Regression Discontinuity: Examining Model Misspecification

Randall E. Schumacker
University of Alabama

The Regression Discontinuity (RD) design looks similar to the non-equivalent group design, which uses analysis of covariance, but assumptions and advantages are much different. The major problem in analyzing data from the RD design is model misspecification. If the regression equation or statistical model does not reflect the data distribution, then biased estimates of the treatment effect will occur. For example, if the true pre-post relationship is curvilinear, but the regression equation only modeled linear regression effects, the treatment effects would be biased. However, a statistical approach is possible using a full model with all terms specified and then test restricted sub-models that omit individual parameters.

The basic RD Design is a two-group pretest-posttest model and is depicted as follows:

C O X O
C O O

The RD design looks similar to the non-equivalent group design, which uses analysis of covariance, but assumptions and advantages are much different (Campbell, 1989; Loftin & Madison, 1991; Schumacker, 1992). The RD design does not have subject selection bias (pre-defined group membership) rather uses a pre-test measure to assign treatment or non-treatment status. The basic RD model would have an intercept term, pre-test measure, and dummy-coded group assignment variable regressed on a post-test measure. The pre-test measure does not have to be the same as the post-test measure.

The major problem in analyzing data from the RD design is model misspecification. If the regression equation or statistical model does not reflect the data distribution, then biased estimates of the treatment effect will occur. For example, if the true pre-post relationship is curvilinear, but the regression equation only modeled linear regression effects, the treatment effects would be biased. Consequently, it is a good idea to visually inspect the pre-post scatter plot to see what type of relationship exists. However, a statistical approach is possible using a full model with all terms specified and then test restricted sub-models that omit individual parameters. This will be illustrated in this paper.

There are five central assumptions when performing an RD analysis. A major concern is the model specification in the pre-post distribution being a polynomial function rather than a logarithmic or exponential function. The five central assumptions are:

1. The cutoff value must be absolute without exception. A subject selection bias is introduced and the treatment effect is biased if incorrect assignment to groups based on the cutoff value occurred (unless it is known to be random).
2. The pre-post distribution is a polynomial function. If the pre-post relationship is logarithmic, exponential or some other function, the model is misspecified and the treatment effect is biased. The data can be transformed to create a polynomial distribution prior to analysis to yield appropriate model specification.
3. There must be a sufficient number of pretest values in the comparison group to estimate the pre-post regression line.
4. The experimental and comparison groups must be formed from a single continuous pretest distribution with the division between groups determined by the cutoff value.
5. The treatment or program intervention must be delivered to all subjects, i.e., all receive the same reading program, amount of training, etc.

Model specification can be identified in three different ways or types: exactly specified, over specified, and under specified RD models. An exactly specified model has an equation that fits the "true" data. So if the "true" data is linear then a simple straight-line pre-post relationship with a treatment effect would yield unbiased treatment effects. The RD equation would include a term for the posttest Y, the pretest X, and the dummy-coded treatment variable Z with no unnecessary terms. When we exactly specify the true model, we get unbiased and efficient estimates of the treatment effect. If the RD equation is over specified it includes additional parameter estimates that are not required, i.e. interaction or curvilinear coefficients, and treatment effect would be inefficient. If the RD equation is under specified it leaves out important parameter estimates and the treatment effect would be biased.

The basic steps being proposed to statistical test the type of model when conducting an RD analyses would be as follows:

1. Subtract the cut-off score from the pretest score ($X_{pre} - X_{cut}$).
2. Visually examine the pre-post scatter plot for type of data relationship.
3. Determine if any higher-order polynomial terms or interactions are present.
4. Estimate the "full" RD regression equation.
5. Modify the RD equation by dropping individual non-significant terms.

Methodology

The "full" RD regression equation with subsequent "modified" or "restricted" regression models permit one to statistically determine the best fitting model for estimating treatment effects. A "full" regression discontinuity model could be as outlined below.

