# The Regression Point Displacement Design

## for Evaluating Community-Based Pilot Programs and Demonstration Projects

William M.K. Trochim

Cornell University

Donald T. Campbell

Lehigh University

(in preparation)

This is a rough draft that probably does not deserve quotation. We thought that early dissemination among friends would enhance the final product and help detect the more serious errors and omissions. Comments are appreciated.

Running Head: REGRESSION POINT DISPLACEMENT

## Abstract

The Regression Point Displacement Design (RPDD) is a pretest-posttest quasi-experimental design that usually involves a single treated group and multiple control groups and is usually used on group-level data rather than on data for individuals. The design may be most appropriate when one has routinely-collected multi-wave data for multiple groups and introduces a treatment for one of them. The estimate of the effectiveness of the treatment is accomplished by comparing the displacement of the treated group posttest from the pre-post regression line of the control groups. Although this design has been used on several occasions (and has probably been independently re-invented many times over), it has primarily been viewed as an analysis strategy rather than a design that one might consciously choose to adopt. Yet the design appears to have potential especially for assessing the impact of pilot programs and demonstration projects, where alternative evaluation designs may not be feasible. The validity issues and statistical analysis for this design are described. Several real and hypothetical examples are presented from medicine, education, and community intervention contexts.

# The Regression Point Displacement Design

## for Evaluating Community-Based Pilot Programs and Demonstration Projects

This paper describes an old but neglected quasi-experimental research design. Figure 1 reprints its most widely distributed exemplar, used by Riecken, Boruch, et al. (1974, p. 115), and Cook and Campbell (1979, pp. 143-146), neither source presenting any statistical analysis. They cite Fleiss and Tanur (1973) from whom we borrow Figure 2, in which is claimed an effect significant at the $p < .05$ level. They in turn cite Smith (1957) and Ehrenberg (1968) as at least partial predecessors. In the examples of Figure 1 and 2, the measures employed on the x- and y-axes are quite different. This option is shared with the Regression-Discontinuity (RD) design (Trochim, 1984), which is related to, but distinguishable from, the design described in this paper. Our Figure 3 illustrates this application, drawn from the same dataset as Figure 1. While generally the interpretability of the design will be substantially greater when the same measure is used before and after, we will not argue that this is so for Figure 3.

---

Insert Figures 1-3 about here

---

We envisage a typical application as employing repeated (e.g., annual) rate measures, for cities or some larger or smaller reporting units, with one (or several) units receiving an intensive ameliorative effort not given in the others. Because of this anticipated usage, we shall refer to the x-axis as the "pretest" and the y-axis as the "posttest" in what follows. Some treatment or program, very often likely to be a community-based pilot program or demonstration project, is administered to the treated unit. While in Figure 1 the treated unit was the most extreme on the pretest, the method is applicable no matter where in the distribution of pretest values the "experimental" unit falls. Indeed, the statistical power and interpretability is likely to be better for mid-distribution demonstration sites. The untreated units we will identify as "control units" or "control groups." The analysis fits a regression line to the control units and tests the significance of the departure of the experimental unit from that regression line.

This design has probably been independently reinvented many times over, but was most likely viewed primarily as an analytic technique rather than as a design strategy to which one might aspire. Riecken, Boruch, et al. (1974) designated it the "Posttest-Only Comparison Group Design," Cook and Campbell (1979) the "Quantified Multiple Control Groups, Posttest Only Design." Campbell (1990, cited in Coyle, Boruch and Turner, pp. 143-146)

called it the "Regression Displacement Design." It is our hope to bring some order into this fragmentary tradition by suggesting yet another name, the "Regression Point Displacement Design" (or RPDD in what follows).[1]

Demonstration programs and pilot projects usually receive only very weak and methodologically suspect evaluation. Often there is only a comparison of rates for that one unit for a time period before the special effort and a time period afterward. Or a single comparison unit (e.g., another city similar to the demonstration one) is employed, inevitably differing in many ways even in the very unlikely event that there had been a "random" choice between the pair as to which would receive the program. The potential power of the RPDD comes from using numerous untreated units as "controls," and from not requiring pretreatment equality between any one of them and the experimental unit.

The RPDD remains very much a quasi-experimental design, for which many of the common threats to validity must be examined on the basis of contextual information not included in the statistical analysis. If a statistically significant displacement is shown, there are many other possible causes that need to be considered, over and above the demonstration program.

The RPDD can be illustrated with a simple hypothetical example. Consider a single site at which a treatment is administered. Furthermore, assume that there are arbitrarily ten other sites that will not get the treatment (control sites) but will be measured. All available persons at each site (treated and control) are measured before the treatment and at some specific time after the treatment. Note that due to normal turnover rates and absenteeism at each site, the persons measured at the pretest may not be the same as those measured at the posttest. The resulting data might look like the simulated values depicted in Figure 4.

---

Insert Figure 4 about here

---

The figure shows the linear regression line of the ten control group pretest-posttest pairs of means. The vertical line indicates the posttest displacement or "shift" of the treatment group from the regression line predicted value. In this case, it is visually clear that the displacement probably exceeds the normal variability one might expect around the regression line and indicates a likely treatment effect.

The central idea of the design is that in the null case, one would not expect the treatment group to differ at greater than chance levels from the regression line of the population. There is evidence for a treatment effect when there is a posttest (vertical) **displacement** of the treated group **point** from the control group **regression** line (thus leading to the name "regression point displacement). Of course, this evidence does not imply that the treatment of interest is what led to or "caused" this vertical regression shift. One must, as always, assess the plausibility of other rival potential causes for such a shift.

The RPDD is characterized by four major features: 1) the use of a single treated unit instead of many; 2) the use of aggregate-level data instead of individual-level; 3) the absence of any need to assure (or attempt to achieve by statistical adjustment) pre-treatment equality between treated and control groups; and 4) the avoidance of "regression artifacts" or underadjustment due to "errors in variables" by employing the observed regression line, rather than using it as a means of adjustment. Feature 4, and to a lesser extent 3, are shared with the Regression-Discontinuity (RD) design. The first two are not absolutely necessary. We would still probably classify a study with two or even three demonstration sites as an RPDD[2]. The key distinction is that in alternative designs there are enough points to allow one to fit the same model to both groups, whereas the RPDD typically does not. For instance, in the RD design, one usually has enough points in both groups to be able to estimate a within-group slope, whereas in the RPDD, that is not possible or justifiable, given the few available treatment group points. The higher aggregation level (e.g., cities) is also a typical RPDD characteristic although it is by no means required. One can envision an RPDD design involving only a single person who receives a treatment, with multiple control persons.

In the exposition that follows, we first treat tests of significance and curve-fitting issues, using the examples of Figures 1 through 4. Next, we relate the RPDD to other experimental designs, making explicit its requirements. Then, we present explorations of applications of the design, using secondary analyses of real data and hypothetical examples. Finally, we deal more systematically with threates to validity not controlled for by tests of significance including those related to the statistical model chosen.

## Statistical Analysis and Power Issues in the RPDD

Traditionally, statisticians have treated data like that of the RPDD as a problem of testing whether an individual point deviates significantly from its regression line prediction, that is, whether its vertical distance from

the regression line is statistically significant. Here, this approach is labelled the "individual point prediction" method. While this is technically correct, it is a cumbersome procedure that requires applying a special t-test formula after computing the control group regression line in order to test the deviation hypothesis. A simpler computational procedure that yields exactly the same result with greater potential generality will be recommended instead and is discussed here as the "analysis of covariance" method. The statistical proof of the equivalence of these two analytic methods is given below.

The Individual Point Prediction (IPP) Method. To analyze the data from a RPDD using the IPP approach, following Fleiss and Tanur (1972, p. 523, Formula 25) one first obtains the regression of posttest means onto pretest means for the control units only. Given a set of pretest values (e.g., means), $X_i$, and corresponding posttest values, $Y_i$, one typically (although not necessarily) fits the linear model[3]

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

This equation can then be used to obtain the predicted posttest value for the treated unit. That is, if the treated unit has pretest value $X_0$, and posttest value $Y_0$, then the predicted posttest value is

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

To test the hypothesis that the observed treatment unit posttest value differs from the predicted one, a simple t-test can be constructed:

$$t = \frac{Y_0 - \hat{Y}_0}{\sqrt{\sigma^2_{Y_i - \hat{Y}_i}\left[1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}\right]}} \qquad (1)$$

where

$Y_0$ = the observed treated unit posttest value

$\hat{Y}_0$ = the predicted treated unit posttest value

$\sigma^2_{Y_i - \hat{Y}_i}$ = the Residual Mean Square

$N$ = the number of control units

$X_0$ = the observed treated unit pretest value

$\bar{X}$ = the mean of the control unit pretest values

$$\Sigma x_i^2 = \Sigma(X_i - \bar{X})^2$$

This t-value can be tested in the usual way with df=N-2. The numerator is simply the treated unit's vertical displacement from the control unit regression line. Note that the treated unit's pretest value enters into the formula through the term in the denominator, $X_0 - \bar{X}$, which is the horizontal distance from the treated unit's pretest value to the pretest mean for all control units. This term increases as the treated unit is more extreme on the pretest. The formula used here is the standard t-test for testing the significance of the prediction of an individual point in a regression analysis. The numerator is simply the difference between the observed and predicted point. The denominator is the formula for the standard error of $\hat{Y}_0$ which is shown in Gujarati (1988, p120, Formula 5.10.6), Johnston (1972, p.42, Formula 2-53), Maddala (1977, p. 82), Draper and Smith (1981, p30, Formula 1.4.11), Mood (1950, p.298, Formula 3) and no doubt elsewhere.

The Analysis of Covariance (ANACOVA) Method. Although the IPP method is statistically correct, it is somewhat computationally inconvenient to first run the regression analysis for the control cases and then construct the post hoc t-test. It turns out that there is a simpler approach that yields identical results. To accomplish the same linear analysis as described above one can fit the following regression model to all of the pre-post values (the control units and the single treatment unit) simultaneously:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i \qquad (2)$$

where:

$Y_i$ = the posttest value for unit i

$X_i$ = the pretest value for unit i

$Z_i$ = 0 if the X,Y pair is for a control unit

1 if the X,Y pair is for the treated unit

$\beta_0$ = intercept term

$\beta_1$ = linear slope

$\beta_2$ = the treatment effect (vertical shift from the regression line)

$e_i$ = the residual

The reader will recognize this as the traditional Analysis of Covariance (ANACOVA) model. The major difference between this application of ANACOVA and the typical use is that in the basic RPDD the treatment group consists

of a single X,Y pair whereas it typically consists of a number of such treatment pairs. The major hypothesis of interest is:

$$H_0: \beta_2 = 0$$

tested against the alternative:

$$H_1: \beta_2 \neq 0$$

when stated as a two-tailed test. This hypothesis is tested using a t-test that is routinely reported in the table of regression coefficients as part of standard regression analysis output (e.g., SAS Institute Inc., 1982, p. 39-83; SPSS Inc., 1983, p. 601-621).

Although both the IPP and ANACOVA models are presented here for only the linear case, the addition of appropriate curvilinear terms into either model would enable testing of treatment effects in the presence of more complex curvilinear pre-post relationships.

Statistical Proof for Equivalence of IPP and ANACOVA Approaches. Here, the equivalence of the two approaches outlined above is shown by statistical derivation. This proof assumes that one regresses a variable, $Y_i$, onto another variable, $X_i$, in a sample of N cases, where the individual point of interest has been excluded from the regression. The hypothesis of interest is whether one specific point $X_0,Y_0$ differs significantly from the N cases. As mentioned earlier, numerous regression texts assert correctly that we can test this question with a post hoc t-test constructed from the results of the regression analysis. The following formula is equivalent to the one given earlier, but expressed slightly differently for convenience:

$$t = \frac{e_0}{\sqrt{MSE\left[1 + \frac{1}{N} + \frac{d^2}{S}\right]}} \qquad (3)$$

where:

$t$ = the test statistic

$e_0$ = $Y_0 - \hat{Y}_0$

MSE = the usual mean squared error in the sample of N cases

$d$ = $X_0 - \bar{X}$

$S$ = $\Sigma x_1^2$

The purpose of this discussion is to show that the t-value found from this formula equals the t-value found by defining a dummy variable, $Z_i$, scored 1 for case $X_0,Y_0$ and 0 for the other N cases, using multiple regression to regress $Y_i$ onto $X_i$ and $Z_i$ in the total sample of N+1 cases, and taking the t-value which tests the partial relationship between Z and Y. This fact does not merely provide an easier way to apply an existing test. It also allows us to generalize formula (3) to any number of covariates (not just a single one) and to the simultaneously testing of more than one point of interest.

In the multiple regression just described, case $X_0,Y_0$ is predicted perfectly because it has its own variable $Z_0$ devoted entirely to that task; the regression formulas will set the partial regression slope, $\beta_1$, to whatever value makes $\hat{Y}_0 = Y_0$. Thus, $\beta_2 = Y_0 - \hat{Y}_0 = e_0$ exactly.

Let $SE(\beta_2)$ denote the standard error of $\beta_2$. Inspection of Formula (3) shows that the problem reduces to proving that

$$SE^2(\beta_2) = MSE\left[1 + \frac{1}{N} + \frac{d^2}{S}\right] \quad (4)$$

Standard regression textbooks (e.g., Darlington, 1990, p. 126) assert:

$$SE^2(\beta_2) = \frac{MSE}{N' * Var(Z_i) * Tol(Z_i)} \quad (5)$$

where:

N' = the size of the sample including case $X_0,Y_0$, so N' = N + 1

$Var(Z_i)$ = the variance of $Z_i$ usinf N' rather than N'-1 in the denominator

$Tol(Z_i)$ = the tolerance of $Z_i$, which equals $1 - r_{ZX}^2$, where $r_{ZX}$ is the correlation between $Z_i$ and $X_i$.

If we define

$$A = N' * Var(Z_i) * Tol(Z_i) \quad (6)$$

then inspection of formulas (4) and (5) shows that the problem reduces to proving that

$$\frac{1}{A} = 1 + \frac{1}{N} + \frac{d^2}{S} \quad (7)$$

since $Z_i$ is a 0/1 dichotomy, its variance equals pq, where p and q are the proportions in the two categories. These proportions are respectively 1/N' and N/N'. Since N' = N+1, we have $Var(Z_i) = N/(N+1)^2$. Thus (6) becomes

$$A = Tol(Z_i) * \frac{N}{N+1} \quad (8)$$

Consider a regression predicting $X_i$ from the dichotomy $Z_i$. $X_0$ is perfectly predicted while the estimated $X_i$ from the other N cases is the mean ($\bar{X}$) of those N cases. Thus the sum of squared errors in this regression is S. Therefore

$$Tol(Z_i) = 1 - r_{ZX}^2 = \frac{S}{S'} \quad (9)$$

where S' denotes the sum of squared deviations of all N' scores from their common mean on X. We leave to the reader the proof that

$$S' = S + d^2\left[\frac{N}{N+1}\right] \quad (10)$$

Substituting (10) into (9) and then (9) into (8) gives

$$A = \left[\frac{N}{N+1}\right] * \frac{S}{S + d^2\left[\frac{N}{N+1}\right]} \quad (11)$$

Thus

$$\frac{1}{A} = \frac{N+1}{N} * \left[1 + d^2 * \frac{N}{(N+1)S}\right] \quad (12)$$

$$= \frac{N+1}{N} + \frac{d^2}{S}$$

$$= 1 + \frac{1}{N} + \frac{d^2}{S}$$

As mentioned in (7), this is the result we sought.

