122-72

CONDITIONS FOR THE PRESENCE OR ABSENCE OF A BIAS IN TREATMENT EFFECT: SOME STATISTICAL MODELS FOR HEAD START EVALUATION

by

Burt S. Barnow

CONDITIONS FOR THE PRESENCE OR ABSENCE OF A BIAS IN TREATMENT EFFECT: SOME STATISTICAL MODELS FOR HEAD START EVALUATION

Burt S. Barnow

The research reported here was supported, in part, by funds granted to the Institute for Research on Poverty at the University of Wisconsin pursuant to the provisions of the Economic Opportunity Act of 1964. The author wishes to thank Glen G. Cain and Arthur S. Goldberger for helpful comments. The views expressed herein are solely those of the author.

April 1972

ABSTRACT

This paper examines the possibilities of carrying out unbiased evaluations for compensatory education programs when group selection is not random and/or when pretests are unavailable. Four models are presented which demonstrate that bias can be avoided if particular selection procedures are used. The models are then considered for their usefulness in evaluating Head Start by using the data collected for the Westinghouse Learning Corporation-Ohio University study of that program. Suggestions are made on how the appropriate model could be selected and how the presence of bias could be determined. In the final section of the paper the importance of correct specification of the model and knowledge of the selection procedures are shown to be crucial for unbiased evaluation of nonrandom experiments and ex post facto analyses.

1. Introduction

In recent years economists have become increasingly interested in evaluating compensatory education programs such as Head Start. Fortunately, psychologists and sociologists have already done extensive work in this area enabling economists to build upon an excellent foundation. In this paper we shall examine the possibilities of carrying out an unbiased evaluation of compensatory programs by using regression analysis. In particular we will examine quasi-experimental situations where random assignment was not used and/or where an ex post facto analysis must be used because there is no pre-treatment information available. This does not imply that quasi-experimental analyses are more desirable than true experiments, but that when a quasi-experimental analysis is the only feasible means of carrying out an evaluation, the analysis may not lead to bias in the estimates of treatment effect. Thus, we shall demonstrate that the following statement by Campbell and Erlebacher (1970; p. 185) is not as general as they assert:

> Evaluations of compensatory educational efforts such as Head Start are commonly quasi-experimental or *ex post facto*. The compensatory program is made available to the most needy, and the "control" groups then sought from among the untreated children in the same community. Often this untreated population is on the average more able than the "experimental" group. In this situation, the ususal procedures of selection, adjustment, and analysis produce systematic biases in the direction of making the compensatory program look deleterious. Not only does matching produce regression artifacts in this direction, but so also do analysis of covariance and partial correlation. These biases of analyses occur both where pretest scores are available and in *ex post facto* studies.

To prove our points we shall present several models with various relationships between the appropriate variables for an evaluation of a compensatory education program. We shall then present an algebraic analysis to determine whether or not regression analysis will lead to an unbiased evaluation. Although regression analysis is used throughout the paper, it should be noted that regression analysis is mathematically equivalent to analysis of covariance so that our results could be expressed equally well in terms of analysis of covariance. In addition to analytical proofs, we have also run Monte Carlo computer simulations of several of the models. These simulations may be helpful to those who prefer a more graphic, empirical demonstration of the points to be made, although we cannot really prove a point by this technique.

Several simplifying assumptions will be made throughout the paper. Relaxation of some of these assumptions can lead to important changes in the models and make the analysis presented inappropriate. Thus, it is dangerous to extrapolate the results found below to models where the basic assumptions are violated. The assumptions common to all of the models presented are:

1. Linearity and additivity of all relationships. We shall not consider models where a compensatory program produces more (or less) gain for more able children.

2. Zero growth rate of ability over time. This assumption assures us that ceteris paribus a child who has received no treatment will have the same expected score on a test given at various times over an interval.

3. The treatment has no effect. This assumption is only used to make the computer simulation plots easier to interpret.

4. The treatment is discrete rather than continuous. We assume that there is only one level of treatment and that a child either receives the treatment or does not receive it. (This assumption is made only for convenience.)

We shall now present four possible models for an evaluation of Head Start. The models presented below represent only a few of the many conceivable ones, and are given only as possibilities. To make comparison of the models presented easy, we shall use the same notation throughout the paper. The variables which are included in the models are:

Y posttest score

X_1^* true ability at the time of the pretest

X_1 pretest score

X_2 socio-economic status

Z dummy variable for treatment defined as

$$Z = \begin{cases} 1 \text{ if received treatment (i.e., in experimental group)} \\ 0 \text{ if did not receive treatment (i.e., in control group)} \end{cases}$$

u disturbance term associated with the pretest

v disturbance term associated with the posttest

2. The Campbell-Erlebacher Two Populations Model

Campbell and Erlebacher (1970) have developed a model to demonstrate that if subjects in the experimental and control groups are selected from two different populations with the control population having a higher mean, regression analysis can produce a spurious negative treatment effect. Although a computer simulation rather than a formal model was used in their paper, it is not difficult to construct the general model that Campbell-Erlbacher deal with implicitly:

Experimental Group

(1E) $X_1 = X_1^* + u$

(2E) $Y = X_1^* + v$

(3E) $X_1^* \sim N(\mu_E, \sigma_*^2)$

(4E) $u \sim N(o, \sigma^2)$

(5E) $v \sim N(o, \sigma^2)$

(6E) $c(u, X_1^*) = c(v, X_1^*) = c(u, v) = 0$

Control Group

(1C) $X_1 = X_1^* + u$

(2C) $Y = X_1^* + v$

(3C) $X_1^* \sim N(\mu_c, \sigma_*^2)$

(4C) $u \sim N(0, \sigma^2)$

(5C) $v \sim N(o, \sigma^2)$

(6C) $c(u, X_1^*) = c(v, X_1^*) = c(u,v)=0$

and where $\mu_c > \mu_E$.

Since many of the features presented in this model are common to the subsequent models presented, it may be helpful to describe some of the implications of the model. Note that since the disturbance terms have zero means, the pretest and posttest scores will be unbiased but fallible measures of true ability. The assumption that u and v are uncorrelated implies that if a child scores higher than his true ability on the pretest we have no a priori knowledge concerning whether he will score higher or lower than his true ability on the posttest. The model excludes the possibility of growth in true ability between the taking of the pretest and the taking of the posttest. (Proportional growth could be introduced into the model by replacing (2E) and (2C) with an equation of the form $Y = aX_1^* + v$, and constant growth could be introduced by making $Y = b + X_1^* + v$.)

We may now solve for $E(Y|X_1)$, the conditional expectation of posttest given pretest, for the experimental group:

(7E) $$E(Y|X_1) = E(Y) - \frac{c(X_1, Y)}{V(X_1)} \cdot E(X_1) + \frac{c(X_1, Y)}{V(X_1)} \cdot X_1$$

(8E) $$= \mu_E - \frac{V(X_1^*)}{V(X_1^*) + V(u)} \cdot \mu_E + \frac{V(X_1^*)}{V(X_1^*) + V(u)}$$

(9E) $$= (1 - P)\mu_E + PX_1$$

where $$P = \frac{V(X_1^*)}{V(X_1^*) + V(u)}$$

Since $0 \leq P \leq 1$, the slope of the conditional expectation is less than unity, so that we have regression towards the mean. Since the only difference between the two populations is in their means, the conditional expectation for the control group is:

(9C) $$E(Y| X_1) = (1-P)\mu_c + PX_1$$

Now define the treatment variable Z as follows:

(10) $$Z = \begin{cases} 1 \text{ if a subject had the treatment} \\ 0 \text{ if a subject did not have the treatment} \end{cases}$$

In the Campbell-Erlebacher example Z corresponds exactly with group membership:

(11) $$Z = \begin{cases} 1 \text{ if a subject was in "Lower" population} \\ 0 \text{ if a subject was in "Higher" population} \end{cases}$$

(9C) and (9E) can be rewritten as:

(12) $$E(Y|X_1, Z = 0) = (1-P)\mu_c + PX_1$$
$$E(Y|X_1, Z = 1) = (1-P)\mu_E + PX_1$$

Since Z takes on only the values of 1 and 0, (12) and (13) can be combined to get:

$$E(Y|X_1,Z) = (1-P)\mu_c + PX_1 + (1-P)\ (\mu_E-\mu_c)Z. \tag{14}$$

Clearly, the coefficient of Z does not measure the effect of the treatment; rather it just reflects the difference in population means. Group membership is serving as a proxy for true ability.

