Empirical Project

Gigi Lin

12-4-2017; updated 9-28-2018

This replication project was assigned during my final year econometrics seminar of undergrad. The aim was to replicate the empirical results of a paper I selected. A summary of the main idea focusing on the empirical strategy and interpretation of the results was required. I chose Thomas Buser's paper "The Effect of Income on Religiousness" because of the interest in the levels of wealth having an impact on one's pursuit of religion. I also wanted to attempt learning regression discontinuity design (RDD) methods as they were not covered in-class, but I believed would prove useful in higher-up econometric courses. Some code has been omitted for sake of length, but the construction and preparation of data analysis has been mostly included.

Part I: Introduction

Thomas Buser uses self-collected survey data to analyze the effects of income shocks on religious behavior. He undertakes a regression discontinuity design (RDD) on the change in eligibility for a government cash transfer on 3 income measures: attendance per month (attendpermonth), identification with Evangelical Christianity (protestant), and self-rated religiousness (religiousness).

There are 3 questions to be answered in this paper:

1. Do people become more or less religious when they become richer?

2. Do people change their religion when they become richer?

3. Do people change attendance frequency when they become richer?

There have been published results that show the relationship of religion on individual behavior, but Buser asserts his paper as the first of its kind to look at the reverse relationship. He uses RDD to analyze the effects of a household being eligible to receive the cash transfer, and those that do not receive it. This approach employs a discontinuous assignment rule that randomizes in determining whether households are placed in the treatment or control group. As Ecuador is a lower middle income country with high poverty and inequality levels, the eligibility of the government cash transfer program may have significant income effects on poorer households. Depending on where a household is initially placed within the wealth index, it asserts their position to receive the transfer.

This paper is interesting as it examines the institutional rule of sorting households into either group, and whether the income shock results in statistically significant changes in attendance, identification, or self-rated religiousness. The idea is to compare households in a small neighbourhood around the cuttoff (above/below). There is not much need for controls as the only difference between the households near the threshold is that one receives the treatment whereas the other does not. The difference in status allows us to determine the treatment effect.

Part II: Data Description

The survey consists of randomly sampled households from poor neighborhoods in 3 urban centers: Guayaquil, Quito, and Santo Domingon. The survey takers collected data on characteristics of households such as: age of the responder, education level of the responder, amount spent on expenditures, their affiliation with religious services, and more. The **SELBEN-II (s2)** survey was conducted around 4-5 years after the initial start. This dataset is relevant as the households were randomly sampled within each city and within the same government parish. This helped ensure the groups were geographically balanced. In addition, in randomly

sampling 2645 households, this decreased the probability of selection bias of households whom received the transfer being different from those whom didn't.

The descriptive statistics are presented in Table 2. Since no data repository was provided, I went through and corresponded frequencies of each variable with their respective meaning (i.e.: denomination of "1" == Catholic). The sample means and standard deviations are also provided.

```
ir_df <- haven::read_dta("20140162_Data.dta")

#Table 2 - religion stats
ir_religion  <- ir_df %>% select(denomination) %>%
  group_by(denomination) %>%
  summarize(Observations = n()) %>%
  mutate(Percent = Observations/nrow(ir_df)*100) %>%
  select(Observations, Percent)

relig_tib <- cbind(tibble(ir_religion$Observations),
                   tibble(ir_religion$Percent))
rownames(relig_tib) <- c("Catholic", "Non-Catholic Christian", "Jewish",
                         "Atheist/none", "Other")
#bind all data together
colnames(relig_tib) <- c("Obsv", "Percent") #add in column names

#Table 2 - attendance stats
ir_attend  <- ir_df %>% select(attendance) %>%
  group_by(attendance) %>%
  summarize(Observations = n()) %>%
  #count how many per categorical group
  mutate(Percent = Observations/nrow(ir_df)*100) %>%
  select(Observations, Percent)

sa_tib <- cbind(tibble(ir_attend$Observations),
                tibble(ir_attend$Percent))
rownames(sa_tib) <- c("Never", "Special Occasions", "< 1 month", "Once/month",
                      "2-3 times/month", "Once/week", "2-3 times/week",
                      "4-6 times/week", "Every day")
colnames(sa_tib) <- c("Obsv", "Percent")

#Table 2 - means and sds
ir_meansd <- ir_df %>% # filter out the selected categories
  select(attendpermonth, religiousness, householdsize, ageresponder,
         schooling_resp, expenditures)

mean <-  data.frame(as.list(apply(ir_meansd[1:6], 2,
                                  mean, na.rm = TRUE)))
sd <- data.frame(as.list(apply(ir_meansd[1:6], 2,
                               sd, na.rm = TRUE)))

meansd <- matrix(c(mean, sd), nrow = 6, ncol = 2,
                 byrow = FALSE)
colnames(meansd) <- c("Means", "SD")
rownames(meansd) <- c("Attendance/mth", "Religiousness", "Household size", "Age",
                      "Years of schooling", "Household expenditure")
```