It is also worth noting that these same values of t are the Studentized residuals or t-residuals that are widely used to test whether a given case is an "outlier" -- that is, whether its Y-value is above or below the range we would expect from its X.[4]

Analytic Example. To illustrate how each analysis operates and to demonstrate their equivalence, the simulated data used to generate Figure 4 are analyzed. The data are given in Table 1. The descriptive statistics for the pretest and posttest means for the ten control units are shown in Table 2. To accomplish the IPP analysis it is first necessary to compute the regression of posttest means onto pretest means for the control groups only. The results of this regression are given in Table 3.

---

Insert Tables 1-3 about here

---

The regression equation is

$$Y_i = 0.923264655 + .991693461X_i$$

and we wish to obtain the predicted $\hat{Y}_0$ value for $X_0 = 51.135$, that is

$$\hat{Y}_0 = 0.923264655 + .991693461(51.135)$$

$$= 51.6335098$$

To test the hypothesis that the observed treatment group's posttest mean differs from this predicted one, the t-test formula (1) is applied:

$$t = \frac{Y_0 - \hat{Y}_0}{\sqrt{\sigma^2\left[1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}\right]}}$$

where

$Y_0$ = is the observed treatment group posttest mean (from Table 1) = 56.622

$\hat{Y}_0$ = is the predicted treatment group posttest mean obtained as shown above using the regression equation in Table 3 = 51.6335098

$\sigma^2$ = is the Residual Mean Square (from Table 3) = 1.806961792

N = is the sample size (i.e., N = 10 control groups)

$X_0$ = is the observed treatment group pretest mean (from Table 1) = 51.135

$\bar{X}$ = is the mean of the ten control group pretest means (from Table 2) = 49.9621648

$\Sigma x_i^2$ = can be obtained by multiplying the variance of the control group pretest means (Table 2) by N-1 = 28.66829164 x 9 = 258.014625

Substituting from the results in Tables 1 to 3 we obtain:

$$t = \frac{56.622 - 51.6335098}{\sqrt{1.806961792\left[1 + \frac{1}{10} + \frac{(51.135 - 49.9621648)^2}{258.014625}\right]}}$$

$$= \frac{56.622 - 51.6335098}{\sqrt{1.99729135}}$$

$$= \frac{4.98849022}{1.41325559}$$

$$= 3.52978631$$

which can then be tested in the usual way where df=(N-2)=8.

Compare this approach with the ANACOVA method. The results for the regression model are shown in Table 4.

---

Insert Table 4 about here

---

The coefficient of interest is associated with the Z term and is exactly the same as the numerator in the t-test above (i.e., 4.988490226). The standard error of this coefficient is equal to the denominator in the individual point prediction analysis. The t-value associated with that coefficient is 3.529786316 and is exactly the t-value reported above. The standard regression analysis printout automatically displays these values. The table further shows that the probability associated with that t-value is .0077, which indicates that the treatment group posttest shift is significantly different from its predicted value if $\alpha = .05$ (two-tailed test) or $\alpha = .025$ (one-tailed test).

The conclusion from this detailed analytic example corroborates the proof provided earlier and demonstrates that either analysis method (IPP or ANACOVA) will yield the exact same result (allowing for some slight rounding error in various calculation routines). There are several reasons for preferring the ANACOVA method. First, as seen here, it is computationally simple to obtain. No post hoc computations need be made. Second, it puts the analysis within a broader framework of general regression modeling instead of construing it as an instance of the special case of individual point prediction. Finally, it enables more complex versions of the RPDD. For instance, it would easily accommodate the case of more than one treatment group whereas the individual point prediction analysis does not. Because of the equivalence of these analyses, all remaining results reported in this paper are from the ANACOVA method.

To illustrate application of the tests of significance and their problems, we now apply them to the examples of Figures 1 through 4.

The Medicaid Regression Point Displacement Design

The first example comes from the Medicaid study discussed in Rieken, Boruch, et al (1974, p. 115) and Cook and Campbell (1979, pp. 143-146) part of which was shown in Figures 1 and 3 above. The original data are shown in Figure 5 (taken from Lohr, 1972; Wilder, 1972, p. 5, Table B).

Insert Figure 5 about here

The figure shows the average number of physician visits per person per year in the United States for the years 1964, 1967 and 1969, broken out by family income ranges. The Medicaid program was introduced during 1964. The legislation mandated that only those families with an annual income under $3,000 were eligible to receive Medicaid. Overall, it appears that the annual average number of physician visits is declining for most income groups with two notable exceptions -- the lowest income group shows an increase over both time intervals and the second lowest group increases between 1967 and 1969.

The central question is whether the introduction of Medicaid is associated with a significant increase in the average number of physician visits per year. Several RPDDs can be constructed from these data. The first is identical to that shown in Riecken, Boruch et al. (1974, p. 115) and Cook and Campbell (1979, pp. 143-146) and displays income group along the horizontal axis and physician visits on the vertical as shown earlier in Figure 1. One problem in analyzing these data concerns the metric for the pretest. We know that income distributions tend to be non-normal and consequently we may need to transform the pretest variable before conducting the analysis. We also see that the highest income group has no upper income limit given. To analyze these data, we decided to use the logarithm of the upper and lower limit for each pretest income interval (with an upper limit for the high pretest group set arbitrarily at $50,000 and the lower limit for the low income group set to $1000) and then use the midpoint between these logs as the pretest value for each group. The transformed data are graphed in Figure 6. The ANACOVA estimate of effect (in log pretest units) is $\beta_2 = .824$ ($t=21.03$, $p=.0002$).

Insert Figure 6 about here

A second RPDD can be constructed from these same data by graphing posttest (1967) physician visits against pretest (1964) ones for each income group as shown earlier in Figure 3. The lowest income group is by definition the Medicaid treatment group indicated by an 'x' on the graph. The other income groups are shown with an 'o' and can be considered comparison or control groups. The Medicaid group had the lowest pretest average number of

physician visits. The question is whether their posttest level is significantly higher than would be predicted given the control group pre-post levels. The ANACOVA estimate of effect is $\beta_2 = .479$ (t=3.63, p=.036).

Medicaid appears to be associated with a significant rise in annual physician visits, but one still cannot conclude that Medicaid is what caused this rise. In order to reach this conclusion, one has to rule out any plausible alternative causal explanations for the observed effect (Cook and Campbell, 1979). Several possibilities suggest themselves. First, it could be that the regression line that is fitted to the data does not accurately reflect the true regression for the population in question. The question is whether the apparent significant effect results from specification of the wrong regression model. This could arise in some contexts because the control groups do not represent the population of interest, an unlikely event here because the control group means include the entire U.S. population income ranges (although the use of a single group to indicate the annual physician visits for all persons with incomes above $15,000 may very well distort the shape of the graph). The more plausible problem is that the control group pre-post relationship is not linear in the population, but is instead quadratic or some other functional form. Following Darlington (1990, p.295), the polynomial regression (including both the linear and quadratic terms in the regression model) is fitted to the data. The resulting equation is:

$$Y_i = -13.56 + 6.6X_i + -.59X_i^2$$

Neither of the X-coefficients is statistically significant (at $p<.05$) a result that is probably attributable to the small number of points and the consequent low statistical power associated with each estimate (note, however, that the linear term alone is significant in the original ANACOVA model). This linear plus quadratic regression line is shown in Figure 7.

---

Insert Figure 7 about here

---

It is clear from the figure that the polynomial regression fits the control group points better than the linear one does. However, it is impossible to know whether it is the better model for the population with so few points available for prediction (remember, than one could fit the observed points perfectly with a fourth-order polynomial model). In judging plausibility one must examine the assumption that the straight line fit is misleading, in that a curvilinear plot would reduce the departure from expectancy. In this example, even if the linear model is not the best, it is clear

from visual inspection that no reasonable model would rule out the sensibility of concluding that there is a significant effect for the Medicaid group. The linear fit reduces the $Y_0 - \hat{Y}_0$ discrepancy from that which any reasonable curvilinear fit would produce. With the polynomial model, the observed treatment group posttest will be even further from the predicted value. We feel that in this case, the linear fit is conservative, rather than misleading. A curvilinear fit should of course be the privileged fit in cases where it reduces the apparent effect. Note that these problems are minimized where the experimental unit falls in the middle of the controls in as much as the $Y_0 - \hat{Y}_0$ discrepancy will vary little as a function of the curvilinear plot chosen.

A second threat to the causal inference is an issue of internal validity. There may have been some other factor affecting only the lowest income group that increased their annual physician visit rate. For instance, if there is another low income subsidy program, such as the WIC nutritional supplement program, it may be that the rise in physician visits is attributable to the increased gynecological care subsidized by that program rather than to the Medicaid subsidies. There is no way to rule out that threat with these data although one could examine it by comparing physician visit rates for WIC versus non-WIC recipients if such data were available. The problem in this context is that there are likely to be many federal or state programs that are targeted to the lowest income group (i.e., those who fall below the official poverty line). Although we would usually argue that an RPDD that gives the treatment to an extreme case is preferable to uncontrolled treatment assignment, it is not preferred when the implicit cutoff is a well-publicized, frequently used value such as the Federal poverty level, which is used as the criterion for assignment in many national programs that constitute alternative potential causes for any observed treatment effect. Thus, in this case, it is impossible to be confident that the observed effect is due to Medicaid alone. Nor does the apparent jump in physician visits for the next highest income group from 1967 to 1970 (evident in Figure 5) solve the problem because it too may very well result from raises in the Medicaid eligibility cutoff values, raises in the official poverty income level (and the consequent eligibility for other Federal programs), or both.

Schizophrenic Reaction Time Study

Fleiss and Tanur (1972) described a version of a Regression Point Displacement analysis that explored reaction time in schizophrenics. The purpose of this analysis was to examine whether clear-cut schizophrenics differ from other groups in their crossmodal-ipsimodal reaction time difference. The data are shown in Figure 2 presented earlier. The ANACOVA estimate of effect is $\beta_2 = 16.11$ ($t=109.59$, $p=.0058$).

Many readers will join us in being surprised that such power can be obtained from an application with just 3 control group points. There are several variables affecting such a p-value. One is the degrees of freedom (in this case, N - 2 = 1 df). A second one is the dispersal of values in the control groups from the fitted line. A third is the magnitude of the departure of the treated group from the center of the fitted line (i.e., the error term is larger the greater the distance of the treatment group pretest mean from the mean of the control group average).

Let us consider the second component. In this case, the linear fit is essentially perfect, producing a very small error term, and hence a very large t-value. But with only 3 points, are not such perfect fits going to happen by chance very frequently? Or, to put it another way, in repeated samplings from the same universe, is not the error-term going to fluctuate widely from replication to replication? Can we be sure that the small-sample values for the t-test are still appropriate for this application, for df=1? Note that Fleiss and Tanur failed to find a significant effect when using within-group variance, in contrast to the Figure 2 analysis using only group means. We would argue that the relative power of an RPDD can never be greater than the power of a more micro-level analysis (e.g., using individual data points instead of group means) on which it is based, even though we may serendipitously find extremely significant estimates in a given RPDD as in this example. Thus, while our presentation here has been inspired in part by Fleiss and Tanur's (1972) seminal paper, we regard the specific application with caution just because the perfect linear fit is so out of line with ordinary experience. Someone interested in applying their finding would be well advised to explore the relationship between the ipsimodal versus crossmodal-ipsimodal differences over a larger number of diagnostic groupings in an effort to get a more plausible error term.

## Requirements for the RPDD

To qualify as an RPDD, there must be multiple comparison or control groups and pre-post measurement. But, given this restriction, there are many alternative versions of the design that are possible. Many of the variations can be described in terms of five major dimensions, where a different RPDD can be constructed for different combinations of these dimensions. The dimensions are:

1. Method of Assignment of Treated Unit. For almost any two-group pre-post design it is possible to construct a RPDD analogue. If the single experimental unit is randomly assigned (from a pool of potential candidate units), the RPDD is analogous to a Randomized Experimental (RE) design. When the experimental unit is chosen because it has the most extreme value (high or low) on the pretest, this

is essentially equivalent to assignment by a cutoff (the cutoff in this case is usually implicit and consists of the pretest values that distinguish this extreme case from the others), making the RPDD analogous to the Regression-Discontinuity (RD) design. Finally, when the treatment group is chosen arbitrarily, for political reasons or personal favoritism, or for any other unspecified reason, we can consider the assignment to be by an unknown or unspecifiable rule, and the RPDD is most analogous to a Non-Equivalent Group Design (NEGD). For most of the experimental or quasi-experimental designs, it is possible to construct RPDD analogues, and useful to do so because consideration of analogous designs and the literatures which have grown around them will help raise validity issues that ought to be considered in the RPDD analogue.

2. The Unit of Measurement. The unit of measurement refers to the entity represented in each pre-post point in a RPDD. Usually, these will be broad units -- states, cities, communities, SES groups, diagnostic groups -- not individual persons. However, a RPDD design can be constructed using either aggregated *individual* data or *group* data. For instance, many readily available databases consist of already aggregated frequencies, rates, proportions or averages across geographically or demographically defined groups. In the Medicaid example of Figures 1 and 3, the average number of physician visits per year per person for six different income groups is used. The RPDD analysis does not require physician visit rates by individual (nor changes in such rates) -- it operates in this case on the group averages. Restricting the data analysed to repeated measures from the same individuals adds power to some statistical analyses. The distinction between this case and that of unmatched persons at the two points in time is not formally made in the RPDD. However, if the group means in each case are based upon the same persons in each year, we might expect a smaller error term (see Cook and Campbell, 1979, p. 115-117). For instance, if a survey is conducted in a number of communities before and after an educational intervention in one of the communities, the people measured on the pretreatment survey are not likely to be the same as those measured afterwards (unmatched). The RPDD is perhaps the strongest design available for studying community-level interventions where different persons are sampled on each occasion. It is also possible to use the RPDD when the pretest and posttest scores are based on the same person. This is illustrated in the reanalysis of the Xanax

clinical drug trial below. There, the same (matched) patients are measured before and after, but for the RPDD, all such patients at a given site (or clinic, hospital) constitute a unit and their average scores are used. This pre-post same-individual RPDD might arise where the same students within a school are measured before and after some treatment is implemented in a single classroom. Classroom average scores could be used in conducting a RPDD analysis or individual scores (grouping all control group cases together) could be used in a more traditional nonequivalent group analysis. One must be careful in using data from separate pre and post samples because of the potential bias that can arise. For instance, if pretest and posttest average scores for a group are used, it is likely that those non-dropouts present on the posttest are unrepresentative of the original pretest group. Contrast this with two different, but repeated random samplings over time from the same community. While even this may be biased, in the sense that the basic demographic structure of the community may have evolved between measurements, it is much less plausible in this case, especially when the time span is relatively short. Random, rather than opportunistic, sampling on each occasion can help to assure some degree of equivalence when repeated measures are not obtained for the same group of people.

3. The Number of Treatment Groups. In the simplest case, the RPDD involves only a single treated group or site. The treatment would be administered in one community or classroom. When the design involves enough treated points that it is reasonable to estimate the same functional form as for the controls, the RPDD essentially transforms into one of the other pre-post designs -- randomized experiment, nonequivalent group design, or regression-discontinuity -- depending on the method used to assign units to treatment or control condition.

4. The Same versus Different Pre-Post Measures. Generally, the same variable is measured before and after the treatment (*matched measures*), but different measures can be used. For instance, if one is looking at the effect of an educational program, it might not make sense to measure content-related performance on the pretest because students would not be expected to know any of the content (and may not even understand the questions). One might use a general measure of prior intelligence or academic achievement (GPA, standardized achievement test scores) as the pre measure, with a treatment content-specific outcome measure. Thus, pretest and posttest in this case are different, or

*unmatched*, measures. In Figure 1, the pre and post measures are still less similar. Whether matched or not, statistical power is likely to be greater when the pre-measure has a strong linear or monotonic relation with the outcome variable.

