Michael Gao

STA 320 Final Project

04/29/2016

Prof. Fan Li

Regression Discontinuity Design in the NCAA Basketball Rankings

1. *Introduction*

Regression discontinuity design is a quasi-experimental approach to determining causal effects that was first introduced by Thistlethwaite and Campbell in 1960s. The basic idea behind regression discontinuity design is that assignment to a treatment is determined by the value of a covariate and its position relative to a threshold value. Imbens and Lemieux (2008) explain that “[this] predictor may itself be associated with the potential outcomes, … so any discontinuity of the conditional distribution of the outcome as a function of this covariate at the cutoff value is interpreted as evidence of a causal effect of the treatment.” Regression Discontinuity Design has recently gained popularity, particularly within the economics literature. Some examples of topics explored using Regression Discontinuity Design within in the past 15 years include “the labor supply effect of welfare, unemployment insurance, and disability programs; the effects of Medicaid on health outcomes; the effect of remedial education programs on educational achievement; the empirical relevance of median voter models; and the effects of unionization on wages and employment” (Lee and Lemieux 2010). In this paper, I discuss the application of Regression discontinuity design to a wholly new field, NCAA rankings and TV ratings.

1. *Hypothesis*

In Regression Discontinuity, the forcing variable is the covariate that determines treatment status. Within this context, the forcing variables is AP Poll Ranking. Every week during the NCAA regular season, the AP Poll releases its Top 25 ranking, which has been adopted by the NCAA. This ranking is then shown next to each team’s name whenever the teams play on television. If a team is not in the top 25, there is no designation placed next to the school name. The hypothesis that this paper attempts to test is that there is a causal effect of being ranked in the top 25 on television ratings for games featuring the ranked teams. In this design, being ranked in the top 25 or not is a binary treatment variable that is dependent on the forcing variable of rank. With this design, it is important to note that the design follows that of a sharp regression discontinuity design, meaning that if a team’s rank is in the top 25, they will receive the treatment (top 25 designation) with 100% probability and 0% probability if they are not in the top 25. Generally, the forcing variable is assumed to be continuous, however it is clearly discrete in this case. However, oftentimes forcing variables are discretized due to data quality (Age being reported in years and months). This problem is further addressed by some novel methods for standard error estimation and a modified set of assumptions, which will be discussed later (Lee and Card 2008).

1. *Data Collection*

The data for this analysis were collected by scraping several websites. Sportsmediawatch.com provided the television rankings for each regular season NCAA game played in the 2013-2014 and 2014-2015 seasons. The top 25 designation and ranks were provided by scraping the ESPN website. Since schools outside of the top 25 are not usually given a ranking, rankings outside the top 25 are determined by the number of votes received. This is essentially filling in the values of the rankings if they were recorded past 25. For example, if a 25th ranked school received 1800 votes to be placed in the top 25 and the next school is listed at 1700 votes, I filled in the rank for that school at 26.

As with any web scraping endeavor, the data was not very well-behaved, with several occurrences of differing formatting and even a mistake in one of the rankings by ESPN (two schools having a ranking of 25). These are all accounted for in the R scripts provided. Because the top 25 are updated every week, the rankings are matched by both school and time to ensure the accuracy of the rankings for each observation. In addition, each game’s rank is determined by the minimum (read: highest) ranking for each game. If only 1 team is ranked, that ranking is used to determine the game’s rank.

