Regression Discontinuity

Francis J. DiTraglia

Econ 224

"Sharp" Regression Discontinuity

Suppose we are interested in learning the causal effect of a binary treatment D on an outcome Y. In some special settings, whether or not a person is treated is a solely determined by a special covariate x, called the *running variable*

$$D_i = \begin{cases} 0 & \text{if } x_i < x_0 \\ 1 & \text{if } x_i \geq x_0 \end{cases}$$

The preceding expression says that D is a *deterministic function* of x: everyone who has $x \geq x_0$ is treated, and no one who has $x < x_0$ is treated. This setting is called a *sharp regression discontinuity design* and it provides us with a powerful tool for causal inference. We'll distinguish this from another kind of regression discontinuity setup called a *fuzzy regression discontinuity design* below. I'll use the shorthand RD to refer to regression discontinuity in these notes.

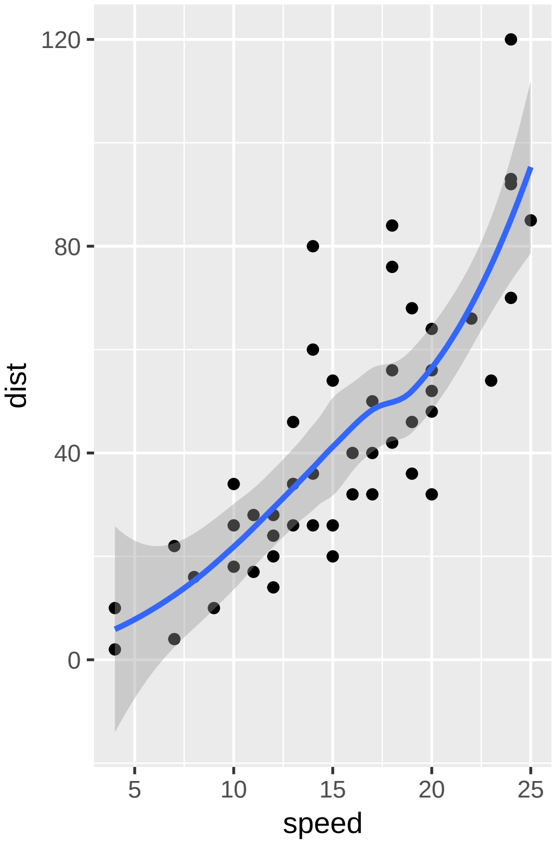
When we previously used regression to carry out causal inference, the idea was to compare two groups of people who had been *matched* using a set of covariates $\mathbf{x}$. One group was treated and the other was not, but both groups had exactly the same values of $\mathbf{x}$. Under the assumption that treatment is "as good as randomly assigned" after conditioning on $\mathbf{x}$, we could learn the causal effect of treatment by comparing the mean outcomes of the two groups.

Sharp RD is very different since there is no way to carry out matching using the running variable x. This is because everyone who has $x < x_0$ is untreated while everyone who has $x \geq x_0$ is treated. Instead of matching people who have the same covariate values, sharp RD *extrapolates* by comparing people with *different* covariate values. The basic idea is very simple: we compare people whose x is close to but slightly *below* the cutoff x_0 to people whose x is close to but slightly *above* the cutoff.

```
library(ggplot2)
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

```
qplot(speed, dist, data=cars) +
    geom_smooth()
```



This is a nice example 2.