Chapter 10 ex 2: Regression Discontinuity

Gertjan Verhoeven

```
library(dagitty)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages ----------------------------------------------------------------------------
```

```
## 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   1.3.1     v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
## -- Conflicts -------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Regression discontinuity analysis:

suppose you are trying to evaluate the effect of a new procedure for coronary bypass surgery that is supposed to help with the postoperative healing process.

The new procedure is risky, however, and is rarely performed in patients who are over 80 years old.

Data from this (hypothetical) example are displayed in Figure 10.10.

(a) Does this seem like an appropriate setting in which to implement a regression discontinuity analysis?

Based on the description given, the answer is NO, because there is no clear assignment mechanism for treatment. Although treatment with the new procedure at age above 80 is rarely performed, it IS performed sometimes. We need to know the decision rule for applying the treatment.

(b) The folder bypass contains data for this example:

- stay is the length of hospital stay after surgery,
- age is the age of the patient,
- and new is the indicator variable indicating that the new surgical procedure was used.

Preoperative disease severity (severity) was unobserved by the researchers, but we have access to it for illustrative purposes. Can you find any evidence using these data that the regression discontinuity design is inappropriate?

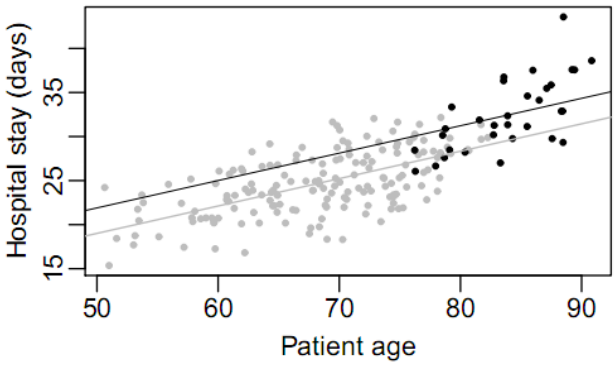


Figure 10.10 *Hypothetical data of length of hospital stay and age of patients, with separate points and regression lines plotted for each treatment condition: the new procedure in gray and the old procedure in black.*

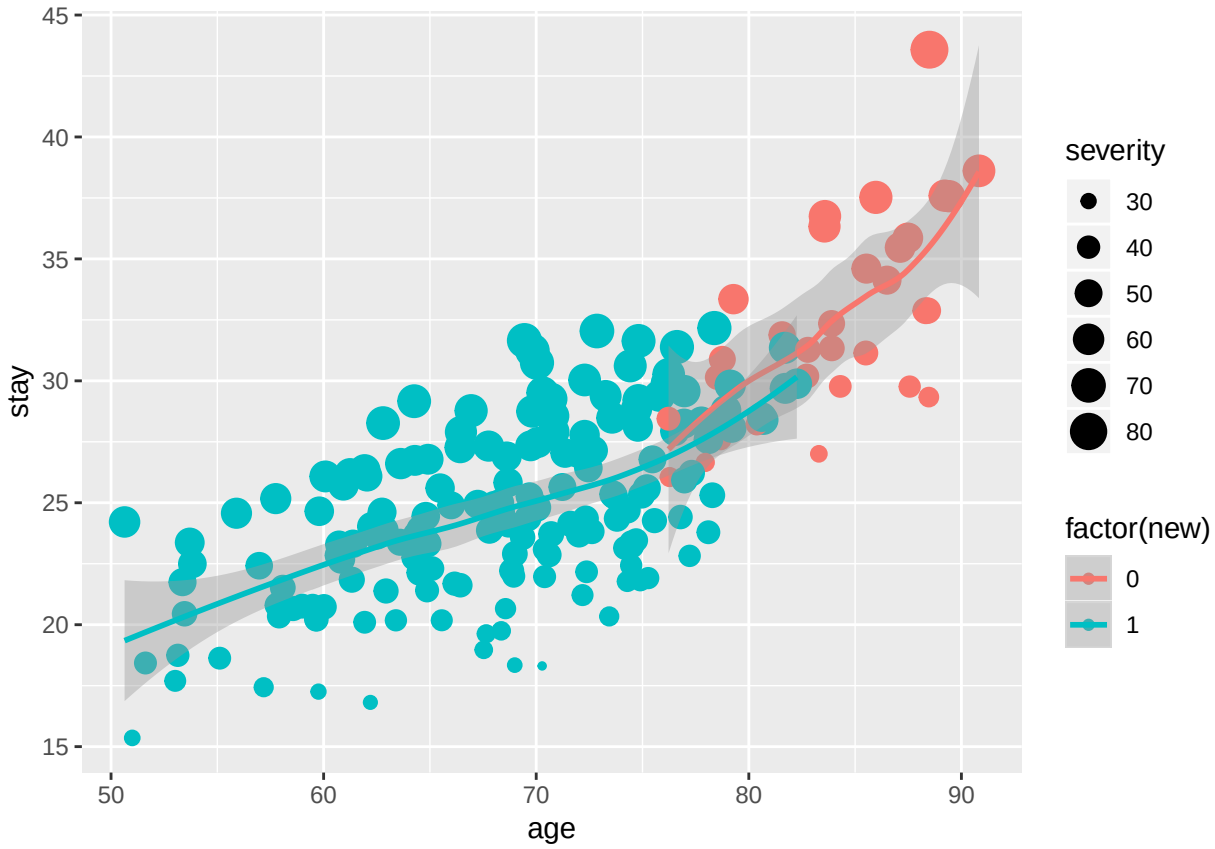
Figure 1:

