Regression Discontinuity Models and the Variance Inflation Factor

Randall E. Schumacker
University of Alabama

The regression-discontinuity design (RD) is a powerful methodological alternative to the quasi-experimental design when conducting evaluations. The RD design involves testing post-test differences between the experimental and comparison group regression lines at the cutoff point for statistical significance. Regression discontinuity models can involve linear, curvilinear, and interaction terms in the model specification, which are not orthogonally specified. Consequently, a variance inflation problem may exist when using regression discontinuity models in evaluation designs. This study investigated the impact of variance inflation on parameters specified in full and restricted regression discontinuity models. It is recommended that VIF be considered when including interaction effects in RD designs.

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. 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.

There are five central assumptions when performing an RD analysis. These 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.

Regression Discontinuity Model Specification

The major concern when analyzing data from the RD design is whether the model or regression equation is correctly specified. If the regression equation or 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.

Three types of model specifications are possible: 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.

RD Modeling Steps

The basic steps to conducting RD analyses would 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.

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.

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

The RD regression equation terms are defined as:

y_i = post test score outcome for i^{th} 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 i^{th} subject
Z_i = group assignment based on cut off score (0 = comparison, 1 = treatment)
e_i = residual score for i^{th} subject.

Variance Inflation Factor

When a full RD regression model is specified, multicollinearity amongst the terms is possible. Multicollinearity can inflate the variance amongst the variables in the model. These inflated variances are problematic in regression because some variables add very little or even no new and independent information to the model (Belsley, Kuh & Welsch, 1980). Although Schroeder, Sjoquist and Stephen (1986) assert that there is no statistical test that can determine whether or not multicollinearity is a problem, there are ways for detecting multicollinearity (Berry and Feldman, 1985).

A recommended approach is to use the Variance Inflation Factor (VIF). VIF measures the impact of multicollinearity among the *X*'s in a regression model on the precision of estimation. It expresses the degree to which multicollinearity amongst the predictors degrades the precision of an estimate. VIF is a statistic used to measured possible multicollinearity amongst the predictor or explanatory variables. VIF is computed as $1/(1-R^2)$ for each of the k–1 independent variable equations. For example, given 4 independent predictor variables, the independent regression equations are formed by using each k-1 independent variable as the dependent variable:

$X_1 = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_4 + e_1$
$X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + e_2$
$X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_3$

Each independent variable model will return an R^2 value and VIF value. The term to exclude in the model is then based on the value of VIF. If X_j is highly correlated with the remaining predictors, its variance inflation factor will be very large. A general rule is that the VIF should not exceed 10 (Belsley, Kuh, & Welsch, 1980). When X_j is orthogonal to the remaining predictors, its variance inflation factor will be 1.

Methods

Data Simulation

The appendix contains an S-PLUS program that generated the simulated data for the study. The rnorm function in S-PLUS generated 100 random normal data points and output nine variables listed in the data command [data <-c(y,x,z,gain,ypost,xc,xz,xsq,xsqz)]. 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.

Regression Discontinuity Models

The least squares regression function, *lm*, was used to run the RD analyses. The S-Plus program includes separate *lm* regression functions for several regression equations. The summary command produced the regression output. The regression discontinuity models begin with a full model followed by a sequence of restricted models. The full regression model and the sequence of restricted models are listed below:

1. Full model: $y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + \beta_5 X_i^2 Z_i + e_i$
2. No Quadratic Interaction: $y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + e_i$
3. No Quadratic Terms: $y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + e_i$
4. Linear Model: $y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + e_i$
5. No Pre-Test Model: $y_i = \beta_0 + \beta_1 Z_i + e_i$

Variance Inflation Factor

The variance inflation factor is computed in several popular statistics packages (S-PLUS, SPSS, and SAS). In this study, the data simulation, regression function, and variance inflation function were all written in S-PLUS. The simulated data generated using the S-PLUS program in the appendix was created and used by the S-PLUS regression and variance inflation functions. A variance inflation function, *vif,* was created and used with the *summary* function following the *lm* regression function for each of the regression equations. The S-PLUS variables were labeled as follows in the full regression equation: ypost = xc + z + xz + xsq + xsqz.

