Scientific Research Publishing

Open Journal of Statistics, 2022, 12, 676-690
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

A Modification to the Fuzzy Regression Discontinuity Model to Settings with Fuzzy Variables

Portia Kuzivakwashe Mafukidze[1], Samuel Musili Mwalili[2], Thomas Mageto[2]

[1]Department of Mathematics, The Pan African University, Institute for Basic Sciences, Technology and Innovation (PAUSTI), Nairobi, Kenya
[2]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
Email: portiamafukidze@gmail.com

How to cite this paper: Mafukidze, P.K., Mwalili, S.M. and Mageto, T. (2022) A Modification to the Fuzzy Regression Discontinuity Model to Settings with Fuzzy Variables. *Open Journal of Statistics*, **12**, 676-690.
https://doi.org/10.4236/ojs.2022.125040

Received: September 24, 2022
Accepted: October 23, 2022
Published: October 26, 2022

Abstract

Despite the fact that fuzzy regression discontinuity designs are growing in popularity, a lot of research takes into account treatment non-compliance difficulties, specifically the fuzziness of the treatment impact. This paper took into account independent and dependent fuzzy factors when creating these designs. Additionally we took into account treatment non-compliance difficulties, specifically the fuzziness of the treatment impact, as other research does. The modified Fuzzy Regression Discontinuity model is preferable for modeling fuzzy data. It enables us to draw improved causal effects accommodating fuzzy variables, not just the fuzziness of the treatment effect as in Fuzzy Regression Discontinuity models. A fuzzy dataset is converted into crisp data by the Centroid method of defuzzification. Once the data is crisp, the traditional least squares methods of approximation are used to estimate the parameters in the model since these parameters are considered crisp whilst the error terms are fuzzy. The Alcohol Use Disorders Identification Test score(AUDIT score) can be used as a cutoff to initiate treatment in this case and can be used to predict the progression of HIV disease and/or AIDS. Counseling helps to lower the use of alcohol in people living with HIV/AIDS (PLWHA) as a result, improving the participants' CD4 counts.

Keywords

Fuzzy Regression Discontinuity, Fuzzification, Centroid Method of Defuzzification, Membership Function

1. Introduction

Regression Discontinuity Design (RDD) is a quasi-experimental pretest-posttest

method for investigating the causal effects of treatments by selecting a cutoff or threshold above or below which the intervention is allocated [1]. Treatment's direct impact on health and other results can be studied when a decision rule allocates therapy, such as antiretroviral drugs, to patients or participants with scores greater (or lower) than a defined cutoff or threshold value on a constantly monitored variable, such as AUDIT scores.

Therapy assignment can be deterministic, which means that every participant on the other side of the threshold receives treatment while every respondent on the other side does not. Probabilistic, on the other hand, means that the chances of obtaining treatment are larger on one side of the cutoff than on the other [2]. The first is referred to as sharp regression discontinuity and the second as fuzzy regression discontinuity.

The likelihood of treatment is discontinuous at the cutoff in fuzzy designs, but not to the extent of a clear 0 to 1 jump. For example, if counseling eligibility is granted to all patients who use alcohol and have an AUDIT score (Alcohol Use Disorders Identification Test score) greater than or equal to a specified threshold, but not all patients undergo treatment, the assignment rule describes treatment status probabilistically but not exactly. As a result, the design is fuzzy. The fuzziness may be due to poor program compliance or poor implementation that handled some non-qualifying units but failed to treat certain eligible units.

Regression discontinuity (RD) approaches have been used and extended extensively since the late 1990s [3]. According to [4], although there is rising interest in these designs, little is known about the degree to which they may be used in medicine. Regression discontinuity designs have been employed in health research and have the potential to be used much more widely. In epidemiology, the designs are gaining popularity as a technique for assessing treatments in the absence of randomized trials. Regression discontinuity designs are appealing because they allow real-world data to be used to evaluate the causal consequences of interventions or exposures. Furthermore, the approach makes use of relatively weak assumptions that may be tested empirically.

There are only a few examples of research which incorporate regression discontinuity designs in clinical studies at the moment. This means RDDs are not widely used in clinical research at the moment. [2] found 32 papers that employed Regression Discontinuity Designs (RDDs) in epidemiology, medicine or public health. As a result, clinical researchers may have more opportunities to apply regression discontinuity designs [5].

Regression discontinuity designs take advantage of a sudden shift in the chance of receiving treatment when a score crosses a certain threshold [6]. When a continuous variable crosses a homogeneous cutoff, the just-treated and just-untreated are homogeneous in regression discontinuity. As a result, RD only recognizes treatment impact at the cutoff while providing local randomization [7].

Although the estimates from the two designs, randomized control trials, and regression discontinuity, are not comparable, [8] argue that their confidence intervals overlap, implying that regression discontinuity could be a viable alterna-

tive to randomized control trials (RCTs). This discovery is particularly striking when an RCT is impractical or inappropriate, as is frequently the case in clinical research. This supports the study's extension and use of regression discontinuity.

The organizational structure of this paper is as follows. Initially, we have a description of the problem statement understudy in Section 2. Secondly, in Section 3, we have an explanation of the procedures and methods employed to conduct this investigation. This section also contains the governing equations for the modified fuzzy regression discontinuity model. Furthermore, the modified fuzzy regression discontinuity model was fitted at the end, using PLWHA data. The results of using the procedures mentioned in Section 3 are discussed and implemented in Section 4. Finally, Section 5 concluded by discussing recommendations and future developments.

