Lecture 10 - Regression Discontinuity Tutorial

Samuel DeWitt

Regression Discontinuity Tutorial

To keep the theme the same (or at least similar) we will be looking at data from a 2009 paper on the minimum legal drinking age and deaths in moving vehicle accidents by Carpenter & Dobkin (AEJ:AP, 1(1): 164-182).

The basic question here is this: does the legal drinking age have a causal effect on motor vehicle mortality?

Carpenter & Dobkin (2009) - Background Information

MLDA Redux - Reagan Admin passed federal legislation (1984) requiring all states to adopt minimum legal drinking age of 21.

Analyses right around this time period possibly subject to bias from unobserved characteristics of states dictating whether they changed MLDA before this or were slow to adopt new law.

Carpenter & Dobkin (2009) use data from 1997 to 2005 from the National Health Interview Survey (NHIS) and mortality data from the National Center for Health Statistics (NCHS)

Carpenter & Dobkin (2009) - Background Information

Using these data, the authors use a RDD to estimate a few interesting effects:

1) What effect does the MLDA have on actual drinking behavior?
2) What effect does the MLDA have on various types of alcohol-related mortality?

RDD Tutorial: First Step - Look at the Data!

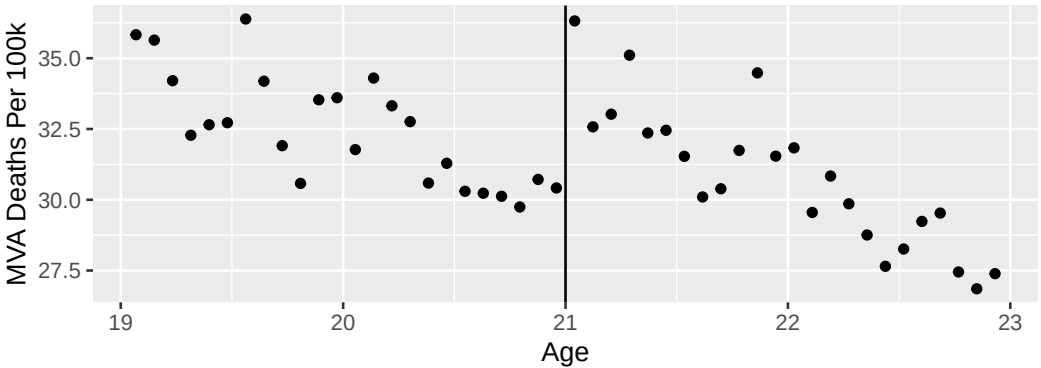
Always a good idea to begin an RDD (and, quite frankly, any research project) with looking at your data.

I will begin with a simple plot of fatalities in moving vehicle accidents by age to see if a discontinuity exists.

Note - The data here are slightly different than those in the paper - they had access to restricted use, individual-level data which I do not. This is an aggregated data set.

RDD Tutorial: First Step - Look at the Data!

```
ggplot(mlda, aes(x=agecell,y=mva))+geom_point()+
  geom_vline(xintercept=21)+
  labs(y="MVA Deaths Per 100k", x="Age")
```



RDD Tutorial: First Step - Look at the Data!

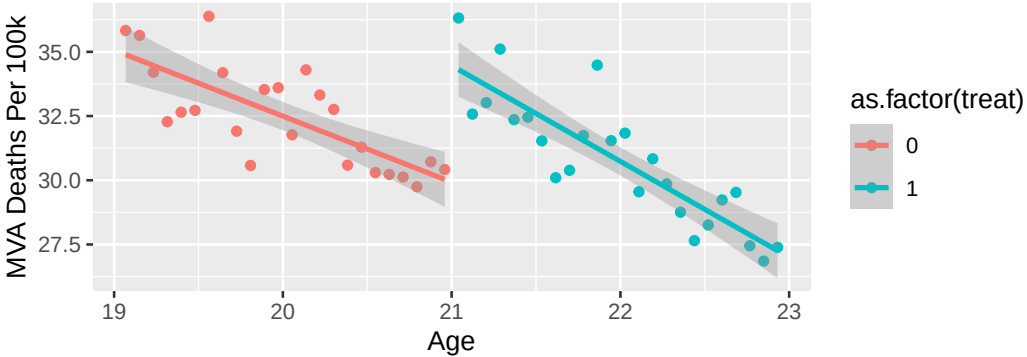
The plot is a fairly simple one - just a scatter plot of age on the x axis and Motor Vehicle Accident Deaths per capita (100k) on the y axis. I add axis labels using the label() function and I add a vertical line at the MLDA (age=21) to highlight where the cutoff of interest is in these data.