The model presented above corresponds exactly to the model developed by Campbell and Erlebacher. We have also prepared computer simulations of the model with 500 and 2000 observations. The values for X_1^* in the experimental group were selected at random from a normal population with a mean of 4.0 and variance of 1.0. The values of X_1^* for the control group were selected at random from a normal population with a mean of 6.0 and a variance of 1.0. For both populations the values for u and v were selected at random from a normal population with a mean of 0.0 and a variance of 1.0. The values for X_1 and Y were computed as specified in the model; i.e. $X_1 = X_1^* + u$. A summary of the statistics for the simulation with 2000 observations is presented in Table I. The results corroborate the analysis of the model--if selection for treatment is based on population membership we will underestimate the effects of the treatment. (In this case a null treatment is estimated to lower posttest scores by .9955 points.)

The simulation with 500 observations was run so that the results could be graphically displayed by a plotter. The pretest-posttest pair for each observation was plotted with a vertical tally used to indicate the points for the experimental group and a horizontal tally used to indicate the points for the control group. The regression equation fitted to the sample of 500 observations is:

$$\hat{Y} = \underset{(13.718)}{3.167} + \underset{(12.713)}{.4694X_1} - \underset{(-8.751)}{1.104Z}, \quad R^2 = .5123$$

where the numbers in parentheses are the t ratios. The theoretical values for the regression equation, that is the values we would obtain from an infinite sample, can be calculated by substituting the population parameters chosen for the simulation into the model. The theoretical equation is:

$$\hat{Y} = 3.00 + .500\ X_1 - 1.00Z$$

The regression equation for the simulation with 500 observations has been drawn in. The vertical distance between the two lines represents the value of the coefficient of Z which could be incorrectly interpreted by an evaluator as the treatment effect.

The models presented in this section and the following section are structured so that the program could be evaluated without bias by using gain scores as the dependent variable in the regression and the dummy variable Z as the sole independent variable. The gain score G is defined as $G = Y - X_1$ for each observation. Campbell and Erlebacher (1970, p. 197) note that the gain scores approach would avoid the bias problem, but they warn that "gain scores are in general such a treacherous quicksand, e.g. are so <u>non</u> comparable [sic] for high versus low scores within any particular sample, that one is reluctant to recommend them for any purpose." This warning has been given by other psychometricians such as Behrnstedt (1969) and Cronbach and Furby (1970). Gain scores produce bias when we introduce complications such as heteroskedasticity and nonzero growth rates into the models. Thus Campbell and Erlebacher have presented one model where regression analysis can lead to a biased estimate of treatment effect in a quasi-experiment.

Table 1

Statistics for Simulation of Campbell-Erlebacher Model
with 2000 Observations

1. Means, Standard Derivations, and Variances

Variable	Mean	Standard Deviation	Variance
Y	4.998	1.704	2.903
X_1	5.024	1.738	3.022
Z	.500	.500	.250

2. Covariance Matrix

$$\begin{array}{cccc} Y & X_1 & Z & \\ 2.903 & 1.956 & -.4855 & Y \\ & 3.022 & -.4823 & X_1 \\ & & .2501 & Z \end{array}$$

3. Correlation Matrix

$$\begin{array}{cccc} Y & X_1 & Z & \\ 1.0000 & .6603 & -.5686 & Y \\ & 1.0000 & -.5548 & X_1 \\ & & 1.0000 & Z \end{array}$$

4. Regression Equation (values in parentheses are t ratios)

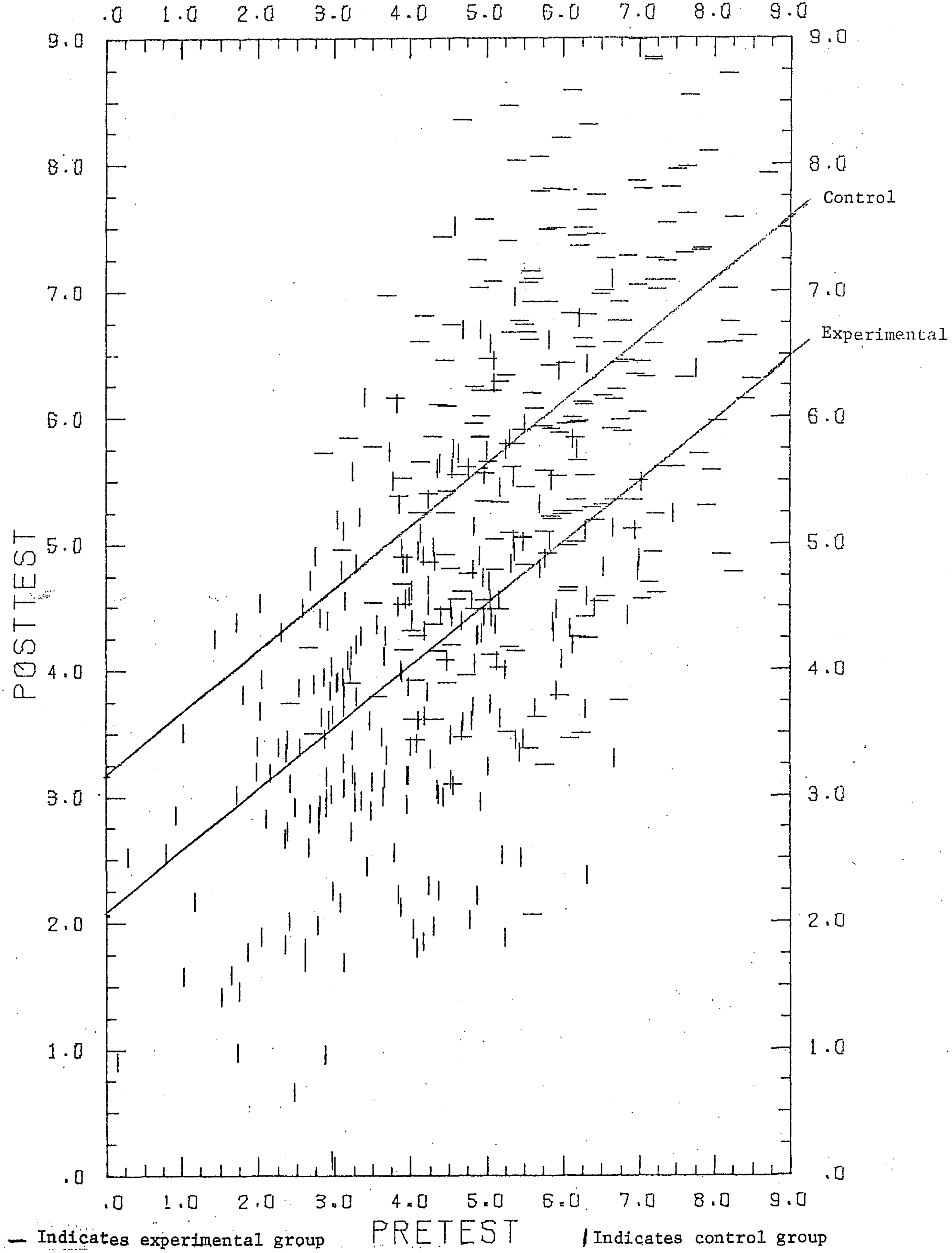
$$\hat{Y} = 3.042 + .4883X_1 - .9955Z$$
$$(25.672)\ (26.073)\ (-15.073)$$

