, though the problem is solvable in its present form. Thus:

$$\begin{aligned} n^2\rho &= E[qX_1+A_1+\ldots+qX_n+A_n][rX_1+B_1+\ldots+rX_n+B_n] \\ &= E[q(X_1+\ldots+X_n) + (A_1+\ldots+A_n)][r(X_1+\ldots+X_n) + (B_1+\ldots+B_n)] \\ &= qr\sum_i\sum_j E(X_iX_j) + q\sum_i\sum_j E(X_iB_j) + r\sum_i\sum_j E(X_iA_j) + \sum_i\sum_j E(A_iB_j) \end{aligned}$$

$$\sum_i\sum_j E(X_iX_j) = C + n, \quad \text{where } C = \sum_{i\neq j}\sum s_{ij}$$

$$E(X_iB_j) = r(1-s_{ij})$$

$$E(X_iA_j) = q(1-s_{ij})$$

Therefore:

$$\rho - (2n^2 - C - n)qr = \sum_i \sum_j E(A_i B_j)$$

Squaring both sides, using Schwarz' inequality, noting that:

$$E(A_i B_j) = \rho - qr \qquad \text{for} \qquad i = j$$

$$E(A_i B_j) = \rho - 2qr + qrs_{ij} \qquad \text{for} \qquad i \neq j$$

and that the Right Hand Side has n^4 terms distributed in the following manner:

number	type	
n^2	$E(A_i B_i)\ E(A_i B_i)$	
$n^2 - n$	$E(A_i B_j)\ E(A_i B_j)$	$i \neq j$
$n^2 - n$	$E(A_i B_j)\ E(A_j B_i)$	$i \neq j$
$2n(n^2-n)$	$E(A_i B_i)\ E(A_j B_k)$	$j \neq k$
$(n^2-n)^2 - 2(n^2-n)$	$E(A_i B_j)\ E(A_k B_\ell)$	$i \neq j$, $k \neq \ell$, and $i \neq k$ or $j \neq \ell$

and letting $D = \sum_{i \neq j}\sum s_{ij}^2$, we obtain:

Theorem 3: $$\frac{5n^2-4n-2C}{3n^2-2n} qr + T_n \geq \rho \geq \frac{5n^2-4n-2C}{3n^2-2n} qr - T_n$$

$$\text{where } T_n = \sqrt{(1-q^2)(1-r^2)+q^2r^2 \ \frac{4C^2 + 4n^2C + (4n-6n^2)D + (2n^3-2n^4)}{(3n^2-2n)^2}}$$

As in the case of one and two mediating variables, the upper and lower bounds have obverse properties. In this section, we shall concern ourselves only with the lower bound, which is shown for selected values of q, r and s in Table III. The lower bound increases as q increases, and this can be proved as follows:

$$\frac{\partial L_n}{\partial q} = \frac{5n^2-4n-2C}{3n^2-2n} r + \frac{q}{T_n} [(1-r^2) - r^2(4C^2+4n^2C+(4n-6n^2)D + 2(n^3-n^4))]$$

(a) First term: $3n^2-2n$ is positive for $n > 0$.

Since C is the sum of the off-diagonal terms of the correlation matrix S, $C \leq n^2-n$ and $5n^2-4n-2C \geq 3n^2-2n > 0$. Therefore, the coefficient of r is positive.

(b) Second term: $r^2 \leq 1$ and $r^2 \geq 0$. Therefore, the second term is non-negative if:

$$t = 4C^2 + 4n^2C + (4n-6n^2)D + 2(n^3-n^4) \leq 0$$

To maximize t, it is necessary to maximize C, given D, since all the coefficients of C are positive. Therefore, using Lagrangean multipliers, it is desired to:

$$\text{Maximize } C = \sum_{i \neq j}\sum s_{ij} \quad , \text{ subject to } \sum_{i \neq j}\sum s_{ij}^2 = D$$

$$\frac{\partial}{\partial s_{ij}} (\Sigma s_{ij} - \lambda(\Sigma s_{ij}^2 - D)) = 0$$

Therefore,

$$s_{ij} = \frac{1}{2\lambda} \; \forall \quad i \neq j$$

Let $s=s_{ij}$. Thus,

$$(n^2-n)s^2 = D$$

$$s = \sqrt{D/(n^2-n)}$$

And C is maximized at $C = \sqrt{(n^2-n)D}$, given D.

