PS630 Lab: RDD and Matching

Haohan Chen

November 30, 2018

Disclaimer: I'm using lots of online resources for this tutorial...

Regression Discontinuity

The paper:

Carpenter, Christopher, and Carlos Dobkin. "The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age." American Economic Journal: Applied Economics 1, no. 1 (2009): 164-82.

Topic: The effect of alcohol consumption on mortality

Software Setup

```r
#---------------
# Package
#---------------
library(rdd)
```

```
## Loading required package: sandwich

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: AER

## Loading required package: car

## Loading required package: carData

## Loading required package: survival

## Loading required package: Formula
```

```r
#---------------
# Package
#---------------
library(readstata13)
AEJfigs = read.dta13("AEJfigs.dta")

#---------------
# Data cleaning
#---------------
```

```r
# All = all deaths
AEJfigs$age = AEJfigs$agecell - 21
AEJfigs$over21 = ifelse(AEJfigs$agecell >= 21,1,0)
```

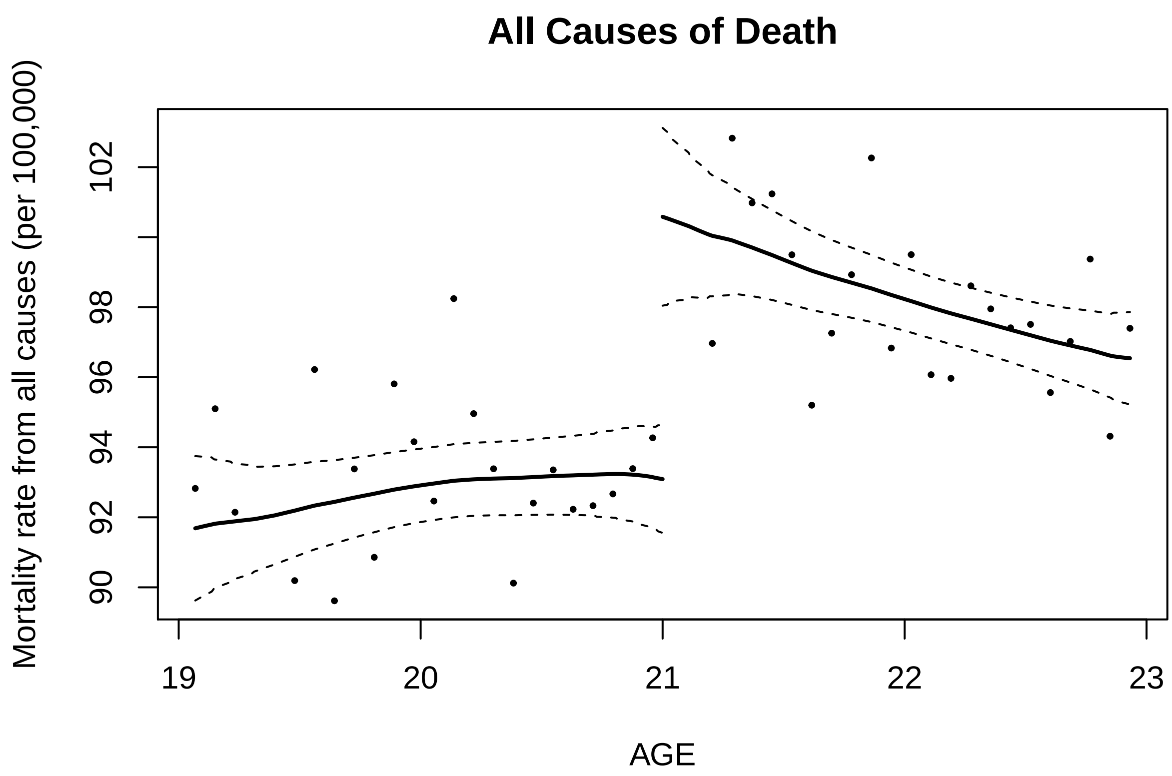
RD Model: Age cutpoint and Overall Mortality

Model Fit

```r
#---------------
# Fit RD model
#---------------
reg.1=RDestimate(all~agecell,data=AEJfigs,cutpoint = 21)
```

Check Assumptions

```r
plot(reg.1)
title(main="All Causes of Death", xlab="AGE",ylab="Mortality rate from all causes (per 100,000)")
```