Table 2: Descriptive Statistics - Religion

```
##                          Obsv     Percent
## Catholic                1971 74.51795841
## Non-Catholic Christian   452 17.08884688
## Jewish                     2  0.07561437
## Atheist/none              53  2.00378072
## Other                    167  6.31379962
```

Table 2: Descriptive Statistics - Service Attendance

```
##                    Obsv   Percent
## Never               192  7.258979
## Special Occasions   310 11.720227
## < 1 month           138  5.217391
## Once/month          324 12.249527
## 2-3 times/month     507 19.168242
## Once/week           738 27.901701
## 2-3 times/week      257  9.716446
## 4-6 times/week      106  4.007561
## Every day            73  2.759924
```

Table 2: Means and Standard Deviations

```
##                         Means    SD
## Attendance/mth        4.318904 6.142649
## Religiousness         6.827599 2.37958
## Household size        4.46276  1.969086
## Age                   42.71871 11.0413
## Years of schooling    7.446388 3.686763
## Household expenditure 297.2648 151.3583
```

Part III: Model and Results: OLS

I first use sharp RDD methods in segregating households based on their s2 scores. In sharp RDD, being below/above the cutoff (score2_1) means the observation (household) is in the treatment or the control group. So Buser randomizes households on eligibility criteria if their $s2 < 0$ is ($eligible = 1$), and ineligible if $s2 > 0$ of ($eligible = 0$). The strict inequality does not change the group composition. When we are running regression analyses, the coefficient on the *eligible* binary indicator will be the estimate of the treatment effect.

```
ir_new <- ir_df %>%  mutate(eligible = case_when(score2_1 >= 0 ~ 0, # not eligible
                                 score2_1 < 0 ~ 1)) # eligible
```

There is an ineligibility/eligibility issue.

```
ir_inelig_col <- ir_new %>%  select(collect_2, score2_1) %>%
  filter(collect_2 == "1" & score2_1 > 0) #38 ineligible, but collected
nrow(ir_inelig_col)
```

```
## [1] 38
```

```
ir_elig_not <- ir_new %>% select(collect_2, score2_1) %>%
  filter(collect_2 =="0" & score2_1 < 0) #193 eligible, but did not collect
nrow(ir_elig_not)
```

```
## [1] 193
```

We find that there were 38 households eligible, but did not collect the transfer, and 193 households ineligible, but still received the transfer. The sample size for all regressions in the paper was still 2645 - the size of the original dataset, so I have decided to bypass losing degrees of freedom by not removing these observations. However, I postulate this is the reason as to why Buser uses fuzzy RDD as household status is not solely determined by their s2. This may also result in differing coefficient estimates when running later regressions, as compared to Buser's results.

Buser uses the model:

$$Y = \alpha + \delta_{eligible} + f(s) + \beta score2_1 + controls$$

Define the variables as:
Y: income measure
eligible: whether the household is eligible or not
$f(s)$: conditioning on the polynomial in $score2_1$
$score2_1$: SELBEN-II score

Let us first try fitting a linear model on the dataset. The asssumptions of no endogeneity and serial correlation seem valid in this model as the zero conditional mean ($E(u|X) = 0$) is satisfied. In our OLS regression, I have decided to include score2_2 (score2_1*eligible) as a way to look at the effects of eligibility on s2. In addition, Buser controls for eligibility before the change (moremoneyold).

The following uses the **1st-order polynomial:**

```
##                 Estimate Std. Error  t value    Pr(>|t|)
## Church_Attend 1.38981721 0.48014072 2.894604 0.003827525
## Evangelical   0.05128881 0.02945199 1.741438 0.081723341
## Religiousness 0.20289451 0.18621082 1.089596 0.275990732
```

We see that at the 1% significance level, recipients spend about 1.39 more days at church services than non-recipients. At the 10% significant level, recipients are 5.12% percentage points more likely to be Evangelical.

Let us conduct further regressions on different order polynomials of the s2 and income measures.

The following uses the **2nd-order polynomial:**

```
##                 Estimate Std. Error  t value    Pr(>|t|)
## Church_Attend 1.37508248 0.47886873 2.871523 0.004117552
## Evangelical   0.05193934 0.02937272 1.768285 0.077128797
## Religiousness 0.20027417 0.18571410 1.078400 0.280953603
```

