An exercise on treatment effects: Reproducing in R, Buser (2015) treatment effect analysis of income on religiousness

Kreshnik Xhangolli

April 20, 2017

Purpose

The purpose of his exercise is to present the treatment effect analysis of income on religiousness conducted by Busser (2015) [1] in R. In addition, this exercise introduces several elements of reproducible research and good analysis such as:

1. Use of one environment for both analysis and report writing. Code and comments are written in `.rnw file` and `knitr` package is employed.
2. Use of user-definied functions for the production of plots when reproducing the same plot with diffrent bin size.
3. Use of formula objects to reproduce the same analysis by adding and dropping covariates, or for different fucntional forms of the score variable.
4. Use of package stargazer to produce publication quality tables. This is a recent package added to R packages.

In terms of analysis we show:

1. Both treatment D and outcome Y variable pass the visual inspection test of conditional mean discontinuity at the at the cutoff point of the score variable S; $E(Y|S), E(D|S)$ discontinuous at cutoff
2. Treatment effect of income on religiousness
   - controling for covariates
   - employing different bandwidths
   - employing up to 3rd order polynomial of the score variable
3. Covariates pass the regression test of no-break at cutoff point

Buser (2015) employes cross-section data from a representative survey (N =2,645) from Ecuador. The fuzzy regression disconuity setting derived from the cash transfer program of theEcuadorian governement, called Bono Desarollo Humano (BDH). Eligibility BDH recipiency is determined by a cutoff percentile on a wealth index. The index is derived from nonlinear principal components analysis on a set of household observable characteristics. A new index, calculated on somewhat differing household characteristic was introduced in 2009, which affected the eligibility of households close to the cutoff point. Some previously non-eligible households gain eligibility, while some previuosly eligible households gained eligibility. Religiousness was measure through three proxies; household self-assessed religiousness, being an evangelical Christian, and attendance to selected religious services. Covariates analyzed are presented in table 1. We recommend reading the paper for the detailed description.

Table 1: List of covariates

Variable	Description	Variable	Description
ciudad	city	parroqui	area
expenditures	total household expenditure	denomination	religion of this household
religiousness	self-assessed religioussness (scale 0-10)	score	normalized, 0 centered wealth index
moremoneyold	transfer recipient before the reform	moremoneynew	transfer recipient after the reform
householdsize	household members	ageresponder	age (years)
schooling$_{r}esp$	Years of schooling	protestant	Protestant
attendance	Religious service attendance	attendpermonth	attendance per month
collect	=moremoneynew		

[1] Buser, Thomas. "The effect of income on religiousness." American Economic Journal: Applied Economics 7.3 (2015): 178-195.

Data Preparation

The first line sets `knitr` options to no warnings, no messages and no double-hash comments for output.

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE, comment = NA)
```

The `foreign::read.dta` function can read only STATA 12 files therefore we first converted the data to this format. After reading the data into a dataframe using relative paths, we inspect data frame column types through a loop checking whether discrete data was converted correctly to factor type. The automatic conversion of discrete values into factors converted also *religiousness (scale 1 to 10)* into a factor. We need this variable as numeric therefore it is converted back to numeric through `base::as.numeric`. The last two lines create two new dataframes, each containing records with either positive or negative score segments. These dataframes will be used in computing fitted lines.

```
remove(list = ls())
library(foreign)
library(ggplot2)
library(gridExtra)
library(dplyr)
library(stargazer)
library(MASS)
library(Formula)

dataName <- "IncRelig_STATA12.dta"
fullPath <- paste("./",dataName, sep="")

df <- read.dta(fullPath)
for (i in names(df)) print(c(i,class(df[[i]])))

df$religiousness <- as.numeric(df$religiousness)

df_pos_sc <- df[df$score>0,]
df_neg_sc <- df[df$score<0,]
```

Exercise 1: Visual Inspection

Plots were build by using `ggplot2` package and the display of multiple plots was done with `gridExtra` package. Different bin sizes were used to calculate $E(Y|S)$. To make code reusable we created a function that would take a bin width, y variable and y axis label as input and would return the desired plot as output. The y axis label is required mainly for better aesthetics. A list of `grobs` was created by applying `base::lapply` to function `visual_inspection` on an input vector of bin widths and the respective inputs for y variable and y axis label. Iteration was possible through `standard evaluation` of `ggplot2`, i.e `aes_string`.