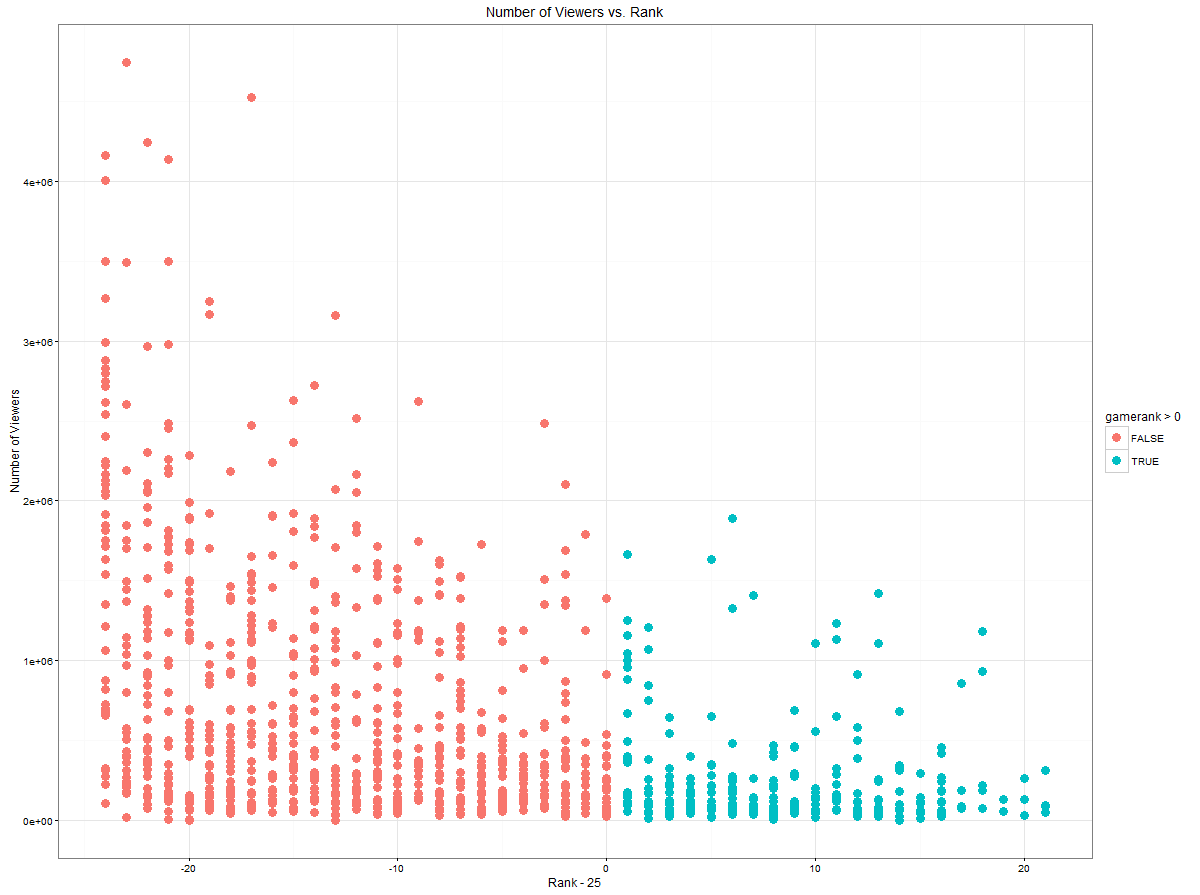


Fig. 1: Number of Viewers (Outcome) vs. Rank (Forcing Variable)

1. *Methods*

In essence, implementation of analysis using Regression Discontinuity Design involves fitting a model for the relationship between the outcome variable and the forcing variable on either side of the threshold value and comparing the estimates at the threshold point. The difference, in theory, should correspond to an estimate for the causal effect at the threshold point. For this particular case, we choose to model the relationship with a nonparametric kernel regression. When performing this type of regression, it is important to choose a bandwidth carefully, as the estimate of the causal effect is extremely sensitive to the bandwidth choice. For this analysis, I use the Imbens-Kalyanaraman Optimal Bandwidth. Conventional cross-validation methods for choosing the bandwidth, *h*, select the bandwidth that is optimal for fitting a curve over the entire support of the data, which may not have the intended effect or properties that are desired. The Imbens-Kalyanaraman Optimal Bandwidth focuses specifically on the case of Regression Discontinuity and attempts to minimize a measure of error at the boundary values (Imbens and Kalyanaraman 2012). The algorithm for selecting the optimal bandwidth is as follows:

1. Estimation of a density and conditional variance σ­­­2(c)

* Calculate the sample variance of the forcing variable
* Use a modified Silverman rule pilot bandwidth given by *h*1 = 1.84 \* *S*x \* *N*-1/5
* Average the outcomes on either side of the threshold
* The new density of Xi at *c* is given by:

Where are the number of units on either side of the threshold of the forcing variable.

