POLS0012 Causal Analysis: Tutorial Exercise 8

For this tutorial, we will analyze Erik Meyersson's 2014 paper, "Islamic Rule and the Empowerment of the Poor and Pious." He asks whether political control by Islamic religious political parties leads to a decrease in women's rights, particularly female education rates. This is difficult to answer using a standard regression, due to selection bias: places that are already favor more traditional gender roles may also be more likely to elect Islamic parties. He looks at Turkey in 1994, where an Islamic party gained political control in many municipalities, and a number of the election results were very close. He uses a regression discontinuity analysis of Islamic control on the rate of secondary school completion by girls, focusing on the Local Average Treatment Effect in areas where Islamic parties barely won or lost their elections.

You'll need to install the `rdd` package for this tutorial. The dataset is called "islamic.csv" and contains the following variables:

- *iwm94*: margin of Islamic party win or loss in 1994, pp: 0 indicates an exact tie

- *hischshr1520f*: secondary school completion rates for ages 15-20 females

- *lpop1994*: log of the locality population in 1994

- *sexr*: sex ratio in locality

- *lareapre*: log of locality area

a) Create a treatment variable, `islamicwin`, that indicates whether or not the Islamic party won the 1994 election.
   Code Hint: A margin of greater than zero means the Islamic party won

   ```
   m$islamicwin<-ifelse(m$iwm94>=0,1,0)
   ```

b) Calculate the difference in means in secondary school completion rates for women between regions where Islamic parties won and lost in 1994. Do you think this is a credible estimate of the causal effect of Islamic party control? Why or why not?
   Code hint: use the option `na.rm=T` so that R knows to ignore missing data

   Code:

   ```
   mean(m$hischshr1520f[m$islamicwin==1],na.rm=T) -
   mean(m$hischshr1520f[m$islamicwin==0],na.rm=T)
   ```

Answer:
The answer is -0.0258. A naive interpretation of this would be that Islamic party control leads to lower female education rates. But this is not a very credible estimate of the causal effect, due to selection bias. The treated and untreated regions are likely to have very different potential outcomes; they are imbalanced. For example, it is likely that places that strongly support Islamic parties are less supportive of female education to begin with.

c) Now we'll start regression discontinuity analysis. First, select optimal bandwidths for testing female high school completion rates using the Imbens-Kalyanaram procedure. Explain what the bandwidth means in this case
Code Hints: Use the `IKbandwidth` function from the `rdd` package. Read the help file to see what the function requires

Code:

```
band <- IKbandwidth(m$iwm94, m$hischshr1520f)
```

Answer:
This gives a bandwidth of 0.239. It means the optimal bias-variance tradeoff is achieved in this RD analysis by only including regions where the vote share margin for the Islamic Party was in the interval $[-0.239, 0.239]$

d) Use your answer to (c) to create a new dataset containing only data within the optimal bandwidth

Code:

```
rdd_est <- m[m$iwm94<band & m$iwm94>-band,]
```

e) Find the Local Average Treatment Effect of Islamic party control on women's secondary school education at the threshold, using the dataset you created in (d) and a simple linear regression that includes the treatment and running variable. How credible do you think this result is?

Code:

```
summary(lm(hischshr1520f~iwm94 + islamicwin, data=rdd_est))
```

Answer:
This gives a LATE estimate of 0.033 with a p value of 0.003, indicating that the effect is strongly statistically significant. This result may not be very credible because it requires that the relationship between female education and the running variable is linear and has the same slope on either side of the cutoff. This is a very strong assumption.

f) Use RD estimation to find the Local Average Treatment Effect of Islamic party control on men's and women's secondary school education at the threshold, using local linear regression estimated with the `RDestimate` function from the `rdd` package. Does the estimate differ from part (e)?

Code Hints: Your code should be of the form `RDestimate(outcome~running.variable,`bw = bandwidth, data=)

Code:

```
rdest <- RDestimate(hischshr1520f~iwm94,cutpoint=0, bw=bandwomen, data=m)
summary(rdest)
```

Answer:
This gives a LATE estimate of 0.0296 with a p value of 0.016, indicating that the effect is clearly significant at the 5% level. The estimate is only slightly lower than in (e)

g) Plot the relationship between the running variable and outcome using local linear regressions. Use your plot to explain why your results do or do not differ strongly between (e) and (f)
Code Hints: You can just use the `plot()` function directly on the object you created in (f). Use the `range` argument to control the x axis

Code:

```
plot(rdest, range=c(-0.6,0.6))
 abline(v=0)
```

Answer:
The plot indicates that the relationship between Y and $\tilde{X}$ does seem to be quite linear on either side of the cutoff. That explains why (f) and (g) gave similar results: linearity is a reasonable fit to the data here. That will not always be the case, though.