5. The Number of Covariates. In the simplest case, the RPDD uses a *single pretreatment variable*. But it is also possible to use *multiple pretreatment variables* that can simultaneously be entered as covariates inthe model. The major problem with multiple pretreatment covariates is that, since each covariate costs one degree of freedom, using multiple covariates requires more control groups. However, the use of multiple covariates when many control groups exist will be an important mechanism for improving the statistical power and efficiency of the treatment effect estimate.

## Illustrations of the RPDD in Secondary Analyses

### The Xanax Randomized Clinical Trial Study

The RPDD can be illustrated using data from studies that were already conducted under different design structures (even though we will not necessarily recommend the RPDD as superior to the original design). Here, data from the Cross-National Collaborative Study (Klerman et al, 1986; Klerman, 1988; Ballenger et al, 1988; Noyes et al, 1988; Pecknold et al, 1988) of the effectiveness of Xanax (Alprazolam) is used to illustrate the RPDD. This was a two-phase multinational, multicenter randomized clinical trial (RCT) to evaluate the drug treatment of panic disorder and associated agoraphobia. Only data from Phase I which compared Xanax with a placebo is used here. It is not the purpose of this reanalysis to assess efficacy across the range of possible outcome measures or even to develop a definitive substantive test of the efficacy hypothesis for these data. Rather, the data enable exploration of the RPDD. Although the original study was ten weeks long and consisted of a screening week, baseline week and eight week study period, the reanalysis reported here is limited to the baseline and first study week. Thus, this reanalysis only examines the immediate (i.e., one-week) effects of Xanax versus Placebo. Only one of the many measures collected in the original RCT is used in this reanalysis. Specifically, the Sheehan Clinician Rated Anxiety Scale (CRAS) (Sheehan et al, 1980), which consists of clinician's ratings of 35 items (e.g., "spells of imbalance", "tires easily", "tension/nervousness/anxiety") rated on a 0 (Absent) to 4 (Very Severe) scale was selected for use in these reanalyses. The average Sheehan CRAS across the 35 items is used here. The pre-post bivariate distribution for the CRAS measure for all patients is shown in Figure 8.

Insert Figure 8 about here

An ANACOVA analysis of these individual data indicated that there was a significant treatment effect of -0.368 CRAS units (SE = 0.037, t=9.82). The full study involved eight participating physician investigators. For each investigator, eligible patients were randomly assigned to either the treatment (Xanax) or to the placebo group. To construct a hypothetical RPDD from these data, the investigators can be treated as separate groups or sites (RPDD units). The pretreatment and posttreatment means for the eight physicians are shown in Figure 9.

Insert Figure 9 about here

The treatment group means are marked with an 'X' while the control groups are shown with an 'O'. The number next to each point identifies the eight investigators. The figure clearly shows the strong overall positive treatment effect, that is, the investigator-averaged CRAS anxiety scores were lower on the posttest for the Xanax groups than for the placebo groups. In the RPDD there is usually one treated site and multiple control groups. To construct such data in this illustrative case, the investigator who had the most patients was arbitrarily designated the RPDD treatment site and the remaining seven were considered control sites. Then, the control group mean for the treated site and the treatment group means for the control sites were discarded. The remaining means which constitute the RPDD are shown in Figure 10.

Insert Figure 10 about here

The figure clearly shows that the treated group posttest mean falls significantly below the control group regression line. The ANACOVA estimate of effect is $\beta_2$ = -.487 (t=-3.069, p=.0278).

In this context, it is worth noting that the treatment is not "pure" in that drug administration is confounded with both site and investigator. Here, for instance, an especially effective psychotherapist might cause significant

reductions in anxiety even though the drug is ineffective. This confounding cannot be disentangled under the one treatment group RPDD and, consequently, it is always preferable to utilize multiple treatment groups where possible. One must be cautious in interpreting treatment effect estimates from the RPDD because of this inherent confounding.

This is, of course, only one of eight such hypothetical RPDDs that can be constructed post hoc from the data in Figure 9. The results for all eight RPDDs are summarized in Table 5.

---

Insert Table 5 about here

---

The first line of the table gives the results for the original ANACOVA on individual-patient data as reported above. We can construct a 95% confidence interval for that estimate as -.368079207±2(.037467215) or -.443013637 to -.293144777. This is followed by the estimates for each of the eight possible investigator-based RPDDs, where the one for Investigator 4 was shown above. All t-values are in the predicted direction and range from -1.391 for Investigator 3 to the largest of -3.069 for Investigator 4 as reported earlier. Four of the estimates fall within the confidence interval of the individual level ANACOVA and four do not. The average of the eight separate RPDD treatment effects is -0.349 and the average t-value is -2.172. The final analysis reported in Table 18 shows that when the entire set of eight pre-post means is entered into an ANACOVA analysis the resulting treatment effect estimate is -.3599 (t=-6.09, p=.0001) which compares quite favorably with the ANACOVA estimate from the patient level data of -0.368 CRAS units (SE = 0.037). It is worth noting that the SEs for the individual RPDDs are five to six times higher than the SE for the individual-level ANACOVA. Thus, the loss in statistical efficiency is appreciable and would lead us to reject the hypothesis in half of the eight possible RPDDs. It is also worth noting that the SE for the ANACOVA on the eight pre-post investigator means is only about twice as great as for the individual-level ANACOVA -- a considerable gain for an addition of seven values to the model.

The Effects of Peer Intervention on Physician Practice Rates

One context in which the RPDD appears to have great potential is in the analysis of health and medical data. There, large-scale medical insurance and health databases are often available describing utilization by geographic areas. Wennberg (1990) has used such databases in conducting "small area analysis" which allows one to compare

surgical or health resource utilization across geographic areas in order to detect variations in medical practice. When certain geographic units (e.g., cities, counties) demonstrate relatively high utilization rates, one might suspect that this results in part from physician preference and practice in that location rather than from higher actual disease incidence rates. Under these circumstances, a key question is whether feedback to physicians regarding the high utilization rates might lead to a reduction due to changes in physicians' discretionary practices.

Wennberg (1990) describes such an example in the context of obstetric and gynecological practice in Maine. Small area analysis indicated that one region of the state had particularly high hysterectomy rates relative to the state average. The Maine Medical Assessment Program (MMAP) was designed to monitor and provide feedback regarding such anomalies in surgical rates. "Beginning in 1979, members of the State association of obstetrics and gynecology met with the physicians practicing in area II to discuss the indications for hysterectomy and present data for their area. Subsequently, the hysterectomy rates dropped close to the State average where they have remained." (Wennberg, 1990, p. 183)

Subsequent to this initial study, the MMAP has continued to collect practice rate data and to intervene with peer physicians where it appeared to be warranted by the rates. An example of such an intervention case is shown in Figure ---. (Continue with example)

It would be a simple matter, in this context, to conduct a RPDD analysis of the data to explicitly test the hypothesis that the MMAP intervention is associated with a significant subsequent decrease in hysterectomy rates. The pretest would consist of hysterectomy rates for the year prior to the intervention; the posttest would be the rates for the following year. The treatment group would be defined as the geographical region that received the feedback about their relatively high surgical rates. Each of the other regions in the State (and perhaps in other States) could be used as control groups.

It is difficult to conceive of way to address the surgical rate hypothesis that would make better use of the readily available data. One could enhance the analysis by including geographically-based demographic covariates (assuming there are a sufficient number of control areas to allow for such an analysis). Or, if the data are available, it would be possible to do an analysis at the physician level, assuming that feedback is only provided for those physicians with significantly higher than average surgery rates.

There are likely to be many instances in which interventions of this type could be studied using the RPDD on readily available data from insurance or health agency databases. Small Area Analysis has already proven useful in identifying unexpected medical practice variations. The RPDD can, in many such contexts, be a useful and efficient method for assessing the effects of interventions designed to ameliorate the more extreme practice cases identified through Small Area Analysis (Hsiao, 1994).

## Hypothetical Regression Point Displacement Examples

To further explore the RPDD several hypothetical examples are constructed. These help illustrate the wide variety of contexts in which the design might prove useful and provide additional cases which point to some of the likely threats to validity that may arise.

### A Hypothetical Educational Program Evaluation

The RPDD has great potential for studying organizational-level treatment interventions. For instance, consider a school district that wishes to make a district-wide change in its educational approach. Assume that the district wishes to increase the number of school days per year from 180 to 220. This would presumably involve a major change in scheduling, teacher contract hours, curriculum redesign, and so on. A key question of interest would be the effects of such a change (or combination of changes) on academic performance of students, teacher turnover rates, student attendance rates, and so on. These outcome measures are routinely collected for this and most other school districts and could be analyzed within a RPDD framework.

The advantage of using an RPDD in this context is that it capitalizes on readily available databases. One could take annual achievement testing data for this school district and any number of other local school districts from the Spring testing of the year immediately preceding the schedule change and for the Spring at the end of that year. At each grade level, a separate RPDD analysis could be constructed using the average test scores for the grade. The RPDD analysis would be easy to compute from reported averages. Contrast this with the more laborious procedure of analyzing individual student-level data. There, one would have to match the pre and posttest data for each student, probably discarding data for the many students who would not have a test on both occasions. This is a time-consuming, difficult process that can result in a significant lowering of sample size due to the high rates at which students move into or out of the district each year, absenteeism on either testing, and so on. The RPDD analysis would include in the average scores all students present on either occasion without requiring that

scores be matched. This might lead to a slight underestimate of the effect of the schedule change because the posttest average for the treated district will have some percentage of students who come into the district during the course of the treatment year. Nevertheless, this potential for bias must be contrasted with the high costs of conducting individual student level analyses that are likely to be at least as equivocal.

Another hypothesis that would be interesting here concerns student attendance rates. A school district that changes to a longer school year might be concerned that the additional school days could be offset by higher student absenteeism rates. Districts routinely collect and report such rates by grade level. Here, an RPDD analysis could be used to examine the hypothesis that there is a greater proportion of missed school days under the new calendar compared with the standard academic calendar. A student-level analysis of attendance rates would be difficult to implement especially where school districts only retain data by classroom or grade level in their databases. The RPDD analysis could be performed on the more readily-obtainable summary data for school districts. The pretest would consist of the percentage of days missed during the year prior to the schedule change (probably by grade level). The posttest would be the same data for the year of the change.

These educational examples indicate some of the key trade-offs with the RPDD. In some situations, it is costly or difficult enough to get adequate matched pre-post data at the individual level that we might turn to the RPDD as a reasonable alternative. This would be the preferred alternative especially as aggregate data has higher reliability, one is able to accumulate larger numbers of control groups, or the treatment effect is expected to be stronger.

A Hypothetical Study of Community-Based AIDS Education Evaluation

Sometimes it is desirable or convenient to implement a social treatment or program throughout an entire community simultaneously. It is difficult and costly to assess the effects of such interventions using traditional pre-post quasi-experimental designs that measure matched individuals. It is also likely that a single-treatment, single-control community design will not provide a strong test of such interventions. To illustrate the RPDD in this context, consider the hypothetical case of a massive comprehensive AIDS Prevention Education campaign instituted in a specific community and including a wide range of activities (e.g., television, radio and newspaper campaigns; special in-school training; community group presentations, parent education, and so on) designed to increase awareness of the disease and its transmission, and to reduce incidence rates in that area (this example is based on

one described in Campbell, 1990). The communities might be limited to cities that are between 100,000 and 500,000 population, each having an independent newspaper, TV and radio area outside of the range of the treated city. In this case, the pretreatment and posttreatment measures could consist of before and after community survey results of knowledge, attitudes and self-reported sexual behaviors of persons in the treated community and several other non-intervention communities. Or, one could use city-level HIV+ rates for the years preceding and immediately following the intervention. The city that would be chosen for intervention is probably high on HIV+ rates, but not necessarily at the very extreme.

To illustrate such an analysis, we took actual HIV+ rate data for 95 U.S. cities for two successive years, 1989 and 1990. To simulate a hypothetical treatment effect resulting from an imagined AIDS education program, we arbitrarily selected one city -- Miami, Florida -- and subtracted 10 points from its observed posttest HIV+ rate. Miami does not evidence the highest HIV+ rates, but is high enough that it might plausibly be selected as a target community for testing an AIDS education program. The resulting data are graphed in Figure 11.

---

Insert Figure 11 about here

---

The blackened circle shows where the Miami posttest value actually was, while the 'X' shows where it was after subtracting the 10 point treatment effect. The plot for the control cities appears to be quite linear, although very steep due to the annual increases in HIV+ rates. Because of the high data aggregation level, the regression line appears to have a small error term, with considerable power because of the large the number of control cities involved. The ANACOVA estimate of effect is $\beta_2 = -13.91$ ($t=-3.03$, $p=.0032$).

The distributions in Figure 11 are markedly skewed by a few cities that have relatively high HIV+ rates. To help minimize this skew, we conducted a second analysis after computing log transformations on both the pre and posttest. The transformed data are plotted in Figure 12.

Insert Figure 12 about here

Here, values from the lower-scoring cities are spread out more evenly. However, the apparent displacement of the Miami data appears to be smaller relative to the variability around the control group regression line as depicted in the figure. The ANACOVA estimate of effect is $\beta_2 = -.178$ (t=-.63, p=.5305).

The analyses of the original and transformed data point out once again the potential frailty of inferences based on the RPDD. In the first case, the artificial 10-point treatment effect stood out clear relative to the variability around the control group regression line. In the second, the displacement was not significant because the transformation increased the relative variability around the regression line for the many low-scoring cities. One would also want to worry about pseudo-effects in such a design. One likely source is the effect of the educational campaign on the demonstration city's records, perhaps increasing the thoroughness and accuracy, and making the treatment look worse.

## Threats to Validity

### Selection Bias

Probably the most important threat to validity in the RPDD is the potential for selection bias that stems from initial between group differences that affect the posttest and are unrelated to the treatment. The plausibility of such a threat rests on the method used for assigning (or selecting) the treated group in the RPDD. The method of assignment determines which traditional multiple unit pre-post designs the specific RPDD is most like. If the RPDD units are randomly assigned, the design is most analogous to a RE. In the RPDD case, however, random assignment is not used to assure probabilistic pretest equivalence as much as to minimize the chance that a unit might be opportunistically chosen because it is well-suited, politically favored, likely to be successful, or any other number of factors that could bring about an apparent effect even if the program is never administered. Random assignment helps to guard against the many pretreatment correlates that might bias the outcome, wittingly or unwittingly.

If the RPDD experimental unit is assigned solely on the basis of its extremity on the pretreatment measure, this is analogous to a RD design because the assignment is by means of an implicit cutoff rule. This is the case with

the Medicaid study as described in Figure 1. There, Congress allocated Medicaid explicitly using an income cutoff rule. Note, however, that this is not the case for other RPDDs constructed from the Medicaid data (shown in Figure 3) where, for instance, 1964 average physician visits constitute the pretest and 1965 values the posttest. Even though in this case the experimental group also turns out to be the lowest in pretest average physician visits, physician visits were not the basis for the allocation of the Medicaid treatment. The RPDD is also analogous to a RD model in the small area analysis of physician practice rates described above, where special training was given to gynecologists in a geographical area identified as having the highest hysterectomy rate. In this case, the RPDD follows the logic of the RD design, with the exception that there is only a single experimental point.

Where the RPDD is structured like the RE or the RD designs, the selection bias problem is largely mitigated by the fact that we know perfectly the rule that determines the assignment to treatment (probabilistic in the RE case and cutoff-based in RD). Just as in those designs, only a factor that correlates perfectly with the known assignment rule poses a legitimate selectivity threat. Of course, as the Medicaid study shows, there can be many such factors because the same implicit cutoff (i.e., the poverty rate) is used to allocate multiple programs.