1. Estimation of the second derivatives   :

* These appeared in Cheng, Fan and Marron (1997) for their paper on boundary correction
* First get a pilot bandwidth . Estimate this using the third derivative of at *c.*
* Take the median of the forcing variable on either side of the threshold and temporarily remove points that are away from the threshold.
* Fit a third order polynomial as follows:

And estimate as .

* Now, using the quantities calculated above, calculate the following pilot bandwidth

The maximum function is to avoid problems in the case that the estimate for happens to be 0, which should not occur in practice.

* Given we can estimate the curvature using a local quadratic fit for the observations with . The curvature  , or the coefficient estimate for the quadriatic terms. Calculate similarly with data that satisfies

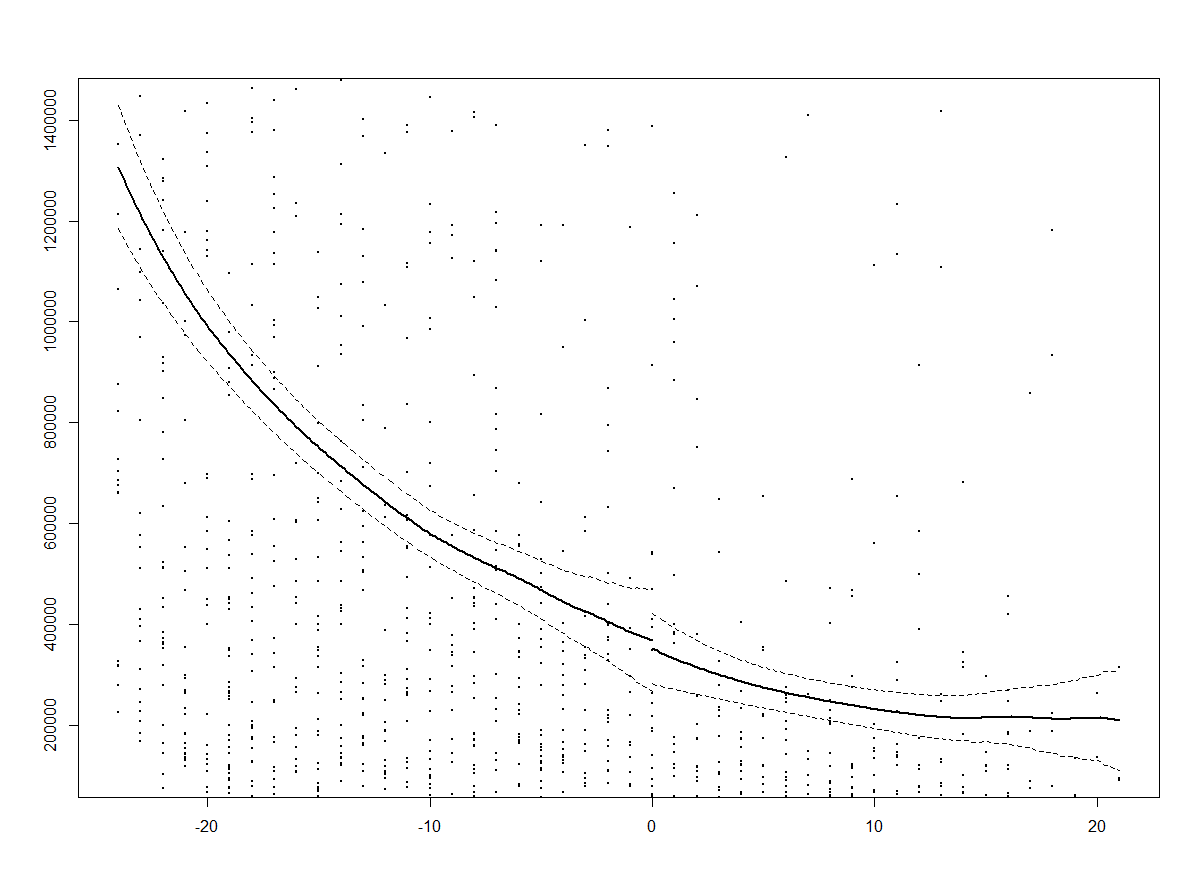
1. Calculation of the Regularization terms and

* These regularization terms are added because the precision with which the second derivatives is low, and so occasionally the calculated optimal bandwidth will be too large.

