Lab Exercise: Discontinuities

This week we conduct the same simulation as usual, but this time using a regression discontinuity to estimate the effect.

1. First, let's generate an 'income' variable for 20,000 people. The data should be drawn randomly from the normal distribution with mean 500 and standard deviation 50.

```
set.seed(54321)
N <- 20000
d <- tibble(income=rnorm(N,500,50))
```

2. Now let's simulate potential outcomes (let's say the outcome is 'attitude to redistribution') for each person that depends on their income. Assume:

$$y_0 = N(10, 2) + \frac{income}{100} + (\frac{income - 600}{50})^2$$

$$y_1 = y_0 + 2$$

So there is a constant treatment effect of 2.

```
d <- d %>% mutate(y_0=rnorm(N,10,2) + income/100 + ((income-600)/50)^2,
                  y_1=y_0+2)
```

3. Actual treatment assignment is not random but deterministic: Imagine it is a poverty-relief program that you receive only if your income is less than 500. Generate this treatment variable.

```
d <- d %>% mutate(D=ifelse(income<500,1,0))
```

4. Now calculate the observed outcome based on the potential outcomes and actual treatment status.

```
d <- d %>% mutate(y_obs=case_when(D==0~y_0,
                                  D==1~y_1))
```

5. Now let's calculate the 'naive' Average Treatment Effect by running a simple OLS regression of the observed outcomes on treatment. What is your estimate of the average treatment effect? How does this compare to the treatment effect we specified earlier?

```
d %>% lm(y_obs ~ D, data=.) %>% stargazer(single.row=T, header=F, title="Q5")
```

6. Our first attempt at a regression discontinuity analysis will be a simple 'non-parametric' difference-in-means comparison either side of the cutoff. Using the data between 480 and 520 on the income scale, perform a difference-in-means test for the efffect of treatment on the outcome.

```
d %>% filter(income<520 & income>480) %>%
  t.test(y_obs~D, data=.) %>%
  tidy() %>% kable()
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	a
-3.396588	18.35812	21.7547	-65.11889	0	6230.504	-3.498839	-3.294337	Welch Two Sample t-test	t

Table 1: Q5

	Dependent variable:
	y_obs
D	7.611*** (0.046)
Constant	17.227*** (0.033)
Observations	20,000
R^2	0.574
Adjusted R^2	0.574
Residual Std. Error	3.280 (df = 19998)
F Statistic	26,916.760*** (df = 1; 19998)
Note:	*p<0.1; **p<0.05; ***p<0.01

7. Next, apply the 'full bandwidth' regression discontinuity method. This just means adding to your regression in Q5 a linear control for the running variable, which in this case is income. Interpret the results. How do they compare to the answer in Q5, Q6 and the treatment effect we specified?

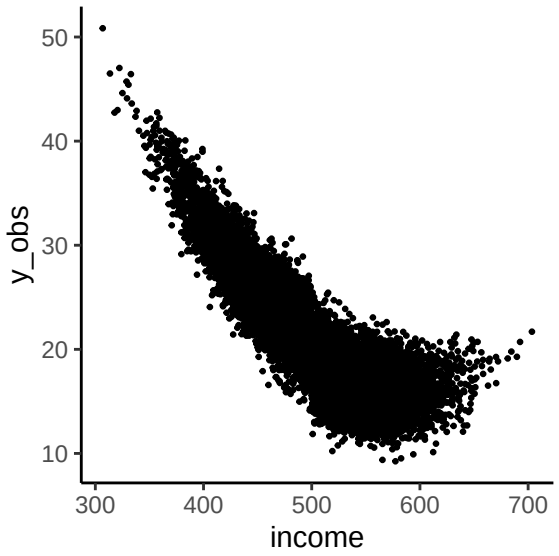
```
d %>% lm(y_obs ~ D + income, data=.) %>% stargazer(single.row=T, header=F, title="Q7")
```

Table 3: Q7

	Dependent variable:
	y_obs
D	1.892*** (0.058)
income	−0.072*** (0.001)
Constant	55.829*** (0.314)
Observations	20,000
R^2	0.758
Adjusted R^2	0.758
Residual Std. Error	2.474 (df = 19997)
F Statistic	31,242.320*** (df = 2; 19997)
Note:	*p<0.1; **p<0.05; ***p<0.01

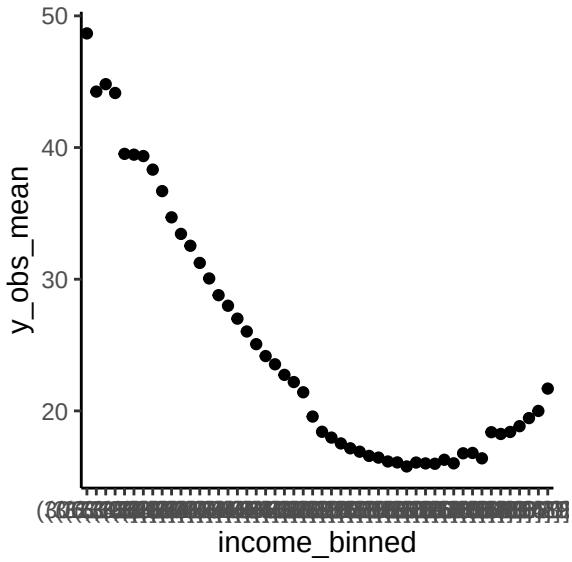
8. Let's try and make a regression discontinuity plot manually to understand what's going on better. First, plot all the data, with the running variable on the x-axis and the observable outcome variable on the y-axis. What can you see in the graph?