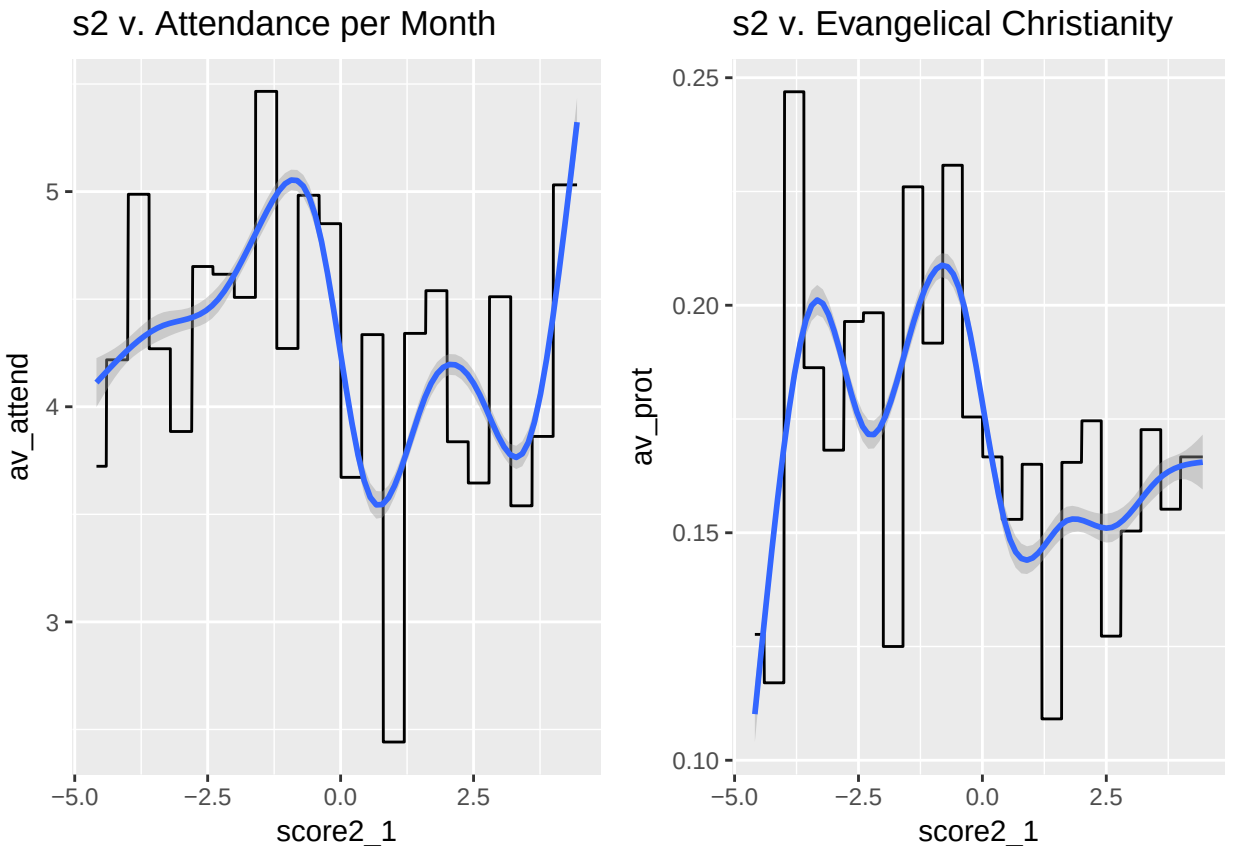
The following uses **3rd-order polynomial:**

```
##                 Estimate Std. Error   t value  Pr(>|t|)
## Church_Attend 1.01581362 0.40059788 2.5357439 0.0112781
## Evangelical   0.03543929 0.02456966 1.4424003 0.1493081
## Religiousness 0.13547804 0.15537629 0.8719351 0.3833231
```

The replication results worsen in approximation to Buser's estimates as the order of the polynomial increases, and the coefficients also weaken in effect. This is most likely due to higher order polynomials becoming more complicated within regression fitting. Let's look at the discontinuity around $s2 = 0$ (which determines household eligibility).

Buser uses a binned approach to segregate households by their s2 scores. Let us try plotting the relationship between
(i) score2_1 & attendpermonth
(ii) score2_1 & protestant

```
(grid.arrange(v1plot, v2plot, ncol=2))
```

