Journal of Educational Statistics
Spring 1984, Volume 9, Number 1, pp. 45–60

MAXIMUM LIKELIHOOD ANALYSIS FOR A GENERALIZED REGESSION-DISCONTINUITY DESIGN

RONALD A. VISSER
and JAN DE LEEUW
Leiden University

KEY WORDS. *Regression-discontinuity, selected groups, quasi-experimentation, truncated regression.*

ABSTRACT. The regression-discontinuity design (RDD) offers the possibility of making inferences about causal effects from observations on selected groups. The quasi-experimental groups are formed by dividing the scores of a premeasurement in two halves. The treatment effect is inferred from the differences between the regression of a postmeasurement on the premeasurement for the two groups. We discuss a generalized form of this design: (a) Apart from parallel shift of the regression lines, differences in variance and covariance are considered; (b) pretest and posttest may be multivariate; and (c) more than two groups may be involved in the design. Data from such a design are considered to have a truncated bivariate distribution. For the RDD, maximum likelihood parameter estimation procedures and tests of hypotheses are presented.

The regression-discontinuity design (RDD) was proposed by Thistlethwaite and Campbell (1960) as an alternative to ex post facto designs. The logic of the design is based on the argument that, under certain circumstances, experiments with selected groups may lead to valid conclusions if the selection mechanism is perfectly known. (For a general discussion see Rubin, 1974, 1977.) In the original RDD, individuals are assigned to one of two groups on the basis of a (quantitative) pretest. Individuals with scores below a fixed cutting score constitute one group; those with higher scores constitute the other. The groups receive different treatments and the treatment effect is evaluated by examining the differences between the regression lines of the posttest on the pretest for the two groups.

Application of the RDD is possible in many field experiments, especially in situations where the screening of individuals is a standard routine. In educational research, for example, students may be selected by means of the results on an examination. The design may then serve to evaluate the effect of a special program for students with poor results. Thus, the RDD can be used as a research method when ethical reasons prohibit experimental evaluation in education. Analogously, in epidemiology the RDD may be used to evaluate the treatment of individuals showing undesirable values on medical screening

tests. Applications of the RDD are found in Seaver and Quarton (1976) and Berk and Rauma (1983).

Thistlethwaite and Campbell (1960) confine the data analysis for the RDD to the study of parallel shift of the regression lines (cf. Campbell & Stanley, 1963; Cook & Campbell, 1979, p. 137). In general, the analysis of data from the RDD can be approached from different points of view. In all cases, the posttest Y depends on the pretest X and a dummy variable Z, indicating group membership. Reichardt (1979, p. 202) proposes an analysis of covariance of Y with X as covariate. The analysis focuses on the distance between the regression lines in the Y direction at a particular point of X. As Reichardt remarks, this may lead to problems in interpretation when the regression lines are not parallel: The effect of the treatments Z interacts with the covariate X, while one would like to eliminate the effect of X.

Another approach is to consider X and Z as stochastic regressors for Y and estimate slope and intercept of the regression line for each group separately using least squares methods. Goldberger (1964, chap. 6) shows that this leads to unbiased and consistent estimates. Differences between regression lines may be tested, and nonparallel regression lines do not cause any particular technical trouble (but, again, interpretation may be difficult).

Both approaches are closely related, using least squares, focusing on the conditional distribution of Y given X and using the between-group independence of observations. Both approaches can be extended to a generalized RDD with multivariate pretest observations and with more than two treatment groups selected by partitioning the pretest scores. Both approaches may provide satisfactory analyses in many situations, particularly when interest is mainly in the direction of the regression lines. A disadvantage of the least squares methods is that X and Z will be highly correlated in many cases, which leads to unstable estimates.

A third approach for the analysis of the RDD considers the bivariate distribution of X and Y. The observations for each treatment group can be viewed as having a truncated bivariate distribution. The truncation is the result of the fact that, by definition of the RDD, each treatment group is confined to a certain subset of pretest values. If the parameter values for the corresponding *un*truncated distribution can be estimated, these may be interpreted as the effect in the case that individuals from the whole range of X-values had been given the same treatment. This removes, for a large part, the interpretational difficulties in case of nonparallel regression lines. Also, we may estimate and test hypotheses about the parameters of the unconditional (and untruncated) distribution of Y. For example, we may study the σ_y^2 or the correlation corresponding to each treatment group, which may be interesting in some situations. In addition, it may be argued that it is quite natural to study the joint distribution of pre- and postmeasurement.

In this paper an analysis is derived according to the last point of view, using the assumption of normal distributions for the unselected groups and applying maximum likelihood theory. It is found that the analysis may be extended to a generalized RDD with multivariate pretest and with several treatment groups (as in the least squares approach) and further to multivariate posttest scores.

First, the joint distribution of the observations is derived from the assumptions, then maximum likelihood estimators and tests of hypotheses are derived and an illustration is presented. The illustration is confined to the two-group problem with univariate pre- and posttest, but not necessarily with parallel regression lines or equal posttest variances. In this univariate case, the maximum likelihood approach leads to results that are as simple as the least squares results. (Given the sample group means and [co]variances, calculations can be done on a pocket calculator.) The mathematical results for the most general problem considered are derived in the Appendix. A further comparison of the various approaches is given in the Discussion.