This should be contrasted with the third assignment strategy -- uncontrolled assignment -- which yields an RPDD most analogous to the NEGD. In this case, the rule for assignment (i.e., selection) of the experimental unit is not explicit or able to be controlled for perfectly in the statistical model. Consequently, one is less sure that the observed treatment effect is attributable to the treatment as opposed to any of the many possible selection factors that might also affect the posttest. For instance, assume a study where there are ten possible treatment sites for some presumably beneficial treatment. Further assume that the selection of the experimental site is intensely political with each site lobbying to be the first to receive the experimental monies (much as city Olympic committees lobby to be selected as the next Olympic site). The city that is ultimately selected is likely to differ in many ways from those that were not. It may be more highly motivated, have greater resources, have more political clout, and so on. If these (and other) factors can affect posttest scores, it will not be possible to say with great confidence whether any observed treatment effect is due to the treatment or to these inherent differences between this city and the controls. Measuring all cities, including the experimental, on a pretest is likely to improve our inferential ability because posttest differences can be adjusted for pretest ones, but our experience with such

adjustments for selection bias warns us that we should be cautious about attaching too much credibility to treatment effect inferences in this case.

Ideally, the demonstration site for a pilot program in a RPDD would be chosen purely at random, perhaps in a public lottery. While with an N of 1 in the experimental group one would not be getting the benefit of plausible pretreatment equalization, one would be reducing the plausibility that a systematic difference on other variables not only determined the choice of the pilot site, but also determined the exceptional departure from expectancy on the outcome variable.

The discussion so far has assumed that the experimental unit was not selected after its eccentricity on the outcome variable was known. Nonetheless, that possibility needs discussion. Consider a case in which 10 cities have measures on HIV+ rates over successive years, and one notices that one of them is exceptionally far below expectancy for the second year. The interpretive problem is that each city has had AIDS prevention programs, all slightly different, so that there is an "experimental program" to be credited with the effect, no matter which city is exceptional, and even if that exceptionality is due to chance.

While such interpretive opportunism is to be discouraged (especially if it is disguised from the reader), the strategy of locating a "truly exceptional" city (or site) first on the basis of posttest scores, and then speculating on what "caused it" should not be entirely prohibited. But in this case, the ordinary p-value for a given t-value cannot be used. Instead, a correction on the order of that for "error-rate experimentwise" is needed. The simplest approach would be to utilize a Bonferoni correction of the p-values (Ryan, 1959, 1960; Dunn, 1961; Darlington, 1990, p. 250-257). If for a specified in-advance site, for a given df (e.g., df=8 for the 10 cities assuming a linear fit) a t-value of 2.306 is required for $p<.05$, when we want a comparable p-value (i.e., $\frac{1}{20}$ for testing the exceptionality of any one of the 10 points from the regression line determined by the 9 others (not specifying which one in advance), we need a t-ratio corresponding to $\frac{1}{20x10}$ or $\frac{1}{200}$, or $p<.005$. For our hypothetical AIDS example, had we not specified the "treated city" in advance, its t-value of 2.306 would have been significant only at $p<.05$ (one-tailed test) and thus would be rejected. Tables for the t-ratios for such small p-values may be hard to locate, although most statistical computing packages report exact p-values in regression summary tables. Since we are focused on the efficacy of therapeutic programs, it would be especially reassuring for attributing significant exceptionality to a

unique program if that exceptionality occurred only in the benefits direction, not in the deterioration direction for that therapeutic context.

The RPDD, unlike other quasi-experiments like the nonequivalent control group design, does not require pretest equivalence between the treated group and the controls. The treated group could theoretically come from anywhere along the pretest continuum. The design rests on the assumption that the treated group posttest mean does not differ significantly from the regression line prediction. Consequently, the traditional concerns about selection bias take on a slightly different form in this context. Here, the key issues are whether the control groups yield an unbiased estimate of the true population regression line and whether the treatment unit is a member of the control group population. This could be assured by randomly sampling control groups from the population, a circumstance which will not be feasible in many situations. If the sample of control groups is not representative of the theoretical population or the regression line is incorrectly estimated, the estimate of the treatment effect will be biased. There is no solution to this problem although it might best be minimized by selecting many control groups with wide pretest variability. For instance, in their study of schizophrenics, Fleiss and Tanur (1972) give this advice for selecting control groups:

> "...a more efficient approach would call for the identification of many of the factors that distinguish schizophrenics from normals: having a mental disorder, being hospitalized, having been treated with drugs some time in the past, and so on. Samples of subjects from groups defined in terms of various combinations of such factors would be drawn and studied. These samples would have one feature in common: they would all consist of subjects who are not schizophrenic." (p. 525).

Measurement Error and Regression Artifacts

The general problem of regression artifacts (or error in independent variables) is taken care of in the RE or RD analogues of the RPDD since such regression is displayed and accounted for by the inclusion of X in the regression analysis (Trochim, 1984; Cappelleri et al., 1991; Trochim et al, 1991) Nonetheless, when choosing the experimental unit involves unknown but systematic variables on which the experimental group differs from the control group in ways that would affect the posttest differentially from the pretest, one might mistakenly conclude

that the treatment was effective when, in fact, the apparent effect should be attributed to measurement error and the resulting regression to the mean.

This is similar to the misleading interpretations that can occur in relation to the "fuzzy" RD design (Campbell in Trochim, 1984, pp. 29-37). If in fact, the choice of the experimental unit had been based upon a latent decision variable related to the pretest by the addition of a pretest random error component, and to the posttest by the addition of a posttest random error component, and if the award of the experiment is based upon extremity on the latent "true score," then a mistaken inference comparable to that graphed in Figure 16 of Trochim (1984, p. 30) is possible.

The problem here is a manifestation of the familiar effect of pretest measurement error in NEGDs as described by Reichardt in Cook and Campbell (1979, Fig. 4.4, p. 161). Reichardt shows that pretest measurement error attenuates the within-group slope. In the RPDD, however, because there is only a single experimental group point, it is impossible to estimate a treatment group slope (and thus, the "slope" cannot be attenuated due to pretest measurement error). However, the true control group regression line would appear to be rotated clockwise slightly (assuming a positive relationship). The further the experimental group is away from the control group pretest mean, the greater will be the deleterious effects of such measurement error -- bias will be greater.

The issue of how the treated unit is chosen is so central to the interpretability of the RPDD that it warrants some belaboring. The central concern is this: can the treated unit be considered a member of the control unit population prior to treatment, or does it come from some different distribution? If we select the treated unit randomly or because of its extremity (implicit cutoff) it is reasonable to infer that the unit is sampled from the control group population. Here, just as in the RE or RD designs, estimates of treatment effect will be unbiased by the measurement error. The regression of the posttest onto pretest accurately describes the amount of regression to the mean expected for all units, treated and control. But when selection of the treated group is not controlled, it is plausible that the treated unit does not come from the control unit population, but rather from some population that differs systematically from the controls. In this case, we must assume that the two populations may differ in their overall pretest averages and, consequently, measurement error would affect the populations differently and there would be regression to different population means, just as in the NEGD.

In any regression analysis, random measurement error on the pretest will attenuate pre-post regression line slopes. This is not likely to be a serious problem in the RPDD, because presumably the group means are less influenced by random error than individual data is. Nevertheless, this needs further investigating. It is likely, for instance, that such an investigation would lead to the conclusion that random measurement error introduces greater bias in treatment effect estimates when the treatment group pretest mean is located further away from the overall pretest mean where the attenuation most affects point predictions. Traditional adjustments for random measurement error have to be modified for the RPDD and may be problematic especially when there are relatively few control points for estimating reliability.

However, we can state unequivocally that the deleterious effects of measurement error will be less manifest in the RPDD than in an individual-level NEGD analysis of the same data (as in the Fleiss and Tanur or Xanax examples below) because the average values used in the RPDD must (by definition) have less variability or error than the individual data on which they are based. We expect that the power and efficiency of the NEGD will be the upper limit for a comparable RPDD (because the loss of degrees of freedom will outweigh the gains in reliability), but that measurement error will be reduced and the bias in estimates due to it will be correspondingly less in the RPDD.

Another regression artifact comes where the choice of the experimental unit is triggered by the error component in the pretest. During 1956 Connecticut endured an extreme crackdown on speeding and subsequently claimed a dramatic reduction in fatalities. But we know that the 1954-55 increase in Connecticut's traffic fatalities was the largest in its history and that the 1954-55 increase caused Governor Ribicoff to initiate the crackdown. Campbell and Ross (1968, 1988) conclude that the purported effects were merely a return to trend, a regression artifact. They also present the effect in the context of other nearby states, as in Figure 13.

---

Insert Figure 13 about here

---

Were one to use the 1955 and 1956 data as an RPDD, one potentially could get a significant pseudo-effect, as plotted in Figure 14.

Insert Figure 14 about here

In this case, there are too few control states to produce significance, but the danger is illustrated (the actual t-value is -1.50, p=.26).

Instrumentation

As for any quasi-experimental design, the range of rival hypotheses should be examined in case a significant effect is found. For pilot studies and demonstration sites, one of the most frequently troubling will be that the program effort has changed uniquely for the experimental unit the measurement process between the pretest and posttest. The pressure on the law enforcement system to show a good effect may lead, for example, to the downgrading of felonies to misdemeanors (e.g., Seidman and Cougens, 1974). Equally frequently, program attention to a problem such as child abuse may lead to increased thoroughness of reporting, and a pseudo-increase (a pseudo harmful effect) specific to the experimental unit.

Statistical Power

Although at first glance it appears that the RPDD design suffers from low statistical power, because of the relatively few pre-post points that are typically used, group means are generally more stable and precise than within-group data. Fleiss and Tanur (1972, see example below) compared a traditional pre-post Analysis of Covariance with the RPDD analysis using the same data and found that the ANACOVA results were not significant while the RPDD results were. They comment,

> "The difference between the analysis of covariance performed at the beginning of this chapter, where significance was not found, and the regression analysis just performed, where significance was found, is that predictability in the former was determined by covariation within groups, whereas predictability in the latter was determined by covariation between groups" (p. 525).

Two major issues related to statistical power need to be investigated. First, what is the power of the RPDD design as it stands? This should be relatively simple to determine and would make it possible to estimate the needed number of control points given some initial estimates of probable treatment effect size and desired $\alpha$ level. An

important factor in statistical power is where the treatment group scores on the pretest continuum. Statistical power will decline as the treated group pretest occurs further from the overall pretest mean. Second, an analysis needs to be done of the power of the RPDD relative to within-group ANACOVA alternatives. This should reveal whether one would ever gain statistical power in the trade-off between within-group variability in the ANACOVA framework and the presumed lower variability in the between-group oriented RPDD.

Violating Assumptions of Statistical Tests

The RPDD may also be subject to violations of the assumptions of the t-test which is used. Fleiss and Tanur (1972) point out that the analysis is technically only valid when the control groups used to estimate the regression are a random sample from a population of such groups. The population in this case would be hypothetical (there are an infinity of potential groups that could be entered as controls) and consequently this assumption can never technically be met. Instead, as Fleiss and Tanur (1972) point out "one must be sure to select groups defined by the presence or absence of enough factors to assure that the variability of their mean responses is high" (p. 525).

One benefit which accrues in the RPDD derives from the usually higher-aggregate values used for the data. For instance, when group means are used (as opposed to individual-level values), we can more reasonably expect that the statistical assumption of normally distributed variables is likely to be met. This is because of the well-known statistical property of the central limit theorem which holds that with sufficient sample sizes, sampling distributions are normally distributed. The advantage, of course, is that one needs to worry less about this distributional assumption which is critical to many statistical tests.

Local History

A key threat to internal validity in this design is "local history" (Cook and Campbell, 1979). Whenever the treatment group consists of persons who are treated together (such as at the same site) and distinct from control group persons, any factor in the setting that affects posttest performance can lead to a pseudo-effect in the data. If treated persons receive multiple treatments, or experience a markedly different setting from controls, or have a change in instrumentation (e.g., clinicians change their implicit judgement standards at the treatment site between pre and posttest, but not at control sites), a pseudo-effect can result. These threats are not as serious an issue for the control groups because, with lots of such groups, setting variability is increased and the potential for systematic bias declines.

External Validity

Finally, the RPDD is not strong in external validity or generalizability. While generalizing to other potential treatment groups that have the same pretest level may be reasonable, it is impossible to know whether any observed treatment effects would hold for groups with other pretest levels. Put another way, it is impossible with this design to study treatment interaction effects. If the treatment effect changes for different pretest levels, it will not be possible to know from the RPDD. Nevertheless, if the treatment group is typical of the potential target treatment group of interest (especially if it is a unit randomly sampled from the population of interest) it will be reasonable to generalize to other similar target groups. One might not be interested in whether the treatment might work for groups having markedly different pretreatment levels (see the Fleiss and Tanur example above).

## Conclusions

The RPDD has a very specific potential range of applicability. It is limited to contexts where one has pre and post measurement and multiple control groups. It is strongest when applied to routinely collected multi-wave administrative data where it is either too costly to match individual cases or or may not be possible. Because such data are widely available and cost constraints a constant factor in our society, it is likely that the RPDD would be widely applicable. In fact, there are probably many instances in which the requirements of the design have already been met and for which a post hoc analysis could be simply constructed.

The design has several important weaknesses that need to be anticipated. Where few control groups are available, one is likely to have low statistical power. It is recommended that a power analysis be routinely reported with any analysis of the RPDD that fails to show significant treatment effects. In terms of internal validity, there are several possible threats that could lead to pseudo-effects in various situations. Care needs to be taken in selecting a heterogeneous set of control groups. In research conducted within organizations, one must look for treatment-related drop-outs that will leave unrepresentative treatment group cases on the posttest. Instrumentation changes due to increased sensitivity to treatment-related issues may distort measurement of the posttest in the treated group. Perhaps the most important threats concern the potential for confounding of the treatment with site or group factors. In this regard, it is important to distinguish the various versions of RPDD by the method used to select the treated unit. Whenever possible, the treated unit should be randomly selected from the population. This will tend to minimize any deliberate selection factors that might threaten internal validity and is also likely to be the variation

that has the greatest statistical power. Failing that, selection of the most extreme case is preferred over arbitrary or convenience-based selection, especially when the same measure is used for before and after measurement. Obviously, where feasible, one should implement the treatment at multiple sites or for multiple groups. Where this cannot be accomplished, the RPDD may still be the strongest method which can feasibly be accomplished in assessing the effects of interventions.

The RPDD has great potential for enhancing our ability to conduct research in natural social contexts. It is relatively inexpensive to apply where appropriate administrative data exist. It is based on well-known statistical models that can be estimated with almost any statistical computing package. It extends our ability to evaluate the effects of community-level programs where other designs are often not readily available. While much work is yet needed to explore the implications and variations of the RPDD, it is clearly a useful addition to the methodological toolkit of the researchers of the experimenting society.

References

Ballenger, J.C., Burrows, G.D., DuPont, R.L., Lesser, I.M., Noyes, R., Pecknold, J.C., Rifkin, A., Swinson, R.P. (1988). Alprazolam in panic disorder and agoraphobia: Results from a multi-center trial, I. Efficacy in short-term treatment. Arch Gen Psychiatry, 45, 413-422.

Campbell, D.T. (1990). Quasi-experimental design in AIDS prevention research. Paper presented at a Panel on Evaluation of AIDS Prevention of the Committee on AIDS Research, CBSSE, National Research Council, January 12-13, Washington, D.C.

Cappelleri, J.C., Trochim, W., Stanley, T.D., and Reichardt, C.S. (1991). Random measurement error doesn't bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. Evaluation Review, 15, 4, 395-419.

Cook, T.D. and Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton-Mifflin.