Results

The descriptive statistics for the RD variables are in Table 1. The intercorrelations amongst the terms in the full RD regression model equation are in Table 2. The RD regression discontinuity results with the VIF values for the full model are in Table 3 for the dependent variable ypost.

Table 1. Descriptive Statistics (*N*=100)

	Mean	Std. Deviation
ypost	25.5852	5.45566
xc	.0899	1.35059
z	.55	.500
xz	11.5740	10.53759
xsq	405.4103	53.68331
xsqz	243.8876	223.08339

Table 2. Pearson Correlation Matrix of Full RD Regression Model

	ypost	xc	z	xz	Xsq	xsqz
ypost	1.000	.821	.971	.971	.821	.969
xc	.821	1.000	.785	.807	.999	.827
z	.971	.785	1.000	.999	.787	.994
xz	.971	.807	.999	1.000	.811	.998
xsq	.821	.999	.787	.811	1.000	.833
xsqz	.969	.827	.994	.998	.833	1.000

Table 3. Full Regression Model and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	-9.772	57.928	-.168	.866		
	xc	-1.909	5.298	-.360	.719	.000	3467.21
	z	-87.861	108.980	-.806	.422	.001	201043.10
	xz	9.978	10.576	.943	.347		841002.30
	xsq	.076	.145	.526	.600	.000	4131.94
	xsqz	-.257	.258	-.993	.323	.001	225640.50

Table 4. Restricted Regression Model (no xsqz) and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	22.580	47.898	.471	.638		
	xc	1.048	4.382	.239	.812	.000	2372.36
	z	19.149	16.345	1.172	.244	.000	4523.18
	xz	-.493	.821	-.601	.549	.000	5067.20
	xsq	-.005	.120	-.039	.969	.000	2825.33

Table 5. Restricted Regression Model (no xsq) and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	20.703	.279	74.128	.000		
	xc	.876	.199	4.401	.000	.202	4.95
	z	19.748	5.807	3.401	.001	.002	576.84
	xz	-.523	.289	-1.810	.073	.002	635.16

Table 6. Restricted Regression Model (no xz) and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	20.436	.240	85.160	.000		
	xc	.628	.146	4.300	.000	.384	2.60
	z	9.259	.395	23.471	.000	.384	2.60

Table 7. Restricted Regression Model (no z) and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	25.287	.314	80.643	.000		
	xc	3.317	.233	14.249	.000	1.000	1.00

Summary & Conclusion

The requirement of a correctly specified RD regression model is linked to multicollinearity of the independent variables in the equation. Table 2 suggests that multicollinearity is present amongst the independent predictors in the RD regression equation, i.e. β_1 = linear pre test regression coefficient (xc); β_2 = mean post test different for treatment group (z); $\beta 3$ = linear interaction regression coefficient between pre and group (xz); β_4 = quadratic regression coefficient for pretest (xsq); and β_5 = quadratic interaction regression coefficient for pre test and group (xsqz). Table 3 indicates that VIF is well beyond the acceptable level of 10 for each of the independent predictor variables in the model. Similar results were

found for the set of independent predictor variables in Table 4, especially note the non-significant treatment effect (z) with an extreme VIF factor. Table 5 indicated that the linear pre test regression coefficient (xc) was acceptable, however, the other independent predictors VIF were too high, i.e., the treatment effect is now significant, but has an extreme VIF factor. In Table 6, a two predictor model with linear pre test and treatment group had both a significant *t*-test value (t = 23.471, p = .0001) and an acceptable VIF factor; thus an acceptable RD model. Table 7, indicated a baseline RD model with linear pre test scores and an expected corresponding VIF = 1.0.