2. Statement of the Problem

The direct effects of alcohol on the progression of HIV and AIDS disease are unknown. Understanding how the determinants of alcohol affect PLWHA is very important in HIV and AIDS progression. Most models use descriptive statistics or simple regression models. Since there is no clear threshold or alcohol cutoff, the modified fuzzy regression discontinuity model is used in this research. The approach that is considered fuzzy has received the majority of attention in research on fuzzy regression discontinuity models. In the same model, fuzzy dependent and independent variables have received little to no attention. This study will examine the case in which both the explanatory and the response variables are fuzzy. This model explains how alcohol use affects PLWHA and how it might speed up the progression of HIV illness by increasing inflammation and impairing immune system performance.

3. Methods and Procedures

Since this study considered both independent and dependent variables as fuzzy data, the observed AUDIT scores, CD4 counts, and viral load counts were first converted from crisp data to fuzzy data using the Intuition method. This involved membership value assignment. The centroid approach was used to convert ambiguous numbers into precise ones. After that, the traditional regression models were fitted into the regression discontinuity designs.

The Intuition method involved membership value assignments based on human intelligence and understanding. CD4 counts in this study were measured in linguistic terms such as normal, compromised, and susceptible; that is, 500 - 1600 cells per cubic millimeter, 200 - 500 cells per cubic millimeter, and 200 or fewer cells per cubic millimeter, respectively. Moreover, the AUDIT score was classified as low-risk, hazardous, or alcohol-dependent.

3.1. Membership Function

Assuming that $\left(\bar{x}_{ijk}, \bar{y}_{ijk}\right)$ $(i=1,2,3,\cdots,n,\ \ j=1,2,3,\cdots,m,\ \ k=1,2,3,\cdots,K)$ is

the observational data of fuzzy inputs and outputs where $\left(\bar{x}_{ijk}, \bar{y}_{ijk}\right) \in \bar{R}$ and it is a fuzzy number set of real number set R. All observations were considered to be triangular fuzzy numbers. $\bar{x}_{ijk} = \left(c_k, d_k, f_k\right)$ and $\bar{y}_{ijk} = \left(g_k, h_k, l_k\right)$ where c_k, d_k, f_k, g_k, h_k and l_k are all real values. The membership functions of $\bar{x}_{ijk}, \bar{y}_{ijk}$ were defined as follows, respectively.

$$\mu_{\bar{x}_{ijk}}\left(x\right) = \begin{cases} \dfrac{x - c_k}{d_k - c_k} & c_k \le x \le d_k \\ \dfrac{x - d_k}{f_k - d_k} & d_k \le x \le f_k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\mu_{\bar{y}_{ijk}}\left(y\right) = \begin{cases} \dfrac{y - g_k}{h_k - g_k} & g_k \le y \le h_k \\ \dfrac{y - h_k}{l_k - h_k} & h_k \le y \le l_k \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

3.2. Centroid Method of Defuzzification

To calculate the intercept and steepness of the slopes, the idea was to convert the fuzzy observations $\bar{x}_{ijk}, \bar{y}_{ijk}$ and use the traditional linear regression [9]. In this study, the ambiguous data $\bar{x}_{ijk}, \bar{y}_{ijk}$ was turned into clear data x_{ijkc}, y_{ijkc} using the Centroid method. Since the observations are continuous, the following formulae were used respectively.

$$x_{ijkc} = \frac{\int_{-\infty}^{+\infty} x\mu_{\bar{x}_{ijk}}\left(x\right)\mathrm{d}x}{\int_{-\infty}^{+\infty} \mu_{\bar{x}_{ijk}}\left(x\right)\mathrm{d}x} = \frac{1}{3}\left(c_k + d_k + f_k\right) \tag{3}$$

$$y_{ijkc} = \frac{\int_{-\infty}^{+\infty} y\mu_{\bar{y}_{ijk}}\left(y\right)\mathrm{d}y}{\int_{-\infty}^{+\infty} \mu_{\bar{y}_{ijk}}\left(y\right)\mathrm{d}y} = \frac{1}{3}\left(g_k + h_k + l_k\right) \tag{4}$$

3.3. Fuzzy Regression Discontinuity Model

The counseling treatment to lead to a reduction in psychological discomfort, which is perceived to lead to an increase in the amount of alcohol drank was the focus of this investigation. The choice of the cutoff to start initiating treatment was also crucial. A fuzzy regression discontinuity model could be useful because not everyone responds to therapy sessions in the same manner, and moving into or out of psychological pain is a dangerous venture.

d_{ij} does not change for all *ij* at $x_{ij} = x$. Instead, the chance of therapy increases by a factor of less than one.

$$\lim_{\mathbf{x}\downarrow\mathbf{x}'} Pr\left(d_{ij} = 1 \mid x_{ij} = x\right) \ne \lim_{\mathbf{x}\uparrow\mathbf{x}'} Pr\left(d_{ij} = 1 \mid x_{ij} = x\right) \tag{5}$$

$E\left[y_{ij}\left(0\right) \mid x_{ij} = x'\right]$ and $E\left[y_{ij}\left(1\right) \mid x_{ij} = x'\right]$ does not depend on d_{ij} and these two are continuous at x'.