F ratio 979.3
R^2 .4951

Y = Posttest
X_1 = Pretest
Z = Dummy variable for treatment

CAMPBELL MODEL

FIGURE 1



3. Selection on True Ability for a One Population Model

It is possible to object to the Campbell-Erlebacher model on the grounds that ability is not necessarily distributed in two separate normal distributions but instead in one larger population. The model developed below modifies the Campbell-Erlebacher model so that there is only one population; to create the conditions similar to the Campbell-Erlebacher model for the computer simulations we can select the experimental and control groups on the basis of true ability. It should be noted that the model is a general one for regression when there are errors in one regressor; we are simply applying the model to a Head Start evaluation.

We suppose the model to be:

(15) $Y = \beta_0 + \beta_1 X_1^* + \beta_2 Z + v$

where the variables are defined in the same way as in the introduction. If Head Start has no effect, as we assume in the simulation, then $\beta_2 = 0$. We further assume that X_1^* is unavailable for the evaluation. What we do have available is the pretest score:

(16) $X_1 = X_1^* + u$

where u is independent of v, X_1^*, and Z. Note that

(17) $c(u,Z) = c(u,v) = c(u,X_1^*) = c(v,X_1^*) = c(v,Z) = 0.$

In addition, we shall assume that u, v, and X_1^* all have normal distributions. We are interested in determining if the regression coefficient of Z will be the same when we regress Y on X_1 and Z rather than Y on X_1^* and Z. Thus when we determine

(18) $E(Y|X_1,Z) = \alpha_0 + \alpha_1 X_1 + \alpha_2 Z,$

will $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$?

First, let us make the following definitions:

(19) $\sigma_{11} = V(X_1^*)$, $\sigma_{ZZ} = V(Z)$, $\sigma_{1Z} = c(X_1^*, Z)$

$$P = \frac{\sigma_{11}}{\sigma_{11}+V(u)}, \qquad r^2 = \frac{(\sigma_{1Z})^2}{\sigma_{11}\sigma_{ZZ}}$$

Note that the variable P in this model is defined as the ratio of the variance of true ability to the variance of the pretests for the entire population, whereas in the Campbell-Erlebacher model P was the ratio of the variances within each group. The variable r^2 is the coefficient of determination we would obtain from a linear regression of X_1^* on Z, and we know that $0 \leq r^2 \leq 1$. We can now use the normal equations found in Johnston (1968, p. 56) to solve for α_1 and α_2 in terms of β_1 and β_2:

$$\alpha_1 = \frac{c(Y,X_1) \cdot V(Z) - c(Y,Z) \cdot c(X_1,Z)}{V(X_1) \cdot V(Z) - c(X_1,Z) \cdot c(X_1,Z)}$$

$$= \frac{(\beta_1\sigma_{11} + \beta_2\sigma_{1Z})\sigma_{ZZ} - (\beta_1\sigma_{1Z} + \beta_2\sigma_{ZZ})\sigma_{1Z}}{(\sigma_{11} + V(u)) \cdot \sigma_{ZZ} - \sigma_{1Z}\sigma_{1Z}}$$

$$= \frac{P\beta_1 \, (\sigma_{11}\sigma_{ZZ} - \sigma_{1Z}\sigma_{1Z})}{\sigma_{11}\sigma_{ZZ} - P\,\sigma_{1Z}\sigma_{1Z}}$$

(20) $$\alpha_1 = \frac{P(1-r^2)}{1 - Pr^2} \cdot \beta_1$$

For α_2 we find:

$$\alpha_2 = \frac{V(X_1) \cdot c(Z,Y) - c(X_1,Z) \cdot c(X_1,Y)}{V(X_1) \cdot V(Z) - c(X_1,Z) \cdot c(X_1,Z)}$$

$$= \frac{\frac{\sigma_{11}}{P}(\beta_1\sigma_{1Z} + \beta_2\sigma_{ZZ}) - \sigma_{1Z}\,(\beta_1\sigma_{11} + \beta_2\sigma_{1Z})}{\frac{\sigma_{11}}{P} \cdot \sigma_{ZZ} - \sigma_{1Z}\sigma_{1Z}}$$

$$= \frac{\beta_1 \left(\frac{\sigma_{1Z}\sigma_{11}}{P} - \sigma_{11}\sigma_{1Z}\right) + \beta_2 \frac{\sigma_{11}\sigma_{ZZ}}{P} - \sigma_{1Z}\sigma_{1Z}}{\frac{\sigma_{11}\sigma_{ZZ}}{P} - \sigma_{1Z}\sigma_{1Z}}$$

$$(21) \quad \alpha_2 = \beta_2 + \frac{\sigma_{1Z}(1-P)}{\sigma_{ZZ}(1-r^2P)} \cdot \beta_1$$

Thus we find that in general $\alpha_1 \neq \beta_1$ and $\alpha_2 \neq \beta_2$.

Since we are trying to evaluate the effects of Head Start, we will be particularly interested in knowing when $\alpha_2 = \beta_2$. One possibility of this is when $P = 1$; if this is the case, however, then $V(u) = 0$ and $X_1 = X_1^*$. Thus $P = 1$ means that we have no measurement error and that we can directly calculate β_1 and β_2. A more interesting case is that $\sigma_{1Z} = 0$ suffices for $\alpha_2 = \beta_2$. But $\sigma_{1Z} = 0$ is equivalent to $E(X_1^*|Z) = E(X_1^*)$ for both values of Z; thus the mean of true abilities must be the same in the experimental and control groups for $\alpha_2 = \beta_2$. This would presumably occur if treatment were determined randomly.