```
bypass <- read.csv("bypass.data.csv", sep = "")

bypass <- as.tibble(bypass)
```

```
ggplot(bypass, aes(x = age, y = stay, group = factor(new), colour = factor(new))) + geom_point(aes(size
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



It is clear that severity explains stay as well.

```
ggplot(bypass, aes(x = age, y = severity, group = factor(new), colour = factor(new))) + geom_point(aes(s
  geom_smooth()
```

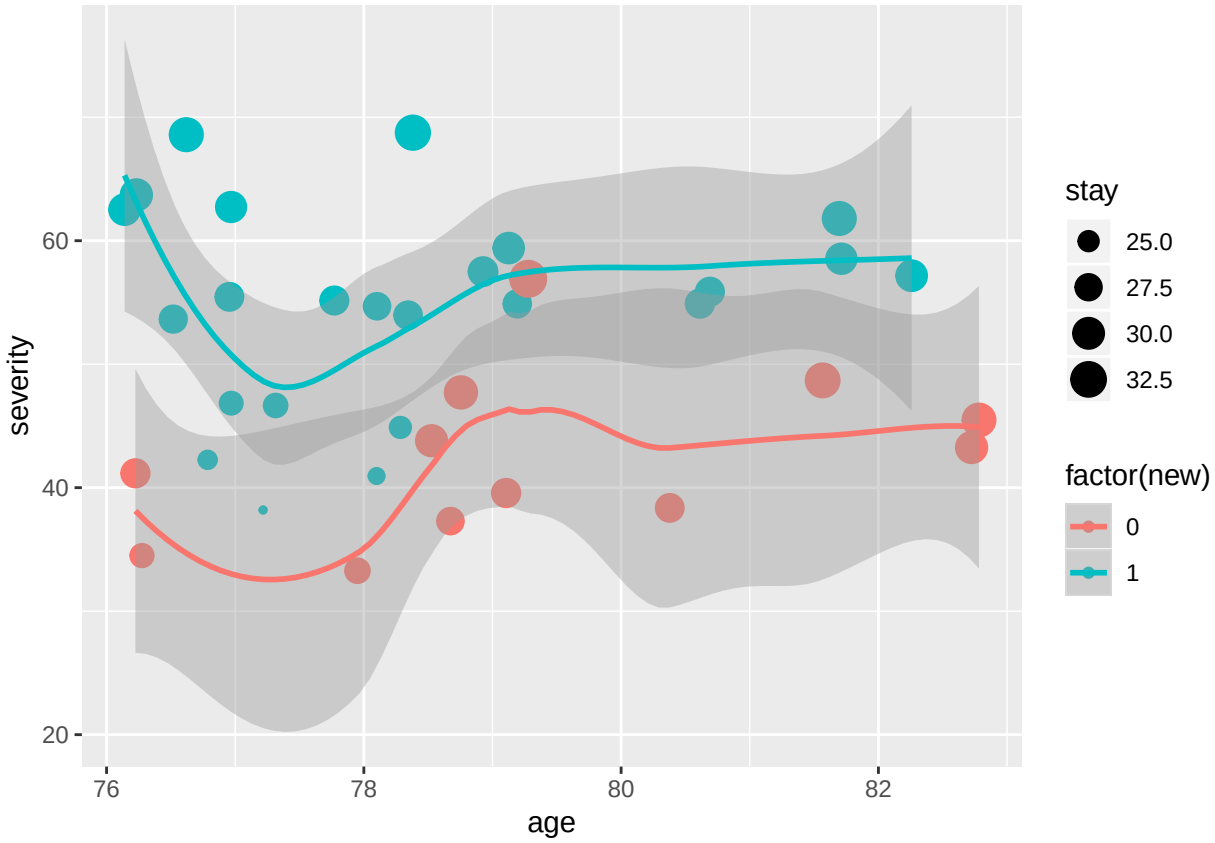
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