Finally,

I use the epanechnikov kernel in this case, for which the value of .

1. *Results*



*Fig. 2: Number of Viewers (Y) vs. Ranking - 25 (x)*

The results of this analysis indicate that there is no significant causal effect of being ranked in the top 25 on the number of viewers that watch the games on television. Our estimate of the LATE (Local Area Treatment Effect) is -25922 with an optimal Imbens- Kalyanaraman bandwidth of 14.187. The standard error of this estimate, when accounting for clustering effects (Lee and Card 2008) is 31435, meaning that we cannot discern whether or not there is a significant causal effect that has occurred. The RDD package also provides estimates under different values of the bandwidth for local linear regression corresponding to half and double the Imbens-Kalyanaraman bandwidth, both of which are also insignificant (Half: -63376 SE: 39542, Double: 36991 SE: 43431). It is important to note that though the LATE is sensitive to choice of bandwidth, none of these effects can be viewed as significant.

1. *Considerations*

There are several considerations that should be taken into account for this analysis. First, the quality of the data may not be extremely accurate. Webscraping for the purposes of data generation relies on the integrity of the websites that are being scraped. Though the author has taken precautions to ensure that the data is cleaned, there may be underlying quality issues that affect the power of this analysis. In addition, there were several approximations and definitions that were chosen subjectively by the author in this study. Of primary concern is that game ranking is a metric which is entirely subjective in its definition. For the purposes of is analysis, game ranking is defined as the highest rank (lowest number) ranking for both of the teams playing in a game. The primary motivation for this metric is to make sure that lower ranked teams’ viewer counts are not inflated by the presence of a higher-ranked team as an opponent. Nevertheless, this may cause some further questions.

Next, the issue of the forcing variable being discrete must be addressed. Under traditional regression discontinuity techniques, the forcing variable must be continuous. However, truly continuous data is often hard to find. Lee and Card (2008) address this in their paper titled, *Regression discontinuity inference with specification error*. Essentially, when a forcing variable is discrete, the idea in regression discontinuity design of checking the values of the outcome “just above” and “just below” the threshold of the forcing variable cannot be done. In fact, without assumptions about the functional form between the outcome and forcing variable, those values are nonidentifiable. The inference procedure described in the paper is to collapse the data to the cell level, while retaining information on the means, variances, and number of observations in each cell (read: value of the discrete forcing variable). Then, one runs the cell size-weighted regression using this data. The value of is then given by the following:

Where J is the number of values of the discrete variable and is the least squares estimator in the regression of on where is the treatment indicator. This estimator is implemented in the RDD package on CRAN.

1. *Reproducibility*

All scripts for webscraping and analysis can be found at github.com/michaelgao8/NCAA-Top-25-Regression-Discontinuity. The RDD package source can be found at https://github.com/ddimmery/rdd

The RDD package is an implementation of methods discussed in references, which are listed below. The analysis can be reproduced by doing the following:

1. Create AllSeasons.Rdata using get\_tvrankings.R
2. Create SeasonRanks.Rdata using get\_top25rankings.R
3. Knit the markdown file in FinalProject.Rmd

Alternatively, a make file has been provided. In order to run the make file, simply open a terminal in the directory and type ‘make’ with no quotations. The Rdata files are available as backup on the github repository also in case of issues with the webscraping scripts (changing websites).

1. *References*

Imbens, G.W., Kalyanaraman, K., 2012. Optimal Bandwidth Choice for the Regression Discontinuity Estimator. Review of Economic Studies. 79 (3): 933-959

Imbens, G.W., Lemieux, T., 2008. Regression discontinuity designs: A guide to the practice. Journal of Econometrics 142, 615-635.

Lee, D.S., Lemieux, T., 2010. Regression Discontinuity Designs in Economics. Journal of Economic Literature 48, 281-355.

Lee, D.S., Card, D., 2008. Regression discontinuity inference with specification error. Journal of Econometrics 142, 655-674.

Scraped websites:

http://espn.go.com/mens-college-basketball/rankings

http://www.sportsmediawatch.com/