From a quick glance, it appears that the rate of motor vehicle deaths jumps upward right after the cutoff and then continues the decline we see on the left half of the figure.

I'll create a new plot on the next slide that shows this a bit better.

Closer Look at Motor Vehicle Fatalities

```
mlda$treat<-with(mlda, ifelse(agecell>=21,1,0))
ggplot(mlda, aes(x=agecell,y=mva, color=as.factor(treat)))+geom_point()+
  geom_smooth(method="lm")+
  labs(y="MVA Deaths Per 100k", x="Age")
```



Closer Look at Motor Vehicle Fatalities

One addition here is that I have created a variable called **treat** that is equivalent to 1 if the observation is 21 years old or older and 0 otherwise. This allows me to use the color=as.factor(treat) option within the aes() code in the ggplot function to color code observations above and below the cutoff drinking age.

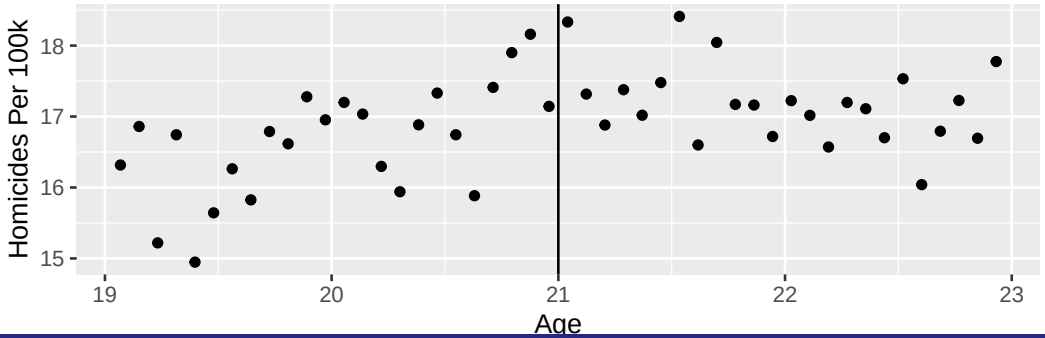
The other addition is that I have added linear best fit lines (with standard errors) to the plot to highlight the trends on each side of the figure.

This allows us to see more clearly just how large the jump is at the discontinuity and what the overall trends are.

RDD Tutorial: First Step - Look at the Data!

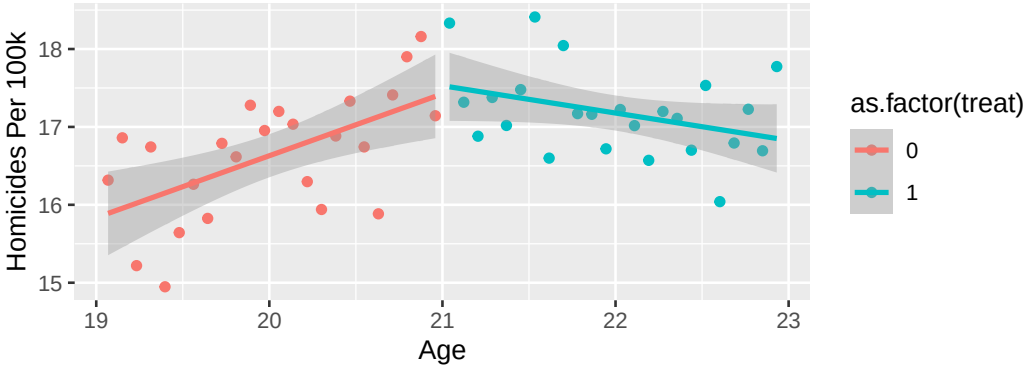
Now, for deaths per capita due to homicide...

```
ggplot(mlda, aes(x=agecell,y=homicide))+geom_point()+
  geom_vline(xintercept=21)+
  labs(y="Homicides Per 100k", x="Age")
```