Consequently, we have

$$E\left[y_{ij} \mid x_{ij}=x\right]=\left[1-Pr\left(d_{ij}=1 \mid x_{ij}=x\right)\right]\times E\left[y_{ij}(0) \mid x_{ij}=x\right] + Pr\left(d_{ij}=1 \mid x_{ij}\right)\times E\left[y_{ij}(1) \mid x_{ij}=x\right] \tag{6}$$

which reduces to

$$E\left[y_{ij} \mid x_{ij}=x\right]=E\left[y_{ij}(0) \mid x_{ij}=x\right]+Pr\left(d_{ij}=1 \mid x_{ij}\right) \times\left(E\left[y_{ij}(1) \mid x_{ij}=x\right]-E\left[y_{ij}(0) \mid x_{ij}=x\right]\right) \tag{7}$$

It follows that

$$\lim_{\mathbf{x}\downarrow\mathbf{x}'} E\left(y_{ij} \mid x_{ij}=x\right)-\lim_{\mathbf{x}\uparrow\mathbf{x}'} E\left(y_{ij} \mid x_{ij}=x\right) \tag{8}$$

gives

$$\left(\lim_{\mathbf{x}\downarrow\mathbf{x}'} Pr\left(d_{ij}=1 \mid x_{ij}=x\right)-\lim_{\mathbf{x}\uparrow\mathbf{x}'} Pr\left(d_{ij}=1 \mid x_{ij}=x\right)\right) \times\left(E\left[y_{ij}(1) \mid x_{ij}=x'\right]-E\left[y_{ij}(0) \mid x_{ij}=x'\right]\right) \tag{9}$$

and

$$\mathrm{ACE}=\delta_{FRD}=\frac{\lim_{\mathbf{x}\downarrow\mathbf{x}'} E\left(y_{ij} \mid x_{ij}=x\right)-\lim_{\mathbf{x}\uparrow\mathbf{x}'} E\left(y_{ij} \mid x_{ij}=x\right)}{\lim_{\mathbf{x}\downarrow\mathbf{x}'} Pr\left(d_{ij}=1 \mid x_{ij}=x\right)-\lim_{\mathbf{x}\uparrow\mathbf{x}'} Pr\left(d_{ij}=1 \mid x_{ij}=x\right)} \tag{10}$$

Running four regressions can allow us to calculate the Average Causal Effect (ACE). Please note that *a* means above the cutoff, while *b* means below the cutoff.

$$\min_{\alpha_{yb}\beta_{yb}} \sum_{ij \mid x'-h<x_{ij}\le x'} \left(y_{ij}-\alpha_{yb}-\beta_{yb}\left(x_{ij}-x'\right)\right)^2 \tag{11}$$

$$\min_{\alpha_{ya}\beta_{ya}} \sum_{ij \mid x'<x_{ij}<x'+h} \left(y_{ij}-\alpha_{ya}-\beta_{ya}\left(x_{ij}-x'\right)\right)^2 \tag{12}$$

$$\min_{\alpha_{db}\beta_{db}} \sum_{ij \mid x'-h<x_{ij}\le x'} \left(d_{ij}-\alpha_{db}-\beta_{db}\left(x_{ij}-x'\right)\right)^2 \tag{13}$$

$$\min_{\alpha_{da}\beta_{da}} \sum_{ij \mid x'<x_{ij}<x'+h} \left(d_{ij}-\alpha_{da}-\beta_{da}\left(x_{ij}-x'\right)\right)^2 \tag{14}$$

After running these equations, to find the Average Causal Effect

$$\delta_{FRD}=\frac{\hat{\alpha}_{ya}-\hat{\alpha}_{yb}}{\hat{\alpha}_{da}-\hat{\alpha}_{db}} \tag{15}$$

With the aid of instrumental measures, this research directly evaluated the effect where $x'-h<x_{ij}<x'+h$ and the outcome equation is given in Equation (16).

$$y_{ij}=\alpha+\beta\left(1-d_{ij}\right)\left(x_{ij}-x'\right)+\gamma d_{ij}\left(x_{ij}-x'\right)+\delta d_{ij}+e_{ij} \tag{16}$$

This research used Z_{ij} and $Z_{ij}\left(x_{ij}-x'\right)$ as instrumental variables for d_{ij} and $d_{ij}\left(x_{ij}-x'\right)$. $Z_{ij}=1$ if $x_{ij}\ge x'$ and 0 otherwise. In this study, to implement instrumental variables, two-stage least squares were implemented. The problem is that d_{ij} and $d_{ij}\left(x_{ij}-x'\right)$ are correlated with e_{ij}.

Step 1: Regress d_{ij} on Z_{ij} and $x_{ij}-x'$ using a model of the form

$$d_{ij}=\omega_0+\omega_1\left(1-Z_{ij}\right)\left(x_{ij}-x'\right)+\omega_2 Z_{ij}\left(x_{ij}-x'\right)+\omega_3 Z_{ij}+u_{ij} \tag{17}$$

where u_{ij} are mean zero error terms for $i=1,2,3,\cdots,n$, $j=1,2,3,\cdots,m$ and Z_{ij} has the following properties:

- Z_{ij} affects d_{ij}.
- Z_{ij} affects y_{ij} only through d_{ij}.
- Z_{ij} is not affected by other factors meaning it is exogenous that is $\operatorname{corr}\left(Z_{ij},u_{ij}\right)=0$.

Step 2: From equation 17 the estimated program participation will be

$$\hat{d}_{ij}=\hat{\omega}_0+\hat{\omega}_1\left(1-Z_{ij}\right)\left(x_{ij}-x'\right)+\hat{\omega}_2 Z_{ij}\left(x_{ij}-x'\right)+\hat{\omega}_3 Z_{ij} \tag{18}$$

Step 3: Substituting the obtained in the main equation that is equation 16 the new model became

$$y_{ij}=\alpha+\beta\left(1-\hat{d}_{ij}\right)\left(x_{ij}-x'\right)+\gamma\hat{d}_{ij}\left(x_{ij}-x'\right)+\delta\hat{d}_{ij}+e_{ij}^{*} \tag{19}$$