```
visual_inspection <- function(bin_input, y_input, y_label_input){
    graph_out <- ggplot(df,aes_string(x="score", y=y_input))+
            geom_point(stat = "summary_bin", fun.y = mean, binwidth = bin_input) +
                  geom_vline(xintercept = 0) +
                    geom_smooth(data = df_pos_sc,method = "lm",se=FALSE) +
      geom_smooth(data = df_neg_sc,method = "lm",se=FALSE) +
      labs(title = paste("bin width",bin_input),
           y = y_label_input,
           x = "score") +
      theme(axis.title = element_text(size=6),
            plot.title = element_text(size=6),
            axis.text = element_text(size=5))
    return(graph_out)
}

bin_vec <- c(0.8,0.4,0.2)

g1 <- lapply(bin_vec,visual_inspection,"attendpermonth","Church month attend")
```

```r
g1[4:6] <- lapply(bin_vec,visual_inspection,"religiousness","Religiousness")
g1[7:9] <- lapply(bin_vec,visual_inspection,"protestant","Protestant")

grid.arrange(grobs = g1,ncol=3)
```

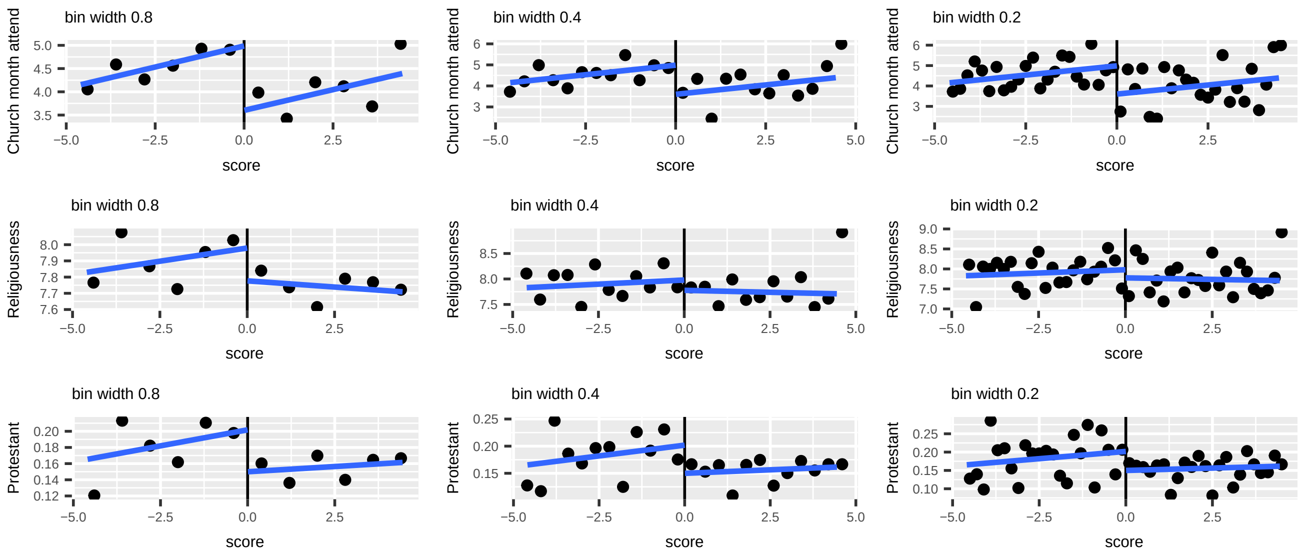

Figure 1: Disconuity visual inspection

From a visual inspection of Fig 1. we can we can assume that a disconuitiy of monthly church attendance at the threshhold of the SELBEN II score exists. This is visible for different bin width used. In a similar manner we can assume the disconuity also from the visual inspection of Protestant affiliation (or likelihood of being Protestant) for all the three bin width presented. As for self perceived religiousness, while we might perceive a disconuity with binwidth 0.8, the disconuity is less clear when we reduce the bin width to 0.4 or 0.2. All the above results are in line with Buser(2015).

Exercise 2: Calculating effects

For each of the dependent variables we will analyze 6 models using OLS. The simplest model measured the transfer effects controling on the forcing variable *score* and previous cash transfer status. More complex models introduced second and third order degree polynomials of the forcing variable and we added complexity by controlling on an additional set of controls (household size, age and education level). We built function `analyze_models` that uses *Formula* objects to reuse code and `stargazer` package to display the results in publication style outputs. The name of the dependent variable and the title for the table are required as input. The latter is required only for presentation purposes. In the beginning we declare the Formula object `overall_model` that has 3 right hand side (RHS) parts and 4 left hand side (LHS) parts. We create the values of attributes `lhs` with `if else` statements. Instead the values for `rhs` are stored in a list. For each iteration on the elements of `rhs` list, we recreate a formula object by specifying `lhs` and `rhs` from `overall_model` and then run a regression employing the temporary formula. The list with regression results for each model is then feeded to `stargazer`.