```
d %>% ggplot() +
  geom_point(aes(x=income, y=y_obs), size=0.5) +
  theme_classic()
```



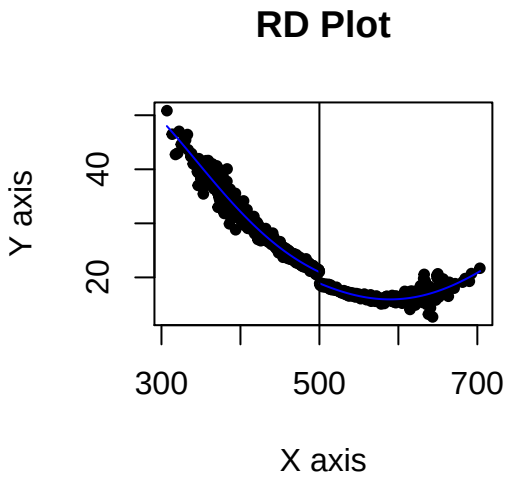
9. The graph in Q8 is difficult to interpret so most regression discontinuity plots 'bin' the data into groups to more easily see the pattern. Bin the income data into 50 groups (try `cut()` in R), then calculate the average observed outcome in each bin and plot the two against each other.

```
d %>% mutate(income_binned=cut(income,50)) %>%
  group_by(income_binned) %>%
  summarize(y_obs_mean=mean(y_obs,na.rm=T)) %>%
  ggplot() +
  geom_point(aes(x=income_binned,y=y_obs_mean))+
  theme_classic()
```



10. An easier way to make a nice regression discontinuity plot is to use the `rdplot` command in the `rdrobust` package. Create this plot and be sure to specify the appropriate cutoff value based on the treatment assignment mechanism described above.

```
library(rdrobust)
rdplot(d$y_obs,d$income, c=500)
```



11. The regression discontinuity plot makes it clear that approximating the data with a straight line is likely to create a bias. Recall that in our original specification of the potential outcomes we made them depend on $income^2$. Do we get a better estimate of the treatment effect if we include a quadratic term in our regression discontinuity? Compare the results to your answer in Q7.

```
d %>% lm(y_obs ~ D + income + I(income^2), data=.) %>%
  stargazer(single.row=T, header=F, title="Q11")
```

Table 4: Q11

	Dependent variable:
	y_obs
D	1.948*** (0.048)
income	−0.476*** (0.004)
I(income^2)	0.0004*** (0.00000)
Constant	155.665*** (1.037)
Observations	20,000
R^2	0.838
Adjusted R^2	0.838
Residual Std. Error	2.024 (df = 19996)
F Statistic	34,395.080*** (df = 3; 19996)
Note:	*p<0.1; **p<0.05; ***p<0.01

12. A third method of performing a regression discontinuity is to perform the 'limited bandwidth' regression approach only on a narrow 'bandwidth' of data close to the cutoff. Subset the data again to between 480 and 520 on the income scale and apply the regression discontinuity you used in Q11. How do the results compare to your full-bandwidth regression discontinuity in Q7 and the difference-in-means estimate in Q6?

```
d %>% filter(income<520 & income>480) %>%
  lm(y_obs ~ D + income + I(income^2), data=.) %>%
  stargazer(single.row=T, header=F, title="Q12")
```

12. One way of easily picking an 'optimal' bandwidth (instead of assuming the range 480 to 520) is to use

Table 5: Q12

	Dependent variable:
	y_obs
D	2.075*** (0.101)
income	−0.788*** (0.216)
I(incomeˆ2)	0.001*** (0.0002)
Constant	232.638*** (54.005)
Observations	6,236
R^2	0.427
Adjusted R^2	0.427
Residual Std. Error	2.021 (df = 6232)
F Statistic	1,549.035*** (df = 3; 6232)
Note:	*p<0.1; **p<0.05; ***p<0.01

the automatic process in the `rdrobust` command of the `rdrobust` package. This method also provides more accurate standard errors. Apply the method, defining the cutoff appropriately and a quadratic running variable, and interpret the results. (Note `rdrobust` assumes treatment is above the threshold, not below, so the sign of your result might be the opposite of what it should be).