It follows that

$$y_{ij}^{*}=\alpha+\beta\left(1-\hat{d}_{ij}\right)\left(x_{ij}^{*}-x'^{*}\right)+\gamma\hat{d}_{ij}\left(x_{ij}^{*}-x'^{*}\right)+\delta\hat{d}_{ij}+e_{ij}^{**} \tag{20}$$

where y_{ij}^{*} and x_{ij}^{*} are fuzzy observations with membership functions $\mu_{\bar{y}_{ij}}(y)$ and $\mu_{\bar{x}_{ij}}(x)$ respectively. Moreover, e_{ij}^{**} is the fuzzy error associated with the model and the parameters of the model are crisp. Letting $x_{ij}^{*}-x'^{*}=x_{ij}^{c}$

$$d_{ij}=\omega_0+\omega_1\left(1-Z_{ij}\right)x_{ij}^{c}+\omega_2 Z_{ij}x_{ij}^{c}+\omega_3 Z_{ij}+u_{ij} \tag{21}$$

and

$$y_{ij}^{*}=\alpha+\beta\left(1-\hat{d}_{ij}\right)x_{ij}^{c}+\gamma\hat{d}_{ij}x_{ij}^{c}+\delta\hat{d}_{ij}+e_{ij}^{**} \tag{22}$$

where $\hat{d}_{ij}=\left(\hat{d}_{1j}\cdots\hat{d}_{nj}\right)^{\mathrm{T}}$. The design matrices Z and X were used to fit the model and are defined as

$$Z=\begin{pmatrix}1 & \left(1-Z_{1j}\right)x_{1j}^{c} & Z_{1j}x_{1j}^{c} & Z_{1j}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & \left(1-Z_{nj}\right)x_{nj}^{c} & Z_{nj}x_{nj}^{c} & Z_{nj}\end{pmatrix}$$

and

$$X=\begin{pmatrix}1 & \left(1-\hat{d}_{1j}\right)x_{1j}^{c} & \hat{d}_{1j}x_{1j}^{c} & \hat{d}_{1j}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & \left(1-\hat{d}_{nj}\right)x_{nj}^{c} & \hat{d}_{nj}x_{nj}^{c} & \hat{d}_{nj}\end{pmatrix}$$

In matrix form, the models are

$$d_{ij}=Z\omega+u_{ij} \tag{23}$$

and

$$y_{ij}^{*}=X\psi+e_{ij}^{**} \tag{24}$$

where $\omega=\left(\omega_0,\omega_1,\omega_2,\omega_3\right)^{\mathrm{T}}$ and $\psi=\left(\alpha,\beta,\gamma,\delta\right)^{\mathrm{T}}$. The least squares estimates

of $\omega = (\omega_0, \omega_1, \omega_2, \omega_3)^{\mathrm{T}}$ and $\psi = (\alpha, \beta, \gamma, \delta)^{\mathrm{T}}$ are $\hat{\omega} = \left(Z^{\mathrm{T}}Z\right)^{-1} Z^{\mathrm{T}} d_{ij}$ and $\hat{\psi} = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}} y_{ij}^{*}$ respectively. Likewise $\hat{d}_{ij} = \left(\hat{d}_{1j} \cdots \hat{d}_{nj}\right)^{\mathrm{T}} = Z\left(Z^{\mathrm{T}}Z\right)^{-1} Z^{\mathrm{T}} d_{ij}$. The causal effect at the threshold x' is given by $\hat{\delta}$. Now $\mathrm{Var}(\hat{\psi}) = \mathrm{Var}\left(y_{ij}^{*}\right)\left(X^{\mathrm{T}}X\right)^{-1} = \sigma_2\left(X^{\mathrm{T}}X\right)^{-1}$. The Local Average Treatment Effect variance will be as a result given by $\left[\sigma_2\left(X^{\mathrm{T}}X\right)^{-1}\right]_{44}$ from the covariance-variance matrix. The estimator of the LATE was to be unbiased and consistent [10]. To increase efficiency of this variance, the estimate $\hat{d}_{ij}$ can be replaced by d_{ij}.

To estimate the fuzzy error terms e_{ij}^{**}, this paper sticks to the idea from [11]. The regression coefficients were determined by using the lowest, maximum, and center points of each observed triangular fuzzy number. The error terms (e_{ij}^{**}) in model 24 can be conceptually stated as follows:

$$e_{ij}^{**} = y_{ij}^{*} - X\psi \tag{25}$$

By dividing the difference in membership values by the actual membership values, it is possible to estimate Equation (26). That is

$$e_{ij}^{**} = \frac{\int_{s_{y_{ij}} \cup s_{\hat{y}_{ij}}} \left|\mu_{y_{ij}}(y) - \mu_{\hat{y}_{ij}}(y)\right| \mathrm{d}y}{\int_{s_{y_{ij}}} \mu_{y_{ij}}(y)\,\mathrm{d}y} \tag{26}$$

4. Main Results

Simulations in this section were used to compare and examine the asymptotic properties of the modified fuzzy regression discontinuity model. The sample sizes of 30, 50, 100, and 200 are first considered. **Table 1** shows the simulated estimates and the values of other parameters are set at $\alpha = 10$, $\beta = 0.5$, $\gamma = 0.3$ and $\delta = 0.3$. These results are based on 1000 replications. If the running variable's value falls below the threshold, $x' = 15$, no treatment is provided. Observations with $x_{ij} \geq 15$ receive therapy with a probability of 0.8, indicating that the running variable and the cutoff only partially determine treatment status.

A similar trend is shown when using sample sizes of 100, 1000, 10,000, and 100,000. The estimators' subsequent curves, which take the highlighted sample sizes into account, exhibit a bell-shaped distribution. Additionally, as the sample size increases, the guesses get closer to the actual population parameters. This curve in **Figure 1** justifies the asymptotic consistency of the estimator.