Darlington, R. (1990). Regression and Linear Models. New York: McGraw Hill.

Draper, N. and Smith, H. (1981). Applied Regression Analysis: Second Edition. New York: Wiley.

Dunn, O.J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56, 52-64.

Ehrenberg, A.S.C. (1968). The elements of lawlike relationships. Journal of the Royal Statistical Society, Series A, 131, 280-302.

Fleiss, J.L. and Tanur, J.M. (1972). The analysis of covariance in psychopathology. In M. Hammer, K. Salzinger and S. Sutton (Eds.). Psychopathology. New York: Wiley.

Gujarati, D.N. (1988). Basic Econometrics. (Second Edition). New York: McGraw-Hill.

Hsiao, V. (1994). Use of the regression point displacement design to evaluate interventions based on small area analysis. Unpublished Doctoral Dissertation, Cornell University, Ithaca, NY.

Johnston, J. (1972). Econometric Methods. (Second Edition). New York: McGraw-Hill.

Klerman, G.L., Coleman, J.H., Purpura, M.D. (1986). The design and conduct of the Upjohn Cross-national Collaborative Panic Study. Psychopharmacology Bulletin , 22, 1 59-64.

Klerman, G.L. (1988). Overview of the Cross-National Collaborative Panic Study. Arch Gen Psychiatry, 45, 407-412.

Lohr, B.W. (1972). An historical of the research on the factors related to the utilization of health services. Duplicated research report, Bureau for Health Services Research and Evaluation, Social and Economic Analysis Division, Rockville, MD, January.

Maddala, G.S. (1977). Econometrics. New York: McGraw-Hill.

Mood, A.M. (1950). Introduction to the Theory of Statistics. New York: McGraw-Hill.

Noyes, R., DuPont, R.L., Pecknold, J.C., Rifkin, A., Rubin, R.T., Swinson, R.P., Ballenger, J.C., Burrows, G.D. (1988). Alprazolam in panic disorder and agoraphobia: Results from a multi-center trial, II. Patient acceptance, side effects and safety, Arch Gen Psychiatry, 45: 423-428.

Pecknold, J.C., Swinson, R.P., Kuch, K., Lewis, C.P. (1988). Alprazolam in panic disorder and agoraphobia: Results from a multi-center trial, III. Discontinuation effects. Arch Gen Psychiatry, 45: 429-436.

Rieken, H.W., Boruch, R.F. Campbell, D.T. Coplan, W. Glennan, T.K., Pratt, J., Rees, A. and Williams, W. (1974). Social Experimentation: A Method for Planning and Evaluating Social Innovations. New York: Academic Press.

Ryan, T.A. (1959). Multiple comparisons in psychological research. Psychological Bulletin, 56, 26-47.

Ryan, T.A. (1960). Significance tests for multiple comparisons of proportions, variances, and other statistics. psychological Bulletin, 57, 318-328.

SAS Institute Inc., (1982). SAS User's Guide: Statistics, 1982 Edition, SAS Institute: Cary, North Carolina.

Seidman, D. and Cougens, M. (1974). Getting the criminal rate down: Political pressure and crime reporting. Law and Society Review, 8, 3, 475-493.

Sheehan, D.V., Ballenger, J.C., Jacobsen, G. (1980). Treatment of endogenous anxiety with phobic, hysterical and hypochondriachal symptoms. Arch Gen Psychiatry, 37,51-59.

Smith, H.F. (1957). Interpretation of adjusted treatment means and regressions in analysis of covariance. Biometrics, 13, 282-308.

SPSS Inc., (1983). SPSS-X User's Guide. McGraw-Hill: New York.

Trochim, W.M.K. (1984). Research design for program evaluation: The Regression-Discontinuity approach. Beverly Hills, CA: Sage.

Trochim, W., Cappelleri, J.C. and Reichardt, C.S. (1991). Random measurement error doesn't bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. Evaluation Review, 15,5, 571-604.

Wennberg, J.E. (1990). Small area analysis and the medical care outcome problem. In L. Sechrest, E. Perrin and J. Bunker (Eds.). Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, Washington, DC.

Wilder, C.S. (1972). Physician visits, volume, and interval since last visit, U.S., 1969. National Center for Health Statistics, July (Series 10, No. 75; DHEW Pub. No. (HSM) 72-1064).

## Footnotes

1 Our rejection of the Regression Displacement Design suggested most recently by Coyle, Boruch and Tanur (1990) is motivated by the need to distinguish its acronym from that of the Regression Discontinuity Design. Adding the term "Point" to the name accomplishes this while conveying the key distinguishing feature of the design.

[2] This is analogous to distinguishing a repeated measures design with many repeated measures from a time series design. There is no hard and fast rule, but with one or a few treated points we prefer RPDD to other quasi-experimental designations.

[3] For simplicity of exposition, and generality, we use the symbol $X_i$ rather than $\bar{X}$ to indicate the mean pretest value for group i. In other words, $X_i$ indicates the pretest value for unit i whether that be a group mean, individual value, frequency, rate, and so on.

[4] In Mystat or in Systat module MGLH you can get all N+1 t-residuals with the commands

```
>model y = constant + x
>save tres
>estimate
```

These commands create a Systat file arbitrarily named 'TRES'; variable STUDENT in that file contains the t-residuals. Thus t-residuals offer another generalization of formula (1) -- a single command computes t for every case, and they are not limited to a single regressor or covariate. However, t-residuals, do not provide the size or standard error of the discrepancy of the discrepency of each point from the regression line, and do not test the hypothesis that several points as a group differ from the line. The dummy coding procedure, on the other hand, can do all of these.

Table 1. Pretest means, posttest means and dummy coded treatment group variable for eleven groups (ten control, one treatment) in the hypothetical simulation example.

| Pretest Mean ($X_i$) | Posttest Mean ($Y_i$) | Group ($Z_i$) |
|---|---|---|
| 51.135 | 56.622 | 1 |
| 50.175 | 51.031 | 0 |
| 53.113 | 55.312 | 0 |
| 43.679 | 42.476 | 0 |
| 45.141 | 46.164 | 0 |
| 49.085 | 49.662 | 0 |
| 42.036 | 43.966 | 0 |
| 55.787 | 57.557 | 0 |
| 48.340 | 47.335 | 0 |
| 58.255 | 57.382 | 0 |
| 54.010 | 53.820 | 0 |

Table 2. Descriptive statistics for the pretest and posttest means of the ten control groups in the hypothetical simulation of the RPDD.

Pretest ($X_i$)

| Mean: | Std. Dev.: | Std. Error: | Variance: | Coef. Var.: | Count: |
|---|---|---|---|---|---|
| 49.9621648 | 5.354277882 | 1.693171333 | 28.66829164 | 10.7166651 | 10 |
| Minimum: | Maximum: | Range: | Sum: | Sum Squared: | # Missing: |
| 42.0357478 | 58.25538854 | 16.2196408 | 499.621648 | 25220.1937 | 0 |

Posttest ($Y_i$)

| Mean: | Std. Dev.: | Std. Error: | Variance: | Coef. Var.: | Count: |
|---|---|---|---|---|---|
| 50.4704168 | 5.458954973 | 1.726273136 | 29.8001894 | 1.0816148E1 | 10 |
| Minimum: | Maximum: | Range: | Sum: | Sum Squared: | # Missing: |
| 42.4762136 | 57.5566133 | 15.0803997 | 504.704168 | 25740.8314 | 0 |

Table 3. Individual Point Prediction (IPP) regression analysis for the regression of posttest means onto pretest means for the control group cases only in the hypothetical simulation example.

Regression Analysis

| Count: | R: | R-squared: | Adj. R-squared: | RMS Residual: |
|---|---|---|---|---|
| 10 | .972677443 | .946101408 | .939364084 | 1.344232789 |

Analysis of Variance Table

| Source | DF: | Sum Squares: | Mean Square: | F-test: |
|---|---|---|---|---|
| REGRESSION | 1 | 253.746010226 | 253.746010226 | 140.426881963 |
| RESIDUAL | 8 | 14.455694333 | 1.806961792 | p = .0001 |
| TOTAL | 9 | 268.201704559 | | |

Beta Coefficient Table

| Variable: | Coefficient: | Std. Err.: | Std. Coeff.: | t-Value: | Probability: |
|---|---|---|---|---|---|
| INTERCEPT | .923264655 | | | | |
| SLOPE | .991693461 | .083685906 | .972677443 | 11.850184892 | .0001 |

Table 4. Analysis of covariance (ANACOVA) regression analysis of the posttest means regressed onto the pretest means and the dummy coded treatment variable for the data from the hypothetical simulation example.

Regression Analysis

| Count: | R: | R-squared: | Adj. R-squared: | RMS Residual: |
|---|---|---|---|---|
| 11 | .975822179 | .952228926 | .940286157 | 1.344232789 |

Analysis of Variance Table

| Source | DF: | Sum Squares: | Mean Square: | F-test: |
|---|---|---|---|---|
| REGRESSION | 2 | 288.147806786 | 144.073903393 | 79.732678388 |
| RESIDUAL | 8 | 14.455694333 | 1.806961792 | p = .0001 |
| TOTAL | 10 | 302.603501119 | | |

Beta Coefficient Table

| Variable: | Coefficient: | Std. Err.: | Std. Coeff.: | t-Value: | Probability: |
|---|---|---|---|---|---|
| INTERCEPT | .923264655 | | | | |
| Pretest | .991693461 | .083685906 | .917936249 | 11.850184892 | .0001 |
| Z | 4.988490226 | 1.413255585 | .273423481 | 3.529786316 | .0077 |

Table 5. ANACOVA results estimating the effect of Xanax (Alprazolam) on average Sheehan Clinician Rated Anxiety Symptoms (CRAS) values for the original individual-patient data (N=517); a separate RPDD analysis for each of the eight investigators/sites; the average estimates across the eight investigator RPDDs; and the ANACOVA analysis of all eight investigator pre-post means for both treatments.

| Analysis | COEFFICIENT | SE | t-value | p-value |
|---|---|---|---|---|
| ANACOVA on individual data (N=517) | -.368079207 | .037467215 | 9.824034305 | .0001 |
| Investigator 1 | -.351058903 | .184474098 | -1.903025450 | .1154 |
| Investigator 2 | -.362404076 | .130436066 | -2.778403923 | .0390 |
| Investigator 3 | -.283931976 | .204046723 | -1.391504707 | .2228 |
| Investigator 4 | -.486888118 | .158630038 | -3.069331151 | .0278 |
| Investigator 5 | -.477622277 | .172091558 | -2.775396320 | .0391 |
| Investigator 6 | -.272003935 | .168467991 | -1.614573392 | .1673 |
| Investigator 7 | -.271218747 | .132104558 | -2.053061243 | .0953 |
| Investigator 8 | -.292493200 | .163115159 | -1.793169944 | .1329 |
| Average for 8 investigators | -.349702654 | .164170774 | -2.172308270 | .1050 |
| ANACOVA on investigator means (N=8) | -.359928881 | .059109426 | 6.0891960090 | .0001 |

Figure Captions

Figure 1. Medicaid example modified from Riecken and Boruch.

Figure 2. Fleiss and Tanur (1973) graph.

Figure 3. Medicaid RPDD with physician visit for both pre and post.

Figure 4. Hypothetical RPDD using simulated data.

Figure 5. Average number of physicians per person per year for the years 1964, 1967 and 1969, by family income ranges (taken from Lohr, 1972; Wilder, 1972, p. 5, Table B).

Figure 6. RPDD for the study of the effects of Medicaid on physician visit rates with logs of income as pretest.

Figure 7. Second-order polynomial regression line for the RPDD data from the study of the effects of Medicaid on physician visit rates.

Figure 8. Bivariate distribution of the Sheehan Clinician Rated Anxiety Scale (CRAS) from the Cross-National Collaborative Study of the effectiveness of Xanax (Alprazolam).

Figure 9. CRAS pretest and posttest means for the eight investigators in the Cross-National Collaborative Study of the effectiveness of Xanax (Alprazolam).

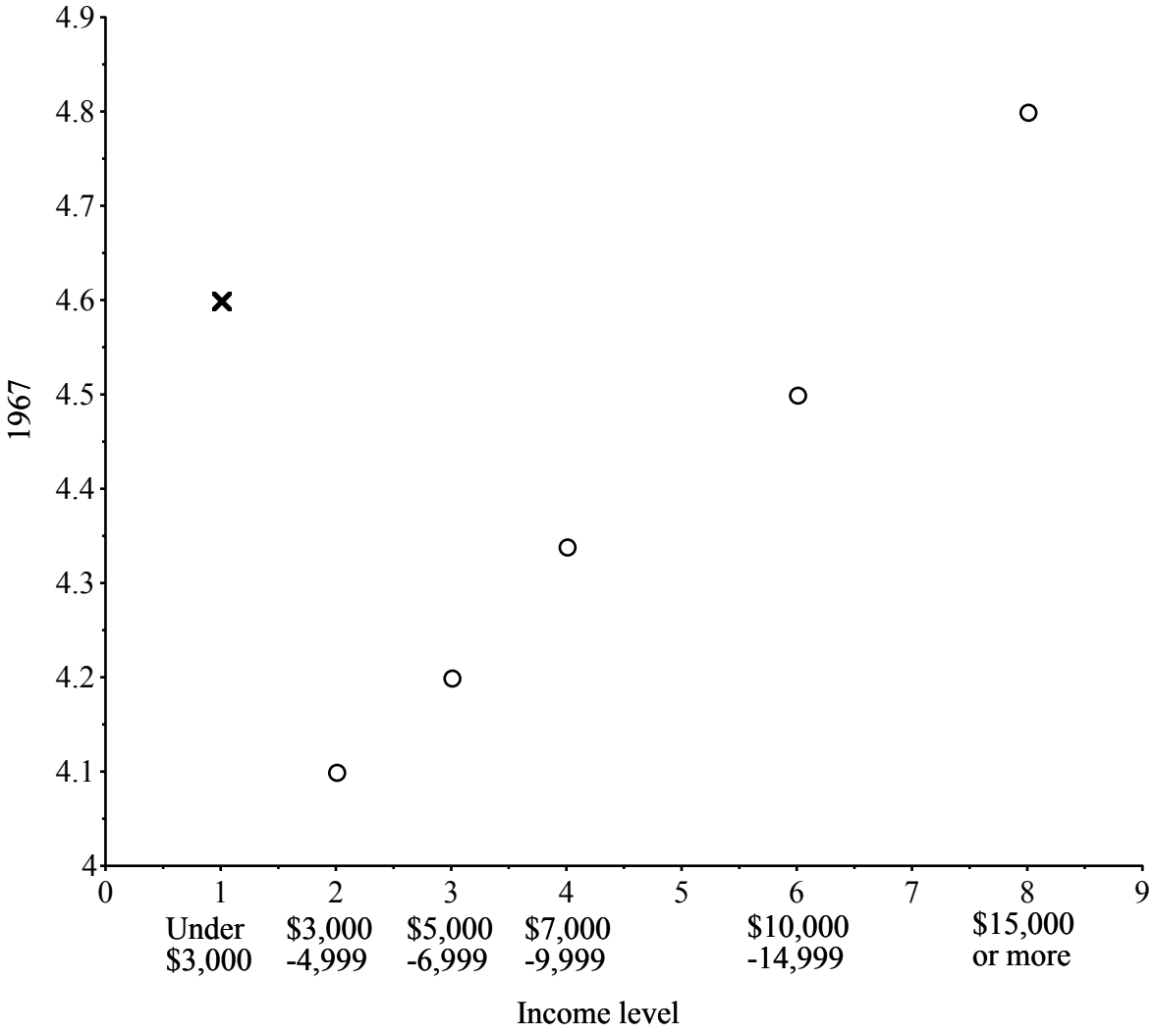
Figure 10. RPDD for a single investigator in the Cross-National Collaborative Study of the effectiveness of Xanax (Alprazolam).