Results

```r
summary(reg.1)
```

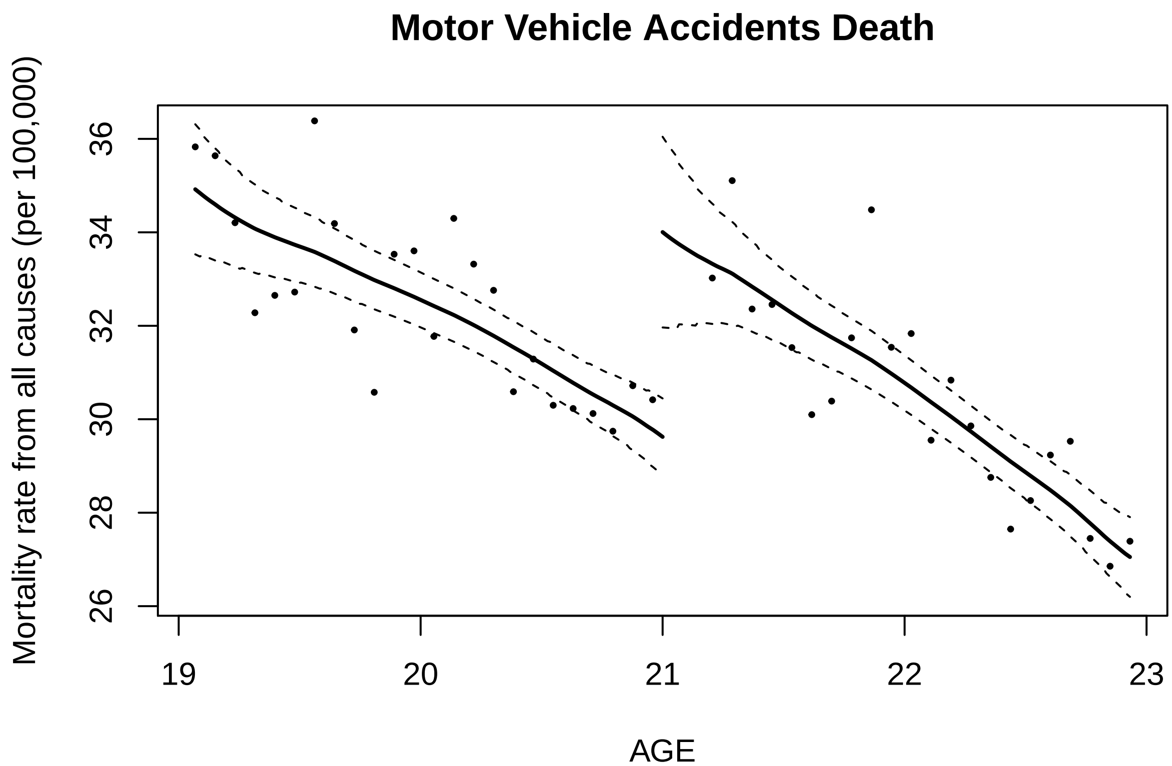
```
##
## Call:
## RDestimate(formula = all ~ agecell, data = AEJfigs, cutpoint = 21)
```

```
##
## Type:
## sharp
##
## Estimates:
##             Bandwidth  Observations  Estimate  Std. Error  z value
## LATE        1.6561     40            9.001     1.480       6.080
## Half-BW     0.8281     20            9.579     1.914       5.004
## Double-BW   3.3123     48            7.953     1.278       6.223
##             Pr(>|z|)
## LATE        1.199e-09  ***
## Half-BW     5.609e-07  ***
## Double-BW   4.882e-10  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## F-statistics:
##             F      Num. DoF  Denom. DoF  p
## LATE        33.08  3         36          3.799e-10
## Half-BW     29.05  3         16          2.078e-06
## Double-BW   32.54  3         44          6.129e-11
```

Other DVs: Death

Motor Vehicle Accident

```r
reg.2=RDestimate(mva~agecell,data=AEJfigs,cutpoint = 21)
plot(reg.2)
title(main="Motor Vehicle Accidents Death", xlab="AGE",ylab="Mortality rate from all causes (per 100,000
```




```r
summary(reg.2)
```

```
##
## Call:
## RDestimate(formula = mva ~ agecell, data = AEJfigs, cutpoint = 21)
##
## Type:
## sharp
##
## Estimates:
##            Bandwidth  Observations  Estimate  Std. Error  z value
## LATE       1.2109     30            4.977     1.0590      4.700
## Half-BW    0.6054     14            4.956     1.3767      3.600
## Double-BW  2.4218     48            4.566     0.7086      6.444
##            Pr(>|z|)
## LATE       2.607e-06  ***
## Half-BW    3.182e-04  ***
## Double-BW  1.162e-10  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## F-statistics:
##            F      Num. DoF  Denom. DoF  p
## LATE       13.32  3         26          3.692e-05
## Half-BW    12.76  3         10          1.879e-03
## Double-BW  26.99  3         44          9.322e-10
```