4.1. Application of the Intuition Method

The scales for the variables CD4 counts, viral loads, and AUDIT score as they were utilized using the intuition technique in this study, are shown in **Table 2**. Please take note that, the corresponding scales were divided into ranges under the World Health Organization and previous research.

The distribution of AUDIT scores for those living with HIV and AIDS is depicted in the pie chart in **Figure 2**. It demonstrates that, in the PLWHA data set,

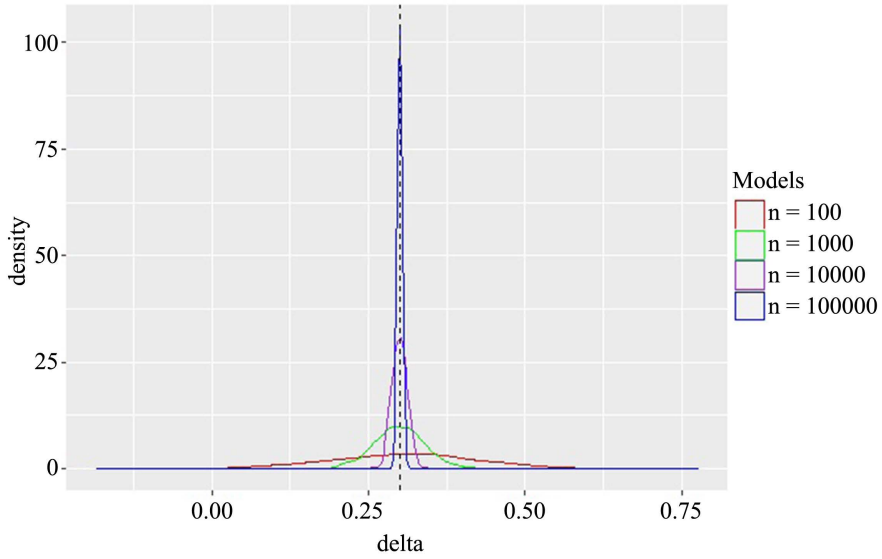

Figure 1. Distribution of fuzzy estimate δ.

Table 1. Simulation: Modified fuzzy RDD.

$n = 30$	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9	ψ_{10}
α	10.05918	10.4296	10.11746	10.03732	10.01225	10.05182	9.915751	10.11658	10.0589	9.926157
β	0.831685	0.078345	0.202603	0.786152	0.226335	0.423697	0.945735	0.441477	0.706616	1.179589
γ	0.189016	0.000149	0.470398	0.262198	0.586674	0.504595	0.439227	0.183835	−0.00332	−0.03192
δ	0.319571	0.86729	−0.01636	−0.15306	0.148632	0.326433	−0.06694	0.654968	0.527264	0.178303
$n = 50$										
α	9.966581	10.16867	9.919504	9.874817	10.09583	10.05389	9.848505	9.918116	10.03086	10.0523
β	0.804199	0.802487	0.583962	0.577569	0.869835	0.475645	0.485317	0.192588	−0.06378	0.163141
γ	−0.26246	0.03434	0.351195	0.137382	0.084462	0.622308	0.230856	0.59359	0.755761	0.474862
δ	0.601342	0.29998	0.143981	0.172199	0.313364	0.282063	0.422161	0.31634	0.27081	0.252742
$n = 100$										
α	9.852781	10.11833	10.04046	10.06843	10.00159	10.02169	10.11082	10.05546	9.991145	10.05531
β	0.494862	0.446694	0.416527	0.529289	0.361199	0.640202	0.413273	0.257557	0.439515	0.68903
γ	0.35931	0.181592	0.301358	0.241737	0.227408	0.102037	0.278026	0.319172	0.487405	0.154375
δ	0.200624	0.35338	0.286409	0.352353	0.604385	0.383008	0.124819	0.326383	0.469301	0.39342
$n = 200$										
α	9.989306	9.979731	10.01844	10.08927	10.05984	9.908219	9.984348	10.11999	9.99771	9.960223
β	0.421749	0.626889	0.473746	0.482122	0.43808	0.391467	0.631968	0.481441	0.482591	0.509286
γ	0.356991	0.400704	0.269958	0.242712	0.49179	0.331533	0.291931	0.385506	0.434324	0.299226
δ	0.185288	0.23846	0.315732	0.372257	0.124594	0.408139	0.272714	0.263936	0.253258	0.300624

Table 2. Scale variables.

AUDIT score			
Low	Medium	High	Very High
0 ≤ AUDIT score < 8	8 ≤ AUDIT score < 15	15 ≤ AUDIT score < 20	20 ≤ AUDIT score < 40
CD4 counts			
Susceptible	Compromised	Normal	
0 ≤ CD4 count < 200	200 ≤ CD4 count < 500	500 ≤ CD4 count < 1600	
Viral loads			
Target not detectable	Target detectable		
≤100	>100		

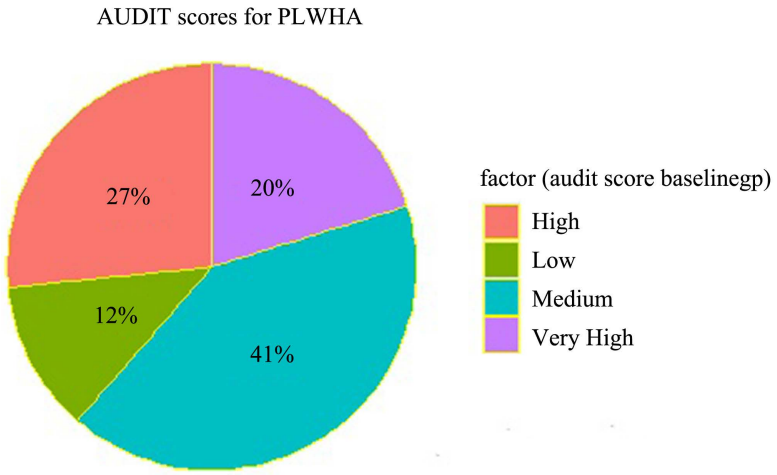

Figure 2. Distribution of AUDIT scores.

the majority of respondents have medium AUDIT scores, followed by high, very high, and low AUDIT scores.