Derivation of the Distribution

From the statistical point of view adopted here, the generalized RDD links a number of truncated multivariate distributions defined on the regions R_j of pretest values, corresponding with the different treatments. It is the truncation that complicates the application of standard methods and makes the formulation of a special model necessary.

From the discussion of the RDD, a model for the joint distribution of the pre- and postmeasurements can be constructed. The sample consists of n independent, identically distributed pairs (X_i, Y_i). X is the premeasurement; Y the postmeasurement. There are m treatments, and for each treatment there is an untruncated density $g_j(x, y)$. The interpretation is that if everybody in the population received treatment j, then $g_j(x, y)$ would be the density of pre- and postmeasurements. Because the treatment is assumed to influence postmeasurements and not premeasurements, it follows from this interpretation that the marginals $g_j(x)$ must be the same for all j. Thus, using conditional densities, we can write $g_j(x, y) = g_j(y|x)g(x)$, with $g(x)$ the common marginal.

We now suppose that individuals are assigned to treatments on the basis of premeasurements. The space of the premeasurements is partitioned into m events R_j. The selected (or truncated) density for treatment j is obtained by restricting $g_j(x, y)$ to R_j. Thus $g_j^*(x, y) = 0$ if $x \notin R_j$, and $g_j^*(x, y) = g_j(x, y)/\pi_j$, otherwise. Here π_j is the probability that X is in R_j. It follows that the joint distribution of X and Y after selection is

$$g^*(x,y) = \sum_{j=1}^{m} \pi_j g_j^*(x,y) = \sum_{j=1}^{m} \mu_j(x) g_j(x,y), \tag{1}$$

where μ_j is the indicator function of R_j, *i.e.*, $\mu_j(x) = 1$ if $x \in R_j$ and $\mu_j(x) = 0$, otherwise. This is the basic RDD model for a single observation. The density for the n independent pairs (X_i, Y_i) is

$$\begin{aligned} g^*(\{x_i,y_i\}) &= \prod_{i=1}^{n} g^*(x_i,y_i) = \prod_{i=1}^{n} \sum_{j=1}^{m} \mu_j(x_i) g_j(x_i,y_i) \\ &= \prod_{j=1}^{m} \prod_{i \in I_j} g_j(x_i,y_i) = \prod_{j=1}^{m} \prod_{i \in I_j} g_j(y_i|x_i) \prod_{i=1}^{n} g(x_i). \end{aligned} \tag{2}$$

(Here, I_j indicates the set of indices i such that $x_i \in R_j$.) For later reference, it is also convenient to define the conditional density

$$g^*(\{y_i\}|\{x_i\}) = \prod_{j=1}^{m} \prod_{i \in I_j} g_j(y_i|x_i). \tag{3}$$

If interpreted as functions of the parameters, for fixed observations, Equations 2 and 3 define the unconditional and conditional likelihoods.

The likelihood function (2) is formally identical to the likelihood function from a sample of n independent observations from m different (untruncated!) distributions. This is a remarkable result, because in the RDD we consider a joint set of truncated distributions. The selection regions enter into the likelihood only through the N_j, which are the number of observations in selection region R_j. In the usual case, the N_j are fixed numbers; in the RDD, the N_j depend on the observations and are stochastic. The similarity between the two cases, however, implies a similarity between the maximum likelihood estimators. We comment on this similarity below. However, the stochastic character of the N_j will influence the properties of the estimators, and in particular, their variances and covariances.

Maximum Likelihood Estimators

Maximum likelihood estimators can be derived from the model (2) when an appropriate form for the untruncated densities $g_j(x,y)$ is substituted. In this paper, it is assumed that the distributions are normal. A derivation of estimators and tests for multivariate X and Y is presented in the Appendix. Here, we discuss the results for univariate X and Y.

For each bivariate normal density $g_j(x,y)$ we have the following parameters: λ and σ^2 for the mean and variance of X (the marginal does not depend on j), η_j and γ_j for the mean and variance of Y in distribution j, and δ_j for the covariance (X,Y) in distribution j. In the derivation, it is convenient to transform the parameters so that $\xi_j = \gamma_j - \delta_j^2/\sigma^2$ and $\psi_j = \delta_j/\sigma^2$. Observe the one-to-one correspondence between (σ^2,δ,γ) and (σ^2,ψ,ξ) (ξ_j is the variance of the

conditional distribution of Y given X in group j). The vector of all parameters is indicated by Θ.

The following abbreviations for sample statistics are used: $\overline{X}_j$ and $\overline{Y}_j$ are the sample means for group j; S_j and Q_j are the sample variances of X and Y for group j; W_j is the sample covariance of group j. For the total sample, the mean of X is $\overline{X}$, and the variance of X is S.