<u>Table III</u>

Lower Bounds for the Case of n Mediating Variables

s #	-.20						0						.5						.7						.8					
q	.6	.6	.6	.7	.7	.8	.6	.6	.6	.7	.7	.8	.6	.6	.6	.7	.7	.8	.6	.6	.6	.7	.7	.8	.6	.6	.6	.7	.7	.8
r	.6	.7	.8	.7	.8	.8	.6	.7	.8	.7	.8	.8	.6	.7	.8	.7	.8	.8	.6	.7	.8	.7	.8	.8	.6	.7	.8	.7	.8	.8
1	-.280	-.151	0.0	-.020	.132	.280	-.280	-.151	0.0	-.020	.132	.280	-.280	-.151	0.0	-.020	.132	.280	-.280	-.151	0.0	-.020	.132	.280	-.280	-.151	0.0	.020	.132	.280
2	-.026	.159	.384	.367	.630	*	-.074	.099	.304	.288	.516	.795	-.184	-.037	.135	.117	.295	.478	-.224	-.085	.77	.059	.224	.389	-.243	-.108	.050	.031	.191	.350
3	.004	.194	.423	.407	.674	*	-.049	.128	.337	.321	.553	.835	-.171	-.016	.159	.141	.322	.508	-.216	-.076	.088	.069	.236	.402	-.238	-.102	.057	.038	.199	.359
n: 5	.021	.214	..445	.430	.697	*	-.034	.145	.356	.341	.574	.853	-.163	-.013	.162	.145	.326	.512	-.211	-.070	.094	.076	.243	.410	-.235	-.098	.061	.043	.204	.364
10	.032	.226	.458	.443	.710	*	-.025	.155	.367	.352	.586	.863	-.158	-.007	.168	.151	.333	.520	-.208	-.007	.098	.079	.247	.415	-.233	-.096	.064	.045	.207	.368
15	.035	.230	.461	.447	.713	*	-.022	.158	.371	.356	.589	.866	-.157	-.006	.170	.153	.335	.522	-.208	-.066	.099	.081	.248	.417	-.232	-.095	.065	.046	.208	.369
∞	.041	.236	.468	.454	.719	*	-.017	.164	.317	.362	.596	.870	-.154	-.003	.174	.157	.339	.526	.206	-.064	.101	.083	.251	.420	-.231	-.094	.066	.047	.209	.370

# Where s_{ij} = s for every i ≠ j

Hence,

$$t \leq 4(n^2-n)D + 4n^2 \sqrt{n^2-nD} + (4n-6n^2)D + (2n^3-2n^4)$$

$$t \leq 2n^2[2\sqrt{(n^2-n)D} - D - (n^2-n)] = 2n^2t'$$

where t' equals the expression in the brackets. To maximize t', subject to D, set:

$$\frac{\partial t'}{\partial D} = \frac{\sqrt{n^2-n}}{\sqrt{D}} - 1 = 0$$

Noting that:

$$\frac{\partial^2 t'}{\partial D^2} = -\frac{1}{2}\sqrt{\frac{n^2-n}{D^3}} \leq 0$$

t' is maximized at $D = n^2-n$, at which value t' equals zero. Therefore $t \leq 0$, and $\partial Ln/\partial q$ is positive when q and r are positive and negative when q and r are negative. Hence, L_n increases as $|q|$ or $|r|$ increases, when q and r have the same sign.

From Table III, it is also evident that the lower bound decreases as s increases.[10] This can be shown formally, as follows:

$$\frac{\partial L_n}{\partial C} = \frac{-2qr}{3n^2-2n} - \frac{q^2r^2}{(3n^2-2n)^2} \quad \frac{4C + 2n^2 + (2n-3n^2)\partial D/\partial C}{T_n}$$

The first term is negative when q and r have the same sign. Moreover, $q^2r^2/(3n^2-2n)^2$ is non-negative. To standardize the result, we assume D is constant, and therefore $\partial D/\partial C$ is zero. For n=1, C is zero and $2n^2 > 0$. For n>1, we note that S is a covariance matrix and therefore positive definite. Hence, $C + n > 0$, and $4C + 2n^2 \geq 2n^2-4n \geq 0$. Given a fixed D, the lower bound increases as C decreases, when q and r have the same sign.

[10] In the limiting case, when $s_{ij} = 1 \ \forall \ i,j$, the bounds take the form of the one mediating variable case.