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + \beta_5 X_i^2 Z_i + e_i$$

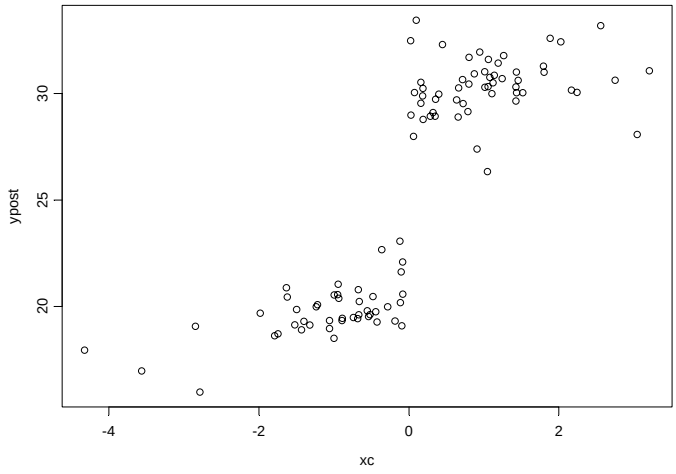
The RD regression equation terms are defined as:

y_i = post test score outcome for ith subject
β_0 = regression coefficient for intercept
β_1 = linear pre test regression coefficient
β_2 = mean post test different for treatment group
β_3 = linear interaction regression coefficient between pre and group
β_4 = quadratic regression coefficient for pretest
β_5 = quadratic interaction regression coefficient for pre test and group
X_i = transformed pre test score for ith subject
Z_i = group assignment based on cut off score (0 = comparison, 1 = treatment)
e_i = residual score for ith subject.

Data Simulation

The S-PLUS program that generated the simulated data and computed results for the RD analysis is footnoted (SPLUS, 2005). The *rnorm* function in S-PLUS generated 100 random normal data points (Chambers, Mallows, & Stuck, 1976). The post test scores (*Y*) and pre test scores (*X*) were created by adding residual error (*ey* or *ex*) to this random normal variable (*true*). Group assignment (*Z*) was determined based on subtracting a cut score of 20 from the pre test score (1–treatment, 0–comparison). This 10 point treatment gain was added to the post test score (*Y*). Optional *print* and *write* statements are included to either view or save the data in a file.

Figure 1. Simulated Regression Discontinuity data



The least squares regression function, *lm*, was used to run the RD analyses where ypost = post test score; xc = transformed pre test score; z = group assignment; xz = linear interaction; xsq = quadratic pre test; and xsqz = quadratic interaction of pre test and group. The sequence of RD regression equations that were tested are as follows:

1. Full model: lm (y_{post} ~ xc + z + xz + xsq + xsqz)
2. No quadratic Interaction: lm (y_{post} ~ xc + z + xz + xsq)
3. No quadratic Interaction: lm (y_{post} ~ xc + z + xz)
4. Linear model: lm (y_{post} ~ xc + z)
5. No pre test model: lm (y_{post} ~ xc)

A visual inspection of the simulated data in Figure 1 indicates that we would expect the best fitting RD model to be the linear model. The scatter plot displays the y_{post} (post test scores) and xc (transformed pre test scores) variables. A ten point treatment effect is visible between the two groups. Recall that the treatment group had a mean of 30 and the comparison group had a mean of 20, which are visually present in the scatter plot.

Results

The full model results indicated that all regression coefficients were non-significant. The model misspecification (*over specified*) further indicated an inefficient treatment effect (*z* = -87.86), which we know is not true given the simulated data.

RD Full Model **(F=384.3, df = 5, 95)**

```
Coefficients:
              Value Std. Error   t value  Pr(>|t|)
(Intercept)  -9.7719   57.9279    -0.1687    0.8664
         xc  -1.9094    5.2984    -0.3604    0.7194
          z -87.8617  108.9808    -0.8062    0.4222
         xz   9.9789   10.5763     0.9435    0.3478
        xsq   0.0766    0.1455     0.5261    0.6001
       xsqz  -0.2570    0.2588    -0.9932    0.3232
```

The RD restricted model that dropped the quadratic interaction between squared pre test and group is still *over specified* because all of the regression coefficients were non-significant. The treatment effect was inefficient and over estimated at 19.15 (Z) compared to the known treatment effect of 10 points.

RD – drop quadratic interaction of pre test and group **(F = 480.2, df = 4, 95)**

```
Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  22.5800  47.8979    0.4714   0.6384
         xc   1.0476   4.3824    0.2390   0.8116
          z  19.1494  16.3454    1.1715   0.2443
         xz  -0.4933   0.8209   -0.6010   0.5493
        xsq  -0.0047   0.1203   -0.0392   0.9688
```

The RD model with both quadratic terms removed is still *over specified* and yielded a larger *F* value, however, the linear interaction (*xz*) between pre test and group was not statistically significant (*t* = -1.81; p = .07). The treatment effect was also inefficient and higher than the known true treatment effect value.