Derivation of maximum likelihood estimators is more or less straightforward, using the formal equality between the density (2) and the joint distribution in the usual case of independent and untruncated distributions. The likelihood function can be split up in a number of additive quadratic forms (see Appendix). From there it follows that $\hat{\eta}_j = \overline{Y}_j - \hat{\psi}_j(\overline{X}_j - \hat{\lambda})$. This has an intuitive appeal; the means η_j in the untruncated distribution are estimated by the group means $\overline{Y}_j$ plus a correction for the truncation. This correction is the difference between the group means $\overline{X}_j$ and the grand mean λ weighted with the direction coefficient ψ of the regression line. The distribution of the pretest X is not truncated and so the observations X_i form a natural basis to extrapolate from $\overline{Y}_j$ to η_j. Now, it follows directly that $\hat{\lambda} = \overline{X}$ and $\hat{\sigma}^2 = S$, as was to be expected from the usual maximum likelihood estimators for normal distributions. Substituting the results thus far and setting the partial derivatives with respect to ψ_j and ξ_j equal to zero, results in $\hat{\psi}_j = W_j / S_j$ and $\hat{\xi}_j = Q_j - W_j^2/S_j$.

Retransformation to the original parameters gives $\hat{\gamma}_j = Q_j + W_j^2(S - S_j)/S_j^2$ and $\hat{\delta} = S\, W_j / S_j$ for the variances and covariances. Again, a form of extrapolation can be recognized.

The maximum likelihood is given through

$$min\ L(\Theta) = n \log S + \Sigma_j N_j \log (Q_j - W_j^2/S_j) + 2n,$$

where $L(\Theta) = -2 \log l(\Theta) - 2n \log (2\pi)$, and $l(\Theta)$ is the likelihood function. It can be concluded that estimation of the parameters in the untruncated distributions from observations in the truncated distributions leads to simple and transparent expressions that are computationally simple once the usual sample statistics are known.

Likelihood Ratio Tests

In applications of the RDD it will usually be desirable not only to estimate the parameters but also to test hypotheses about the differences between the treatment groups. The maximum likelihood approach makes it possible to use likelihood ratio tests for several such hypotheses. Computation of the test statistics is equivalent to the minimization of $L(\Theta)$ under constraints corresponding to the null hypothesis. We will discuss three examples of such hypotheses.

The most restrictive hypothesis states that both groups come from the same underlying distribution: $\eta_1 = \eta_2$, $\delta_1 = \delta_2$, $\gamma_1 = \gamma_2$. Under these restrictions, the

estimators are equal to the usual (one sample) maximum likelihood estimators for the bivariate normal distribution; details are found in many places in literature.

A second hypothesis of interest could be that the two regression lines are parallel while the variances γ_j are equal. Notice that the parameter ψ_j is the direction coefficient of the regression line. Consequently, the hypothesis corresponds to the minimization of $L(\Theta)$ under the constraint $\psi_1 = \psi_2$ and $\xi_1 = \xi_2$. Substitution of these restrictions in (2) and minimizing again gives $\hat{\psi} = \tilde{W}/\tilde{S}$ and $\hat{\xi} = (\tilde{Q} - \tilde{W}^2/\tilde{S})/n$, where $\tilde{W}$ is a weighted sum $\tilde{W} = N_1W_1 + N_2W_2$, and $\tilde{S}$ and $\tilde{Q}$ are defined analogously. In this case

$$\min L(\Theta) = n \log S + n \log ((\tilde{Q} - \tilde{W}/\tilde{S})/n) + 2n.$$

A third hypothesis could be one of parallel regression lines without the requirement that $\gamma_1 = \gamma_2$. This corresponds to the minimization of $L(\Theta)$ under the restriction $\psi_1 = \psi_2$ but not necessarily $\xi_1 = \xi_2$. Substitution in (2) and subsequent minimization results in the equations

$$\hat{\xi}_j = \hat{\psi}^2 S_j - 2\hat{\psi} W_j + Q_j \text{ and}$$

$$\hat{\psi} = \frac{(N_1W_1/\hat{\xi}_1) + (N_2W_2/\hat{\xi}_2)}{(N_1S_1/\hat{\xi}_1) + (N_2S_2/\hat{\xi}_2)}.$$

Substitution of $\hat{\xi}_j$ in the expression for $\hat{\psi}$ results in a third degree equation

$$a_3\psi^3 + a_2\psi^2 + a_1\psi + a_0 = 0,$$

with

$$a_3 = nS_1S_2;$$

$$a_2 = -S_1W_2\,(2N_1 + N_2) - S_2W_1(N_1 + 2N_2);$$

$$a_1 = 2n\,W_1W_2 + N_1S_1Q_2 + N_2S_2Q_2;$$

$$a_0 = -N_1W_1Q_2 - N_2W_2Q_1.$$

This equation may be solved numerically.

Although the joint distribution in (2) differs in some respects from the usual form of joint densities, these differences do not involve the regularity conditions for the application of maximum likelihood theory. So, the maximum likelihood estimates and the corresponding chi-square statistics have the usual asymptotic distributions.

A Small Monte Carlo Study

The estimated parameter values for the untruncated distributions are extrapolations from selected groups to a whole population. It is to be expected that the choice of the selection regions will influence the accuracy of the estimators. However, the sample distributions of $\hat{\eta}_j$, $\hat{\gamma}_j$, and $\hat{\delta}_j$ are more com-

plicated than in the case where the conditional distribution of Y given X is studied. We will not enter into the sample distribution problem, but merely indicate what to expect from the estimators by performing a small Monte Carlo study.