Internal Cause of Death

```
reg.3=RDestimate(internal~agecell,data=AEJfigs,cutpoint = 21)
plot(reg.3)
title(main="Internal Causes of Death", xlab="AGE",ylab="Mortality rate from all causes (per 100,000)")
```


```
summary(reg.3)
```

```
##
## Call:
## RDestimate(formula = internal ~ agecell, data = AEJfigs, cutpoint = 21)
##
## Type:
## sharp
##
## Estimates:
##            Bandwidth  Observations  Estimate  Std. Error  z value
## LATE       0.8809     22            1.4128    0.8206      1.722
## Half-BW    0.4405     10            1.8691    1.0203      1.832
## Double-BW  1.7618     42            0.7652    0.6179      1.239
##            Pr(>|z|)
## LATE       0.08513   .
## Half-BW    0.06698   .
## Double-BW  0.21553
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## F-statistics:
##            F       Num. DoF  Denom. DoF  p
## LATE        6.830  3         18          5.734e-03
## Half-BW     1.765  3          6          5.068e-01
## Double-BW  22.695  3         38          2.750e-08
```

Further reading: https://rpubs.com/cuborican/RDD

Matching

Example:

"Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programs," Journal of the American Statistical Association, Vol. 94, No. 448 (December 1999), pp. 1053-1062.

Topic: The effect on trainee earnings of an employment program

Software Setup

```
#---------------
# load packages
#---------------
library(MatchIt)
# Political scientists package
# https://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf

# Alternative packages: Matching, designmatch
# https://cran.r-project.org/web/packages/Matching/Matching.pdf
# https://cran.r-project.org/web/packages/designmatch/designmatch.pdf

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 1.2.

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.0     v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------ tidyverse_conflicts(
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
```

```
#---------------
# load dataset
#---------------
data("lalonde")
names(lalonde)
```

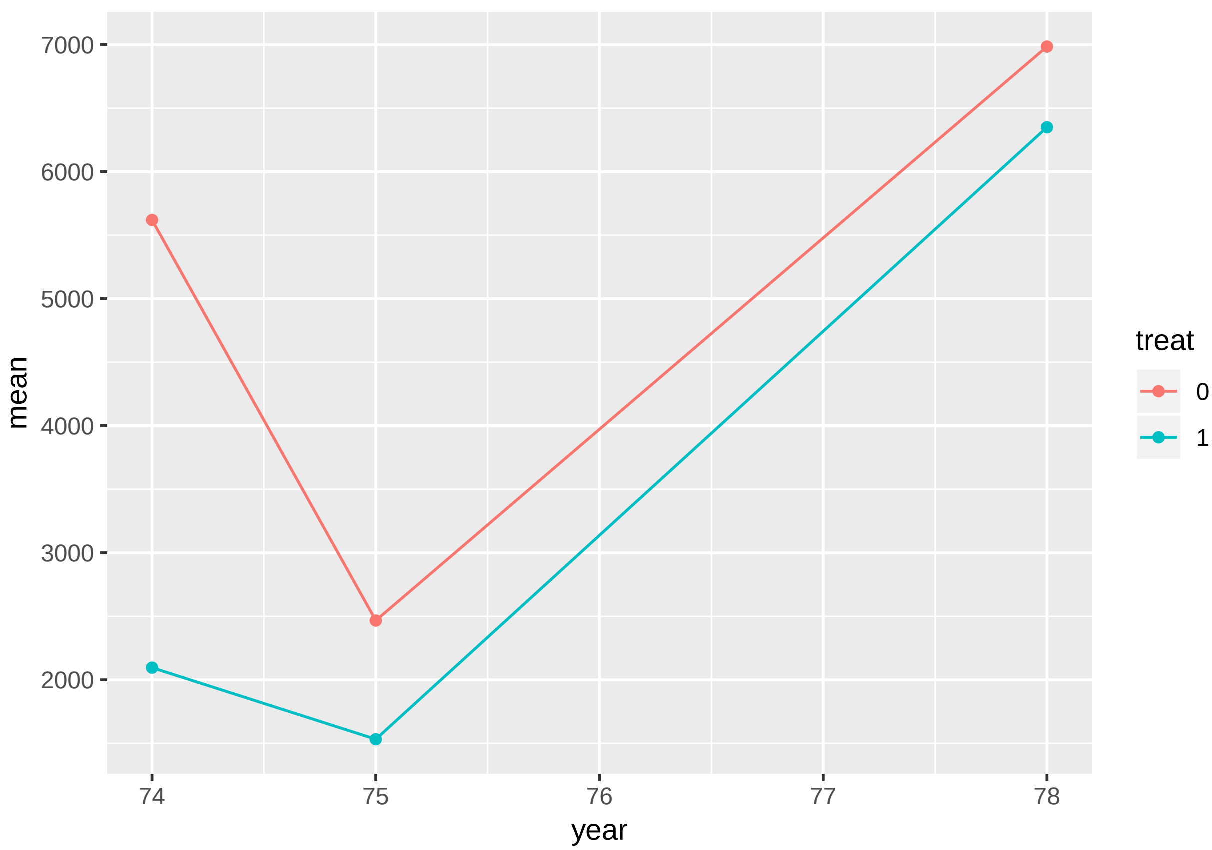
```
##  [1] "treat"    "age"      "educ"     "black"    "hispan"   "married"
##  [7] "nodegree" "re74"     "re75"     "re78"
```