Closer Look at Deaths by Homicide Per 100k

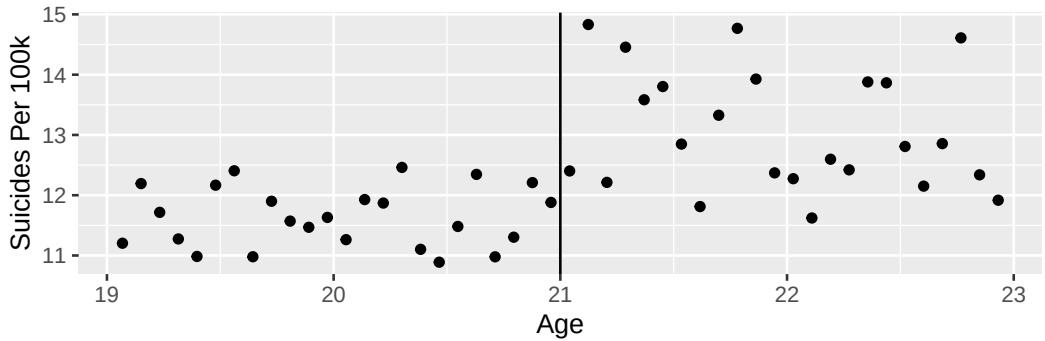
```
ggplot(mlda, aes(x=agecell,y=homicide, color=as.factor(treat)))+geom_point()+
  geom_smooth(method="lm")+
  labs(y="Homicides Per 100k", x="Age")
```



RDD Tutorial: First Step - Look at the Data!

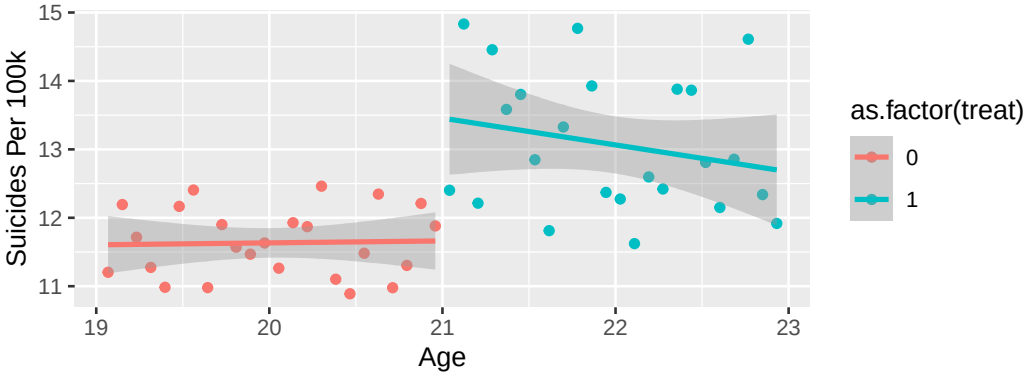
Finally, for deaths per capita due to suicide...

```
ggplot(mlda, aes(x=agecell,y=suicide))+geom_point()+
  geom_vline(xintercept=21)+
  labs(y="Suicides Per 100k", x="Age")
```



Closer Look at Deaths by Suicide

```
ggplot(mlda, aes(x=agecell,y=suicide, color=as.factor(treat)))+geom_point()+
  geom_smooth(method="lm")+
  labs(y="Suicides Per 100k", x="Age")
```



Beginning the RDD Procedure

We will be using a package called “rdd” to estimate the Regression Discontinuity analysis.

Luckily, we can skip the step where we look for manipulation, since age is a variable not prone to manipulation. We will also skip the covariate balance step here for ease of explanation.