This looks a bit as if severity-age relationship is different for treated vs untreated. But maybe not so important.

```
ggplot(bypass %>% filter(age > 76 & age < 83), aes(x = age, y = severity, group = factor(new), colour =
  geom_smooth()
```

```
## Warning: package 'bindrcpp' was built under R version 3.5.2
```

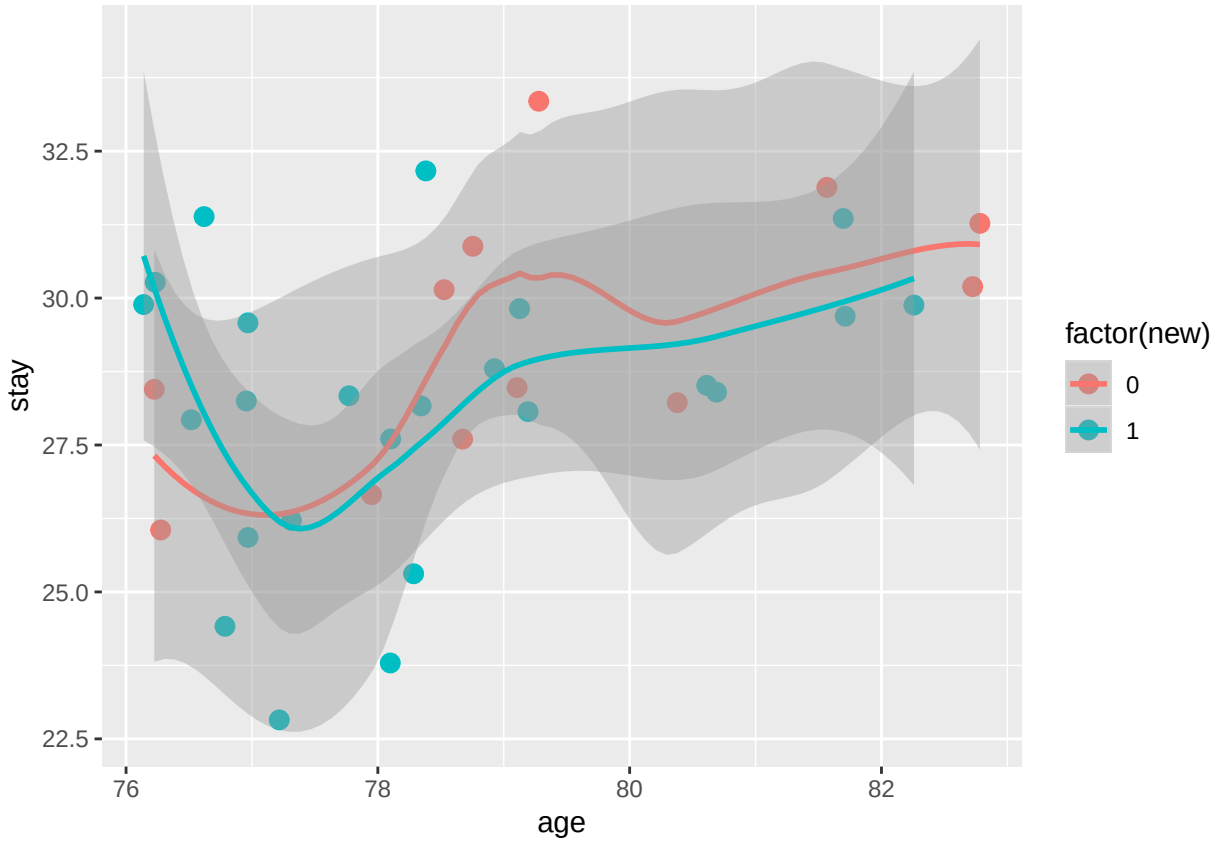
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