Difference in Mean

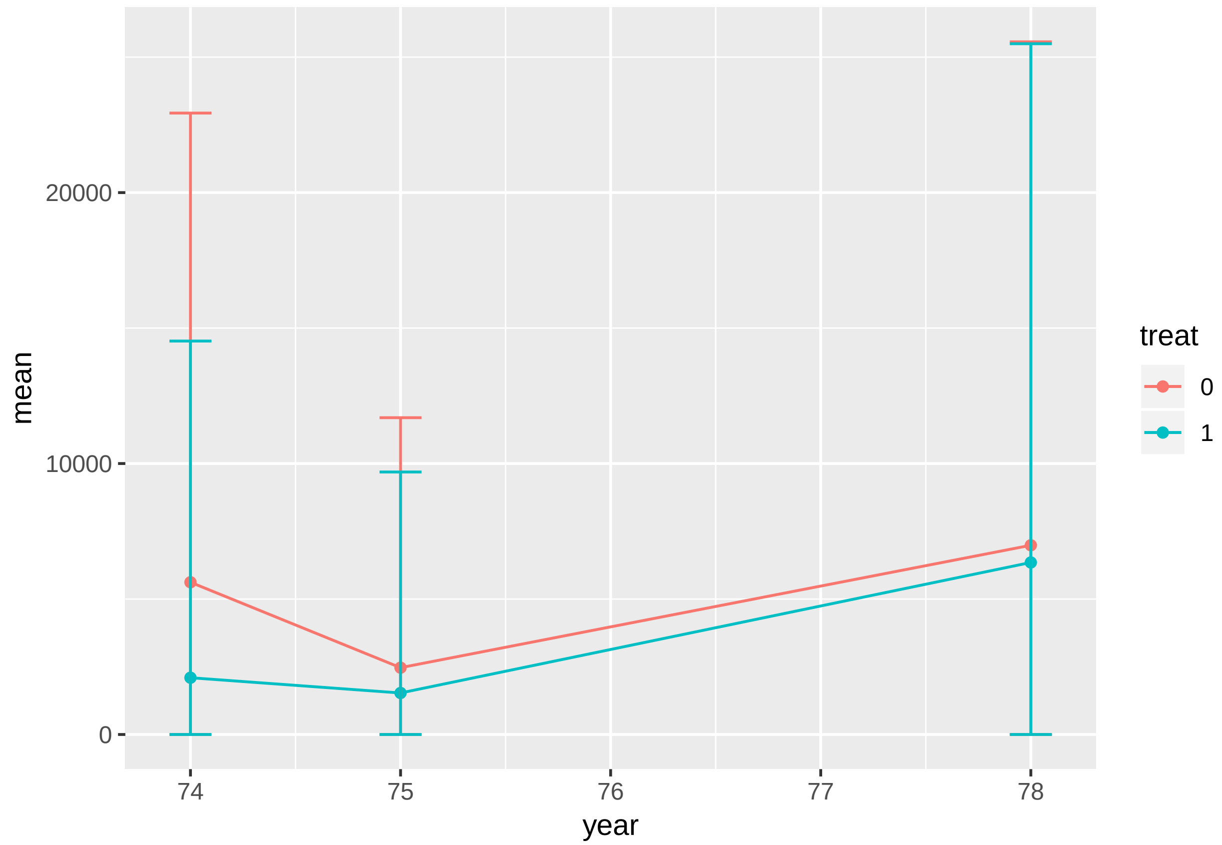
```r
# Table
lalonde %>%
  group_by(treat) %>%
  select(treat, re74, re75, re78) %>%
  summarise_all(funs(mean, sd))
```

```
## # A tibble: 2 x 7
##   treat re74_mean re75_mean re78_mean re74_sd re75_sd re78_sd
##   <int>     <dbl>     <dbl>     <dbl>   <dbl>   <dbl>   <dbl>
## 1     0     5619.     2466.     6984.   6789.   3292.   7294.
## 2     1     2096.     1532.     6349.   4887.   3219.   7867.
```

```r
# Plot (without error)
lalonde %>%
  group_by(treat) %>%
  select(treat, re74, re75, re78) %>%
  summarise_all(funs(mean, lo = quantile(., 0.025), hi = quantile(., 0.975))) %>%
  gather(key, value, -treat) %>%
  mutate(year = as.integer(substr(key, 3, 4)),
         stat = substr(key, 6, 10),
         treat = as.factor(treat)) %>%
  select(-key) %>%
  spread(stat, value) %>%
  ggplot(aes(x = year, y = mean, color = treat)) + geom_point() + geom_line()
```



```r
# Plot (with variance)
lalonde %>%
  group_by(treat) %>%
  select(treat, re74, re75, re78) %>%
  summarise_all(funs(mean, lo = quantile(., 0.025), hi = quantile(., 0.975))) %>%
  gather(key, value, -treat) %>%
  mutate(year = as.integer(substr(key, 3, 4)),
         stat = substr(key, 6, 10),
         treat = as.factor(treat)) %>%
  select(-key) %>%
  spread(stat, value) %>%
  ggplot(aes(x = year, y = mean, color = treat)) + geom_point() + geom_line() +
  geom_errorbar(aes(ymin = lo, ymax = hi), width=0.2)
```