Estimating a Basic RDD - Motor Vehicle Deaths Per Capita

```
rdd_mva<-RDestimate(mva~agecell+as.factor(treat), cutpoint=21,
            bw=c(0.5,1,2), data=mlda)
```

Summarizing the RDD

```
summary(rdd_mva)
```

```
##
## Call:
## RDestimate(formula = mva ~ agecell + as.factor(treat), data = mlda,
##     cutpoint = 21, bw = c(0.5, 1, 2))
##
## Type:
## fuzzy
##
## Estimates:
##       Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
## [1,]  0.5        12            4.891     1.4968      3.268    0.0010839891321
## [2,]  1.0        24            5.181     1.1578      4.475    0.0000076379678
## [3,]  2.0        48            4.583     0.7465      6.139    0.0000000008286
##
## [1,]  **
## [2,]  ***
## [3,]  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## F-statistics:
##       F      Num. DoF  Denom. DoF  p
## [1,]  10.45  3          8          0.003841306805
## [2,]  13.48  3         20          0.000048703216
```

Interpreting the Results - Motor Vehicle Deaths

Bandwidths: These values represent how far below and above the cutoff we allow a score on the running variable to be in order to be included in the estimate for the treatment effect.

In this example, I include bandwidths of .5, 1.0, and 2.0. This means that I include people 6-months before/after their 21st birthday in the first estimate, 1 year before/after for the second estimate, and 2 years before/after for the third estimate.

Interpreting the Results - Motor Vehicle Deaths

Observations: This represents the total number of cases included within the bandwidth. This includes cases above *and* below the bandwidth.

Interpreting the Results - Motor Vehicle Deaths

Estimate: This is an estimate for the "treatment" effect of being 21 on the rate of motor vehicle accident deaths per 100,000.

A value of 4.891 suggests that individuals who between 21 to 21.5 years old have a motor vehicle accident death rate that is 4.891 persons per 100,000 higher than the rate for individuals who are between 20.5 and 21 years old.

Interpreting the Results - Motor Vehicle Deaths

Std. Error: This is an estimate for the standard error of the treatment effect estimate. Because our sample is but one of many potential samples, we assume there is sampling error in our estimate of the true treatment effect.

For the treatment effect of 4.891, the standard error for this effect is 1.4968, so treatment effect estimates within 1 standard error of this point estimate would range from 3.3942 to 6.3878.

If you recall from earlier in the semester (the Stats tutorials) a standard error represents uncertainty around a point estimate due to sampling error and is generally used to understand a plausible range of values where the true population effect should lie.

Interpreting the Results - Motor Vehicle Deaths

z value: This is the standardized score for the treatment effect estimate. It is obtained by dividing the estimate value by its standard error:

$$z = \frac{\text{Estimate}}{\text{Standard Error}} = \frac{4.891}{1.4968} = 3.268$$

Interpreting the Results - Motor Vehicle Deaths

The z value represents how far (in standard error units) our estimate of the treatment effect is away from the expected population estimate.

In this example, we assume that cases above/below the threshold have the same (or very similar) rate of motor vehicle deaths, so this value is 0 (i.e., the **expected** treatment effect estimate is 0 because we assume the groups are the same w/respect to motor vehicle deaths).

Interpreting the Results - Motor Vehicle Deaths

Pr(>|z|): This is a probability value (or p-value). It reflects the probability of observing a treatment effect estimate as or more extreme than the one we have observed **if the null hypothesis is true**.

That last part is quite important - the null hypothesis we test against here assumes that the treatment effect will be 0, so the probability value we obtain assumes that to be true and reflects the likelihood we would observe our treatment effect estimate if the true difference in the population is 0.

Interpreting the Results - Motor Vehicle Deaths

A p-value of 0.00108 tells us that the probability of observing a treatment effect estimate of 4.891 or greater in a sample of this size assuming that the true effect is zero is about 1 time in a thousand.

NOTE: This does not automatically mean that we are right about this estimate! A value of 4.891 exists along a distribution of potential group differences below/above this threshold which assumes the average difference to be 0. The p-value merely tells us that observing a value like this is very improbable if our assumption about the true population value (here, 0) is true.