In the Monte Carlo study both X and Y are univariate. The two regressions differ by a shift only. X is from $N(0,1)$, and Y is from $N(\pm 1,1)$; the correlation between X and Y is $+\frac{1}{2}$. There are 15 conditions in the Monte Carlo experiment. In each condition the pretest values are divided in two selection regions by a cut-off c, as in the original RDD. Five different values of c ($c = 0$, $\frac{1}{2}$, 1, 3/2, 2) and three different sample sizes ($N = 100$, 500, 1000) are studied. Each condition is replicated 10 times. In Tables I, II, and III means and standard deviations over replications are collected.

The important conclusion from the Tables is that estimation of η_2, δ_2, and γ_2 deteriorates rapidly with increasing c. This is basically because estimates of these parameters are effectively based on a small number of observations in

TABLE Ia
Results for n = 100, 10 Replications, Averages of Estimates

	$c = 0.0$	$c = 0.5$	$c = 1.0$	$c = 1.5$	$c = 2.0$
$\hat{\pi}$	0.5000	0.6800	0.8300	0.9230	0.9640
$\hat{\lambda}$	−0.0143	0.0004	0.0020	0.0080	0.0718
$\hat{\sigma}^2$	0.9854	1.0101	1.0386	1.0588	1.0258
$\hat{\eta}_1$	−0.9980	−1.0215	−0.9812	−1.0322	−0.9253
$\hat{\eta}_2$	0.9357	0.8081	1.1153	0.5723	−0.7310
$\hat{\delta}_1$	0.4352	0.4544	0.5551	0.5479	0.4640
$\hat{\delta}_2$	0.6342	0.6916	0.4389	0.6683	1.1994
$\hat{\gamma}_1$	1.0576	0.9610	1.0540	1.0800	0.9497
$\hat{\gamma}_2$	1.1815	1.1011	0.9986	2.3129	25.3917

TABLE Ib
Results for n = 100, 10 Replications, Standard Deviations

	$c = 0.0$	$c = 0.5$	$c = 1.0$	$c = 1.5$	$c = 2.0$
$\hat{\pi}$	0.0544	0.0480	0.0173	0.0300	0.0128
$\hat{\lambda}$	0.1125	0.0943	0.0708	0.0840	0.1107
$\hat{\sigma}^2$	0.0589	0.0630	0.1146	0.1537	0.1485
$\hat{\eta}_1$	0.3421	0.1372	0.0910	0.0923	0.0863
$\hat{\eta}_2$	0.2583	0.2502	0.7189	2.1839	11.2850
$\hat{\delta}_1$	0.3110	0.1352	0.1027	0.0863	0.1366
$\hat{\delta}_2$	0.2279	0.2119	0.4243	1.1572	4.6650
$\hat{\gamma}_1$	0.2401	0.1928	0.1500	0.0956	0.1838
$\hat{\gamma}_2$	0.2619	0.2918	0.6888	2.4279	49.9567

the "tail" of the bivariate distribution, if c is large. If n is increased, this effect is counterbalanced. On the whole, the results are quite encouraging. If $c = 1$, for example, a total sample of 100 already gives quite decent estimates.

Illustration

As an illustration, data are analyzed from the so-called COPIH project (Bonjer, Van der Lee, & Jonkers, 1981; Van der Lee, 1983). The purpose of the project is to educate employees about their life-style in relation to the risk of heart diseases. A total of 15,274 subjects were screened by occupational health services. If the values of some indicators were beyond certain bounds, the subjects were given advice about their life-style, especially their eating and smoking habits.

After a few years the subjects were screened again and the values of the indicators were compared to evaluate the effect of the advice. In this example

TABLE IIa
Results for n = 500, 10 Replications, Averages of Estimates

	$c = 0.0$	$c = 0.5$	$c = 1.0$	$c = 1.5$	$c = 2.0$
$\hat{\pi}$	0.4966	0.6912	0.8336	0.9302	0.9762
$\hat{\lambda}$	0.0054	0.0104	0.0203	0.0209	−0.0043
$\hat{\sigma}^2$	1.0139	1.0009	0.9943	1.0196	1.0157
$\hat{\eta}_1$	−0.9898	−0.9971	−0.9831	−1.0010	−1.0028
$\hat{\eta}_2$	1.0238	1.0378	0.8911	0.7937	1.1226
$\hat{\delta}_1$	0.4946	0.4937	0.4897	0.5125	0.5233
$\hat{\delta}_2$	0.5108	0.4956	0.5630	0.6048	0.4847
$\hat{\gamma}_1$	0.9874	1.0164	1.0072	1.0158	1.0167
$\hat{\gamma}_2$	1.0082	1.0294	1.0270	1.0882	2.2334