Two-sample t-test

```r
# Earning Growth
#----------------
t.test(re78 - re75 ~ treat, data = lalonde)
```

```
##
##  Welch Two Sample t-test
##
## data:  re78 - re75 by treat
## t = -0.43138, df = 299.89, p-value = 0.6665
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1665.248  1066.442
## sample estimates:
## mean in group 0 mean in group 1
##        4517.685        4817.088
```

```r
# Remember what we did?
```

Checking Balance in Covariates

```r
# Get a sense of the imbalance
lalonde %>%
  group_by(treat) %>%
  select(age, educ, black, hispan, married, nodegree) %>%
  summarise_all(funs(mean)) %>%
  gather(key, value, -treat) %>%
  spread(treat, value) %>%
  setNames(c("Covariate", "Control", "Treated")) %>%
  mutate(`T - C` = Treated - Control)
```

```
## Adding missing grouping variables: `treat`
```

```
## # A tibble: 6 x 4
##   Covariate Control Treated `T - C`
##   <chr>       <dbl>   <dbl>   <dbl>
## 1 age        28.0    25.8    -2.21
## 2 black       0.203   0.843   0.640
## 3 educ       10.2    10.3     0.111
## 4 hispan      0.142   0.0595 -0.0827
## 5 married     0.513   0.189  -0.324
## 6 nodegree    0.597   0.708   0.111
```

```r
# A statistical summary of their difference
vars = c("age", "educ", "black", "hispan", "married", "nodegree")
ttests = apply(lalonde[, vars], 2, function(x) t.test(x ~ lalonde$treat))

# T statistics
sapply(ttests, function(x) x$statistic["t"])
```

```
##      age.t     educ.t    black.t   hispan.t  married.t nodegree.t
##   2.991074  -0.546756 -19.344264   3.409136   8.596140  -2.712695
```

```r
# p value
sapply(ttests, function(x) x$p.value)
```

```
##          age         educ        black       hispan      married
## 2.914274e-03 5.847977e-01 1.205890e-58 7.042071e-04 1.461278e-16
##     nodegree
## 6.982225e-03
```