Estimating a Basic RDD - Homicide Deaths Per Capita

Why might we also want to know if there are differences in the homicide rate above/below the age of 21?

Doing so will allow us to better interpret the meaning behind any treatment effect we estimate for motor vehicle deaths. If there's not also an effect for the homicide rate (and the suicide rate, below) this tells us that the increase in mortality among youth just below/above the age of 21 is better explained by the increased availablity of alcohol coupled with operating motor vehicles.

Estimating a Basic RDD - Homicide Deaths Per Capita

```
rdd_homicide<-RDestimate(homicide~agecell+as.factor(treat), cutpoint=21,
            bw=c(0.5,1,2), data=mlda)
```

Summarizing the RDD - Homicide Deaths Per Capita

```
summary(rdd_homicide)
```

```
##
## Call:
## RDestimate(formula = homicide ~ agecell + as.factor(treat), data = mlda,
##     cutpoint = 21, bw = c(0.5, 1, 2))
##
## Type:
## fuzzy
##
## Estimates:
##       Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
## [1,]  0.5        12             0.2909   0.7694       0.3781  0.7053
## [2,]  1.0        24            -0.0657   0.5523      -0.1190  0.9053
## [3,]  2.0        48             0.1409   0.3836       0.3673  0.7134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## F-statistics:
##       F      Num. DoF  Denom. DoF  p
## [1,]  1.020  3          8          0.4332166
## [2,]  2.147  3         20          0.1261490
## [3,]  7.322  3         44          0.0004373
```

Interpreting the Results - Homicide

Although all of the treatment effect estimates are not exactly 0, we cannot reject the null hypothesis that each is any different from 0 (see the p-value column).

Estimating a Basic RDD - Suicide Deaths Per Capita

```
rdd_suicide<-RDestimate(suicide~agecell+as.factor(treat), cutpoint=21,
            bw=c(0.5,1,2), data=mlda)
```

Summarizing the RDD - Suicide Deaths Per Capita

```
summary(rdd_suicide)
```

```
##
## Call:
## RDestimate(formula = suicide ~ agecell + as.factor(treat), data = mlda,
##     cutpoint = 21, bw = c(0.5, 1, 2))
##
## Type:
## fuzzy
##
## Estimates:
##       Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
## [1,]  0.5        12            0.950     1.0079      0.9425   0.345928
## [2,]  1.0        24            1.565     0.8278      1.8906   0.058674  .
## [3,]  2.0        48            1.802     0.5522      3.2632   0.001102  **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## F-statistics:
##       F       Num. DoF  Denom. DoF  p
## [1,]   2.996  3          8          0.0953535191
## [2,]   7.525  3         20          0.0014645368
## [3,]  15.161  3         44          0.0000006465
```

Interpreting the Results - Suicide

There are some significant findings here - namely, that the homicide rate is significantly different above/below the cutoff when using bandwidths of 1.0 (p-value<.10) and 2.0 (p-value<.01).

So, what does this tell us? It tells us that there is a separate impact of alcohol accessibility on the suicide rate, but this is only evident with larger bandwidths. All else equal, the larger the bandwidth, the more likely that our estimate is biased.

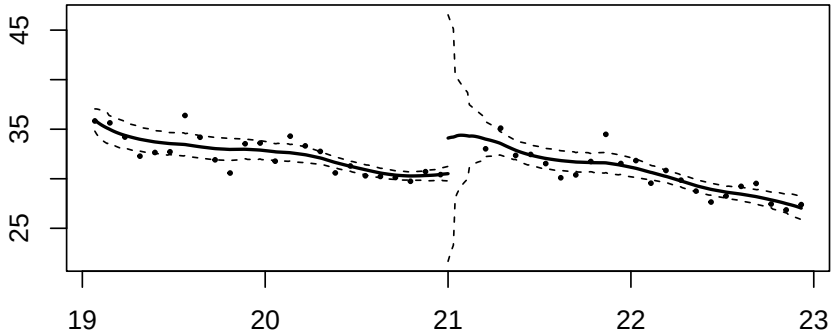
I'd consider these estimates suggestive of an effect of MLDA on the suicide rate, but would be cautious about that because it's only apparent with larger bandwidths.