To prepare the computer simulations for this model, values of X_1^* were selected at random from a normal population with a mean of 5.0 and a variance of 2.0. The values of u and v were selected at random from a normal population with a mean of 0.0 and a variance of 1.0. The true effect of Head Start, β_2, was assumed to be 0.0, as was the constant β_0. The observations were assigned to the experimental and control groups on the basis of true ability as follows: Those who were in the upper half of the sample on the basis of true ability were placed in the control group; those who were in the lower half were placed in the experimental group which is analogous to the Campbell-Erlebacher model where treatment depended on true ability. Simulations were run with 2000 and 500 osbservations. The results of the simulation with 2000 observations are contained in Table II. The theoretical values for the regression equation,

Table 2

Statistics for Simulation of One Population-Selection on Ability Model with 2000 Observations

1. Means, Standard Deviations, and Variances

Variable	Mean	Standard Deviation	Variance
Y	5.032	1.711	2.928
X_1	5.030	1.696	2.878
Z	.500	.500	.250

2. Covariance Matrix

$$\begin{array}{cccl} Y & X_1 & Z & \\ 2.928 & 1.925 & -.556 & Y \\ & 2.878 & -.562 & X_1 \\ & & .250 & Z \end{array}$$

3. Correlation Matrix

$$\begin{array}{cccl} Y & X_1 & Z & \\ 1.000 & .663 & -.650 & Y \\ & 1.000 & -.662 & X_1 \\ & & 1.000 & Z \end{array}$$

4. Regression Equations (values in parentheses are t ratios)

$$\hat{Y} = 3.572 + .418X_1 - 1.284Z$$
$$(26.649)\ (19.983)\ (-18.103)$$

F ratio 1075.6

R^2 .5186

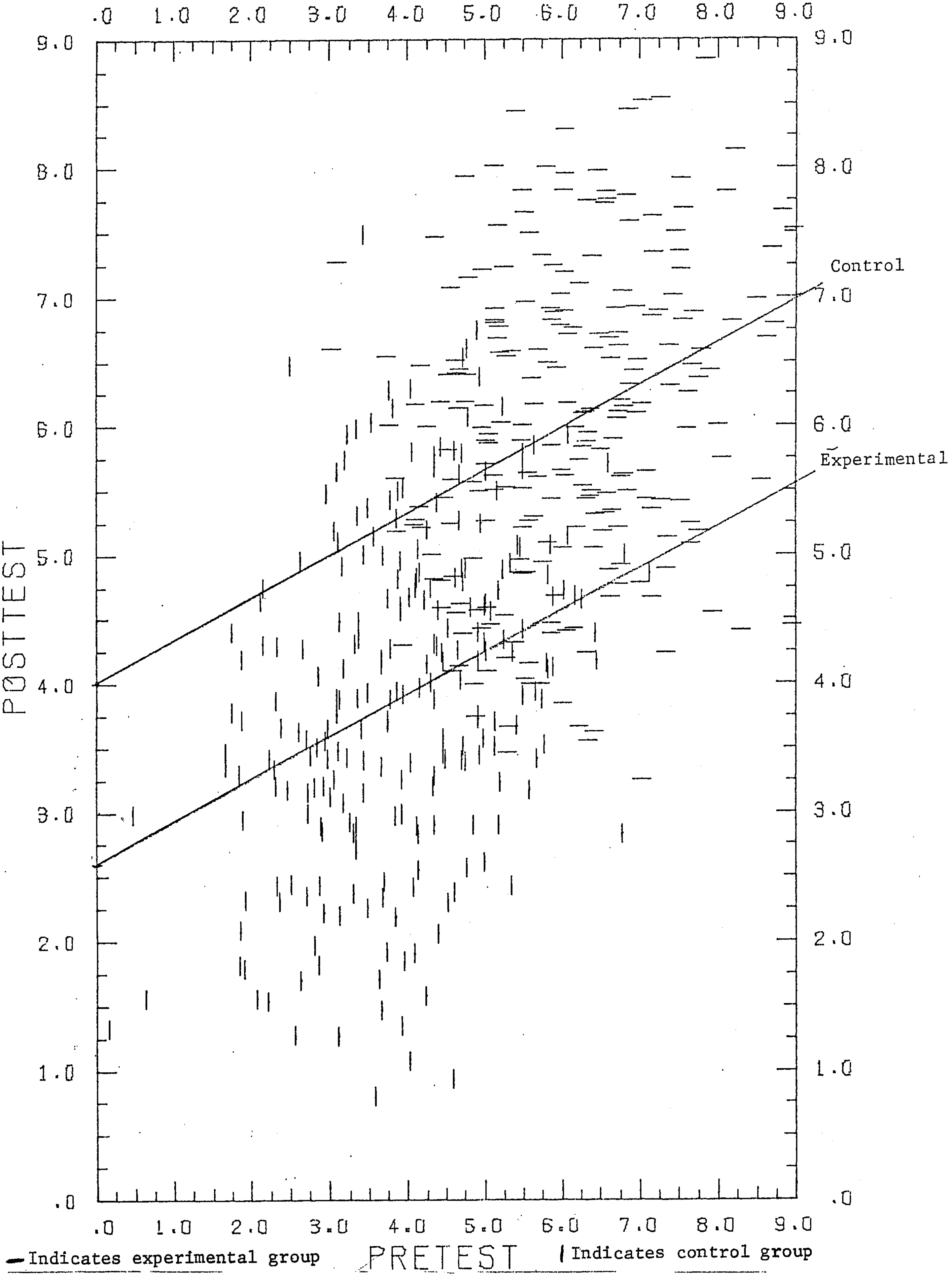
Y = Posttest

X_1 = Pretest

Z = Dummy variable for treatment

SELECT ON TRUE ABILITY

FIGURE 2



calculated from the formulas in Goldberger (1972) are:

$$\hat{Y} = 3.57 + .42X_1 - 1.31Z$$

A similar simulation with 500 observations was also run, and the results are plotted in Figure 2 with the regression lines drawn in. The equation produced from this simulation was:

$$\underset{(14.14)}{\hat{Y} = 4.00} + \underset{(7.46)}{.334X_1} - \underset{(-9.66)}{1.40Z}, \quad R^2 = .4673$$

which is similar to both the simulation with 2000 observations and the theoretical equations. As in the Campbell-Erlebacher model, if the Head Start program is made available to those who are most needy, an evaluation which did not take this selection procedure into account would lead to biased estimates of the treatment effect. Thus we have a second case where nonrandom selection will lead to bias in the evaluation.

Note that if we select members for the experimental and control groups on the basis of pretests (X_1), we cannot say whether the above procedures will produce biased results even though $c(X_1^*,Z) \neq 0$. The model does not apply since $c(u,Z) \neq 0$ and assumption (17) is violated.

4. Selection on Pretests for a One Population Model

We now consider a model where a child is assigned to either the Head Start or the control group on the basis of his pretest score rather than his true ability. Since pretest scores are correlated (although not perfectly) with true ability, when we select on the former we are also selecting, in a sense, on ability. Lord and Novick (1968, p. 141) refer to these two methods of selction as explicit selection and incidental selection. Keeping in mind that we are dealing with a single population, we may specify our model as follows:

(22) $X_1 = X_1^* + u$

(23) $Y = X_1^* + v$

(24) $X_1^* \sim N(\mu,\sigma_*^2)$

(25) $u \sim N(o, \sigma^2)$

(26) $v \sim N(o,\sigma^2)$

(27) $c(u,X_1^*) = c(v,X_1^*) = c(u,v) = 0$

If we solve for $E(Y|X_1)$ as we did in (7E) - (9E) we find:

(28) $E(Y|X_1) = (1-P)\mu + PX_1$

where once again $P = \dfrac{V(X_1^*)}{V(X_1^*) + V\ (u)}$.

Lord and Novick (1968, p. 143) point out that "since the true regression of Y on X is linear, the ... regression function is not affected by explicit selection on X; hence the regression coefficients will be the same in the unselected [entire] group and in any selected group." Thus, suppose that those who score in the lower half on a pretest are assigned to the experimental (Head Start) group, and those who score in the upper half are assigned to the control group. Assuming once again that the treatment has no effect, we find:

(29) $E(Y|X_1) = (1-P)\mu + PX_1$ for the experimental group, and

(30) $E(Y|X_1) = (1-P)\mu + PX_1$ for the control group.