```r
analyze_models <- function(dependent_inp, title_inp){

    overall_model <- as.formula(paste0("attendpermonth | religiousness | protestant ",
                               " ~ moremoneynew + moremoneyold + score   | ",
                               "householdsize + ageresponder + schooling_resp | ",
                               "I(score^2) | ", "I(score^3)"))

    overall_model <- Formula(overall_model)

    if (dependent_inp == "attendpermonth") model_lhs <- 1
    else if (dependent_inp == "religiousness") model_lhs <- 2
    else if (dependent_inp == "protestant") model_lhs <- 3

    models_rhs <- list()
```

```r
    models_rhs[["simple"]] <- 1
    models_rhs[["poly 2"]] <- c(1,3)
    models_rhs[["poly 3"]] <- -2
    models_rhs[["controls"]] <- c(1,2)
    models_rhs[["controls poly 2"]] <- -4
    models_rhs[["controls poly 3"]] <- c(1,2,3,4)

    list_models <- list()

    for (i in names(models_rhs)){
      temp_formula <- as.formula(terms(overall_model,
                                       lhs = model_lhs,
                                       rhs = models_rhs[[i]]))
      list_models[[i]] <- lm(temp_formula,data=df)
    }

    return(
      stargazer(list_models,
                title = paste0("Regresion Resuls: Effects on ", title_inp),
                column.sep.width = "2pt", no.space = TRUE, header = FALSE,
                notes = paste0("Controls include household size, age ",
                               "and years of schooling of the respondent"),
                omit.stat = c("f","ser","adj.rsq"),
                covariate.labels = c("Transfer now","Transfer before",
                                     "Score", "Score^2", "Score^3",
                                     "Hh size", "Age", "Education"),
                dep.var.labels.include = FALSE, dep.var.caption = "")
    )
}
```

Overall the significance of the transfer effects is in line with Buser(2015). The cash transfereffects have a positive effect on monthly church attendance and Protestant affilition, but do not affect self perceived religiousness. We note that our estimations of the transfer effects differ from those of Buser (2015). In our simplest model we estimated an increase of 1.4 church attendances monthly conditional on receiving the transfer, compared to an increase of 1.7 from Buser(2015). Also we estimated that a cash transfer would increase the likelihood of being Protestant by 5.3 percentage points, while Buser (2015) estimate was of 6.6 percentage points. These differences were consistent also for the more complex models. We have no hypothesis why the differences occur.