TABLE IIb
Results for n = 500, 10 Replications, Standard Deviations

	$c = 0.0$	$c = 0.5$	$c = 1.0$	$c = 1.5$	$c = 2.0$
π	0.0201	0.0199	0.0135	0.0083	0.0061
$\hat{\lambda}$	0.0491	0.0391	0.0537	0.0558	0.0255
$\hat{\sigma}^2$	0.0610	0.0478	0.0498	0.0489	0.0521
$\hat{\eta}_1$	0.1082	0.0925	0.0493	0.0402	0.0359
$\hat{\eta}_2$	0.0724	0.1776	0.3267	0.5346	2.7800
$\hat{\delta}_1$	0.0528	0.0593	0.0379	0.0528	0.0621
$\hat{\delta}_2$	0.1008	0.1584	0.2239	0.2617	1.1899
$\hat{\gamma}_1$	0.0796	0.0790	0.0493	0.0743	0.0823
$\hat{\gamma}_2$	0.1179	0.1680	0.2709	0.4105	1.6657

one of the indicators, the amount of serum cholesterol in the blood, is analyzed. A plot of mean posttest values against mean pretest values is shown in Figure 1. Sufficient statistics, intermediate results, and estimated values of the parameters are represented in Table IV. To test the hypothesis of parallel regression lines, the minimum value of $L(\Theta)$ was computed under $H_0: \psi_1 = \psi_2$, $\xi_1 = \xi_2$. This gives the following estimated values

$$\hat{\psi} = .61, \qquad \hat{\delta} = 89.8,$$
$$\hat{\xi} = 56.0, \qquad \hat{\gamma} = 110.9,$$
$$\hat{\eta}_1 = 63.4, \qquad \hat{\eta}_2 = 62.5.$$

The likelihood ratio statistics, i.e., the difference between min $L(\Theta)$ and min $L(\Theta)$ under H_0 equals 312.3. Under H_0 this statistic has asymptotically a χ^2 distribution with $df = 2$, the difference in number of parameters. Clearly, the result is significant, as could be expected by looking at Figure 1.

TABLE IIIa
Results for n = 1000, 10 Replications, Averages of Estimates

	$c = 0.0$	$c = 0.5$	$c = 1.0$	$c = 1.5$	$c = 2.0$
$\hat{\pi}$	0.4972	0.6878	0.8371	0.9326	0.9774
$\hat{\lambda}$	0.0012	0.0101	0.0029	0.0177	0.0065
$\hat{\sigma}^2$	0.9909	1.0039	1.0063	0.9941	1.0075
$\hat{\eta}_1$	−1.0054	−0.9895	−0.9828	−0.9893	−0.9877
$\hat{\eta}_2$	0.9560	0.9551	1.0390	0.8647	1.8915
$\hat{\delta}_1$	0.4992	0.5135	0.5024	0.4787	0.5127
$\hat{\delta}_2$	0.5302	0.5346	0.4839	0.5797	0.1472
$\hat{\gamma}_1$	1.0044	1.0156	0.9735	0.9768	1.0209
$\hat{\gamma}_2$	1.0247	1.0446	0.9720	1.1831	0.9806

TABLE IIIb
Results for n = 1000, 10 Replications, Standard Deviations

	$c = 0.0$	$c = 0.5$	$c = 1.0$	$c = 1.5$	$c = 2.0$
$\hat{\pi}$	0.0107	0.0085	0.0124	0.0048	0.0058
$\hat{\lambda}$	0.0237	0.0248	0.0258	0.0244	0.0313
$\hat{\sigma}^2$	0.0288	0.0507	0.0490	0.0361	0.0322
$\hat{\eta}_1$	0.0960	0.0578	0.0312	0.0238	0.0293
$\hat{\eta}_2$	0.0629	0.1136	0.1696	0.3636	1.3600
$\hat{\delta}_1$	0.0647	0.0648	0.0430	0.0290	0.0293
$\hat{\delta}_2$	0.0430	0.0942	0.1220	0.1759	0.5768
$\hat{\gamma}_1$	0.1049	0.0637	0.0537	0.0472	0.0496
$\hat{\gamma}_2$	0.0469	0.0774	0.1255	0.2375	0.3327

The results may be summarized as follows. As a general effect the mean and variance have decreased between pretest and posttest for both groups. Although the sample mean of Y in the intervention group (70.8) is much higher than in the control group (59.8), there are no differences when the results from both groups are extrapolated to the whole range of X. In the intervention group, the regression line has a smaller slope (there is a smaller correlation). An interpretation could be that, although the intervention would not give an overall effect in the mean for the whole population, it has a positive effect for at least some people in the extreme intervention group. The rotation of the regression line is a positive effect in the range of the intervention group. The decrease of the correlation, and subsequent increase of the conditional variance (ξ_j), may possibly be the result of a differential effect of the intervention. In the analysis it is seen that it can be misleading to study the conditional variance of Y only. The extrapolated variance (γ) in the intervention group is even smaller than in the control group.

TABLE IV
Summarizing Statistics and Estimates for the COPIH Blood Serum Cholesterol Data (unit: mmol/10 l)

		Group 1 (control)	Group 2 (intervention)
Basic statistics			
N_j		10243	5031
ΣX		619855	406548
ΣX^2		38044839	33174464
$\Sigma X.Y$		37131850	28957419
ΣY		607706	356285
ΣY^2		36791194	25669959
Intermediate results			
$\bar{X}$		60.52	80.81
$\bar{Y}$		59.33	70.82
S_j		52.16	63.98
Q		71.92	87.18
W		34.81	33.10
Parameter estimates			
mean y	η	63.79	63.78
	ψ	0.67	0.52
	ξ	48.69	70.05
variance y	γ	114.16	109.40
covariance	δ	98.11	76.06
correlation	ρ	0.76	0.60

Note. $n = 15274$; $\bar{X} = 67.20$; $S = 147.03$.