4.2. Fuzzification

An assumption that all of the observations are triangular fuzzy integers was made. Therefore, the following R function code was used for the independent variable. The respective graphical presentation is also shown in **Figure 3**.

```
mfunction < - function(x) {
if (x < = 8) {
return((x-c_k)/(d_k-c_k))
} else
if (x > 8 & x < 15)
return((x-d_k)/(f_k-d_k))
} else {
return(0) } } }
```

The same procedure can be performed in R for the dependent variable.

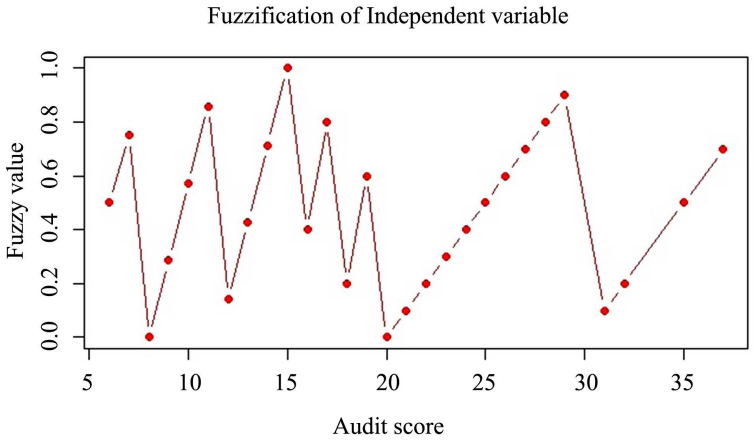

Figure 3. Fuzzified AUDIT scores graph.

4.3. Implementation of the Modified Fuzzy Regression Discontinuity Model

Due to the available data set, this study chose to utilize an AUDIT score of 15 for analysis purposes. Due to imperfect compliance, it is relatively challenging to quantify the magnitude of the leap at the cutoff under this approach. People who received treatment included those who scored below the cutoff as well as those who scored above the cutoff but were not meant to. It was a fuzzy regression discontinuity since it looked like this (**Figure 4**).

4.3.1. Check for Discontinuity in the Running Variable around the Cutoff Point in Fuzzy RDD

In this case, a manipulation test would be a test of the continuity of the density at the cutoff point x'.

$$H_0 : \lim_{x \uparrow x'} f\left(x_{ij}\right) = \lim_{x \downarrow x'} f\left(x_{ij}\right)$$

versus

$$H_1 : \lim_{x \uparrow x'} f\left(x_{ij}\right) \neq \lim_{x \downarrow x'} f\left(x_{ij}\right)$$

Now considering **Figure 5**, taking note of how the confidence intervals in the plot overlap. The overlap's p-value is 0.465517, which is significantly greater than 0.05. The McCrary test demonstrates that a formal test cannot rule out the null hypothesis that there is no discontinuity in the density at the cutoff. This means that there is lack of solid proof that the two lines differ significantly from one another. Based on the plot in **Figure 5**, one can reasonably state with certainty that there is no manipulation or bunching.

4.3.2. Check for Discontinuity in the Outcome Variable across Running Variable in Fuzzy RDD

Now that it was known and understood that this was a fuzzy regression discontinuity design and that there was no manipulation of AUDIT scores around the 15-point threshold, it was possible to ascertain whether there was a discontinuity in final results depending on participation in the counseling program. **Figure 6**

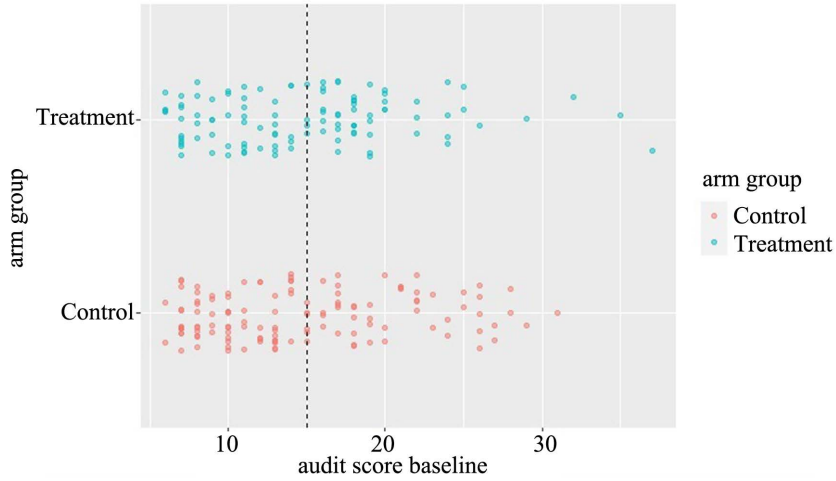

Figure 4. Noncompliance around a cutoff.