```r
analyze_models("attendpermonth","church attendance")
```

```r
# Output in Table 2
```

```r
analyze_models("religiousness","self perceived religiousness")
```

```r
# Output in Table 3
```

```r
analyze_models("protestant","Protestant affiliation")
```

```r
# Output in Table 4
```

```r
# This chunk of code computes a backward stepwise regression on monthly
# church attendance. Results for this chunk are verbose and therefore
# not presented. Summary of results is presented by calling "anova"
# attribute of mode_step

fmla_all_contr <- as.formula(paste0("attendpermonth ~ moremoneynew + moremoneyold + ",
                                    "score  + I(score^2) + I(score^3)+",
                                   "householdsize + ageresponder + schooling_resp + ",
```

Table 2: Regresion Resuls: Effects on church attendance

	(1)	(2)	(3)	(4)	(5)	(6)
Transfer now	1.391***	1.389***	1.149*	1.509***	1.504***	1.233**
	(0.477)	(0.479)	(0.632)	(0.474)	(0.476)	(0.628)
Transfer before	0.074	0.074	0.074	0.206	0.206	0.206
	(0.239)	(0.239)	(0.239)	(0.238)	(0.238)	(0.238)
Score	0.181*	0.181*	0.061	0.203**	0.202**	0.067
	(0.093)	(0.093)	(0.226)	(0.092)	(0.093)	(0.225)
Score^2		−0.001	−0.001		−0.002	−0.003
		(0.020)	(0.020)		(0.020)	(0.020)
Score^3			0.007			0.007
			(0.011)			(0.011)
Hh size				0.032	0.032	0.032
				(0.061)	(0.061)	(0.061)
Age				0.066***	0.066***	0.066***
				(0.012)	(0.012)	(0.012)
Education				0.054	0.054	0.054
				(0.036)	(0.036)	(0.036)
Constant	3.558***	3.566***	3.692***	0.056	0.075	0.231
	(0.305)	(0.342)	(0.404)	(0.808)	(0.822)	(0.856)
Observations	2,645	2,645	2,645	2,630	2,630	2,630
R^2	0.004	0.004	0.004	0.016	0.016	0.016

Note: *p<0.1; **p<0.05; ***p<0.01

Controls include household size, age and years of schooling of the respondent

Table 3: Regresion Resuls: Effects on self perceived religiousness

	(1)	(2)	(3)	(4)	(5)	(6)
Transfer now	0.217	0.206	0.236	0.224	0.215	0.264
	(0.185)	(0.186)	(0.245)	(0.185)	(0.185)	(0.245)
Transfer before	0.005	0.006	0.006	0.045	0.045	0.045
	(0.093)	(0.093)	(0.093)	(0.093)	(0.093)	(0.093)
Score	0.011	0.009	0.024	0.010	0.009	0.034
	(0.036)	(0.036)	(0.088)	(0.036)	(0.036)	(0.088)
Score^2		−0.005	−0.005		−0.004	−0.004
		(0.008)	(0.008)		(0.008)	(0.008)
Score^3			−0.001			−0.001
			(0.004)			(0.004)
Hh size				−0.039	−0.038	−0.038
				(0.024)	(0.024)	(0.024)
Age				0.021***	0.020***	0.021***
				(0.005)	(0.005)	(0.005)
Education				0.016	0.016	0.016
				(0.014)	(0.014)	(0.014)
Constant	7.714***	7.753***	7.737***	6.861***	6.892***	6.864***
	(0.118)	(0.133)	(0.157)	(0.315)	(0.321)	(0.334)
Observations	2,645	2,645	2,645	2,630	2,630	2,630
R^2	0.001	0.001	0.001	0.010	0.010	0.010

Note: *p<0.1; **p<0.05; ***p<0.01

Controls include household size, age and years of schooling of the respondent

Table 4: Regresion Resuls: Effects on Protestant affiliation

	(1)	(2)	(3)	(4)	(5)	(6)
Transfer now	0.053*	0.051*	0.026	0.057*	0.056*	0.028
	(0.029)	(0.029)	(0.039)	(0.029)	(0.029)	(0.039)
Transfer before	−0.010	−0.010	−0.010	−0.008	−0.008	−0.008
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Score	0.005	0.005	−0.008	0.006	0.006	−0.008
	(0.006)	(0.006)	(0.014)	(0.006)	(0.006)	(0.014)
Score^2		−0.001	−0.001		−0.001	−0.001
		(0.001)	(0.001)		(0.001)	(0.001)
Score^3			0.001			0.001
			(0.001)			(0.001)
Hh size				0.004	0.004	0.004
				(0.004)	(0.004)	(0.004)
Age				0.001	0.001	0.001
				(0.001)	(0.001)	(0.001)
Education				−0.001	−0.001	−0.001
				(0.002)	(0.002)	(0.002)
Constant	0.148***	0.153***	0.167***	0.107**	0.113**	0.129**
	(0.019)	(0.021)	(0.025)	(0.050)	(0.051)	(0.053)
Observations	2,645	2,645	2,645	2,630	2,630	2,630
R^2	0.002	0.002	0.003	0.003	0.003	0.004
Note:						*p<0.1; **p<0.05; ***p<0.01
	Controls include household size, age and years of schooling of the respondent					