It is also apparent from Table III that the lower bound increases as n increases. This can be proved for the special case when $s_{ij} = s \; \forall \; i \neq j$. In this case, $C = (n^2-n)s$, $D = (n^2-n)s^2$, and:

$$L_n = \frac{(5-2s)n-(4-2s)}{3n-2} qr - \sqrt{(1-q^2)(1-r^2) - \frac{2q^2r^2(1-s)^2(n^2-n)}{(3n-2)^2}}$$

Therefore:

$$\frac{\partial L_n}{\partial n} = \frac{2-2s}{(3n-2)^2} qr - \frac{q^2r^2(1-s)^2(n-2)}{(3n-2)^3 \sqrt{(1-q^2)(1-r^2) - \frac{2q^2r^2(1-s)^2(n^2-n)}{(3n-2)^2}}}$$

$$= \frac{qr(1-s)}{(3n-2)^2} [2 - \frac{(1-s)qr(n-2)}{\sqrt{(1-q^2)(1-r^2)(3n-2)^2 - 2q^2r^2(1-s)^2(n^2-n)}}$$

$qr(1-s)/(3n-2)^2$ is non-negative when q and r have the same sign. Therefore, for n=1,

$$\frac{\partial L_1}{\partial n} = qr[2 + \frac{qr}{\sqrt{(1-q^2)(1-r^2)}}] \geq 0$$

and for n=2,

$$\frac{\partial L_2}{\partial n} = \frac{2qr(1-s)}{(3n-2)^2} \geq 0$$

For $n \geq 3$, we note that, from Lemma 1, $s \geq 2q^2-1$ and $s \geq 2r^2-1$. Hence, $q^2 \leq (s+1)/2$, $r^2 \leq (s+1)/2$ and

$$\frac{\partial L_n}{\partial n} \geq \frac{2qr(1-s)}{(3n-2)^2} [1 - \frac{qr(n-2)}{\sqrt{(3n-2)^2 - 8q^2r^2(n^2-n)}}$$

The expression in the brackets is at a minimum whfn q = r = 1. Therefore,

$$\frac{\partial L_n}{\partial n} \geq \frac{2qr(1-s)}{(3n-2)^2} \left[1 - \frac{n-2}{\sqrt{n^2-4n+4}}\right] = 0$$

The lower bound thus increases as n increases, given s, when q and r have the same sign.

In the limit, as n approaches infinity:

$$\text{Lim } L_n = \frac{5-2s}{3} qr - \sqrt{(1-q^2)(1-r^2) - \frac{2(1-s)^2 q^2 r^2}{9}}$$

Limiting cases are shown in Table III.

The range $R \leq 2T_n$, since U_n may exceed 1 and L_n may fall below -1. From the arguments presented above, given their special assumptions, it is apparent that T_n decreases as $|q|$ increases, $|r|$ increases, C decreases, or n increases.

V. Conclusion

The theoretical discussion presented in this paper provides a guide for the construction of a viable match:

(i) In the case of one mediating variable it was shown that the range of the correlation coefficient between the mediated variables decreases sharply as either q or r approaches one. (In the case where q or r equals one, ρ is determined with certainty.) Moreover, the range of ρ in the case of n mediating variables can be no greater than the range in the case of one variable. Therefore, X variables should be chosen that are highly correlated with either the Y or Z variable.

(ii) The upper bound on the range of ρ declines as the sum of the correlation coefficients between the X variables declines. From Tables II and III it is evident that the lower bound on ρ is very sensitive to the value of the s parameter. Therefore considerable gain in the accuracy of the match can be achieved by choosing X variables that are uncorrelated or even negatively correlated.

(iii) The upper bound on the range of ρ also declines, given certain strong assumptions, as the number of X variables increases. Table III shows that there is a large gain from increasing the number of mediating variables from 1 to 5 but minimal gain from increasing it beyond 5. Therefore, at least five X variables should be chosen in engineering a match.

BIBLIOGRAPHY

Alter, Horst, "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970", Annals of Economic and Social Measurement, Vol. 3, No. 2, April, 1974.

Budd, Edward, "The Creation of a Microdata File for Estimating the Size Distribution of Income", Review of Income and Wealth, Series 17, No. 4, December, 1971.

Okner, Benjamin, "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File", Annals of Economic and Social Measurement, Vol. 1, No.3, July, 1972.

Ruggles, Nancy and Richard, "A Strategy for Merging and Matching Microdata Sets", Annals of Economic and Social Measurement, Vol. 3, No.2, April, 1974.

Sims, Christopher, "Rejoinder", Annals of Economic and Social Measurement, Vol. 1, No. 3, July, 1972.