Defining a variable Z such that $Z = \begin{cases} 1 \text{ for experimental group} \\ 0 \text{ for control group} \end{cases}$

(29) and (30) can be written as:

(31) $E(Y|X_1,\ Z=1) = (1-P)\mu + PX_1$

(32) $E(Y|X_1,\ Z=0) = (1-P)\mu + PX_1$

so that

(33) $E(Y|X_1,Z) = (1-P)\mu + PX_1 + 0 \cdot Z$

Thus we find that explicit selection on X_1, the fallible indicator of X_1^*, does not produce a biased treatment effect. Even though the control group is more able, there is no bias in the estimates of the effects of Head Start; the dummy variable Z provides no information about the child's ability that is not already contained in the pretest score (which determines the value of Z). Campbell (1969) refers to this type of evaluation as a "regression discontinuity design" and recommends its use when evaluators are unwilling to use random selection procedures. Goldberger (1972) has demonstrated that this type of evaluation is about 9/25 as efficient as a random selection experiment with equal sample size.

Computer simulations with 2000 and 500 observations were conducted with this model. The data used in the simulations are the same for the simulations for the previous model except that group membership (Z) was determined by pretests rather than true ability. The theoretical values for the regression equations are

$$\hat{Y} = 1.67 + .67X_1 + 0.0Z$$

The results of the simulation with 2000 observations are summarized in Table III. The points from the simulation with 500 observations are plotted in Figure 3 with the fitted regression line drawn in. The fitted equation for the simulation with 500 observations is

$$\underset{(5.10)}{\hat{Y} = 2.04} + \underset{(9.65)}{.596X_1} - \underset{(-.486)}{.098Z}\ , \quad R^2 = .3675$$

Both simulations verify the analysis and show that an evaluation conducted in this manner will lead to unbiased estimates of the effects of Head Start. Thus, if the "creaming" or "scraping" techniques were

Table 3

Statistics for Simulation of One Population-Selection on Pretests Model with 2000 Observations

1. Means, Standard Deviations, and Variances

Variable	Mean	Standard Deviation	Variance
Y	5.032	1.711	2.928
X_1	5.030	1.696	2.878
Z	.500	.500	.250

2. Covariance Matrix

Y	X_1	Z	
2.928	1.925	-.453	Y
	2.878	-.679	X_1
		.250	Z

3. Correlation Matrix

Y	X_1	Z	
1.000	.663	-.529	Y
	1.000	-.800	X_1
		1.000	Z

4. Regression Equation (values in parentheses are t ratios)

$$\hat{Y} = 1.637 + .673X_1 + .019Z$$

$$(8.864)\ (23.869)\ (.197)$$

F ratio	783.2
R^2	.4396

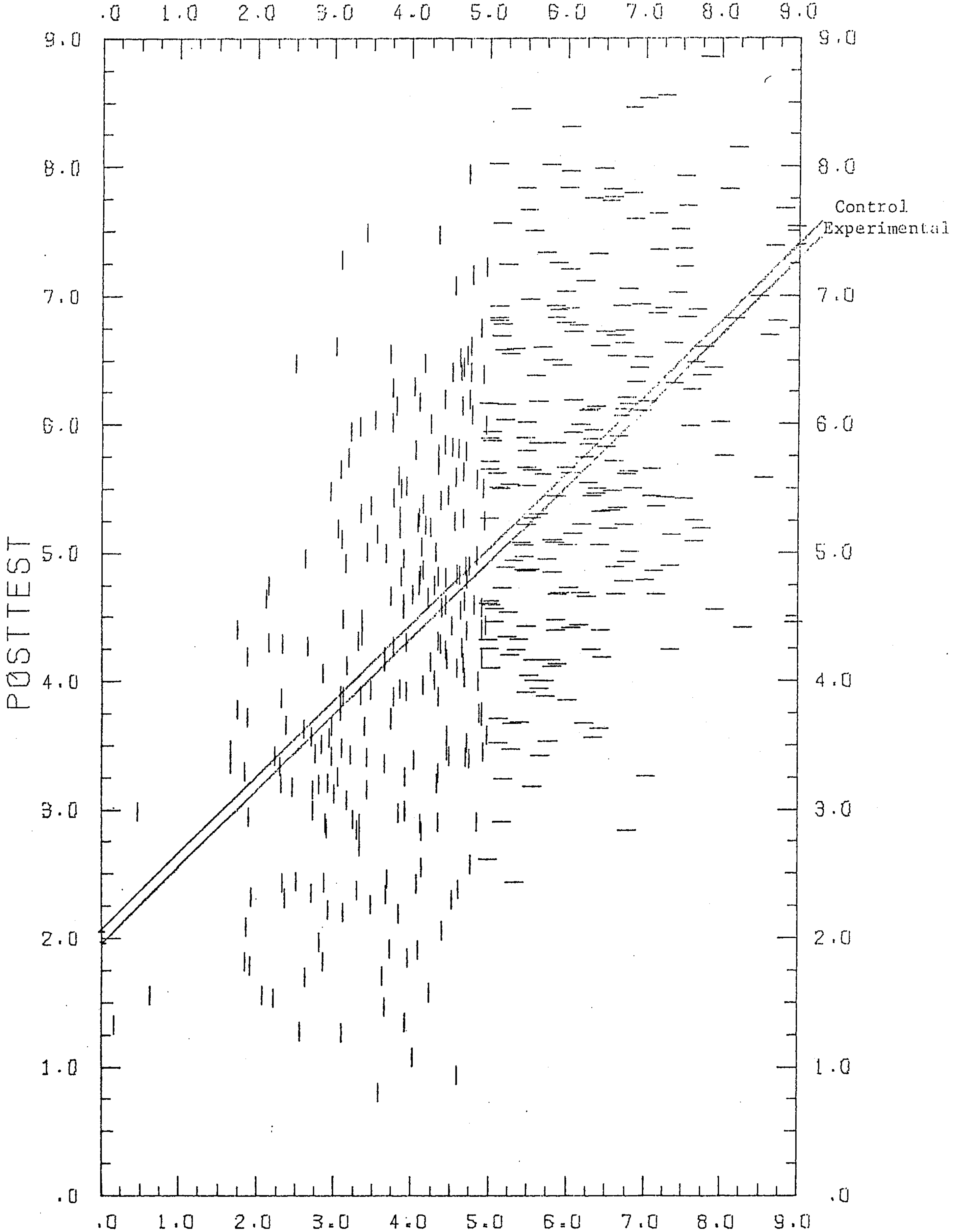
Y = Postest

X_1 = Pretest

Z = Dummy variable for treatment

SELECT ØN PRETESTS

FIGURE 3



—Indicates experimental group

| Indicates control group

used for group selection on the basis of pretests, we can still carry out an unbiased evaluation.