Plotting the RDD

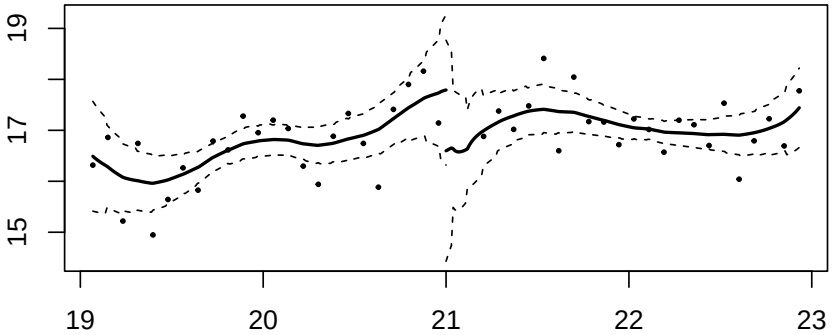
There are simple ways to plot an RDD object, but they result in less appealing plots than using the ggplot() function. I'll show you how to do it anyway.

UPDATE - I have to omit the code from the following slides (formatting nightmare situation) but it's simply: plot(nameofrddobject) - e.g., plot(rdd_mva).

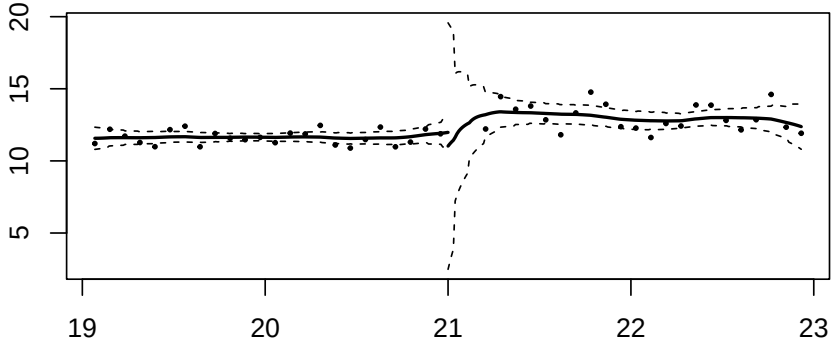
Plotting the RDD - Motor Vehicle Deaths



Plotting the RDD - Homicides



Plotting the RDD - Suicides



Another Example - Access to Head Start

Background: Head Start is an early intervention program targeted toward parents and children. It provides a variety of benefits, including preschool, nutritional, and medical services.

Head Start was established in 1965 and implementing the program was the responsibility of the Office of Economic Opportunity (OEO). In implementing the program, OEO created a discontinuity in grant-writing assistance for Head Start at the county level by targeting the 300 poorest US counties - this allowed Ludwig & Miller (2007) to compare counties just above and just below the cutoff to evaluate the impact of Head Start access on a variety of relevant outcomes.