The regression discontinuity approach to analyzing evaluation data is more robust to violations than the corresponding quasi-experimental design that is commonly used in state and federal grant data analysis. However, model misspecification can result in erroneous conclusions regarding program gains. Correspondingly, if the variance inflation factor is not considered along with model specification, then multicollinearity amongst the predictor variables can inflate the variance leading to misinterpretation of the R-squared values and treatment gain. A visual presentation of overlap by the independent variables is also possible (Stine, 1995). It is therefore recommended that model specification along with the variance inflation factor be checked when using regression discontinuity.

References

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression Diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley.

Berry, W. D. & Feldman, S. (1985). *Multiple regression in practice*. London: Sage Publications.

Schroeder, L. D., Sjoquist, D. L. & Stephan, P. E. (1986). *Understanding regression analysis*. Beverly Hills, CA: Sage Publications.

Stine, Robert, A. (1995, February). Graphical Interpretation of Variance Inflation Factors. *The American Statistician, 49*, 53-56.

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

APPENDIX
S-PLUS Program

```
#
# Data for Normal Distribution
# Pretest X cutoff score is 20 (mean X)
# Program Gain is 10
# Mean Posttest Y is 30
# XC is Pre test minus cut score to center at 0 point
# ex and ey add residual error to true score
#

seed <-1357
set.seed(seed)              # same seed value so results can be reproduced

true <- rnorm(100,20,1)
ex   <- rnorm(100,0,1)
ey   <- rnorm(100,0,1)

x <- true + ex              # create y and x scores with residual error
y <- true + ey

z <- ifelse(x >= 20, 1, 0)  # assign treatment group using pretest cutoff score

gain <- (10 * z)            # add 10 point to treatment group (z = 1)

ypost <- y + gain           # add 10 points to post test score

xc<- (x-20)                 # subtract cut score from pre test

xz<-x*z                     # linear interaction pre test and group

xsq<-x*x                    # quadratic interaction

xsqz<-xsq*z                 # quadratic interaction pre test and group
```

```
data<-c(y,x,z,gain,ypost,xc,xz,xsq,xsqz)

RD.data<-matrix(data,nrow=100,byrow=F)
dimnames(RD.data)
dim(RD.data)  #100 rows 9 columns
variables<-c("y","x","z","gain","ypost","xc","xz","xsq","xsqz")
dimnames(RD.data)<-list(NULL,variables)

#print(RD.data)
#save generated data in ASCII file
write.table(RD.data, file = "RD.txt", sep=",", append=F)

#
#Variance Inflation Factor Function
#

vif <- function(object, ...)
UseMethod("vif")

vif.default <- function(object, ...)
stop("No default method for vif.  Sorry.")

vif.lm <- function(object, ...)

{
  V <- summary(object)$cov.unscaled
  Vi <- crossprod(model.matrix(object))
        nam <- names(coef(object))
  if(k <- match("(Intercept)", nam, nomatch = F)) {
                v1 <- diag(V)[-k]
                v2 <- (diag(Vi)[-k] - Vi[k, -k]^2/Vi[k,k])
                nam <- nam[-k]
        } else {
                v1 <- diag(V)
                v2 <- diag(Vi)
                warning("No intercept term detected.  Results may surprise.")
        }
        structure(v1*v2, names = nam)
}

#
#RD Regression models with Variance Inflation Factor
#
#Sequence of RD equations
#

fit <- lm (ypost~xc + z + xz + xsq + xsqz)
summary(fit)
vif(fit)

fit <- lm (ypost~xc + z + xz + xsq)
summary(fit)
vif(fit)

fit <- lm (ypost~xc + z + xz)
summary(fit)
vif(fit)

fit <- lm (ypost~xc + z)
summary(fit)
vif(fit)

fit <- lm (ypost~xc)
summary(fit)
vif(fit)
```