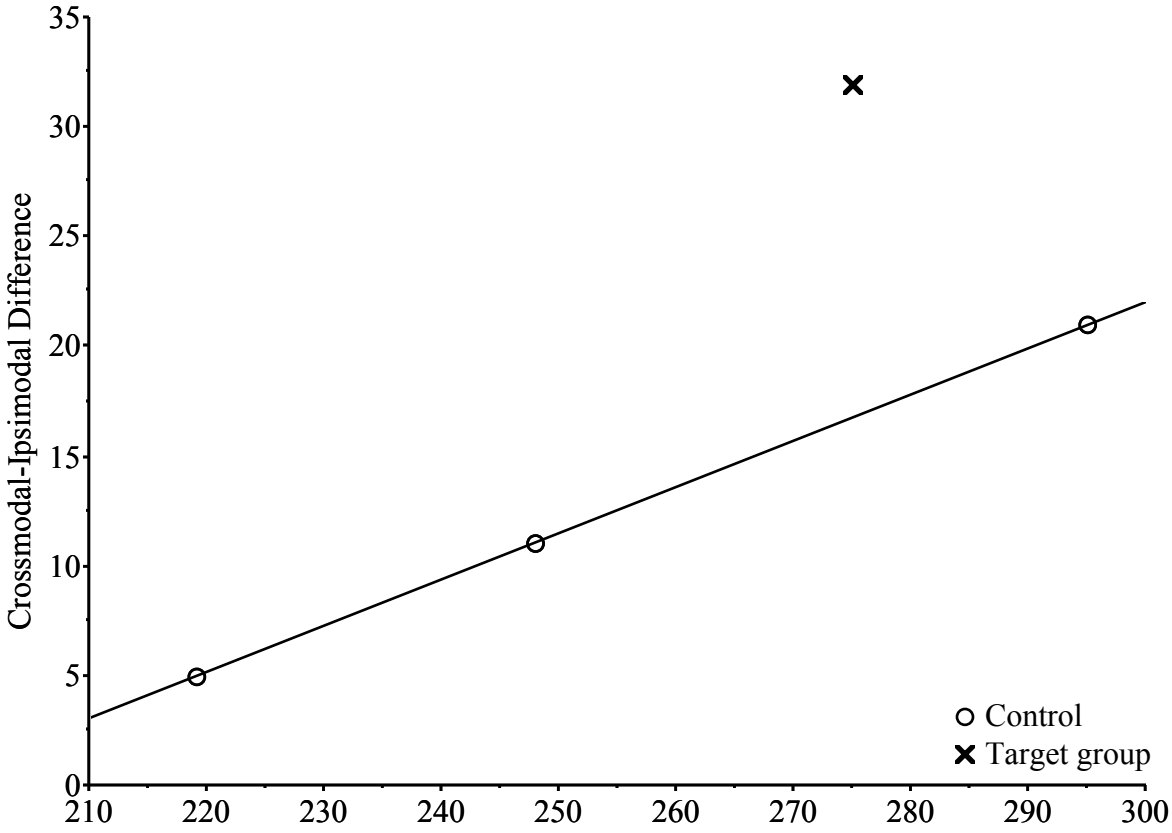
Figure 11. Hypothetical example of a RPDD study of the effects of a massive city-wide AIDS education program in Miami.

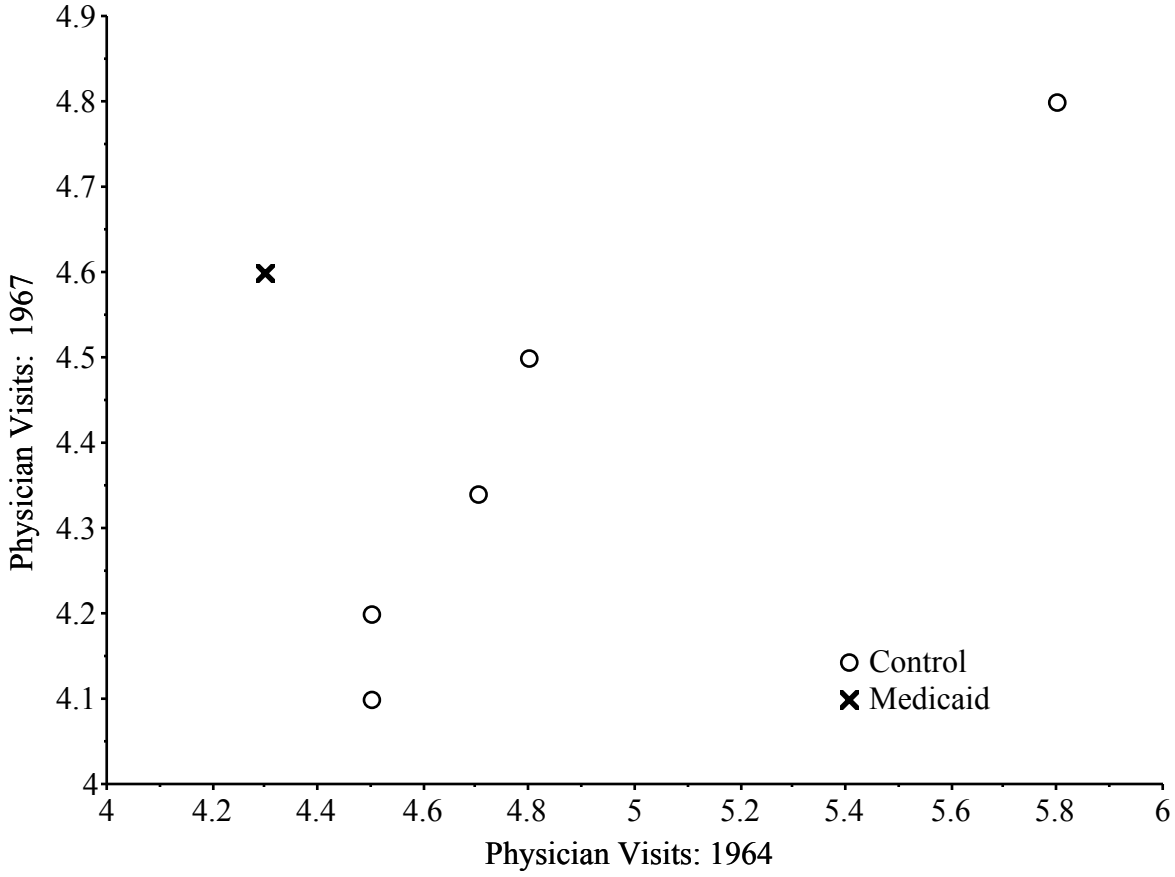
Figure 12. Hypothetical example of a RPDD study of the effects of a massive city-wide AIDS education program in Miami. A log transformation has been applied to both the pre and posttest data.

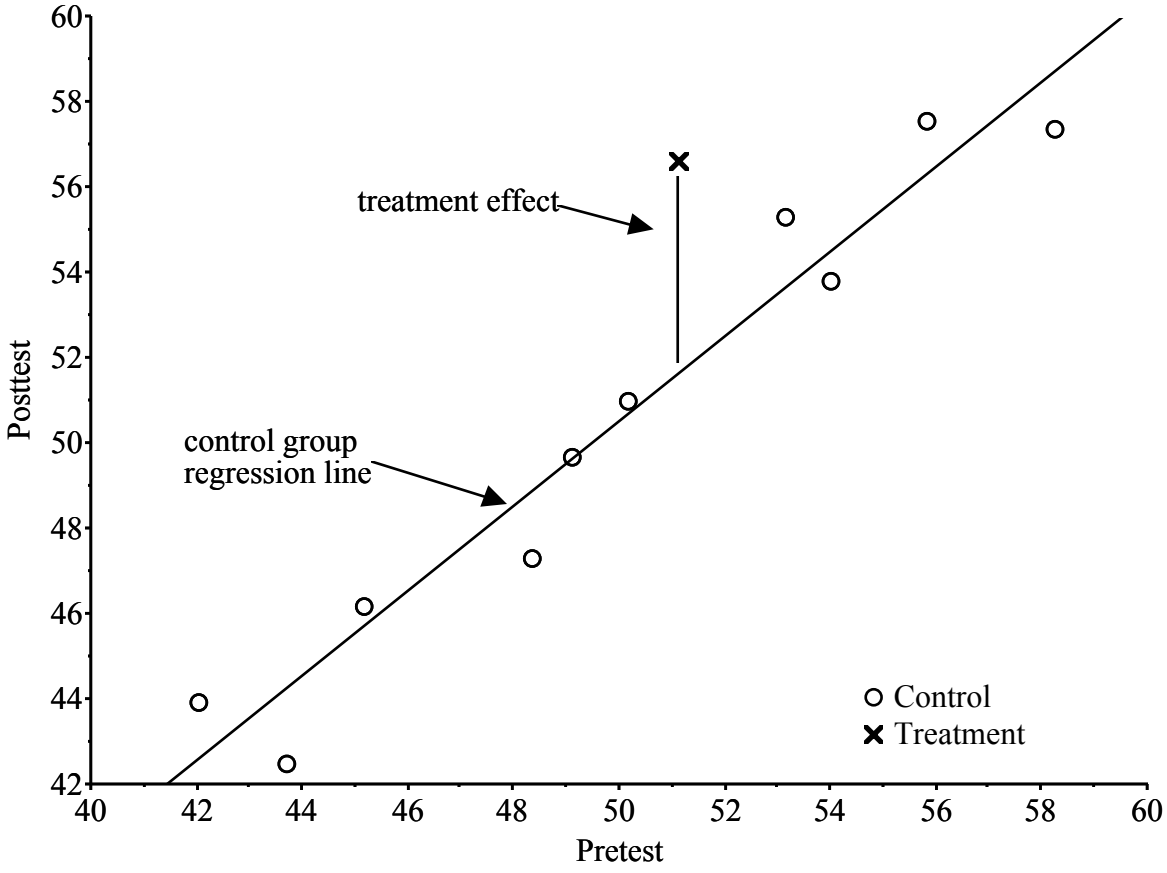
Figure 13. Traffic fatalities for 5 Northeastern states, 1951-1959.

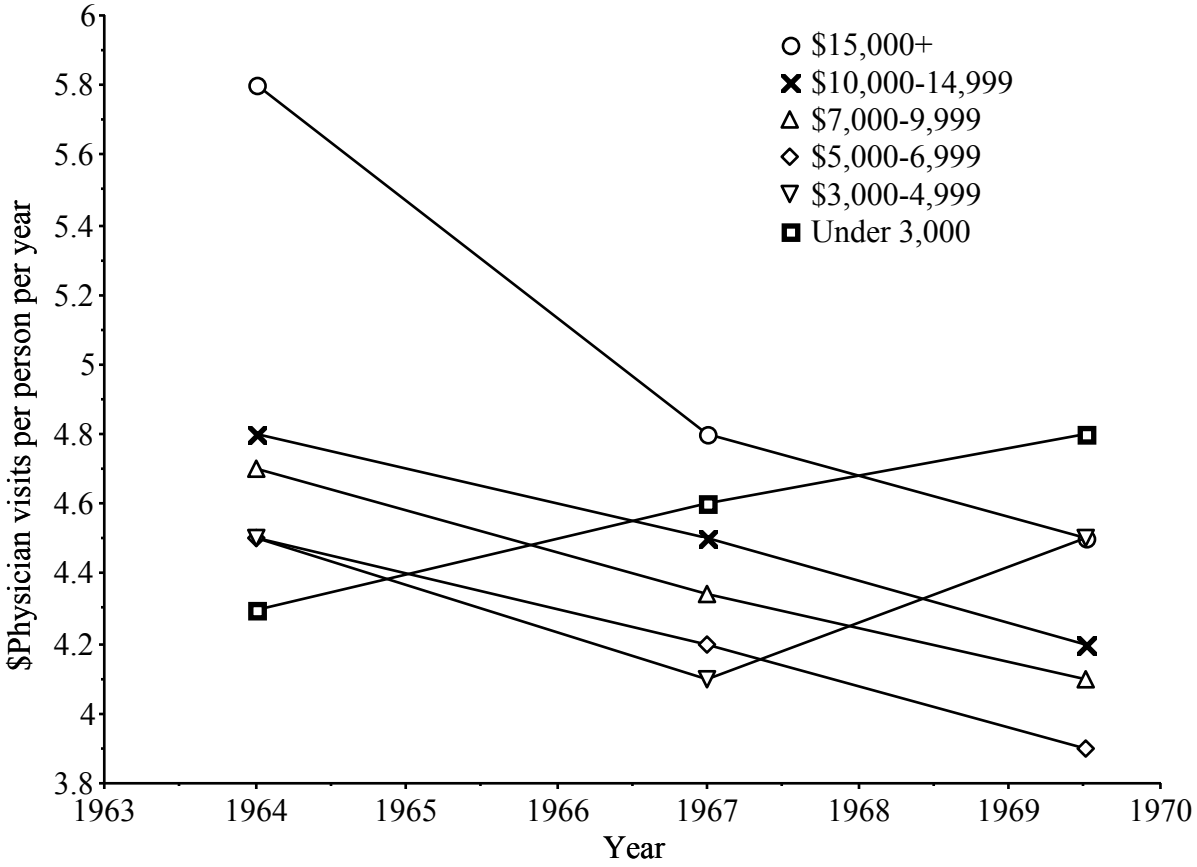
Figure 14. Connecticut Speeding Crackdown, RPDD.