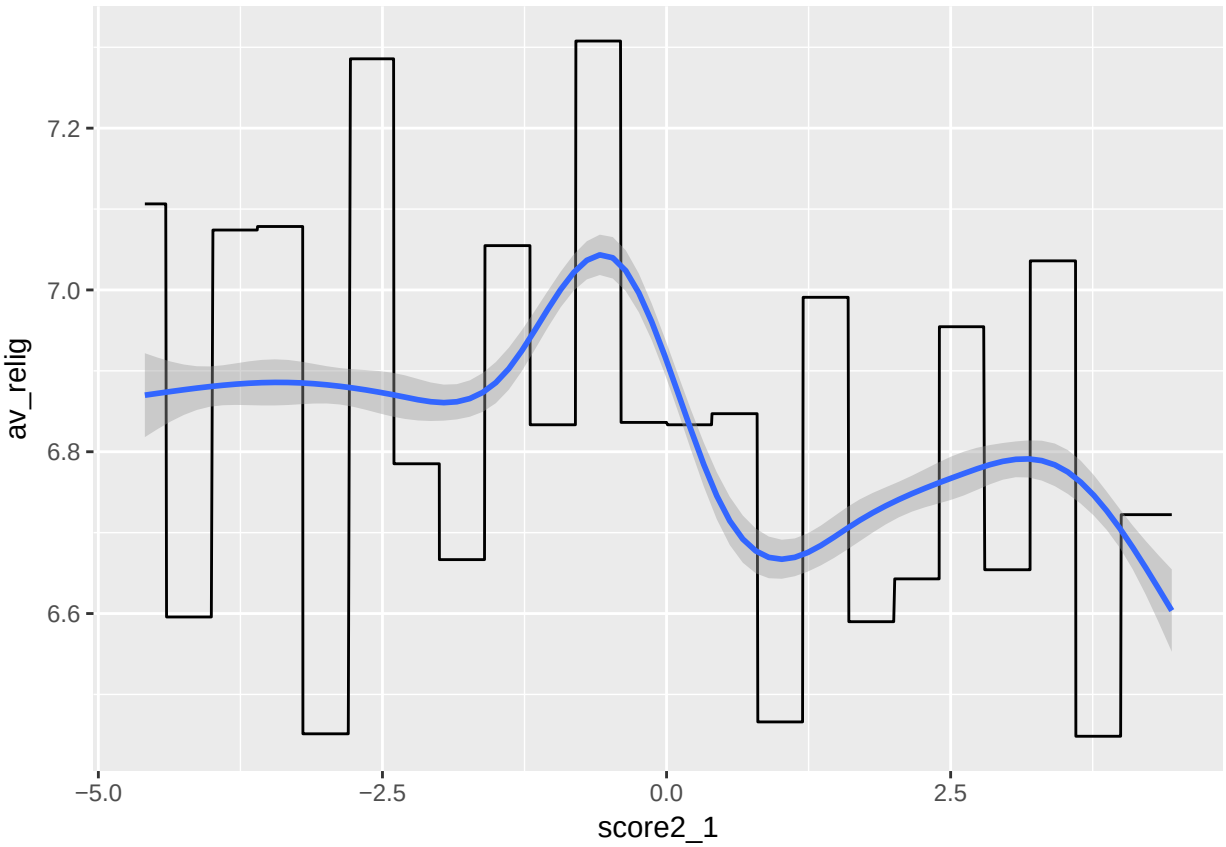



```
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

We see that by using the auto-fit line method, it exhibits a polynomial relationship between s2 and attendpermonth, similarly with s2 and protestant. There is a sharp drop in the slope of the best-fit line around s2=0. This could potentially mean a higher order polynomial regression will fit the relationship better.

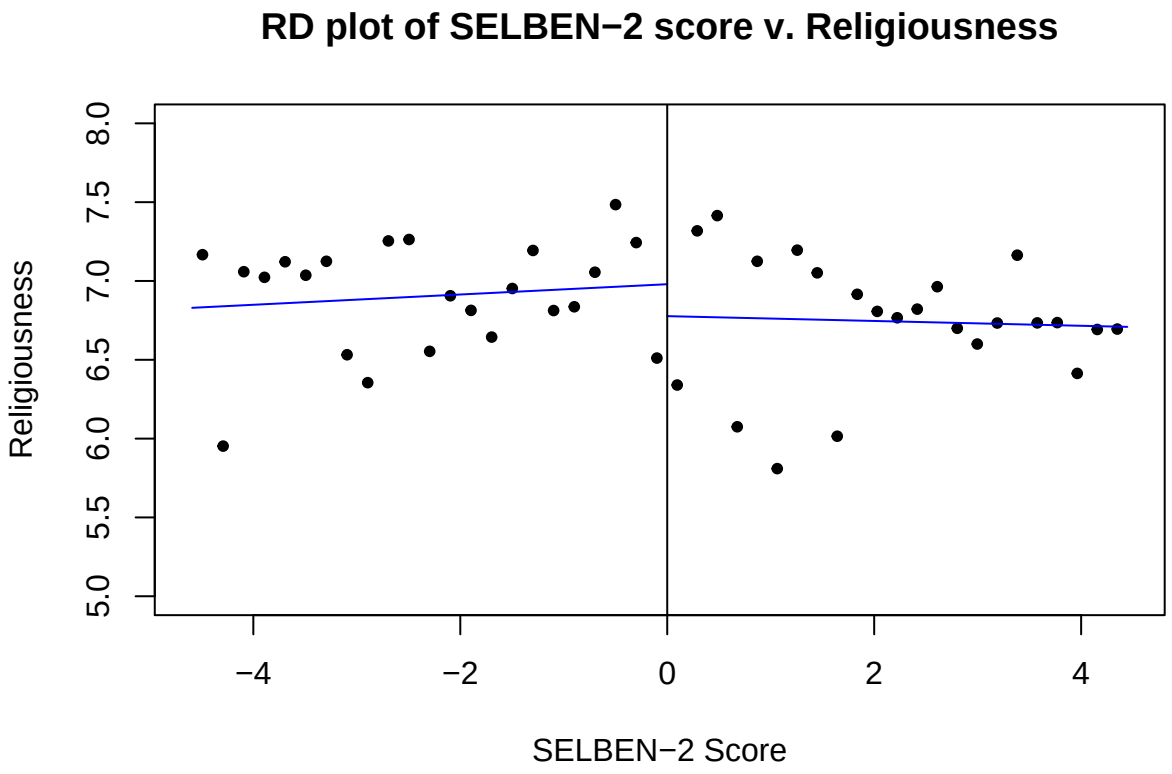
(iii) score2_1 & religiousness