h) Perform placebo tests to check that the relationship between the running variable and outcome is not fundamentally discontinuous, by estimating RD estimates at placebo cutoffs of -0.1, -0.05, 0.05 and 0.1. What do you conclude?
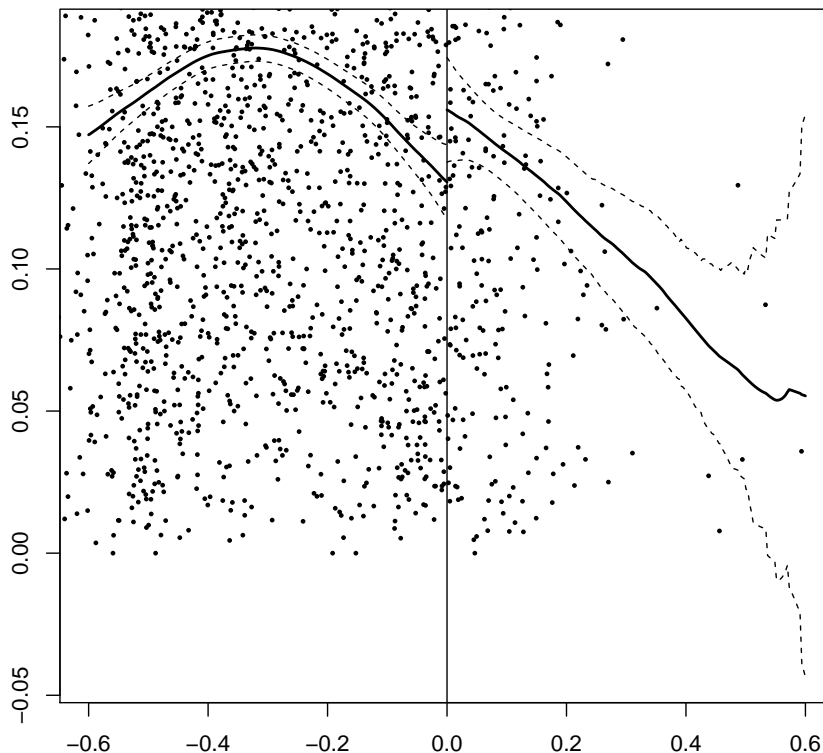
Code:

```
summary(RDestimate(hischshr1520f~iwm94,bw=band,cutpoint=-0.1,data=m))
summary(RDestimate(hischshr1520f~iwm94,bw=band,cutpoint=-0.05,data=m))
summary(RDestimate(hischshr1520f~iwm94,bw=band,cutpoint= 0.05,data=m))
summary(RDestimate(hischshr1520f~iwm94,bw=band,cutpoint= 0.1,data=m))
```

Answer:
At -0.05, 0.05 and 0.1 there is no evidence of a discontinuous relationship (the p-values are

Figure 1: **RD Plot for Question (g)**

i) Perform a robustness check for local randomisation at the threshold by estimating RD esti-
mates in the same way as (g) for the three background covariates *sexr*, *lop1994* and *lareapre*.
What do you conclude?

```
summary(RDestimate(sexr~iwm94,bw=band, data=m))
summary(RDestimate(lpop1994~iwm94,bw=band, data=m))
summary(RDestimate(lareapre~iwm94,bw=band, data=m))
```

j) Perform a McCrary test: another way to check for sorting at the thesthold. Plot and interpret the results.

Code Hints: Use the `DCdensity` function in the `rdd` package with the option `verbose=TRUE`

Code:

```
DCdensity(islamic$iwm94, verbose =TRUE)
```

Answer:

The p-value for the resultant test statistic is 0.51, meaning that we cannot reject the null hypothesis of no discontinuous jump in the density of the running variable. Visually, you can also see that there is no strong evidence for this in the plot.

k) [**Challenging Question**] Examine the sensitivity of the main RD result to the choice of bandwidth by calculating and plotting RD estimates and their associated 95% confidence intervals for a range of bandwidths from 0.05 to 0.6. To what extent do the results depend on the choice of bandwidth?

Hints: Begin by creating a vector of thresholds such as `thresholds <- seq(from=0.05,to=0.6, by=0.005)`. Then use a for loop. You can extract the estimate and standard error from an RD estimate named `rdest` with the code `rdest$est[1]` and `rdest$se[1]`

Code:

```
 rdests=rdci.up=rdci.down <- c()
  thresholds <- seq(from=0.05,to=0.6,by=0.005)

  for(i in 1:length(thresholds)){
    rdest <- RDestimate(hischshr1520f~iwm94,bw=thresholds[i], data=m)
    rdests[i] <- rdest$est[1]
    rdci.up[i] <- rdests[i] + 1.96*rdest$se[1]
    rdci.down[i] <- rdests[i] -1.96*rdest$se[1]
  }

  pdf("thresholds.pdf")
  plot(rdests,
       type="l",
       lwd=2,
       ylim=c(-0.1,0.1),
       xaxt="n",
       xlab="Threshold",
```

```
    ylab="Estimate")
axis(1,
     at=c(1,31,51,71,91,111),
     labels=c(0.05,0.2,0.3,0.4,0.5,0.6))
abline(h=0)
lines(rdci.up,
      lty=3)
lines(rdci.down,
      lty=3)
legend("topright",
       c("RD Estimate","95\% Confidence Interval"),
       lty=c(1,3))
dev.off()
```

Answer:

The plot shows a fairly typical pattern in RD analysis. At very low bandwidths, very little data is used in estimation, making the variance very high. Unsurprisingly therefore, the confidence interval is very wide at low bandwidths: recall that the optimal IK bandwidth is chosen with the twin goals in mind of reducing bias *and* variance. As the threshold (and therefore the effective sample size) increases, the confidence interval decreases.

The size of the RD estimate does not change much as the bandwidth increases, which is reassuring: we would not want our result to be extremely sensitive to the bandwidth choice. Its statistical significance does change somewhat as the threshold changes. It is significant at the 95% level for bandwidths between approximately 0.15 and 0.55, becoming statistically indistinguishable from 0 again at higher bandwidths. Again, this is reasurring as the main result holds across a wide range of bandwidths, and it would not be very defensible in the first place to do RD estimation with a bandwidth as high as 0.55, where we are using data that is very far indeed from the cutoff. Overall, if you saw this plot in a published paper you should feel reasurred that the author's results are not driven primarily by the choice of threshold.

Figure 2: **Sensitivity of RD Result to the Choice of Bandwidth**