```
rdrobust(d$y_obs,d$income, c=500, p=2) %>% summary()
```

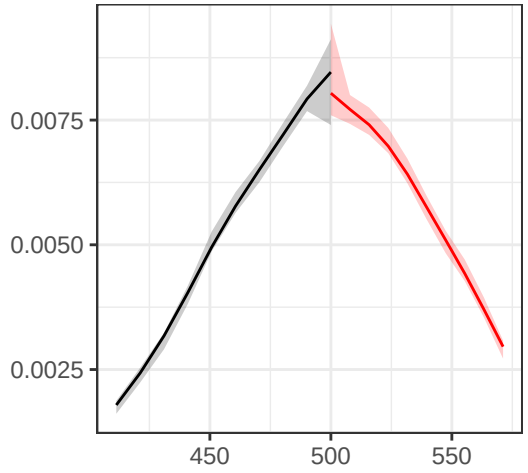
```
## Call: rdrobust
##
## Number of Obs.                 20000
## BW type                        mserd
## Kernel                     Triangular
## VCE method                         NN
##
## Number of Obs.                 10051        9949
## Eff. Number of Obs.             6899        6802
## Order est. (p)                     2           2
## Order bias  (p)                    3           3
## BW est. (h)                   50.583      50.583
## BW bias (b)                   70.044      70.044
## rho (h/b)                      0.722       0.722
##
## ===============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ===============================================================================
##   Conventional    -2.097     0.104   -20.167     0.000    [-2.301 , -1.893]
##         Robust         -         -   -17.841     0.000    [-2.329 , -1.868]
## ===============================================================================
```

13. Just to check our assumptions: We know there is no sorting around the cutoff in our model because we specified the treatment to be precisely based on the income cutoff and to not allow for any self-selection. But anyway let's run the standard test for sorting - the McCrary density test using the `rddensity` package. Also make a nice graph with the `rdplotdensity` command.

```
library(rddensity)
density_test <- rddensity(d$income, c=500)
summary(density_test)
```

```
##
## RD Manipulation Test using local polynomial density estimation.
##
## Number of obs =       20000
## Model =               unrestricted
## Kernel =              triangular
## BW method =           comb
## VCE method =          jackknife
##
## Cutoff c = 500        Left of c           Right of c
## Number of obs         10051               9949
## Eff. Number of obs    4435                3638
## Order est. (p)        2                   2
## Order bias (q)        3                   3
## BW est. (h)           29.558              23.724
##
## Method                T                   P > |T|
## Robust                0.3917              0.6953
```

```
rdplotdensity(density_test, d$income)$Estplot[[1]]
```




```
## list()
## attr(,"class")
## [1] "waiver"
```

15. Now let's change how our potential outcomes are defined. Our y_0 stays the same, but this time let's abandon our constant treatment effect and assume the treatment effect itself varies depending on the level of income:

$$y_0 = N(10, 2) + \frac{income}{100} + (\frac{income - 600}{50})^2$$

For those with income near the threshold, between 490 and 510, the treatment effect is:

$$y_1 = y_0 + 10$$

For everyone else, the treatment effect is actually negative:

$$y_1 = y_0 - 3$$

```
d <- d %>% mutate(y_0=rnorm(N,10,2) + income/100 + ((income-600)/50)^2,
                  y_1=case_when(income<510 & income>490~y_0+10,
                                TRUE~y_0-3))
```

16. Calculate the observed outcomes again and run the optimal-bandwidth regression discontinuity analysis using `rdrobust`. How do you interpret the results? For which people is the treatment effect estimated here?

```
d <- d %>% mutate(y_obs=case_when(D==0~y_0,
                                  D==1~y_1))

rdrobust(d$y_obs, d$income, c=500, p=2) %>% summary()
```

```
## Call: rdrobust
##
## Number of Obs.                20000
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.               10051        9949
## Eff. Number of Obs.           1057        1067
## Order est. (p)                   2           2
## Order bias  (p)                  3           3
## BW est. (h)                  6.487       6.487
## BW bias (b)                 21.079      21.079
## rho (h/b)                    0.308       0.308
##
## ================================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ================================================================================
##   Conventional   -10.237     0.274   -37.404     0.000   [-10.773 , -9.701]
##         Robust         -         -   -36.715     0.000   [-10.657 , -9.577]
## ================================================================================
```