5. A One Population Omitted Variable Model.

We now examine a more complex model where socio-economic status (SES) is allowed to have an effect on posttest scores and where no pretests are available; since true ability is unobserved, we have an omitted variable problem rather than the errors in variable problem encountered in the previous models. The model to be considered is

(34) $Y = \beta_0 + \beta_1 X_1^* + \beta_2 X_2 + \beta_3 Z + v$

where Y = posttest score

X_1^* = true ability

X_2 = SES

Z = treatment variable

v = disturbance term

We shall assume that X_1^*, X_2, and v have normal distributions and that v is independent of X_1^* and X_2. Presently we are interested in the value of β_3 in (34) which measures the effectiveness of the treatment. If we had observations on Y, X_1^*, X_2, and Z we could run an ordinary least squares regression to get an unbiased estimate of β_3. However, we shall assume that X_1^* is unobservable, and we must determine if the regression

(35) $Y = \alpha_0 + \alpha_2 X_2 + \alpha_3 Z$

will yield an $\alpha_3 = \beta_3$. For convenience we define the following terms:

(36)
$$\begin{array}{lll} \sigma_{11} = V(X_1^*) & \sigma_{12} = c(X_1^*,X_2) & \sigma_{1Z} = c(X_1^*,Z) \\ \sigma_{22} = V(X_2) & \sigma_{2Z} = c(X_2,Z) & \sigma_{ZZ} = V(Z) \\ \sigma_{1Y} = c(X_1^*,Y) & \sigma_{2Y} = c(X_2,Y) & \sigma_{ZY} = c(Z,Y) \end{array}$$

We now write the normal equations for (34):

(37) $\sigma_{11}\beta_1 + \sigma_{12}\beta_2 + \sigma_{1Z}\beta_3 = \sigma_{1Y}$

(38) $\sigma_{12}\beta_1 + \sigma_{22}\beta_2 + \sigma_{2Z}\beta_3 = \sigma_{2Y}$

(39) $\sigma_{1Z}\beta_1 + \sigma_{2Z}\beta_2 + \sigma_{ZZ}\beta_3 = \sigma_{ZY}$

Similarily, the normal equations for (35) are:

(40) $\sigma_{22}\alpha_2 + \sigma_{2Z}\alpha_3 = \sigma_{2Y}$

(41) $\sigma_{2Z}\alpha_2 + \sigma_{ZZ}\alpha_3 = \sigma_{ZY}$

We now solve (40) and (41) for α_3 and use equations (37), (38), and (39) to express the answer in terms of the β's:

$$\alpha_3 = \frac{\sigma_{22}\sigma_{ZY} - \sigma_{2Z}\sigma_{2Y}}{\sigma_{22}\sigma_{ZZ} - \sigma_{2Z}\sigma_{2Z}}$$

$$= \frac{\sigma_{22}(\sigma_{1Z}\beta_1 + \sigma_{2Z}\beta_2 + \sigma_{ZZ}\beta_3) - \sigma_{2Z}(\sigma_{12}\beta_1 + \sigma_{22}\beta_2 + \sigma_{2Z}\beta_3)}{\sigma_{22}\sigma_{ZZ} - \sigma_{2Z}\sigma_{2Z}}$$

$$(42) \qquad = \beta_3 + \frac{\sigma_{1Z}\sigma_{22} - \sigma_{12}\sigma_{2Z}}{\sigma_{22}\sigma_{ZZ} - \sigma_{2Z}\sigma_{2Z}} \cdot \beta_1$$

This is the standard result of running a regression without including a relevant variable. It can be demonstrated that

$$(43) \qquad \frac{\sigma_{1Z}\sigma_{22} - \sigma_{12}\sigma_{2Z}}{\sigma_{22}\sigma_{ZZZ} - \sigma_{2Z}\sigma_{2Z}} = b_{1Z\cdot 2}$$

where $b_{1Z\cdot 2}$ is the partial regression coefficient of Z when X_1^* is regressed on Z and X_2. (This is a mechanical relationship and does not depend on a causal model where X_1^* is determined by Z and X_2.)

Thus in general $\alpha_3 \neq \beta_3$ with the extent and direction of the inequality depending upon β_1 and the variances and covariances of X_1^*, X_2, and Z. There are several cases where $\alpha_3 = \beta_3$ and the omitted variable presents no problem. One such case is when $\beta_4 = 0$, but if $\beta_4 = 0$ true ability has no effect on test scores and the test would be worthless. Another case is that when $\sigma_{1Z} = 0$ and either $\sigma_{12} = 0$ or $\sigma_{2Z} = 0$ then $\alpha_3 = \beta_3$. In an experiment where children were randomly assigned to the two groups the sample covariances for σ_{1Z} and σ_{2Z} would presumably be 0, and an ordinary least squares regression would produce an α_3 that is an unbiased estimate of β_3. Note that in this model $\sigma_{2Z} = 0$ is not a sufficient condition for eliminating bias. Since partial regression coefficients always have the same signs as the partial correlation coefficients, $r_{1Z\cdot 2} = 0$ is a sufficient condition for $b_{1Z\cdot 2} = 0$ and hence for $\alpha_3 = \beta_3$. Thus, if children at any given SES level are assigned randomly with regard to true ability to either the experimental or control group, a regression analysis will produce no spurious treatment effect. To clarify this point, consider the following example. Suppose that children are stratified by SES into three groups--high, middle, and low. Further assume that the administrators of a Head Start program are primarily interested in helping those children who are most disadvantaged. Then assume that 90% of the low SES group were selected at random and assigned to the Head Start group with the remaining 10% assigned to the control group. For the middle SES group we shall assume that 50% of the children were selected randomly and assigned to the Head Start group with the remainder assigned to the control group. Finally we assume that for the high SES children 10% were selected randomly and assigned to the Head Start group and the other 90% were assigned to the control group. If this selection procedure is used we know that r_{1Z}, the simple correlation of ability and treatment, will not be 0 since most of the abler children will be assigned to the control group (assuming that ability and SES are positively correlated). The random selection within each SES group assures us that $r_{1Z \cdot 2}$, the partial correlation between ability and treatment controlling for SES, will be equal to 0. Thus if we run a regression of Y on X_2 and Z for a sample that was selected in this manner the α_3 obtained will contain no spurious treatment effect.

Computer simulations of the model with the selection procedure described above were carried out with 2100 and 600 observations. The true ability scores were selected randomly from a population with a mean of 5.0 and a variance of 1.0. The SES values for each observation were set equal to the true ability score plus a disturbance term selected randomly from a normal population with a mean of 0.0 and a variance of 1.0. To determine the treatment group for each observation the data were sorted by SES with the upper third classified as "high SES", the middle third as "middle SES" and the lowest third as "low SES". A random number for each observation was then selected from a rectangular (uniform) distribution with a range of 0.0 to 1.0. This random number was used to determine whether the observation was placed in the experimental group or the control group. For "low SES" observations, the child was assigned to the experimental group if the random number was greater than .10; otherwise he was assigned to the control group. Children in the "middle SES" classification were placed in the experimental group if the random number was greater than .50, and children in the "high SES" group were assigned to the experimental group if the random number was greater than .90. The values of v were selected at random from a normal population with a mean of 0.0 and a variance of 1.0. Posttest scores were then determined from equation (34) where the β values were assigned as $\beta_0 = 0.0$, $\beta_1 = 1.0$, $\beta_2 = .20$, and $\beta_3 = 0.0$; once again the simulation assumes that the true effect of the treatment is nil. A summary of the simulation with 2100 observations appears in Table IV. The pretest-posttest points from the simulation with 600 observations are shown in Figure 4 with the regression lines drawn in. The fitted equation for this simulation is

$$\hat{Y} = \underset{(10.6)}{2.48} + \underset{(16.6)}{.692}\, X_2 + \underset{(.482)}{.037}\, Z, \quad R^2 = .3847.$$

The computer simulations confirm the conclusion that so long as group assignment is random within each SES group there will be no spurious treatment effect.

Although three SES groups were used in this example, there would be no spurious treatment effect regardless of the number of SES groups; the spurious treatment effect is avoided so long as group assignment is random within each SES group.

