**Category**

* [Advanced Modeling](https://datascienceplus.com/category/advanced-modeling/)

**Tags**

* [Linear Regression](https://datascienceplus.com/tag/linear-regression/)
* [Logistic Regression](https://datascienceplus.com/tag/logistic-regression/)
* [Multiple Regression](https://datascienceplus.com/tag/multiple-regression/)
* [R Programming](https://datascienceplus.com/tag/rstats/)

Regression analysis is one of the most requested machine learning methods in 2019. One group of regression analysis for measuring effects and to evaluate a policy program is Regression Discontinuity Design. This method is well suited for benchmarking and finding improvements for optimization in organizations. It can, therefore, be used to design organizations so they generate more value for employees and customers. In this article, you learn how to do Regression Discontinuity Design in R.

**Program Evaluation Definition**

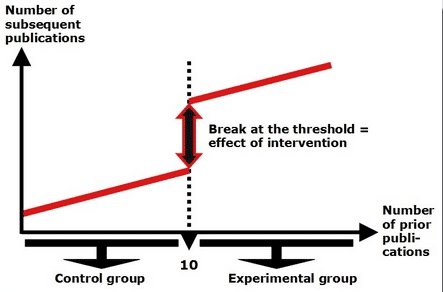
Program evaluation is the selection of outcomes of a program or policy, evaluated by a set of standards as a means to investigate if the program or policy contributes to improvement in a desired goal or result. When beginning an evaluation, program people therefore often want the answer questions like:

* Does the program work? And how can it be improved?
* Is the program worthwhile?

The main goal of program evaluation is, therefore, to investigate if the policy change contributes to the improvement of the program. Program evaluation can help improve program services, in particular:

* Planning the evaluation
* Determining the evaluation questions (Determining the goal of the evaluation: What is the evaluation question, what is the evaluation to find out?)
* Answering evaluation questions with evaluation methods (What methods will be used?)
* Making the results useful by determining how will the results be reported so that they can be used by the organization to make improvements

**Regression Discontinuity Design**

Regression Discontinuity Design (RDD) requires a continuous eligibility score on which the population of interest is ranked and a clearly defined cutoff point above or below which the population is determined eligible for a program.  
The purpose of RDD is therefore to evaluate before- and after effect of a policy program as plottet in the below graph:  
[](https://i0.wp.com/datascienceplus.com/wp-content/uploads/2019/08/p1-2.png?ssl=1)

The eligibility index must be continuous around the cutoff point and the population of interest around the cutoff point must be very similar in observable and unobservable characteristics.  
RDD estimates local average treatment effects around the cutoff point, where treatment and comparison units are most similar. It provides useful evidence on whether a program should be cut or expanded at the margin. However, it does not answer whether a program should exist or not: in this case, the average treatment effect provides better evidence than the local average treatment effect.

RDD is a key method in the toolkit of any applied researcher interested in unveiling the causal effects of policies. The use of RDD has increased exponentially in the last few years. Researchers have used it to evaluate electoral accountability; SME policies; social protection programs such as conditional cash transfers; and educational programs such as school grants.

In RDD, assignment of treatment and control is not random but rather based on some clear-cut threshold (or cutoff point) of an observed variable such as age, income, and score. Causal inference is then made comparing individuals on both sides of the cutoff point.

**Conditions and Assumptions in RDD**

Two main conditions are needed in order to apply a regression discontinuity design:

**Conditions in RDD**

A continuous eligibility index: a continuous measure on which the population of interest is ranked (i.e. test score, poverty score, age).  
A clearly defined cutoff point: a point on the index above or below which the population is determined to be eligible for the program. For example, students with a test score of at least 80 of 100 might be eligible for a scholarship, households with a poverty score less than 60 out of 100 might be eligible for food stamps, and individuals age 67 and older might be eligible for a pension. The cutoff points in these examples are 80, 60, and 67, respectively. The cutoff point may also be referred to as the threshold.