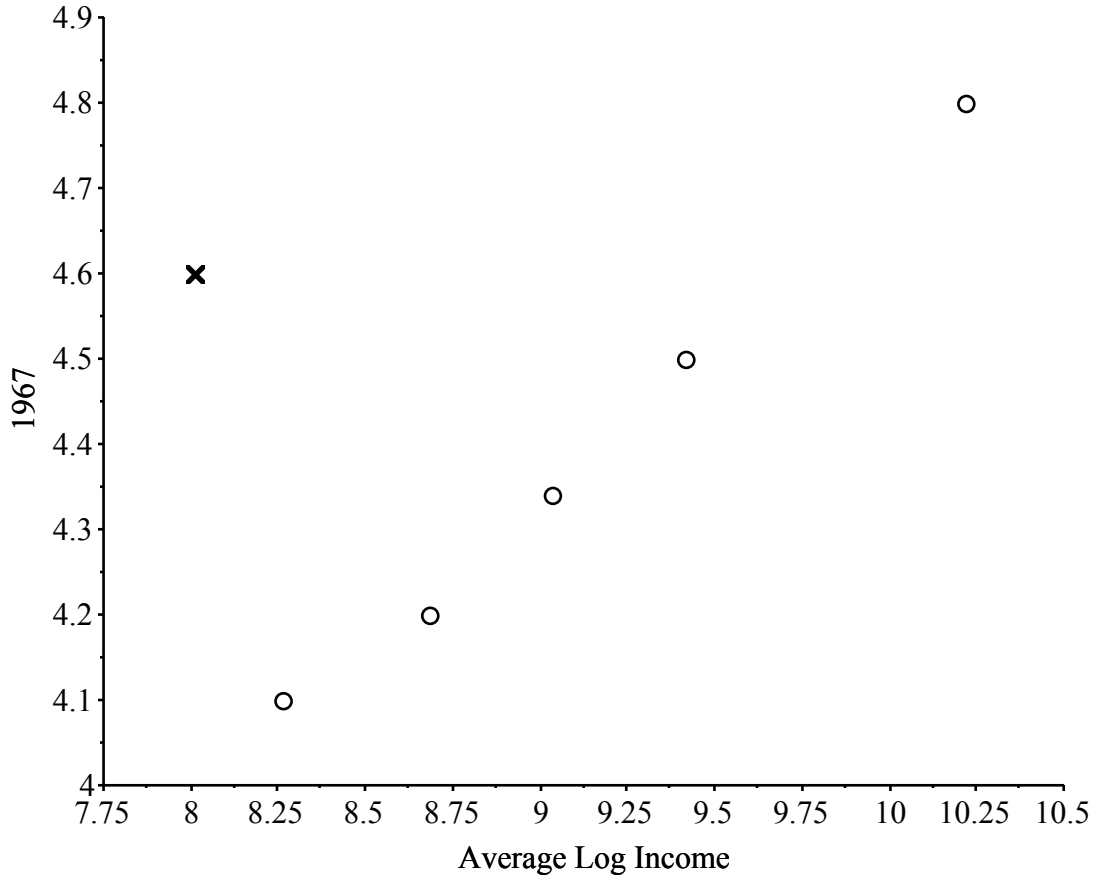

4.9
4.8
4.7
4.6
4.5
4.4
4.3
4.2
4.1
4
1967
0
1
2
3
4
5
6
7
8
9
Under $3,000
$3,000 -4,999
$5,000 -6,999
$7,000 -9,999
$10,000 -14,999
$15,000 or more
Income level

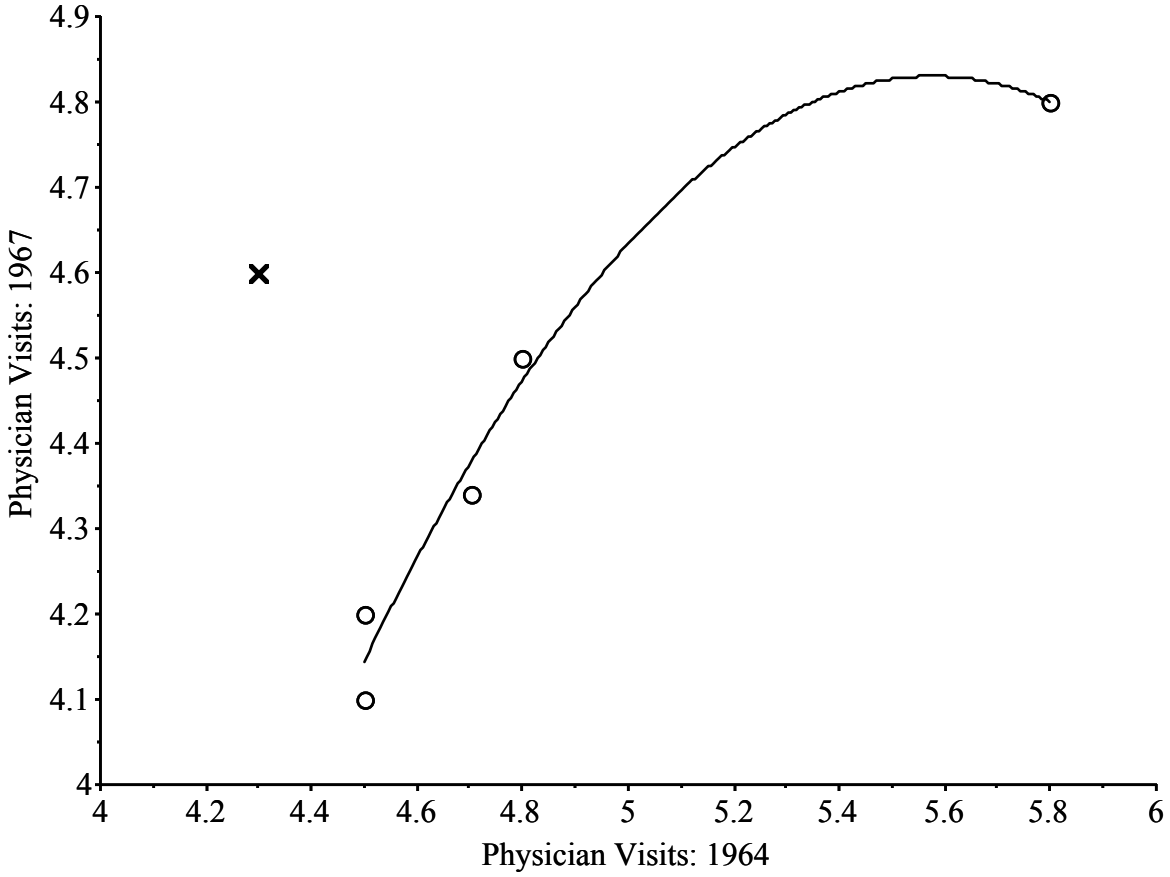

Crossmodal-Ipsimodal Difference
35
30
25
20
15
10
5
0
210
220
230
240
250
260
270
280
290
300
Control
Target group

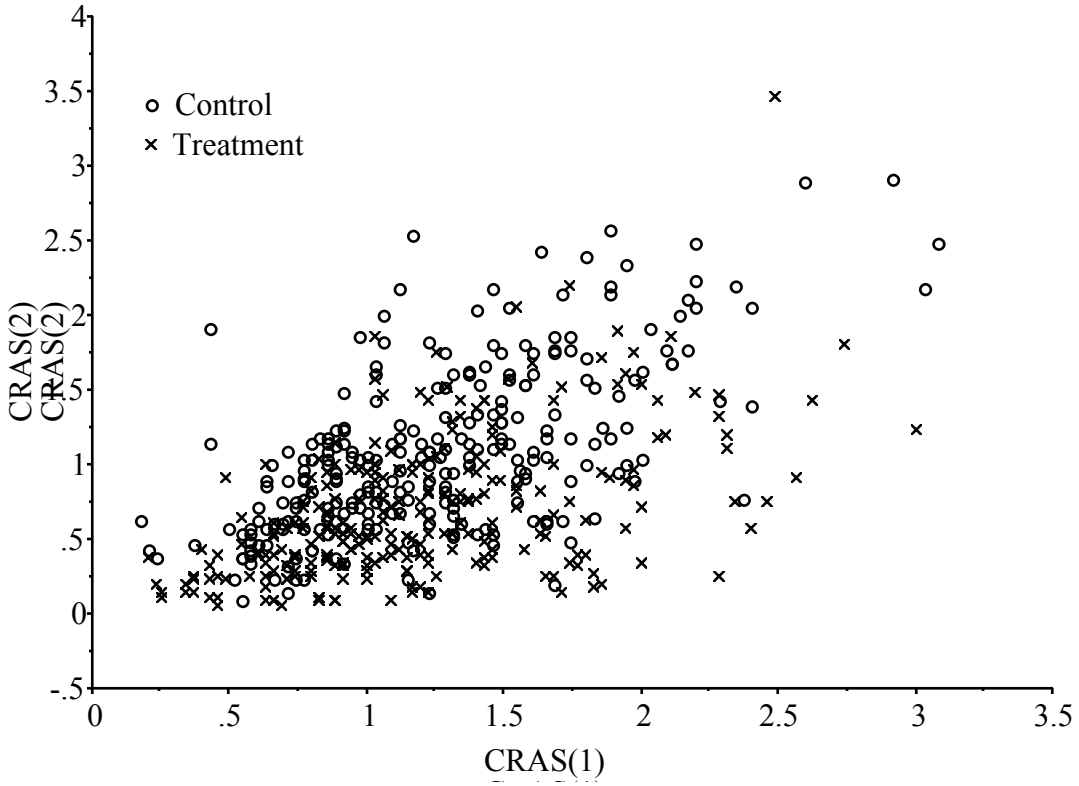

4.9
4.8
4.7
4.6
4.5
4.4
4.3
4.2
4.1
4
Physician Visits: 1967
4
4.2
4.4
4.6
4.8
5
5.2
5.4
5.6
5.8
6
Physician Visits: 1964
Control
Medicaid

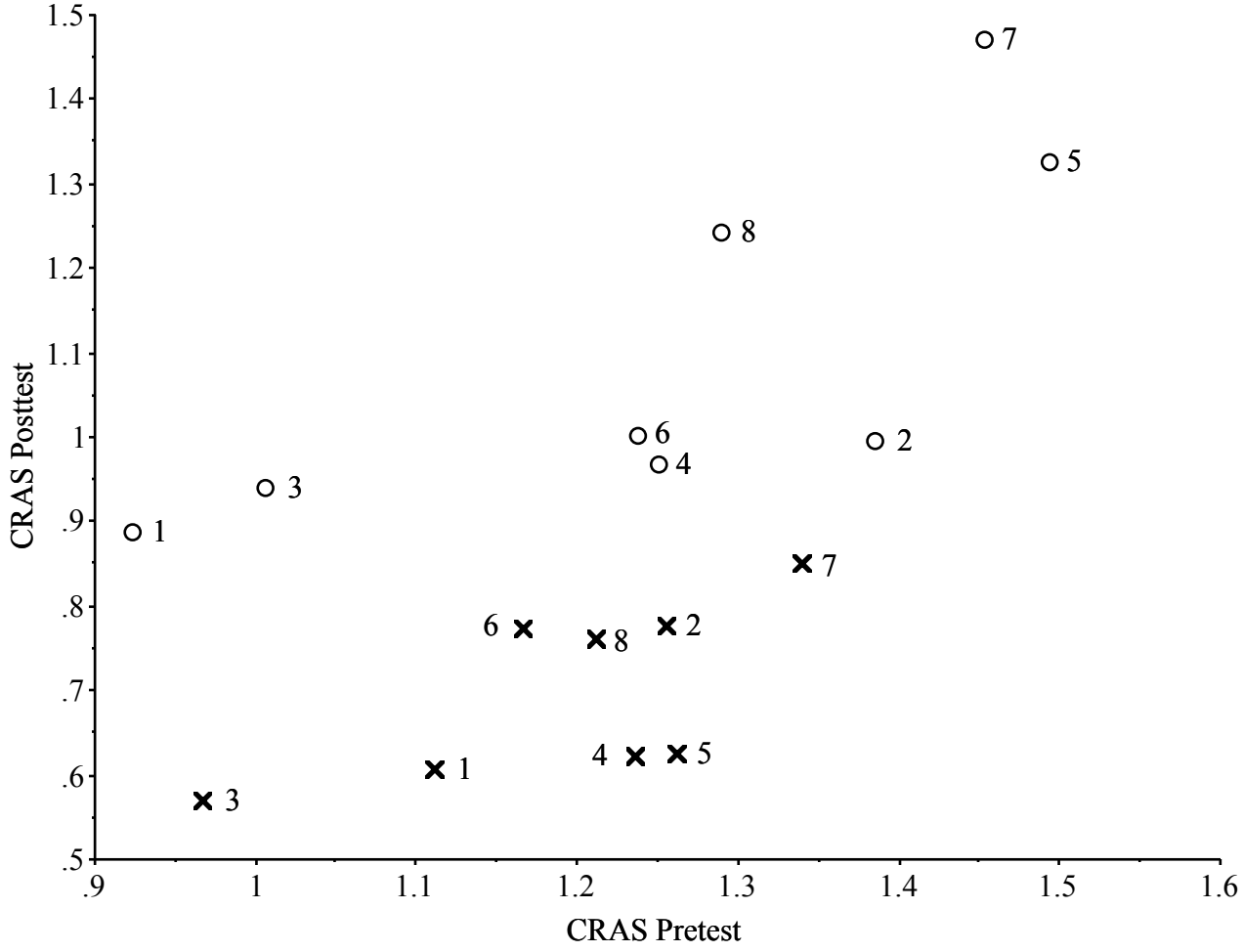

60
58
56
54
52
50
48
46
44
42
Posttest
40
42
44
46
48
50
52
54
56
58
60
Pretest
treatment effect
control group
regression line
Control
Treatment

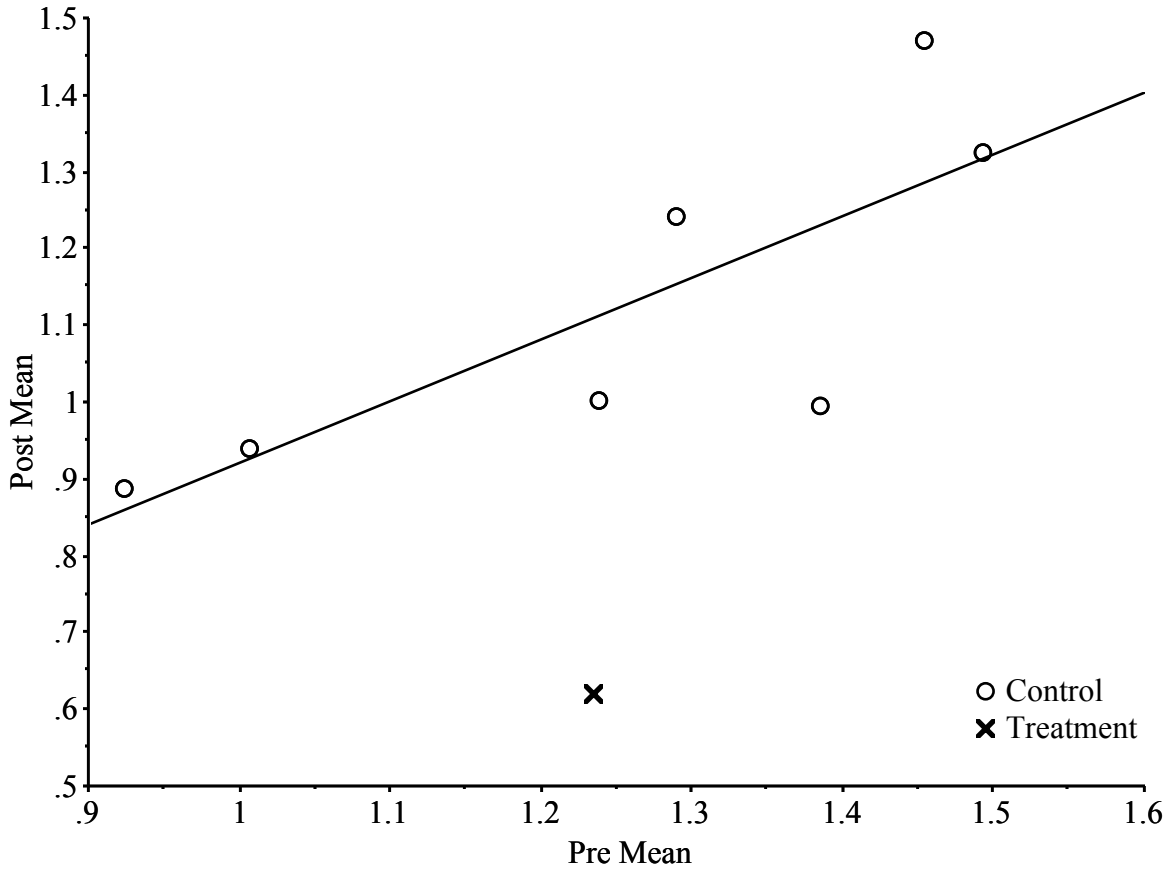

$15,000+
$10,000-14,999
$7,000-9,999
$5,000-6,999
$3,000-4,999
Under 3,000
$Physician visits per person per year
Year
6
5.8
5.6
5.4
5.2
5
4.8
4.6
4.4
4.2
4
3.8
1963
1964
1965
1966
1967
1968
1969
1970