```
## `geom_smooth()` using method = 'gam'
```

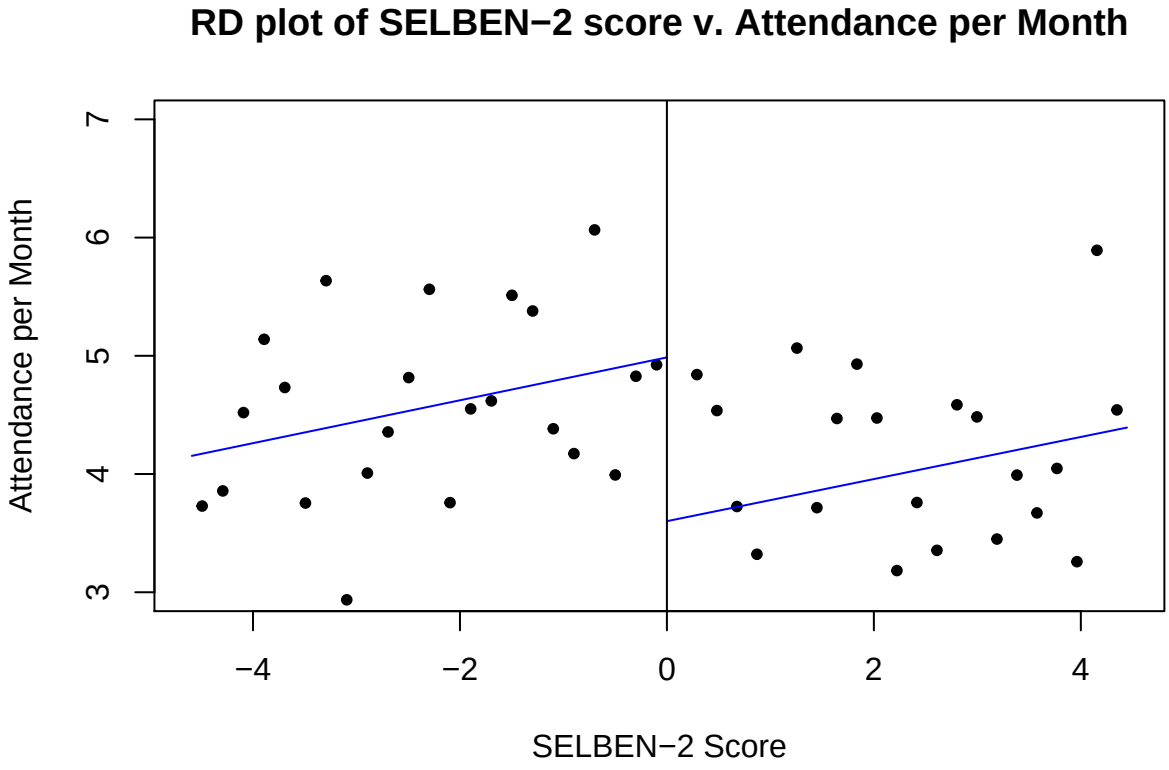


Getting similar results to (i) and (ii) from above. Using OLS is not plausible as linear regression is not a good estimation technique to analyze the effects of eligibility for cash transfer as around the cutoff, there is a noticeable difference in the slopes; a best-fit line with a constant slope cannot properly depict the effect of eligibility.

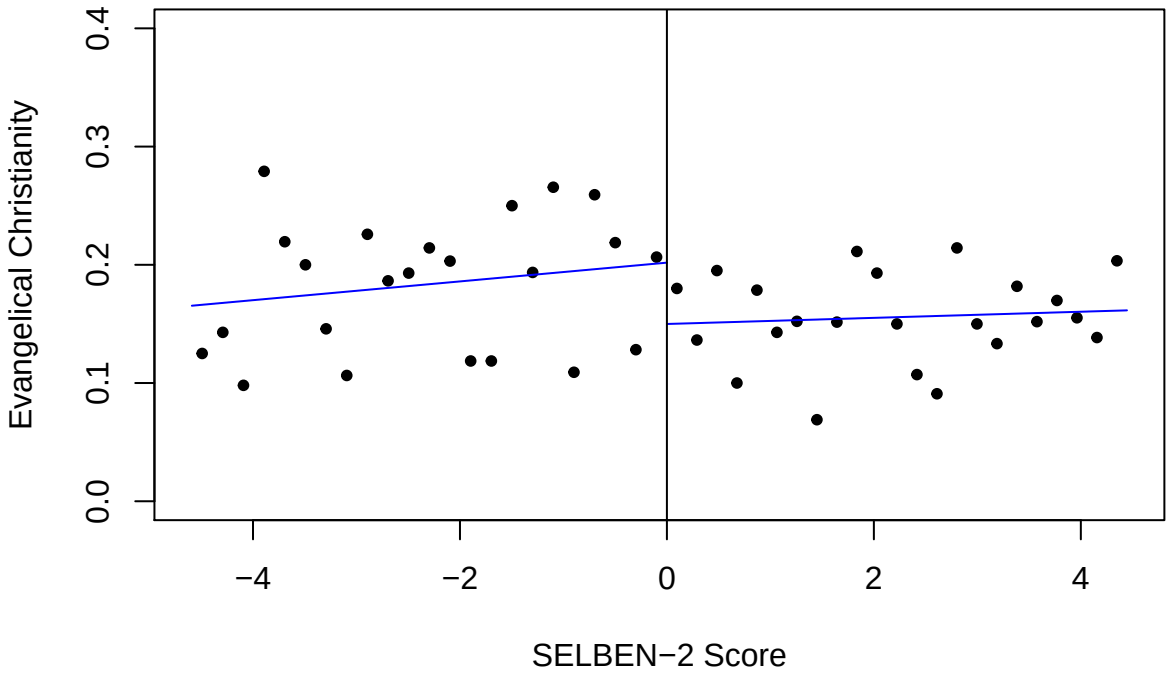
Let's examine the relationship by plotting some Regression Discontinutiy (RD) plots between the income measure (Y) and the $s2$:




```
## Call: rdplot
##
## Number of Obs.                2645
## Kernel                         uni
##
## Number of Obs.                1344           1301
## Eff. Number of Obs.           1344           1301
## Order poly. fit (p)              1              1
## BW poly. fit (h)             4.591          4.446
## Number of bins scale             0              0
```

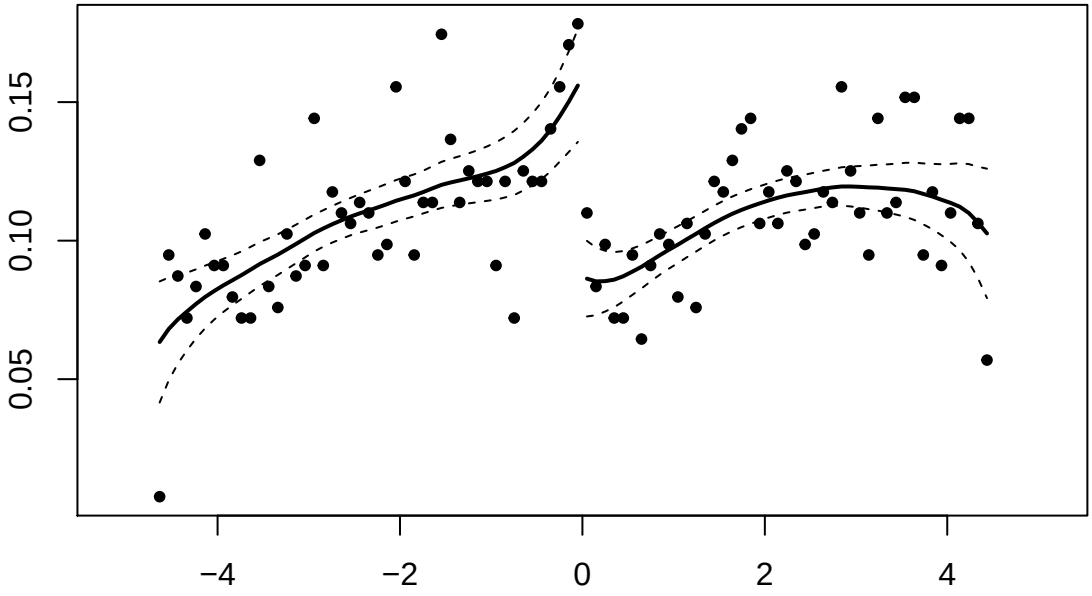



```
## Call: rdplot
##
## Number of Obs.                2645
## Kernel                         uni
##
## Number of Obs.                1344           1301
## Eff. Number of Obs.           1344           1301
## Order poly. fit (p)              1              1
## BW poly. fit (h)             4.591          4.446
## Number of bins scale             0              0
```




```
## Call: rdplot
##
## Number of Obs.                2645
## Kernel                         uni
##
## Number of Obs.               1344          1301
## Eff. Number of Obs.          1344          1301
## Order poly. fit (p)             1             1
## BW poly. fit (h)            4.591         4.446
## Number of bins scale            0             0
```

We see now that the slopes between the treatment and control groups are noticeably different. RDD would be the better estimation technique. As a final closing in this section in using the OLS model, we look at the McCrary 2008 Sorting test. This test examines the density of observations in the assignment variable, s2.