```
                                " expenditures + ciudad + parroqui"))

mod_long <- lm(fmla_all_contr, data = df[is.na(df$expenditures)==FALSE,])
mod_step <- stepAIC(mod_long, direction = "backward")
```

We conducted stepwise selection using `MAAS::stepAIC` function. As the dependent variable we check monthly church attendance. The long model included current and previous cash transfers, 3rd degree polynomial of the scoring variable and controls on age, education, household size, expenditures, city and area. `Backward` stepwise selection was used. The final model based on AIC criteria includes cash transfer age, education and the third monomial of the scoring variable. The results of stepwise selection are surprising because as seen in Table 1, education and the third monomial are not significant in regression of model 6. We might be overfitting the data because of the significance of the third degree monomial of the scoring variable.

```
mod_step$anova

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
attendpermonth ~ moremoneynew + moremoneyold + score + I(score^2) +
    I(score^3) + householdsize + ageresponder + schooling_resp +
    expenditures + ciudad + parroqui

Final Model:
attendpermonth ~ moremoneynew + I(score^3) + ageresponder + schooling_resp


             Step Df     Deviance Resid. Df Resid. Dev      AIC
1                                      2584   95063.92 9522.710
2        - ciudad  0 2.910383e-11      2584   95063.92 9522.710
3      - parroqui 35 1.507734e+03      2619   96571.65 9494.079
4 - householdsize  1 1.477077e-01      2620   96571.80 9492.083
5    - I(score^2)  1 4.091488e-01      2621   96572.21 9490.094
6         - score  1 2.162418e+00      2622   96574.37 9488.153
```

```
7  - moremoneyold  1 3.190124e+01      2623   96606.27 9487.022
8  - expenditures  1 5.419493e+01      2624   96660.47 9486.496
```

Exercise 3: Breaks in controlling covariates

For the visual inspection for breaks on covariates we use the same approach as in Exercise 1. Using `standard evaluation` of `ggplot2` we create a function for generating the desired graph. This function is looped through the list of desired covariates. The final result is presented through `gridExtra`. A visual inspection of the covariates household size, age, education and previous cash transfers does not hint at any disconuity at the threshold of the SELBEN II score.

```
break_inspection <- function(y_input, y_label_input){
  graph_out <- ggplot(df,aes_string(x="score", y=y_input))+
    geom_point(stat = "summary_bin",fun.y = mean, binwidth = 0.2) +
    geom_vline(xintercept = 0) +
    geom_smooth(data = df_pos_sc,method = "lm",se=FALSE) +
    geom_smooth(data = df_neg_sc,method = "lm",se=FALSE) +
    labs(y = y_label_input, x = "score")+
    theme(axis.title = element_text(size=8), axis.text = element_text(size=6))
  return(graph_out)
}

check_list <- c("householdsize","ageresponder","schooling_resp","moremoneyold")
y_label_list <- c("HH size", "Age", "Education", "Former recipient")

g2 <- list()

for (i in 1:length(check_list)){
  g2[[check_list[i]]] <- break_inspection(check_list[i],
                                          y_label_list[i])
}