**Assumptions in RDD**

The eligibility index should be continuous around the cutoff point. There should be no jumps in the eligibility index at the cutoff point or any other sign of individuals manipulating their eligibility index in order to increase their chances of being included in or excluded from the program. The McCrary Density Test tests this assumption by checking the eligibility index density function for discontinuities around the cutoff point.  
Individuals close to the cutoff point should be very similar, on average, in observed and unobserved characteristics. In the RD framework, this means that the distribution of the observed and unobserved variables should be continuous around the threshold. Even though researchers can check the similarity between observed covariates, the similarity between unobserved characteristics has to be assumed. This is considered a plausible assumption to make for individuals very close to the cutoff point, that is, for a relatively narrow window.

RDD estimates local average treatment effects around the cutoff point, where treatment and comparison units are most similar. The units to the left and right of the cutoff look more and more similar as they near the cutoff. Given that the design meets all assumptions and conditions outlined above, the units directly to the left and right of the cutoff point should be so similar that they lay the groundwork for comparison as well as does randomized assignment of the treatment.

Because the RDD estimates the local average treatment effects around the cutoff point, or locally, the estimate does not necessarily apply to units with scores further away from the cutoff point. These units may not be as similar to each other as the eligible and ineligible units close to the cutoff. RDD’s inability to compute an average treatment effect for all program participants is both a strength and a limitation, depending on the question of interest. If the evaluation primarily seeks to answer whether the program should exist or not, then the RDD will not provide a sufficient answer: the average treatment effect for the entire eligible population would be the most relevant parameter in this case. However, if the policy question of interest is whether the program should be cut or expanded at the margin, then the RDD produces precisely the local estimate of interest to inform this important policy decision.

**RDD Methodology**

The estimation of the treatment effects can be performed parametrically as follows:

$$ y\_i=\alpha+\delta X\_i+h(Z\_i)+\varepsilon\_i $$

where \(y\_i\) is the outcome of interest of individual \(i\), \(X\_i\) is an indicator function that takes the value of 1 for individuals assigned to the treatment and 0 otherwise, \(Z\_i\) is the assignment variable that defines an observable clear cutoff point, and \(h(Z\_i ) \) is a flexible function in Z. The identification strategy hinges on the homogeneity of Z at the threshold. It is standard to center the assignment variable at the cutoff point. In this case, one would use \( h(Z\_i-Z\_0 ) \) instead of with \( Z\_0 \) being the cutoff. With that assumption, the parameter of interest, \( \delta \), provides the treatment effect estimate.

In the case of a sharp design with perfect compliance, the parameter \( \delta \) identifies the average treatment effect on the treated (ATT). In the case of a fuzzy design, \( \delta \) corresponds to the intent-to-treat effects – i.e. the effect of the eligibility rather than the treatment itself on the outcomes of interest. The LATE can be estimated using an IV approach. This could be done as follows:

First stage:  
$$ P\_i =\alpha+\delta X\_i+h(Z\_i)+\varepsilon\_i $$

Second stage:  
$$ y\_i=\mu+\delta\hat{P\_i}+h(Z\_i)+u\_i $$

Where P\_i is a dummy variable that identifies actual participation of individual \( i \) in the program/intervention. Notice that with a parametric specification, you should specify \( h(Z\_i ) \) the same way in both regressions.

**Regression Discontinuity Design in R**

Let us start making a RDD in R. We will try three models with the AEJfigs.dta [dataset](http://masteringmetrics.com/resources/)

**RDD Model #1**

# Loading packages in R

library(AER)

Loading required package: car

Loading required package: lmtest

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

as.Date, as.Date.numeric

Loading required package: sandwich

Loading required package: survival

library(foreign)

library(rdd)

Loading required package: Formula

library(stargazer)

Please cite as:

Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

R package version 5.2. <http://CRAN.R-project.org/package=stargazer>

# Read Data and do data management in R

AEJfigs=read.dta("AEJfigs.dta")

All = all deaths

AEJfigs$age = AEJfigs$agecell - 21

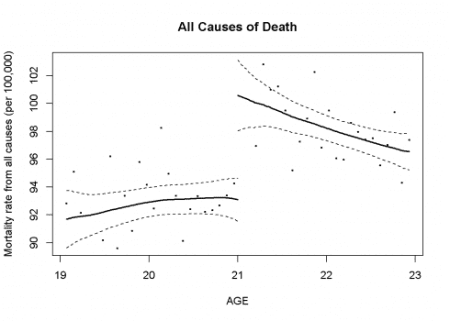
AEJfigs$over21 = ifelse(AEJfigs$agecell >= 21,1,0)