RD – drop both quadratic interaction effects **(F = 646.9, df = 3, 96)**

```
Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  20.7031   0.2793   74.1277   0.0000
         xc   0.8761   0.1991    4.4010   0.0000
          z  19.7477   5.8067    3.4008   0.0010
         xz  -0.5234   0.2891   -1.8104   0.0734
```

The RD model with all interaction terms removed is an *exactly specified* model. This RD analysis modeled the "true" nature of the linear relationship between pre and post scores and yielded an intercept value of 20, which is close to the comparison group mean and a treatment effect of 9.26, which is close to the known treatment effect of 10 points, given the introduction of random error.

RD – drop linear interaction **(F = 946.5, df = 2, 97)**

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 20.4362  0.2400   85.1603  0.0000
         xc  0.6279  0.1460    4.2995  0.0000
          z  9.2592  0.3945   23.4706  0.0000
```

The RD model without the pre test score term removed is an *under specified* model. This RD analysis yielded a biased treatment effect that overestimated the "true" effect of 10 points. The F value is inflated and a key variable, the pre test score was omitted. Recall that *under specified* models leave out important variables, hence affect the model validity.

RD No pre test Model (F = 1591, df = 1, 98)

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  19.7607   0.1969   100.3519   0.0000
          z  10.5900   0.2655    39.8842   0.0000
```

Conclusions

The RD design was one of three designs approved for program evaluation by the Department of Education many decades ago, yet the technique is not widely used (Thistlethwaite & Campbell, 1960; McNeil, 1984; Trochim, 1984). The regression discontinuity design uses a least-squares equation to yield an intercept (baseline measure) and regression weight (treatment effect measure) in assessing program effectiveness. A positive or negative regression weight determines gain or loss due to treatment or intervention effect, which is also tested for statistical significance. However, if the regression model is misspecified then treatment effects are inefficient and biased estimates.

RD is a powerful alternative to using quasi-experimental designs with distinct advantages. Regression discontinuity has fewer assumptions in comparison to not meeting assumptions in quasi-experimental designs that use analysis of covariance, i.e., random sampling; normality of treatment levels; homogeneity of variance; independence of variance estimates; linear regression assumption; and homogeneity of regression lines. The analysis of covariance assumptions are seldom met, thus leading to erroneous interpretations of treatment effects (Campbell, 1989; Loftin & Madison, 1991).

The RD normal distribution assumption is not problematic and can be handled by robust regression methods or probit data transformation. The cut-off score misspecification is usually not a problem because state agencies or school districts mandate a cut-off score for high-stakes testing. The model misspecification can also be examined by including linear, polynomial, and interaction terms in the RD equation and then dropping non-significant terms. Other advantages include RD designs being able to explore treatment effect differences at different cutoff points, use different pre-test measures than post-test measures, do not require matching of subjects, and can use multiple comparison groups with different cutoff scores.

Educational researchers should therefore make increased use of the regression-discontinuity technique for program evaluation because you can use a different pre-test measure for the cut-off value, use different regression models that reflect the distribution of the data (linear, curvilinear, and interaction), and do not have to meet all of the assumptions in ANCOVA to yield stable estimates of treatment effects.

References

Chambers, J. M., Mallows, C. L. and Stuck, B. W. (1976). A Method for Simulating Stable Random Variables. *Journal of the American Statistical Association*, *71*, 340-344.

Campbell, K.T. (1989). *Dangers in using analysis of covariance procedures*. ERIC Document # ED 312298 (http://eric.ed.gov).

Loftin, L., & Madison, S. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), (1991). *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 133-148). Greenwich, CT: JAI Press. (International Book Sellers Number: 1-55938-316-X)

McNeil, Keith (April, 1984). *Random Thoughts on Why the Regression Discontinuity Design Is Not Widely Used*. Paper presented at the American Educational Research Association Annual Meeting. New Orleans, LA.

Schumacker, Randall E. (April, 1992). *Factors Affecting Regression-Discontinuity*. Paper presented at the American Educational Research Association Annual Meeting. San Francisco, CA.

S-PLUS (2005). *S-PLUS 6 User's Guide for Windows*. Insightful, Inc., Seattle, WA.

Thistlethwaite, D.L. & Campbell, D.T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Educational Psychology*, *51*(*6*), 309-317.

Trochim, William M. K. (1984). *Research Design for Program Evaluation, the Regression Discontinuity Approach*. Sage Publications: Beverly Hills, CA.

Send correspondence to: Randall E. Schumacker
University of Alabama
Email: rschumacker@ua.edu