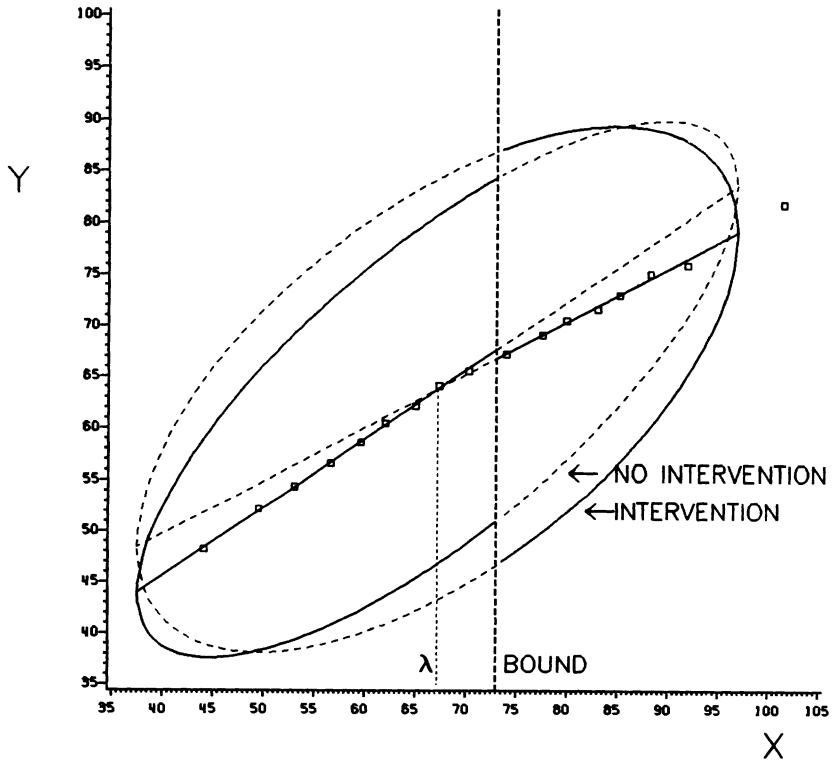
To check for possible nonlinearity in the regression (which could also explain the above results), the conditional means of Y for a number of X values are plotted in Figure 1. There is no strong indication for deviations from linearity.

Discussion

The approach toward the RDD chosen in this paper is based on the idea that selection on the pretest implies truncation of the joint distribution of X and Y. Maximum likelihood methods make it possible to estimate parameters in such a model. This approach is related to work by Birnbaum, Paulson, and Andrews (1950) on selection and to Anderson (1957) on missing data in multivariate analysis.

The main advantage of this approach over the least squares approach is that it facilitates the comparison between nonequivalent groups without using essentially more complicated results. The estimated parameter values may be

FIGURE 1. Plot of the mean posttest values against pretest values and estimated regression lines for the COPIH blood serum cholesterol data (unit: mmol/10 l).



viewed as values that would appear when individuals from the whole range of X had been given the same treatment. So, not only means and direction coefficients can be compared for the different groups, but also (co)variances and correlations. The illustration presented above shows the danger of interpreting only conditional variances when one is interested in the effect of treatments on dispersion. In other situations, the correlation between pre- and posttest can be of interest. The use of selected X values may cause a serious bias in the estimation of ρ (Carroll, 1961). The extrapolation formulas derived here give a correction for this bias. In conclusion, the maximum likelihood approach uses the knowledge of the selection mechanism in the RDD in a more consequent way than the least square approach, where the marginal distribution of X is discarded.

A clear disadvantage of the maximum likelihood approach is the inability to make a simple generalization toward nonlinear regression, as is possible in the least squares approach. The assumption of multivariate normality implies the linearity.

In general, we do not wish to overemphasize the difference between the approaches. The purpose of the RDD is to compare nonequivalent groups. In some situations the least squares approach will do this. In other situations it is convenient to use the "natural" extrapolation formulas derived here. In this paper we show that these extrapolations may be viewed as the consequence of multivariate normality. But of course these formulas can be used on a more heuristic basis without the assumptions (as long as linear regression is appropriate).

A few remarks remain about the method itself. The observation that the likelihood function is formally equal to the likelihood function obtained if we sample fixed numbers from untruncated distributions has an important practical consequence. It means that we can use programs such as LISREL (Jöreskog & Sörbom, 1981) for the parameter estimation. Of course, in the cases above, where closed form estimates are available, this is quite unnecessary. Observe, however, that the statistical properties of the estimates will be different from those in the fixed sample case, which means that we cannot use the estimated standard error from LISREL. In general, it is to be expected that the random N_j will increase the standard errors.

The danger exists that under unfavourable circumstances the method is sensitive to deviations from the assumptions. The method is based on the extrapolation from a (possibly small) part of a distribution to the whole distribution. A small deviation from normality may then lead to relatively inaccurate estimates, particularly when the selection regions coincide with the tail of a distribution. Therefore, it is advisable not to use too extreme parts of the range of the pretest as selection regions in practice.