Propensity Score Estimation

Fit PS Model

```r
m_ps = glm(treat ~ age + educ + black + hispan + married + nodegree,
           family = binomial(link="logit"), data = lalonde)
summary(m_ps)
```

```
##
## Call:
## glm(formula = treat ~ age + educ + black + hispan + married +
##     nodegree, family = binomial(link = "logit"), data = lalonde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7709  -0.4606  -0.2963   0.7766   2.6384
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.67874    1.02120  -4.582 4.61e-06 ***
## age          0.01030    0.01329   0.775  0.43857
## educ         0.15161    0.06568   2.308  0.02098 *
## black        3.12657    0.28514  10.965  < 2e-16 ***
## hispan       0.99947    0.42191   2.369  0.01784 *
## married     -0.92969    0.27128  -3.427  0.00061 ***
## nodegree     0.78719    0.33507   2.349  0.01881 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.49  on 613  degrees of freedom
## Residual deviance: 494.70  on 607  degrees of freedom
## AIC: 508.7
##
## Number of Fisher Scoring iterations: 5
```

Predict PS Score

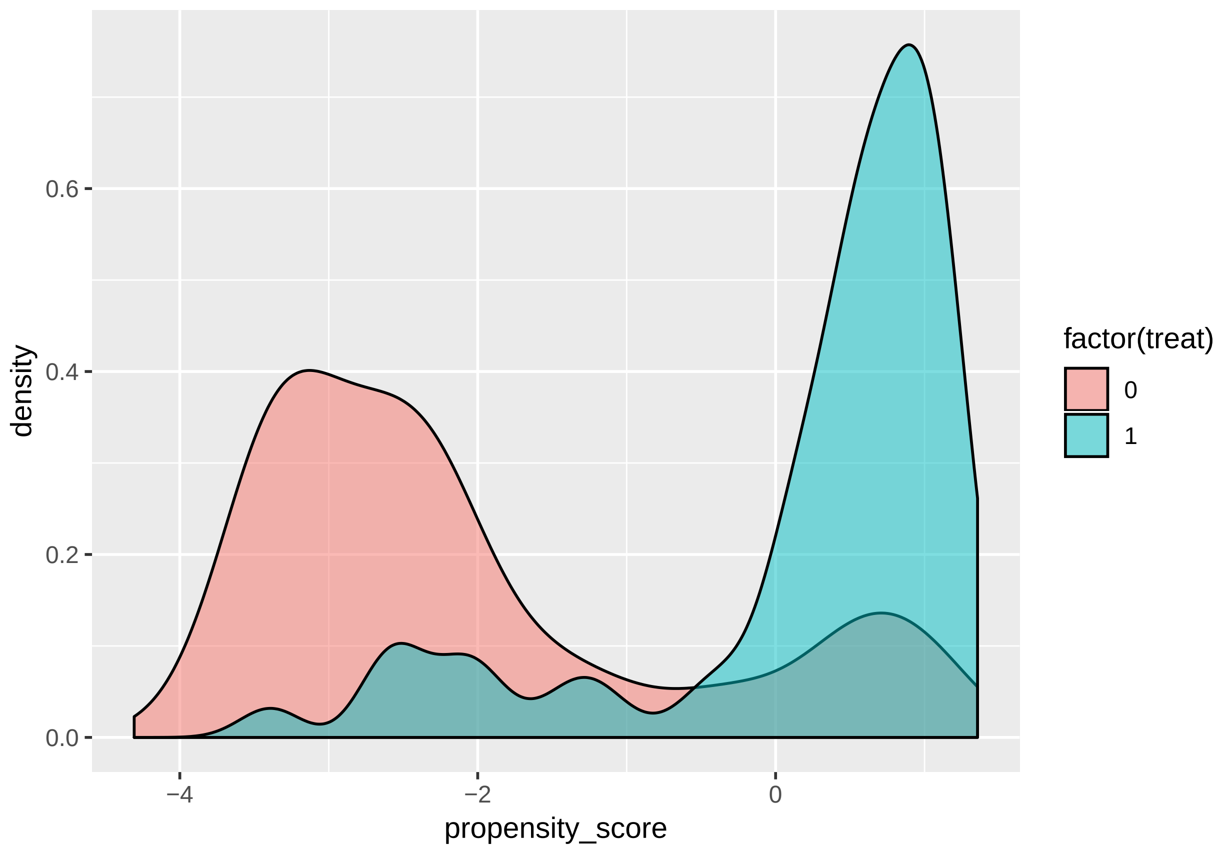
```r
?predict
```

```
## starting httpd help server ... done
```

```r
ps = predict(m_ps)
lalonde_ps = lalonde %>% mutate(propensity_score = ps)
```