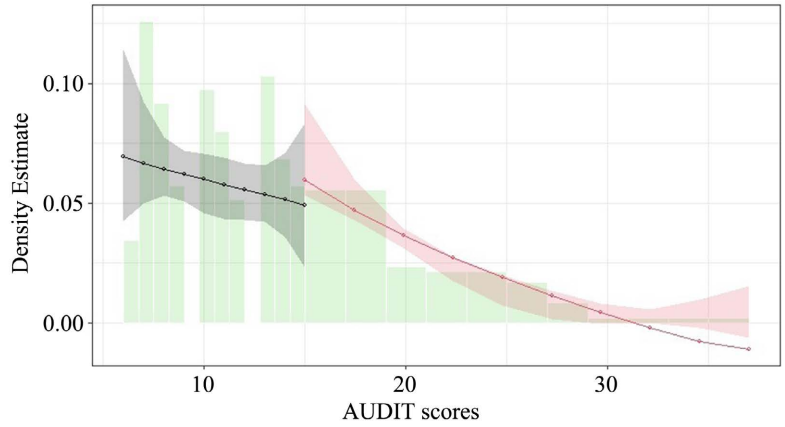

Figure 5. Manipulation test graph.

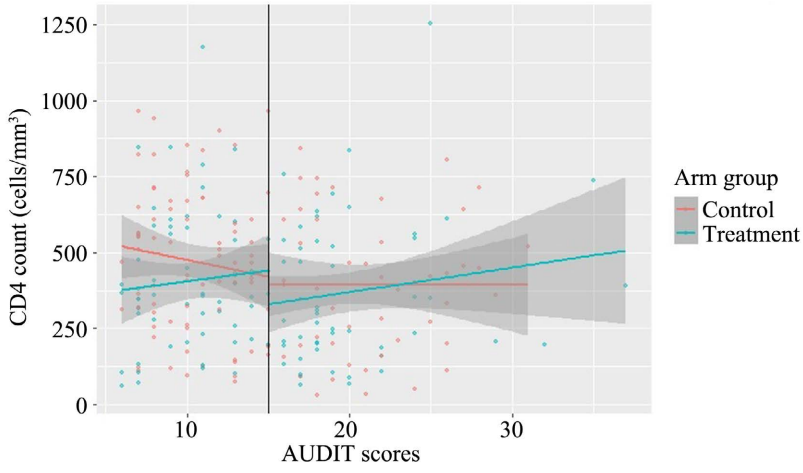

Figure 6. Discontinuity checks.

was obtained by plotting the running variable, AUDIT scores on the x-axis, the outcome variable, CD4 count (cells/mm^3) on the y-axis, and coloring the points according to whether or not they took part in the counseling program.

Figure 6 shows that there were distinct discontinuities for those who comply and those who do not. It appears that taking part in the counseling program increased the final CD4 counts. Since AUDIT scores for PLWHA are used to predict the progression of HIV/AIDS, Figure 7 shows the AUDIT scores before and after the PLWHA received counseling. The curve depicted in Figure 7, shows that the AUDIT scores in the control group increased after 6 months in comparison to the baseline AUDIT scores.

4.3.3. Estimating the Size of the Effect in Fuzzy RDD

In practice, the bandwidth size is determined by the availability of data. In this study, treatment effects are estimated using a modified fuzzy regression discontinuity model with a bandwidth of ±10. With the use of instrumental variables robust, which is a two-stage least squares estimation of an instrumental variable regression, the outcomes are displayed in Table 3.

The effect size was used to comprehend the findings of the research. It is a standardized measure, used to assess the magnitude of an intervention's effect, which aids in determining how effective intervention was. The table above shows that the causal effect of counseling for compliers in the ±10 bandwidth was approximately equal to 0.237 which is a small effect [12] in the positive direction. This coefficient is the most important as this is the shift in intercept when counseling is true or the difference between scores at the threshold.

The average AUDIT score over six months was displayed at the 15-point cutoff. At six months, participants with baseline AUDIT scores of 14.9999 scored

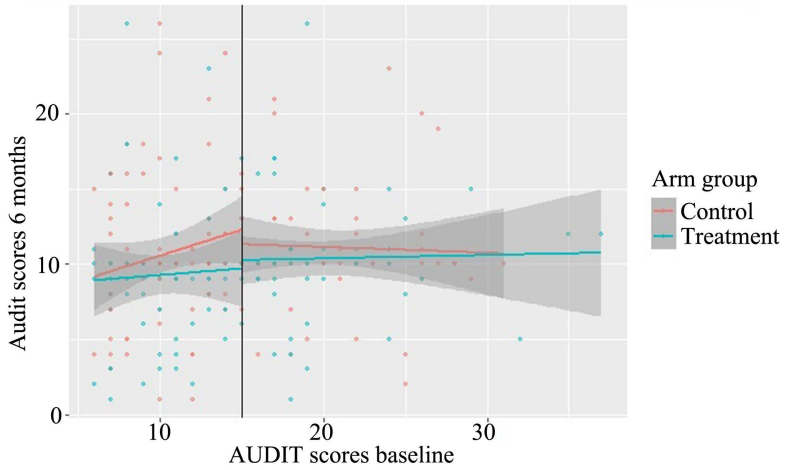

Figure 7. AUDIT scores comparison.

Table 3. Parametric estimation.

term	estimate	standard error	statistic	p. value	conf. low	conf. high
(Intercept)	10.24623623	0.60648725	16.8943966	4.01E−42	9.0512275	11.4412449
audit centered	0.05264287	0.08106509	0.6493901	5.17E−01	−0.1070859	0.2123717
counseling true	0.23715658	1.00934232	0.2151466	8.30E−01	−1.7716286	2.2059418

an average of roughly $\alpha = 10.25$. This demonstrates that there is a possibility that, if treatment is started, the AUDIT scores will decrease over time, having a favorable impact on the CD4 counts and viral loads of PLWHA.

Now, taking into account the AUDIT-centered coefficient, for each baseline score greater than or equal to 15, respondents' scores increased by 0.053 points at the halfway point. The size of the effect is what matters most, so this number is not much of concern.