grid.arrange(grobs = g2, ncol = 2)
```

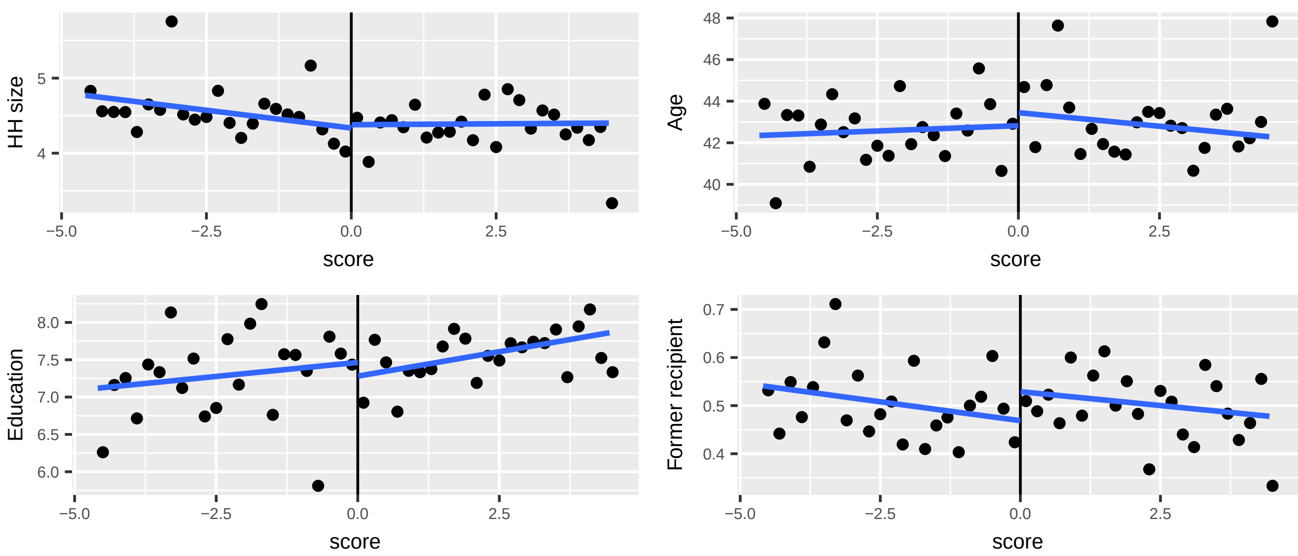

Figure 2: Disconuity visual inspection of covariates

The hypothesis that thera are no breaks in the covariates is supported by the regression results in Table 4.

```
overall_cov_model <- as.formula(paste0("householdsize | ageresponder | ",
                                       "moremoneyold | schooling_resp ~  ",
                                        "moremoneynew +  score + I(score^2) + I(score^3)"))
```

```r
overall_cov_model <- Formula(overall_cov_model)

list_cov_models <- list()

for (i in 1:4){
      temp_cov_formula <- as.formula(terms(overall_cov_model, lhs = i))

      list_cov_models[[i]] <- lm(temp_cov_formula,data=df)
}


stargazer(list_cov_models,
          title = "Regresion Resuls: Effects on Other Covariates",
          column.sep.width = "4pt", no.space = TRUE,
          omit.stat = c("f","ser","adj.rsq"),
          covariate.labels = c("Cash Transfer","Score","Score^2", "Score^3"),
          dep.var.labels = c("Hh size","Age","Education","Previous Transfer"),
          dep.var.caption = "")
```

Table 5: Regresion Resuls: Effects on Other Covariates

	Hh size	Age	Education	Previous Transfer
	(1)	(2)	(3)	(4)
Cash Transfer	−0.025	−1.447	−0.061	0.033
	(0.203)	(1.137)	(0.051)	(0.380)
Score	−0.030	−0.507	−0.014	0.032
	(0.072)	(0.407)	(0.018)	(0.136)
Score^2	0.007	−0.029	0.001	0.002
	(0.006)	(0.036)	(0.002)	(0.012)
Score^3	−0.001	0.024	0.00003	0.004
	(0.004)	(0.021)	(0.001)	(0.007)
Constant	4.434***	43.662***	0.531***	7.406***
	(0.123)	(0.691)	(0.031)	(0.231)
Observations	2,645	2,645	2,645	2,630
R^2	0.003	0.001	0.001	0.003
Note:				*p<0.1; **p<0.05; ***p<0.01