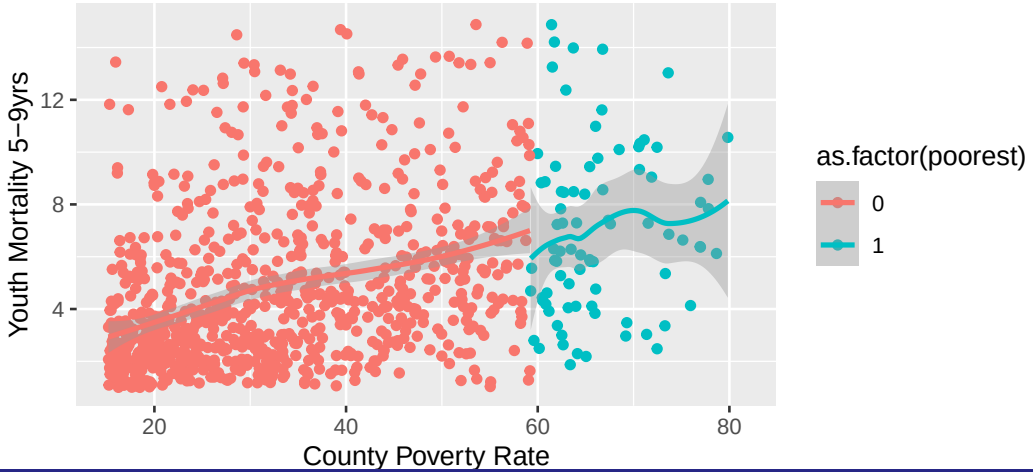
Another Example - Access to Head Start

The following series of slides will guide you through a replication of the primary results from Ludwig & Miller (2007) which indicate that Head Start had multiple positive impacts for the health and educational attainment of youth in those counties. I load the data in the r setup code chunk at the beginning of this lecture using the haven package. I was only able to locate a Stata data file so I could not use the reular load() function. The data file is named *headstart* and I will make it available for you to download.

Head Start - Plotting the Data

First, we want to see what the data look like around the cutoff for some relevant outcomes. What follows will be a few plots including a vertical line for the poverty rate cutoff (how they defined the 300 poorest counties) with linear trends on both sides of the cutoff for treated and untreated counties.

Head Start - Plotting the Data



Head Start - Plotting the Data

From the looks of that plot, it is apparent that there might be a small drop just after the cutoff for mortality among those who were 5 to 9 years old in communities that received Head Start grant application assistance, at least as compared to those counties that did not and fell right under the 300 poorest counties cutoff (where the poverty rate in 1960 was 59.1984 or higher).

However, it looks like the confidence intervals overlap a bit around the cutoff - we want to run a formal model to see if there is a real effect there.

Head Start - Estimating the RDD model

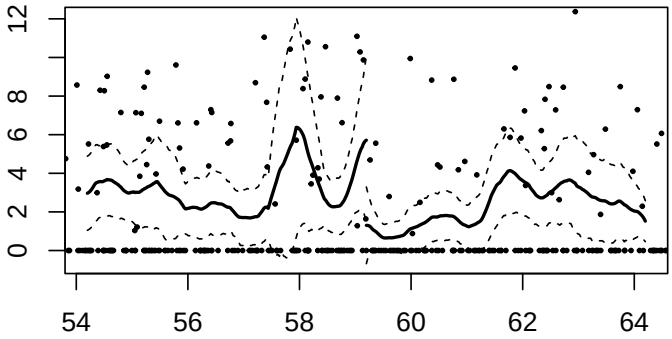
```
##
## Call:
## RDestimate(formula = mort_age59_related_postHS ~ povrate60 +
##     as.factor(poorest), data = headstart, cutpoint = 59.1984,
##     bw = c(0.5, 1, 2))
##
## Type:
## fuzzy
##
## Estimates:
##       Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
## [1,]  0.5         34           -4.414    2.880       -1.532   0.12540
## [2,]  1.0         64           -3.966    2.136       -1.857   0.06334   .
## [3,]  2.0        124           -2.366    1.661       -1.424   0.15450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## F-statistics:
##       F      Num. DoF  Denom. DoF  p
## [1,]  2.418  3          30         0.08573
## [2,]  3.326  3          60         0.02548
## [3,]  2.092  3         120         0.10481
```

Head Start - Estimating the RDD model

My model is a little bit simpler than that of Ludwig & Miller (2007) but I still find some slight evidence for a discontinuity using a bandwidth of 1.0. My interpretation of the estimate is that, as compared to counties who were 1 point lower on the poverty rate than the cutoff of 59.1984, counties 1 point higher than the cutoff experienced a youth mortality rate that was 3.966 points lower, on average.

In total, there are 64 counties used for this comparison. That effect is statistically significant at $p<.10$ (slightly higher than we typically use - $p<.05$ - but still potentially meaningful).

Head Start - Plotting the RDD model



Headstart - Model Sensitivity Checks: Covariate Balance

Headstart - Model Sensitivity Checks: Placebo Outcomes

The End