# RDD model no. 1

reg.1=RDestimate(all~agecell,data=AEJfigs,cutpoint = 21)

plot(reg.1)

title(main="All Causes of Death", xlab="AGE",ylab="Mortality rate from all causes (per 100,000)")

The above code gives us the following plot:  
[](https://i0.wp.com/datascienceplus.com/wp-content/uploads/2019/08/p1-1.png?ssl=1)

Let us also see RDD model no. 1 with coefficients:

summary(reg.1)

*Call:*

*RDestimate(formula = all ~ agecell, data = AEJfigs, cutpoint = 21)*

*Type:*

*sharp*

*Estimates:*

*Bandwidth Observations Estimate Std. Error z value*

*LATE 1.6561 40 9.001 1.480 6.080*

*Half-BW 0.8281 20 9.579 1.914 5.004*

*Double-BW 3.3123 48 7.953 1.278 6.223*

*Pr(>|z|)*

*LATE 1.199e-09 \*\*\**

*Half-BW 5.609e-07 \*\*\**

*Double-BW 4.882e-10 \*\*\**

*---*

*Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1*

*F-statistics:*

*F Num. DoF Denom. DoF p*

*LATE 33.08 3 36 3.799e-10*

*Half-BW 29.05 3 16 2.078e-06*

*Double-BW 32.54 3 44 6.129e-11*

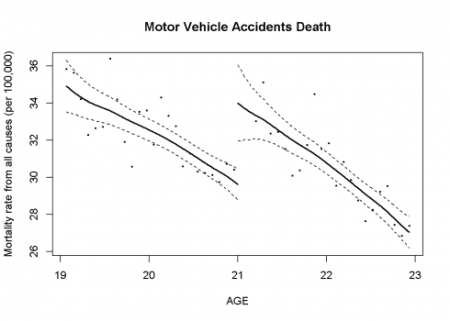
**RDD Model #2**

Let us now make RDD model no. 2 in R:

reg.2=RDestimate(mva~agecell,data=AEJfigs,cutpoint = 21)

plot(reg.2)

title(main="Motor Vehicle Accidents Death", xlab="AGE",ylab="Mortality rate from all causes (per 100,000)")

The above code gives us the following plot:  
[](https://i1.wp.com/datascienceplus.com/wp-content/uploads/2019/08/p2-1.png?ssl=1)

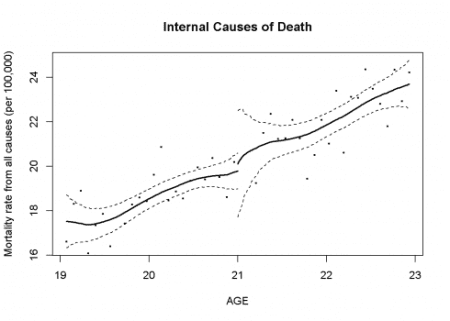
**RDD Model #3**

At last let us try making RDD model no. 3 in R:

reg.3=RDestimate(internal~agecell,data=AEJfigs,cutpoint = 21)

plot(reg.3)

title(main="Internal Causes of Death", xlab="AGE",ylab="Mortality rate from all causes (per 100,000)")

The above code gives us the following plot:  
[](https://i1.wp.com/datascienceplus.com/wp-content/uploads/2019/08/p3-1.png?ssl=1)

Let us also make a summary of RDD model no. 3 in R:

summary(reg.3)

*Call:*

*RDestimate(formula = internal ~ agecell, data = AEJfigs, cutpoint = 21)*

*Type:*

*sharp*

*Estimates:*

*Bandwidth Observations Estimate Std. Error z value*

*LATE 0.8809 22 1.4128 0.8206 1.722*

*Half-BW 0.4405 10 1.8691 1.0203 1.832*

*Double-BW 1.7618 42 0.7652 0.6179 1.239*

*Pr(>|z|)*

*LATE 0.08513 .*

*Half-BW 0.06698 .*

*Double-BW 0.21553*

*---*

*Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1*

*F-statistics:*

*F Num. DoF Denom. DoF p*

*LATE 6.830 3 18 5.734e-03*

*Half-BW 1.765 3 6 5.068e-01*

*Double-BW 22.695 3 38 2.750e-08*