```
## [1] 6.898781e-06
```

We see that there is a clear discontinuity in the density of s2, and this suggests that possibly some households were able to manipulate their treatment status. Manipulation as in some households were able to determine their position on the SELBEN-II index and reported a lower income status in order to receive the cash transfer. If this is true, it violates a major assumption of RDD that individuals cannot manipulate their x value (i.e. households cannot manipulate their s2 to fall within the eligibility treatment group).

Some possible explanations of manipulation may be that households in previous years received the cash transfer, and knowing the SELBEN-II index may be changed, may bias their income estimates downwards in order to ensure they receive it.

Part IV: Advanced OLS & Results

From Part III we see that using linear regression of Y (income measure) on X with other control variables can be severely biased for the causal effect of X on Y due to endogeneity. Since there may be different impacts based on being in the treatment or cuttoff group, we want instead to only examine the treatment effect around s2=0. To do this, we can use RDD to look at the effects of the treatment based on the running (forcing) variable of interest.

RDD consists of sharp and fuzzy RDD. Sharp RDD is where all units below or above a threshold become treated, whereas the other half do not. Fuzzy RDD is a bit more complex in that the running variable being above/below a threshold only increases the probability of being treated, with other factors also influencing whether you actually get treated or not. We can think of sharp RDD as a selection-on-observables situation, whereas fuzzy RD is an IV-type setup.

Some assumptions in RDD are that:

- Assignment to treatment and control isn't random, but whether individual observation is treated is assumed to be random
- Buser assumed that households cannot perfectly manipulate their s2
- As a result, whether a household becomes eligible or not is random

I want to augment my analysis of the linear form by using an Instrumental Variable approach and trying a different model specification than the one Buser uses in the paper (within reason).

1. Instrumental Variable Regression

Fuzzy RDD means there is a higher probability of being in the treatment group. It can happen when there is a publicly suggested cutoff that's not enforced. We see this where some whom were eligible, did not collect - and the converse also being true. Let us estimate the causal effect of receiving the cash-benefit on attendance by 2SLS and instrumenting collect_2 (Z) with eligible while controlling for the running variables. This will allow us to investigate whether collect_2 was a valid variable in Buser's regression model above.

2SLS
First stage: regress X on Z
Second stage: regress Y on fitted values (Y-hat)
IV model: Y (income measure) = b0 + b1*eligible + score2_1 + Z

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 3.9992125  0.1763587 22.676579 3.571257e-104
## varX        0.7111715  0.2887285  2.463115  1.383690e-02
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.15500156 0.01081519 14.331839 6.392155e-45
## varX        0.03534136 0.01770626  1.995981 4.603847e-02
```

```
##              Estimate Std. Error   t value   Pr(>|t|)
## (Intercept) 6.7349484 0.06834419 98.544557 0.00000000
## varX        0.2061071 0.11189081  1.842038 0.06558156
```

In using IV, we find that households whom collect cash benefits increase the number of church attendances by 0.71 days per month. This is the average effect for people who collect the benefit because they are eligible (the compliers). This seems likely and consistent with Buser's other results as a positive income shock allows for increased time to attend services and a higher status.

2. Misspecified model:

Buser may not have included all variables that have explanatory power on Y. By including more explanatory variables that appear to explain Y, we can control for the X's and hopefully decrease omitted variable bias. As Ecuador is a relatively poor country, an increase in income could result in increased expenditure for the household. Usually, most of it is spent on children to close the gap caused by poverty read more here. I want to look at level-log form to see if Y can be better explained by a change in the model. I will log the 3 variables below and look at their changes in p-value and coefficient estimate: - householdsize - education - expenditures

```
##
## Call:
## lm(formula = attendpermonth ~ schooling_resp + householdsize +
##     expenditures + expenditures + eligible + score2_1, data = ir_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2382 -3.6832 -1.5624 -0.0838 26.5262
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.8254888  0.4941842   7.741 1.42e-14 ***
## schooling_resp -0.0385527  0.0362187  -1.064   0.2872
## householdsize  -0.0187031  0.0687971  -0.272   0.7858
## expenditures    0.0006499  0.0008823   0.737   0.4615
```

```
## eligible        1.2316352  0.4850509   2.539   0.0112 *
## score2_1        0.1580728  0.0950050   1.664   0.0963 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.085 on 2499 degrees of freedom
## Multiple R-squared:  0.003587,   Adjusted R-squared:  0.001593
## F-statistic: 1.799 on 5 and 2499 DF,  p-value: 0.1097
```

```
##
## Call:
## lm(formula = attendpermonth ~ leduc + lhhs + lexp + eligible +
##     score2_1, data = ir_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5342 -3.6724 -1.5614 -0.0625 26.5472
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.88122    1.43802   2.699   0.0070 **
## leduc       -0.37348    0.21103  -1.770   0.0769 .
## lhhs        -0.12713    0.32538  -0.391   0.6960
## lexp         0.11671    0.28063   0.416   0.6775
## eligible     1.23200    0.48486   2.541   0.0111 *
## score2_1     0.16045    0.09492   1.690   0.0911 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.083 on 2499 degrees of freedom
## Multiple R-squared:  0.004304,   Adjusted R-squared:  0.002312
## F-statistic: 2.161 on 5 and 2499 DF,  p-value: 0.05579
```