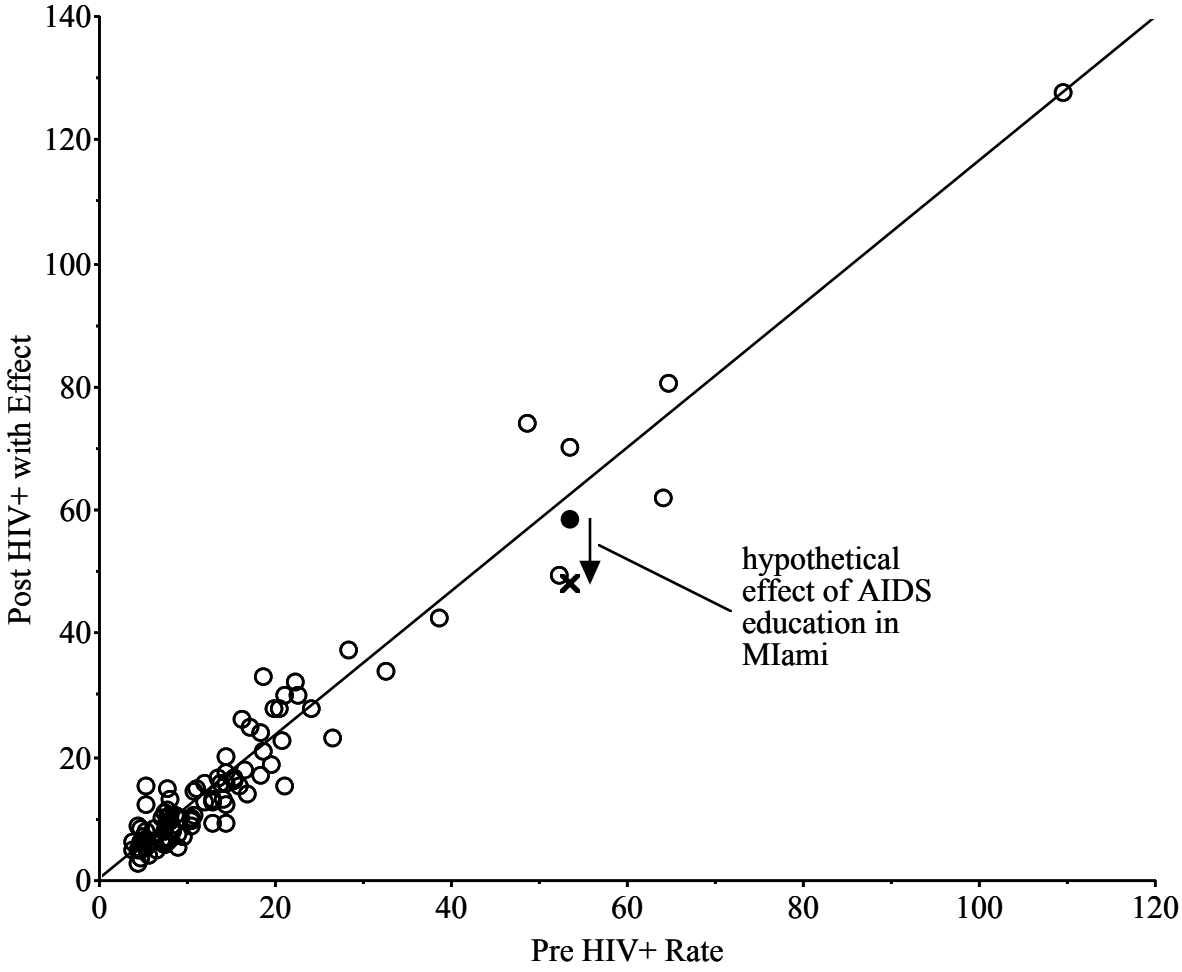

4.9
4.8
4.7
4.6
4.5
4.4
4.3
4.2
4.1
4
1967
7.75
8
8.25
8.5
8.75
9
9.25
9.5
9.75
10
10.25
10.5
Average Log Income

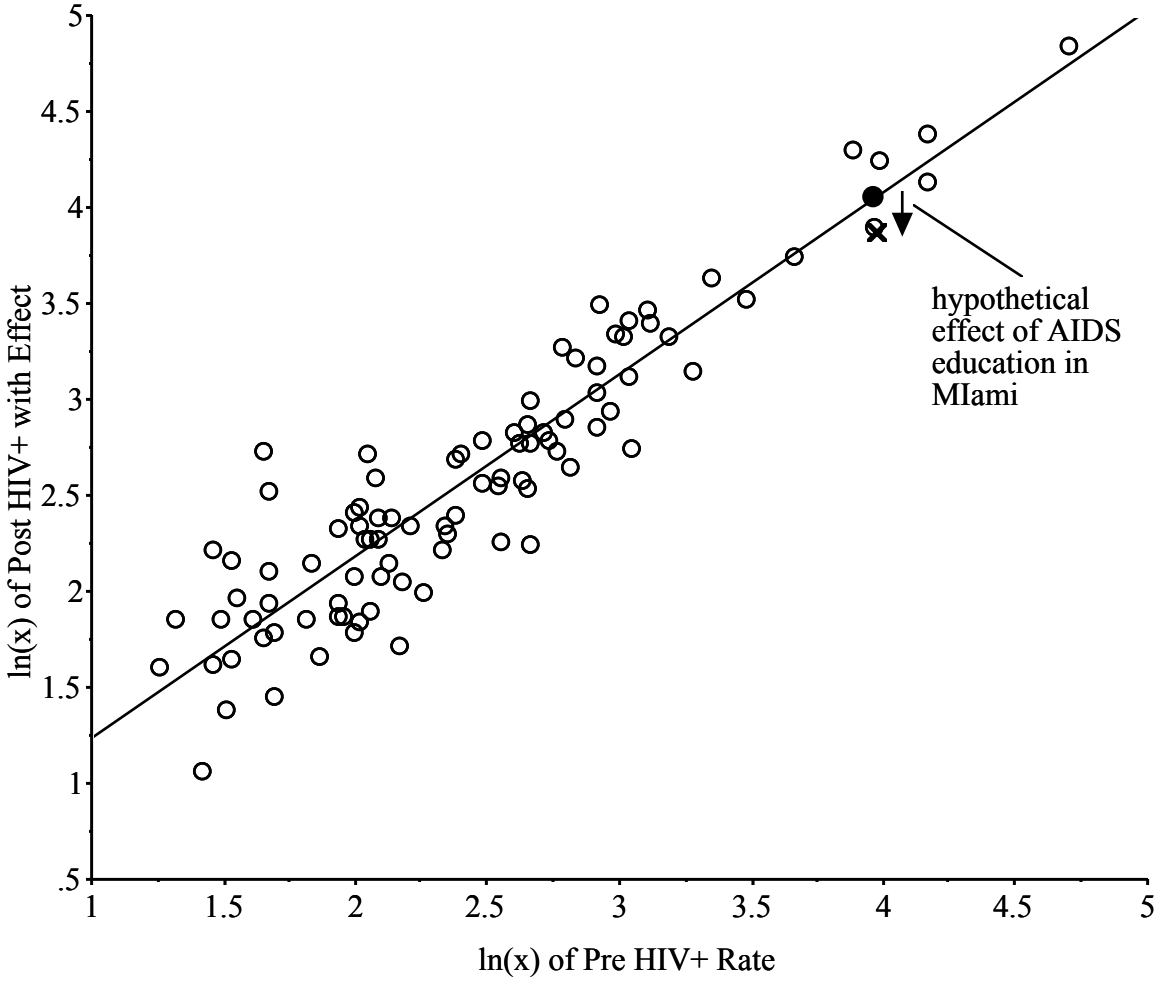

4.9
4.8
4.7
4.6
4.5
4.4
4.3
4.2
4.1
4
Physician Visits: 1967
4
4.2
4.4
4.6
4.8
5
5.2
5.4
5.6
5.8
6
Physician Visits: 1964

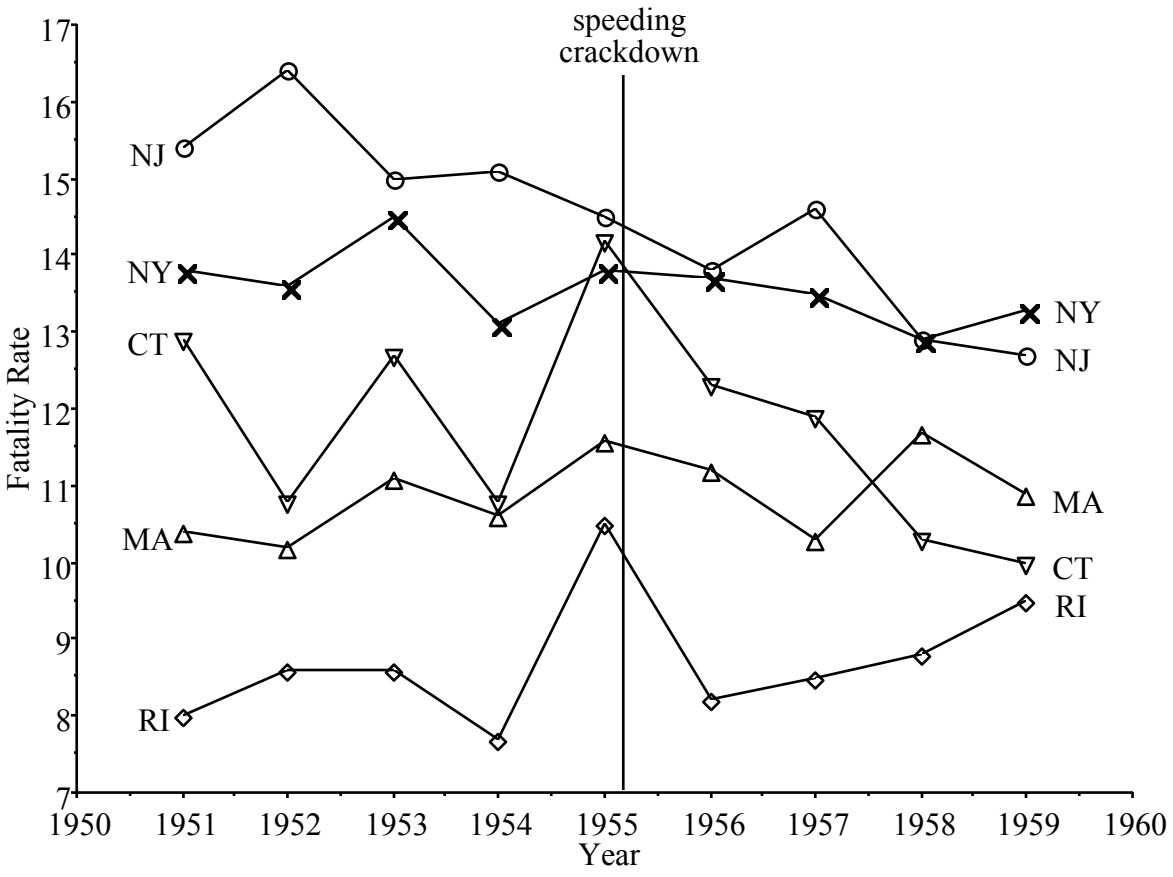

Control
Treatment
CRAS(2)
CRAS(2)
CRAS(1)
4
3.5
3
2.5
2
1.5
1
.5
0
-.5
0
.5
1
1.5
2
2.5
3
3.5

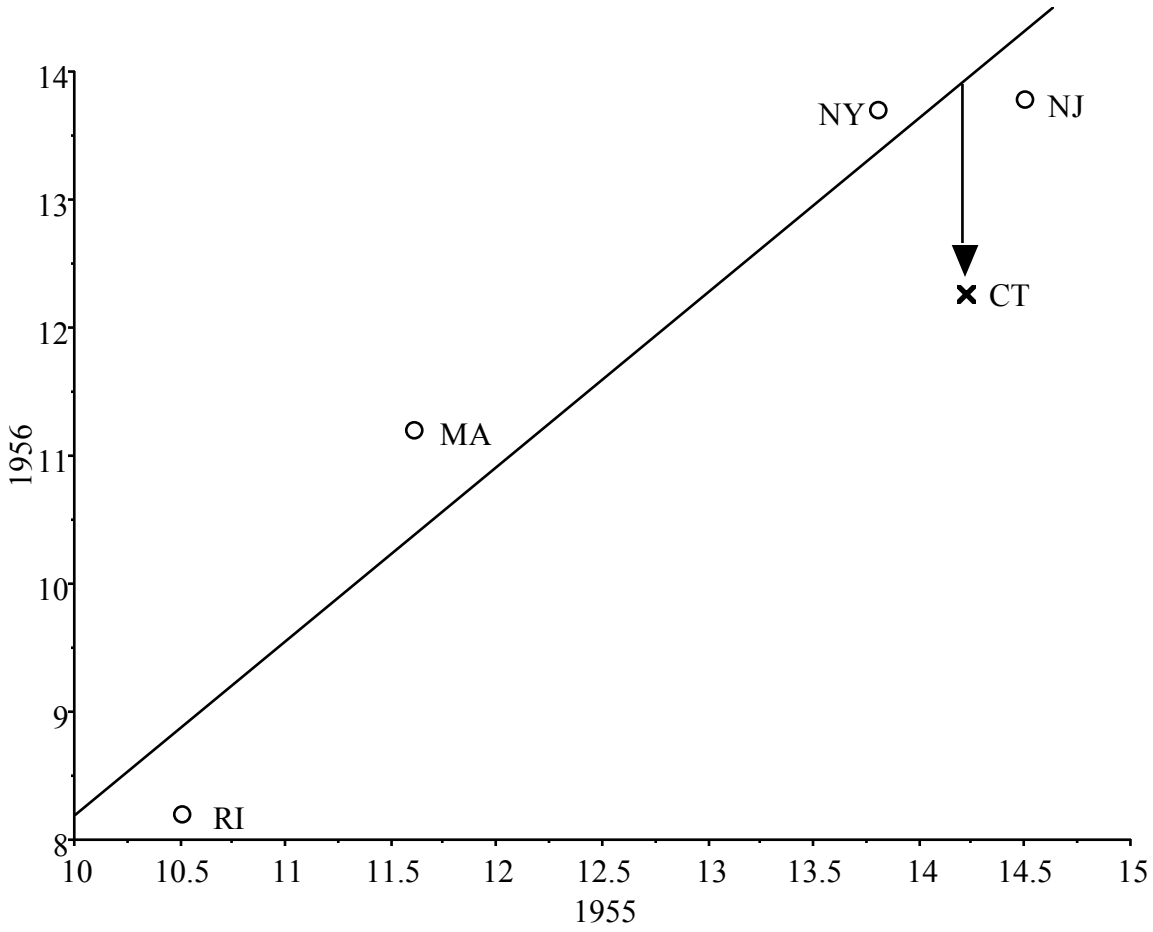

CRAS Posttest
CRAS Pretest
1.5
1.4
1.3
1.2
1.1
1
.9
.8
.7
.6
.5
.9
1
1.1
1.2
1.3
1.4
1.5
1.6
7
5
8
6
4
2
3
1
7
6
8
2
4
5
1
3


1.5
1.4
1.3
1.2
1.1
1
.9
.8
.7
.6
.5
Post Mean
.9
1
1.1
1.2
1.3
1.4
1.5
1.6
Pre Mean
Control
Treatment


140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
Post HIV+ with Effect
Pre HIV+ Rate
hypothetical
effect of AIDS
education in
MIami


5
4.5
4
3.5
3
2.5
2
1.5
1
.5
1
1.5
2
2.5
3
3.5
4
4.5
5
ln(x) of Post HIV+ with Effect
ln(x) of Pre HIV+ Rate
hypothetical
effect of AIDS
education in
MIami


speeding
crackdown
NJ
NY
CT
MA
RI
NY
NJ
MA
CT
RI
Fatality Rate
Year
17
16
15
14
13
12
11
10
9
8
7
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960


14
13
12
11
10
9
8
1956
NY
NJ
CT
MA
RI
10
10.5
11
11.5
12
12.5
13
13.5
14
14.5
15
1955