APPENDIX

Derivation of ML Estimators in the General Case

The Density Function

Suppose X is a p-dimensional random vector of pretreatment variables and Y is a q-dimensional random vector of posttreatment variables. Both X and Y are defined on the same probability space. The range of X, that is, $\Gamma\boldsymbol{R}^p$, is partitioned into m measurable subsets $R_1, \ldots, R_m$, the selection regions, for which $P(X \epsilon R_j) > 0, j = 1, \ldots, m$. Notice that R_j is not necessarily connected, nor are there any restrictions on the smoothness of the boundaries. We consider m multinormal distributions on $\Gamma\boldsymbol{R}^{p+q}$, with densities $g_j(x,y)$, mean vector (λ,η_j), and dispersion matrix

$$D_j = \begin{pmatrix} \Sigma & \Delta_j \\ \Delta_j' & \Gamma_j \end{pmatrix},$$

which is nonsingular for each $j = 1, \ldots, m$. It follows that the marginal density $g_j(x,\cdot)$ of X, depending on λ and Σ only, is independent of j. The multivariate normal densities are substituted in (2). The vector of all parameters λ, η_j, Σ, Δ_j, Γ_j for $j = 1, \ldots, m$, is indicated by Θ. Then, the likelihood function is given by

$$l(\Theta) = \Pi_{j=1}^{m} \Pi_{i \epsilon I_j} g_j(x_i, y_i),$$

where I_j is the set of indices i for which X_i in R_j (I_j is a "random set" with a distribution over all subsets of $\{1,2,\ldots n\}$). N_j is the number of elements in I_j; N_j are multinomially distributed.

Derivation of the ML Estimators

In the sequel it will be understood that Σ_j means summation over $j = 1, \ldots, m$, and Σ_i means summation over all $i \epsilon I_j$. Maximization of $l(\Theta)$ is equivalent to minimization of

$$L(\Theta) = \Sigma_j N_j \log|D_j| + \Sigma_j \Sigma_i \left(\begin{pmatrix} X_i \\ Y_i \end{pmatrix} - \begin{pmatrix} \lambda \\ \eta_j \end{pmatrix} \right)' D_j^{-1} \left(\begin{pmatrix} X_i \\ Y_i \end{pmatrix} - \begin{pmatrix} \lambda \\ \eta_j \end{pmatrix} \right), \tag{A1}$$

which is $-2\log l(\Theta)$ plus a constant. $L(\Theta)$ can be simplified and split up into eight additive components. We use the following notation:

$$\begin{aligned}
&\bar{X}_j = (1/N_j)\Sigma_i X_i \text{ and } \bar{Y}_j = (1/N_j)\Sigma_i Y_i; \\
&S_j = (1/N_j)\Sigma_i (X_i - \bar{X}_j)(X_i - \bar{X}_j)'; \\
&Q_j = (1/N_j)\Sigma_i (Y_i - \bar{Y}_j)(Y_i - \bar{Y}_j)'; \\
&W_j = (1/N_j)\Sigma_i (X_i - \bar{X}_j)(Y_i - \bar{Y}_j)'; \\
&\Xi_j = \Gamma_j - \Delta_j' \Sigma^{-1} \Delta_j; \\
&\psi_j = \Sigma^{-1} \Delta_j.
\end{aligned}$$

Observe that the parameterization (Σ,Γ,Δ) for the dispersion matrices D corresponds one-to-one with the parameterization (Σ,Ξ,ψ). Thus, we may minimize $L(\Theta)$ as a function of $(\lambda,\Sigma,\eta_j,\Xi_j,\psi_j)$. Applying some well-known results on partitioned determinants and partitioned inverses to the D_j matrices, we may write

$$\begin{aligned}
L(\Theta) = {} & n \log|\Sigma| + \\
& + \Sigma_j \; N_j \log|\Xi_j| + \\
& + \Sigma_j \, N_j \operatorname{tr} \Sigma^{-1} S_j +
\end{aligned}$$

$$
\begin{aligned}
&+ \Sigma_j N_j \operatorname{tr} \Xi_j^{-1} Q_j + \\
&+ \Sigma_j N_j \operatorname{tr} \psi_j \Xi_j^{-1} \psi_j' S_j + \\
&- 2\Sigma_j N_j \operatorname{tr} \psi_j \Xi_j^{-1} W_j' + \\
&+ \Sigma_j N_j (\overline{X}_j - \lambda)' \Sigma^{-1} (\overline{X}_j - \lambda) + \\
&+ \Sigma_j N_j (\overline{Y}_j - (\zeta_j + \psi_j' (\overline{X}_j - \lambda)))' \Xi_j^{-1} (\overline{Y}_j - (\zeta_j + \psi_j (\overline{X}_j - \lambda))). \qquad \text{(A2)}
\end{aligned}
$$

The value of λ for which $L(\Theta)$ is minimized is found by equating the partial (vector) derivatives equal to zero. Specifically, $\hat{\lambda} = \Sigma_{i=1}^{n} X_i/n = \overline{X}$, as was to be expected. Then, the mean vectors $\hat{\eta}_j$ can be found by observing that they occur only in the last part of Equation A2, which is minimized by choosing $\hat{\eta}_j = \overline{Y}_j - \hat{\psi}_j (\overline{X}_j - \overline{X})$, making this part equal to zero. Now, there are only two parts left containing Ψ_j. From this we find $\hat{\psi}_j = S_j^{-1} W_j$. The solution for Σ and Ξ_j is obtained as in the usual maximum likelihood estimates for multinormal distributions. After substitution of all estimators found thus far, and writing $U_j = Q_j - W_j' S_j^{-1} W_j$, we find

$$
\begin{aligned}
&\hat{\Xi}_j = U_j; \\
&\hat{\Sigma} = (1/n) \Sigma_{i=1}^{n} (\overline{X}_i - \overline{X})(X_i - \overline{X})'.
\end{aligned}
$$