6. Applications of the Models to the Westinghouse Head Start Evaluation

The primary intention of this paper has been to demonstrate that under some circumstances ex post facto and other quasi-experimental compensatory education evaluations can lead to unbiased estimates of treatment effects. In this section various models are considered for their usefulness in carrying out an evaluation of Head Start based on the data collected for the Westinghouse Learning Croporation-Ohio University study of Head Start (Cicarelli, et al., 1969). The Westinghouse study

Table 4

Statistics for Simulation of One Population-Random Selection Within Each SES Group with 2000 Observations

1. Means, Standard Deviations, and Variances

Variable	Mean	Standard Deviation	Variance
Y	6.01	1.533	2.350
X_2	5.04	1.430	2.044
Z	.494	.500	.250

2. Covariance Matrix

Y	X_2	Z	
2.350	1.361	-.217	Y
	2.044	-.428	X_2
		.250	Z

3. Correlation Matrix

Y	X_2	Z	
1.000	.621	-.354	Y
	1.000	-.599	X_2
		1.000	Z

4. Regression Equation (values in parentheses are t ratios)

$$\hat{Y} = \underset{(18.08)}{2.527} + \underset{(29.81)}{.683X_2} + \underset{(1.26)}{.082Z}$$

F ratio	658.9
R^2	.3859

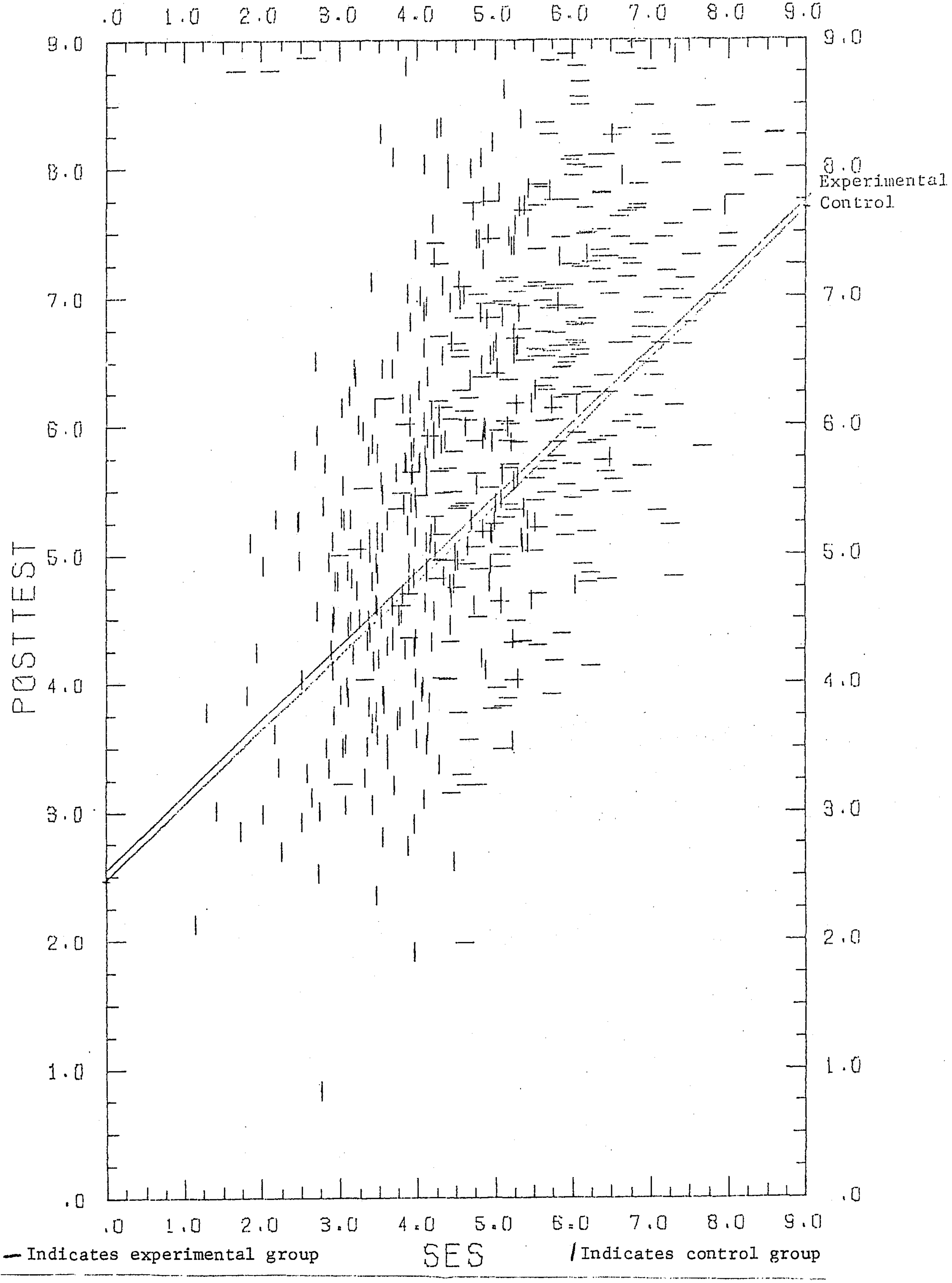
Y = Posttest

X_2 = SES

Z = Dummy variable for treatment

PARTIAL R IS O MODEL

FIGURE 4



was an ex post facto analysis carried out for the Office of Economic Opportunity in 1969 to determine the cognitive and psychological benefits children received from Head Start. The main conclusion of the Westinghouse study was that Head Start had virtually no effect on cognitive test scores. The conclusions of the Westinghouse study have been controversial, with criticisms coming from government officials and academic social scientists. It is beyond the scope of this paper to present a major critique of the study, but some suggestions will be offered describing how the models presented in this paper might be applied to the Westinghouse data.

To carry out the Westinghouse study, a random sample of 300 Head Start centers was selected from a list of all Head Start centers in the country. From the original sample a subsample of 104 centers was selected for analysis. The final sample contained 75 centers with summer Head Start programs and 29 with full-year programs. Eight alumni were selected at random from each center who were in the first, second, and third grades at the time the study was carried out. Then an equal number of control students was selected from the same school by a matching procedure. Each control student was matched to a Head Start participant on the criteria of age, race, sex, grade and kindergarten attendance. Each student was then given a battery of tests to measure cognitive development and attitudes. The parents of the children were interviewed to collect information on the socio-economic and demographic characteristics of the children. In the following discussion we shall only consider models that can be applied to the data collected for evaluating summer Head Start for children who were in the first grade at the time of the study; this minimizes the problems of growth rates and interactions between Head Start alumni and their peers.

Since the Westinghouse data was collected ex post facto, there are no pretests available. If several assumptions are made, however, we can interpret the SES information collected as a type of pretest. The assumptions that must be made to view either a composite index of SES (such as the Hollingshead Index) or a vector of SES variables as a measure of true ability prior to Head Start are: (1) that the SES variable or vector of variables is a function of ability; and (2) that exposure to Head Start does not affect the SES of a child. The first assumption can be stated mathematically as $X_2 = a_0 + a_1X_1^* + u$ in the notation of this paper. If these two assumptions are made we can determine how the models presented in sections 2, 3, and 4 can be applied.

The Campbell-Erlebacher model has the property of being unrefutable ; it assumes that the children in the Head Start group were selected form a less able population than the control group. There is no way of testing such an assertion when participation in the program is voluntary rather than random. This leads Campbell and Erlebacher to conclude that only randomization can avoid bias.