Check overlap

```r
lalonde_ps %>%
  ggplot(aes(x = propensity_score, fill = factor(treat))) +
  geom_density(alpha = 0.5)
```



Matching with the propensity score

Some more complication. Luckily, we have an one-stop solution!

One-Stop Solution

```
m_matchit =
  matchit(treat ~ age + educ + black + hispan + nodegree + married + re74 + re75,
        data = lalonde, method = "nearest", distance = "logit")

summary(m_matchit)
```

```
##
## Call:
## matchit(formula = treat ~ age + educ + black + hispan + nodegree +
##     married + re74 + re75, data = lalonde, method = "nearest",
##     distance = "logit")
##
## Summary of balance for all data:
##            Means Treated Means Control SD Control  Mean Diff   eQQ Med
## distance          0.5774        0.1822     0.2295     0.3952    0.5176
## age              25.8162       28.0303    10.7867    -2.2141    1.0000
## educ             10.3459       10.2354     2.8552     0.1105    1.0000
## black             0.8432        0.2028     0.4026     0.6404    1.0000
## hispan            0.0595        0.1422     0.3497    -0.0827    0.0000
## nodegree          0.7081        0.5967     0.4911     0.1114    0.0000
```

```
## married         0.1892        0.5128     0.5004    -0.3236    0.0000
## re74         2095.5737     5619.2365  6788.7508 -3523.6628 2425.5720
## re75         1532.0553     2466.4844  3291.9962  -934.4291  981.0968
##           eQQ Mean   eQQ Max
## distance    0.3955    0.5966
## age         3.2649   10.0000
## educ        0.7027    4.0000
## black       0.6432    1.0000
## hispan      0.0811    1.0000
## nodegree    0.1135    1.0000
## married     0.3243    1.0000
## re74     3620.9240 9216.5000
## re75     1060.6582 6795.0100
##
##
## Summary of balance for matched data:
##          Means Treated Means Control SD Control Mean Diff  eQQ Med
## distance        0.5774        0.3629     0.2533    0.2145   0.1646
## age            25.8162       25.3027    10.5864    0.5135   3.0000
## educ           10.3459       10.6054     2.6582   -0.2595   0.0000
## black           0.8432        0.4703     0.5005    0.3730   0.0000
## hispan          0.0595        0.2162     0.4128   -0.1568   0.0000
## nodegree        0.7081        0.6378     0.4819    0.0703   0.0000
## married         0.1892        0.2108     0.4090   -0.0216   0.0000
## re74         2095.5737     2342.1076  4238.9757 -246.5339 131.2709
## re75         1532.0553     1614.7451  2632.3533  -82.6898 152.1774
##          eQQ Mean    eQQ Max
## distance   0.2146     0.4492
## age        3.3892     9.0000
## educ       0.4541     3.0000
## black      0.3730     1.0000
## hispan     0.1568     1.0000
## nodegree   0.0703     1.0000
## married    0.0216     1.0000
## re74     545.1182 13121.7500
## re75     349.5371 11365.7100
##
## Percent Balance Improvement:
##          Mean Diff.   eQQ Med eQQ Mean  eQQ Max
## distance    45.7140   68.1921  45.7536  24.7011
## age         76.8070 -200.0000  -3.8079  10.0000
## educ      -134.7737  100.0000  35.3846  25.0000
## black       41.7636  100.0000  42.0168   0.0000
## hispan     -89.4761    0.0000 -93.3333   0.0000
## nodegree    36.9046    0.0000  38.0952   0.0000
## married     93.3191    0.0000  93.3333   0.0000
## re74        93.0035   94.5880  84.9453 -42.3724
## re75        91.1508   84.4891  67.0453 -67.2655
##
## Sample sizes:
##           Control Treated
## All           429     185
## Matched       185     185
## Unmatched     244       0
```

```
## Discarded       0       0
```

Further reading: https://sejdemyr.github.io/r-tutorials/statistics/tutorial8.html