4.3.4. Validation and Robustness Checks of the Model

A simple test for the dependability of a regression discontinuity design that evaluates whether ratings are changed by looking at the distribution of observations was offered by [13]. If ratings were not altered, the density of observations should vary continuously with ratings at the cut-point. **Figure 5** demonstrates that there was no scoring manipulation. Additionally, the results relating to baseline characteristics and results support the idea that potential outcomes might be comparable between the two groups.

5. Conclusions and Suggestions

The modified model under study has high levels of internal validity. Since the data gathering at times may not be precise, a regression discontinuity design can be used rather than a randomized one in clinical research. The underlying appeal of the modified fuzzy RD model is that they let us target the program or therapy at the people who need it or deserve it the most and address non-compliance issues. Since people who may require the program or therapy should not be excluded, as in randomized experiments, this design's real allure is ethical.

Since attention was on a modified fuzzy regression discontinuity model with fuzzily categorized dependent and independent variables, the model parameters were treated as crisp. In truth, these do occasionally tend to be fuzzy, so one can take into consideration situations in which the model has fuzzy parameters. Since the parameters in this investigation were considered to be clear, least-squares approximation techniques were used.

Future Research

Future research employing the modified fuzzy regression discontinuity model must carefully evaluate scenarios where the parameters are crisp, as we did not do in our work. Furthermore, the fuzzy regression discontinuity under parametric estimation was expanded in this study. Future research could look into this modified fuzzy regression discontinuity with non-parametric estimation. Because we used PLWHA data, future research can use the same model on a number of different health-related data sets.

Acknowledgements

I want to thank my supervisors, Prof. Samuel Musili Mwalili and Dr. Thomas Mageto, for helping me with this project and letting me give it my all. It was

going to be a challenging voyage for me without them. Thank you for your continuous support, family, and friends.

Most importantly, I wish to thank my loving and supportive husband, Bothwell Blessed I Kutadza, and my wonderful daughter Kaelyn Kupa Kutadza, who provide unending inspiration.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Herlands, W., McFowland III, E., Wilson, A.G. and Daniel, B. (2018) Automated Local Regression Discontinuity Design Discovery. *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1512-1520. https://doi.org/10.1145/3219819.3219982

[2] Moscoe, E., Boe, J. and Bärnighausen, T. (2015) Regression Discontinuity Designs Are Underutilized in Medicine, Epidemiology, and Public Health: A Review of Current and Best Practice. *Journal of Clinical Epidemiology*, **68**, 132-143. https://www.sciencedirect.com/science/article/pii/S0895435614003990 https://doi.org/10.1016/j.jclinepi.2014.06.021

[3] Imbens, G.W. and Lemieux, T. (2008) Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics*, **142**, 615-635. https://www.sciencedirect.com/science/article/abs/pii/S0304407607001091 https://doi.org/10.1016/j.jeconom.2007.05.001

[4] Boon, M.H., Craig, P. and Thomson, H. (2021) Regression Discontinuity Designs in Health: A Systematic Review. *Epidemiology*, **32**, 87-93. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7707156/ https://doi.org/10.1097/EDE.0000000000001274

[5] Sasabuchi, Y. (2022) Introduction to Regression Discontinuity Design. *Annals of Clinical Epidemiology*, **4**, 1-5. https://doi.org/10.37737/ace.22001 https://www.jstage.jst.go.jp/article/ace/4/1/4_22001/_article/-char/ja/

[6] Bartalotti, O., Bertanha, M. and Calonico, S. (2021) Regression Discontinuity Designs in Policy Evaluation. In: Hashimzade, N. and Thornton, M.A., Eds., *Handbook of Research Methods and Applications in Empirical Microeconomics*, Edward Elgar Publishing, Cheltenham. https://www.elgaronline.com/view/edcoll/9781788976473/9781788976473.00020.xml

[7] Lee, M. (2016) Matching, Regression Discontinuity, Difference in Differences, and beyond. Oxford University Press, New York. https://doi.org/10.1093/acprof:oso/9780190258733.001.0001

[8] Maas, I.L., Nolte, S., Walter, O.B. and Berger, T. (2017) The Regression Discontinuity Design Showed to Be a Valid Alternative to a Randomized Controlled Trial for Estimating Treatment Effects. *Journal of Clinical Epidemiology*, **82**, 94-102. https://www.sciencedirect.com/science/article/abs/pii/S0895435616306825 https://doi.org/10.1016/j.jclinepi.2016.11.008

[9] Zhang, D., Chen, M.H., Ibrahim, J.G. and Boye, M.E. (2017) Bayesian Model Assessment in Joint Modeling of Longitudinal and Survival Data with Applications to Cancer Clinical Trials. *Journal of Computational and Graphical Statistics*, **26**, 121-133. https://www.tandfonline.com/doi/abs/10.1080/10618600.2015.1117472

https://doi.org/10.1080/10618600.2015.1117472

[10] Hahn, J., Todd, J. and Van der Klaauw, W. (2001) Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, **69**, 201-209. https://www.jstor.org/stable/2692190

[11] Kim, B. and Bishu, R.R. (1998) Evaluation of Fuzzy Linear Regression Models by Comparing Membership Functions. *Fuzzy Sets and Systems*, **100**, 343-352. https://www.sciencedirect.com/science/article/abs/pii/S0165011497001000 https://doi.org/10.1016/S0165-0114(97)00100-0

[12] Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, NJ, Routledge, 20-26. https://doi.org/10.4324/9780203771587

[13] McCrary, J. (2008) Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. *Journal of Econometrics*, **142**, 698-714. https://www.sciencedirect.com/science/article/abs/pii/S0304407607001133 https://doi.org/10.1016/j.jeconom.2007.05.005