We see that in using the level-log form, leduc seems effective at capturing some of the effects on Y, as the p-value has decreased by 75%. We can run a regression to see the relationship of leduc on attendance.

```
##
## Call:
## lm(formula = attendpermonth ~ ir_test$leduc + score2_1, data = ir_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1487 -3.6561 -1.6499 -0.1596 26.2659
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.98138    0.42061  11.843   <2e-16 ***
## ir_test$leduc -0.36364    0.20946  -1.736   0.0827 .
## score2_1      -0.04628    0.04759  -0.973   0.3309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.087 on 2502 degrees of freedom
## Multiple R-squared:  0.001634,   Adjusted R-squared:  0.0008357
## F-statistic: 2.047 on 2 and 2502 DF,  p-value: 0.1293
```

Part V: Conclusion

Summarize approach and findings:
Based on the results I obtained using linear regression, the coefficient values are similar when compared to Table 4. Buser concludes that a difference in income does affect families and their frequency of church services and on the type of church they attend, and no effect on how religious people rate themselves to be. Although I was not able to perform a fuzzy RDD properly or determine the threshold treatment effect difference at s2=0, I obtained similar findings.

Strengths and weaknesses:
Buser attempts to uncover the causality between income shocks and behaviors relating to religious association. By random sampling, Buser is assumed to achieve exogeneity in the error terms. In addition, as supported in tables within the Online Appendix, he validates that the F-statistics are high and well above the required thershold. A discontinuity density test was also performed on the controls as dependent variables, none varied greatly around the cutoff being conditinoal on score2_1. This implies thatHowever, there are some weaknesses. Both in Buser's and my results, we see that higher order polynomials showed a weakening effect in the explanatory variables. This may be as researched by Gelman and Imbens in their paper on [**"Why High-Order Polynomials Should Not be Used in Regression Discontinuity Designs"**](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.569.2504&rep=rep1&type=pdf), due to:

i. Noisy weights: weights have unattractive properties and the higher the coefficient estimate, the more sensitive the estimate is to the order of the polynomial. We see this in the coefficient of `attendpermonth`, as the standard errors almost increase by 2-fold in polynomial order 1 vs. polynomial order 3 (Buser's Table 4).
ii. Estimates become highly sensitive to the degree of the polynomial: this is partly due to reason (i) above, and the confidence intervals become larger than necessary.
iii. Inferences are not able to achieve nominal coverage: the global polynomial approximations are not precise enough so that bias can be ignored when estimating the treatment effect. Although my coefficient estimates were similar to Buser's, my standard errors decrease as the poly order increases. This is also inconsistent with the findings of Gelman and Imbens.

```
##                 Estimate Std. Error  t value    Pr(>|t|)
## Church_Attend 1.38981721 0.48014072 2.894604 0.003827525
## Evangelical   0.05128881 0.02945199 1.741438 0.081723341
## Religiousness 0.20289451 0.18621082 1.089596 0.275990732
```

```
##                 Estimate Std. Error  t value    Pr(>|t|)
## Church_Attend 1.37508248 0.47886873 2.871523 0.004117552
## Evangelical   0.05193934 0.02937272 1.768285 0.077128797
## Religiousness 0.20027417 0.18571410 1.078400 0.280953603
```

```
##                 Estimate Std. Error   t value  Pr(>|t|)
## Church_Attend 1.01581362 0.40059788 2.5357439 0.0112781
## Evangelical   0.03543929 0.02456966 1.4424003 0.1493081
## Religiousness 0.13547804 0.15537629 0.8719351 0.3833231
```

In addition, randomization is not perfect as above. When the change in eligibilty came into effect, some households possibly did not collect their cash transfer as they did not know. This biases the estimates on the income measures as the household could have potentially been a part of the treatment/control, than not. As a result, this blurry line dividing households is instead captured within the error terms, and exogeneity may fail. If possible, it would be an interesting condition to query households on the amount they tithe at their church. In this way, it would allow for control of one of the reasons Busers attributes higher attendance religiously when income is low. If the tithe is low regardless of receiving the transfer or not, it would negate Buser's assumptions that receiving the transfer equates to more church attendances due to having higher status and extra income.

There are also many qualitative factors that should be considered before estimating the causal effect of income

shocks on religious participation. Buser speaks about the rise in Evangelical Christianity around the time the survey was conducted. As faith and religion depends on an intrinsic connection to the church, for some households - it may have also been the time that they were introduced to religion, and decided to attend. This would have been a correlation, rather than a causality. Furthermore, one could look at the effects on the household's children years after they started receiving the transfer. For in the article, households whom have increased income spend more on their children. This could also have an effect on the choices the children makes later in terms of education or religion. By surveying the same household over different time periods (young mother vs. mid-age mother), it will allow for extrapolation of the causality of income effects on the choices made religiously.

References:

1. Angirst & Pischke, 2009
2. Gelman, Imbens
3. Pustejovsky, James
4. https://economics.mit.edu/files/32
5. Article: The myth of how families in poverty spend their money
6. Buser, Thomas - The Effect of Income on Religiousness
7. R Markdown Basics
8. RPubs - Regression Discontinuity
9. Stiggler, Matthieu