Transformation to Δ and Γ results in

$$
\begin{aligned}
\hat{\Delta}_j &= \hat{\Sigma} S_j^{-1} W_j; \\
\hat{\Gamma}_j &= Q_j - W_j' S_j^{-1} (I - \hat{\Sigma} S_j^{-1}) W_j,
\end{aligned}
$$

while

$$
\min L(\Theta) = n \log|\hat{\Sigma}| + \Sigma_j N_j \log|U_j| + n(p + q).
$$

ML Estimators Under Constraints

To compute statistics for likelihood-ratio tests, it is necessary to minimize $L(\Theta)$ under constraints corresponding to the null hypothesis. The most restrictive hypothesis is that all m distributions are equal. This may be translated into minimization under the restriction that η_j, Δ_j, and Γ_j are constant over the m regions. Of course, this problem leads to the usual (one sample) estimators for multinormal distributions.

A second hypothesis could be that all regression surfaces are parallel hyperplanes, that is, only the η_j differ. This leads to

$$
\begin{aligned}
\hat{\psi}_j &= S_*^{-1} W_*; \\
\hat{\Xi} &= Q_* - W_*' S_*^{-1} W_*,
\end{aligned}
$$

and

$$
\min L(\Theta) = n \log|\hat{\Sigma}| + n \log|Q_* - W_*' S_*^{-1} W_*| + n(p + q),
$$

where

$$
\begin{aligned}
Q_* &= \Sigma_j N_j Q_j/n; \\
W_* &= \Sigma_j N_j W_j/n; \\
S_* &= \Sigma_j N_j S_j/n.
\end{aligned}
$$

Transforming the parameters, this corresponds to

$$
\begin{aligned}
\hat{\Delta}_j &= \hat{\Sigma} S_*^{-1} W_* \\
&= W_* + B S_*^{-1} W_*,
\end{aligned}
$$

where $B = \Sigma_j N_j(\bar{X}_j - \bar{X})\ (\bar{X}_j - \bar{X})'/n$ is the between-regions dispersion, because $\hat{\Sigma} = S_* + B$, and

$$\hat{\Gamma}_j = Q_* - W_*'S_*^{-1}BS_*W_*.$$

There is one restricted problem left of some practical importance. This corresponds to the hypothesis of parallel regression surfaces without the restriction of equal within-region covariance matrices, that is, the ψ_j are all equal, but not necessarily the Ξ_j. There is no simple closed form for the ML estimators in this case. Thus, some numerical approximation procedure will be necessary to test such a hypothesis.

Acknowledgment

We wish to thank two referees for suggestions that substantially improved the presentation, and Arnold van der Lee for making the data and the graphics available.

References

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association, 52,* 200–203.

Berk, R. A., & Rauma, D. (1983). Capitalization on nonrandom assignment to treatment: A regression-discontinuity evaluation of a crime control program. *Journal of the American Statistical Association, 78,* 21–38.

Birnbaum, Z. W., Paulson, E. & Andrews, F. C. (1950). On the effect of selection performed on some coordinates of a multi-dimensional population. *Psychometrika, 15,* 191–204.

Bonjer, F. H., Van der Lee, A. P. M., & Jonkers, A. H. (1981). How to measure the effectiveness of intervention trials. In D. A. B. Lindberg & P. L. Reichertz (Eds.), *Lecture notes in medical informatics, Vol. II: Medical Information Europe 81* (pp. 731–736). Berlin: Springerverslag.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand-McNally.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26*, 347–372.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation.* Chicago: Rand-McNally.

Goldberger, A. S. (1964). *Econometric theory.* New York: Wiley.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V, Users Guide*. Uppsala: Department of Statistics, University of Uppsala.

Reichardt, C. S. (1979). The statistical analysis of data from non-equivalent group design. In T. D. Cook & D. T. Campbell, *Quasi-experimentation* (Chap. 4). Chicago: Rand-McNally.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2,* 1–26.

Seaver, W. B., & Quarton, R. J. (1976). Regression-discontinuity analysis of dean's list effects. *Journal of Educational Psychology, 68,* 459–465.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51,* 309–317.

Van der Lee, A. P. M. (1983). Het schatten van een interventie effect zonder controlegroep. Leiden: Research Report Subfaculty of Psychology, Leiden University.

Authors

RONALD A. VISSER, Professor, Department of Research Methodology in Psychology, Hooigracht 15, 2312 KM Leiden, The Netherlands. *Specializations:* Statistics, longitudinal research.

JAN DE LEEUW, Professor, Department of Data Theory, FSW/RUL, Middelstegracht 4, 2312 TW Leiden, The Netherlands. *Specializations:* Multivariate analysis, multidimensional scaling.