The models presented in sections 3 and 4 can be tested to some extent. To determine if group selection was done on the basis of true ability, a discriminant analysis could be used to determine if the experimental and control groups differed significantly on some relevant measure of SES (either a composite index or a vector of SES variables). The point made by Cicarelli (1970, p. 213) that a small but significant difference in SES leads to a small artifact (bias) is incorrect; to make a statement of this type we must know the structural relationship between SES and ability. It is not difficult to construct a model where the bias would be large. The statement by Evans and Schiller (1970, p. 217) that "Regression artifacts ... would seem to be at a minimum when the matching was carried out on such variables as sex, race, and kindergarten attendance" is incorrect, too. The analysis in section 3 shows that matching on SES as well as their demographic variables would have been preferable to matching on the qualitative variables alone.

The models presented in sections 4 and 5 demonstrate that analysis of the data, collected for the Westinghouse study, will provide unbiased results if particular selection procedures were used, even if there is a significant difference in SES between the Head Start and control groups. Thus, if a discriminant analysis shows a significant difference between the two groups, the evaluator must be prepared to closely reexamine Head Start selection procedures. Murphy (1969) argues that selection within SES groups was not random; however, her assertions are based on personal experience and may not apply to the Head Start population, in general, and to the Westinghouse study, in particular. Clearly more research must be done before any conclusions can be drawn. Campbell has suggested that even if SES was used to determine group membership as in section 4, the SES index used in a regression analysis may be different; in this case the bias that would remain depends upon the reliability of the two SES measures.

7. Conclusions

We have demonstrated that ex post facto analysis and other quasi-experiments can be valuable tools for program evaluation under some circumstances. In general these tools have the advantages of being less costly than random experiments and enable the evaluation to be carried out quickly. For a nonrandom experimental evaluation to be unbiased, however, the evaluators must understand the selection procedures used for the experimental group and choose the control group properly; there are some cases where an ex post facto analysis cannot be undertaken without introducing bias. In defense of the Westinghouse study it should be noted that the study was commissioned by OEO to determine quickly and cheaply a rough idea of how well Head Start was working; OEO has also undertaken a more complete longitudinal study of Head Start which has not yet been completed. The major shortcoming of the Westinghouse study is that the researchers failed to specify the model with which they were implicitly dealing. Campbell and Erlebacher presented one model which showed how the Westinghouse data could lead to biased results. In this paper we have presented several additional models, some which would not lead to bias. It is the responsibility of any evaluator to explicitly state and defend his model; only in this way can we ascertain the absence or presence of bias in evaluations.

BIBLIOGRAPHY

Althauser, Robert P. and Donald Rubin. "Measurement Error and Regression to the Mean in Matched Samples." Social Forces Vol. 50, No. 2 (1971): 206-214.

Bohrnstedt, George W. "Observations on the Measurement of Change," in E.F. Borgatta, ed., Sociological Methodology 1969. New York: Jossey-Bass Inc., 1969.

Breiter, Carl. "Some Persisting Dilemmas in the Measurement of Change," in C.W. Harris, ed., Problems in Measuring Change. Madison: University of Wisconsin Press, 1967.

Cain, Glen G. and Robinson G. Hollister. "The Methodology of Evaluating Social Action Programs." Madison: Institute for Research on Poverty Discussion Paper 42-69.

Campbell, Donald T. "Reforms as Experiments." American Psychologist Vol. 24, No. 4 (1969): 409-429.

Campbell, Donald T. and Albert Erlebacher. "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful," in J. Hellmuth, ed., Compensatory Education: A National Debate, Vol. III of The Disadvantaged Child. New York: Brunner/Mazel, 1970.

Campbell, Donald T. and Albert Erlebacher. "Reply to the Replies," in J. Hellmuth, ed., Compensatory Education: A National Debate, Vol. III of The Disadvantaged Child. New York: Brunner/Mazel, 1970.

Campbell, Donald T. and Julian C. Stanley. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally & Co., 1963.

Cicarelli, Victor G. "Head Start: Brief of the Study," in David G. Hays, ed., Britannica Review of American Education Vol. I. Chicago: Encyclopedia Britannica, 1969.

Cicarelli, Victor G. "The Relevance of the Regression Artifact Problem to the Westinghouse-Ohio Evaluation of Head Start: A Reply to Campbell and Erlebacher," in J. Hellmuth, ed., Compensatory Education: A National Debate, Vol. III of The Disadvantaged Child. New York: Brunner/Mazel, 1970.

Cicarelli, Victor G., John W. Evans, and Jeffry S. Schiller. "The Impact of Head Start: A Reply to the Report Analysis." Harvard Educational Review Vol. 40, No. 1 (1970): 105-129.

Cicarelli, Victor G., et al. The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development Vol. I and Vol. II. A report presented to the Office of Economic Opportunity pursuant to contract B89-4536, June 1969. Westinghouse Learning Corporation, Ohio University.

Cochran, W.G. "Errors of Measurement in Statistics." Technometrics Vol. 10, No. 4 (1968): 637-666.

Cronbach, Lee J. and Lita Furby. "How We Should Measure 'Change'-Or Should We?" Psychological Bulletin Vol. 74, No. 1 (1970): 68-80.

Evans, John W. "Head Start: Comments on the Criticisms," in David G. Hays, ed., Britannica Review of American Education Vol. I. Chicago: Encyclopedia Britannica, 1969.

Evans, John W. and Jeffry Schiller. "How Preoccupation with Possible Regression Artifacts Can Lead to a Faulty Strategy for the Evaluation of Social Action Programs: A Reply to Campbell and Erlebacher," in J. Hellmuth, ed., Compensatory Education: A National Debate Vol. III of The Disadvantaged Child. New York: Brunner/Mazel, 1970.

Garfinkel, Irwin and Edward M. Gramlich. "A Statistical Analysis of the OEO Experiment in Educational Performance Contracting." Unpublished mimeograph, Office of Economic Opportunity, 1972.

Goldberger, Arthur S. "Econometrics and Psychometrics: A Survey of Communalities." Psychometrika Vol. 36, No. 2 (1971): 83-107.

Goldberger, Arthur S. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Madison: Institute for Research on Poverty Discussion Paper, 123-72.

Johnston, J. Econometric Methods. New York: McGraw-Hill Book Co., 1963.

Kmenta, Jan. Elements of Econometrics. New York: The Macmillan Company, 1971.

Lord, Fredrick M. "Elementary Models for Measuring Change," in C.W. Harris, ed., Problems in Measuring Change. Madison: University of Wisconsin Press, 1967.

Lord, Fredrick M. "Large-Sample Covariance Analysis When the Control Variable is Fallible." Journal of the American Statistical Association Vol. 55 (1960): 307-321.

Lord, Fredrick M. "A Paradox in the Interpretation of Group Comparisons." Psychological Bulletin Vol. 68 (1967): 304-305.

Lord, Fredrick M. and Melvin R. Novick. Statistical Theory of Mental Tests. Reading, Pennsylvania: Addison-Wesley, 1968.

Madow, William G. "Head Start: Methodological Critique," in David G. Hays, ed., Britannica Review of American Education Vol. I. Chicago: Encyclopedia Britannica, 1969.

Murphy, Lois Barclay. "Statement to the House Committee on Education and Labor on the Subject of Head Start." Unpublished manuscript, April 1969.

Smith, Marshall S. and Joan S. Bissell. "Report Analysis: The Impact of Head Start." Harvard Educational Review Vol. 40, No. 1 (1970): 51-104.

Webster, Harold and Carl Breiter. "The Reliability of Changes Measured by Mental Test Scores," in C.W. Harris, ed., Problems in Measuring Change. Madison: University of Wisconsin Press, 1967.

Wholey, Joseph S., et al. Federal Evaluation Policy: Analyzing the Effects of Public Programs. Washington, D.C.: The Urban Institute, 1971.