(c) Estimate the treatment effect using a regression discontinuity estimate (ig- noring) severity.

```
ggplot(bypass %>% filter(age > 76 & age < 83), aes(x = age, y = stay, group = factor(new), colour = fac
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Here the treatment effect appears roughly zero (minus 1 days, but high std err)

```
lmfit <- lm(stay ~ new, data = bypass %>% filter(age > 76 & age < 83))

summary(lmfit)
```

```
##
## Call:
## lm(formula = stay ~ new, data = bypass %>% filter(age > 76 &
##     age < 83))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3676 -1.3657  0.1771  1.6446  3.9729
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.4316     0.6777  43.431   <2e-16 ***
## new          -1.2408     0.8300  -1.495    0.144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.347 on 34 degrees of freedom
## Multiple R-squared:  0.06168,    Adjusted R-squared:  0.03408
## F-statistic: 2.235 on 1 and 34 DF,  p-value: 0.1441
```

(d) Estimate the treatment effect in any way you like, taking advantage of the information in severity.

```r
lmfit <- lm(stay ~ age + severity + new, data = bypass)

summary(lmfit)
```

```
##
## Call:
## lm(formula = stay ~ age + severity + new, data = bypass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72596 -0.20764  0.02638  0.20951  0.74147
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.723896   0.267189   2.709  0.00734 **
## age          0.201153   0.003185  63.158  < 2e-16 ***
## severity     0.302302   0.002164 139.706  < 2e-16 ***
## new         -4.903423   0.076432 -64.154  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2975 on 196 degrees of freedom
## Multiple R-squared:  0.9959, Adjusted R-squared:  0.9959
## F-statistic: 1.599e+04 on 3 and 196 DF,  p-value: < 2.2e-16
```

The effect of the new procedure is clearly effective in reducing length of stay. Effect is minus five days.

This remains if we subset the data on the region of overlap, and ignore age

```r
lmfit <- lm(stay ~ severity + new, data = bypass %>% filter(age > 76 & age < 83))

summary(lmfit)
```

```
##
## Call:
## lm(formula = stay ~ severity + new, data = bypass %>% filter(age >
##     76 & age < 83))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00459 -0.30434 -0.05088  0.31558  1.13558
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.82893    0.51702   32.55   <2e-16 ***
## severity     0.29653    0.01163   25.49   <2e-16 ***
## new         -4.93525    0.23518  -20.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5239 on 33 degrees of freedom
## Multiple R-squared:  0.9546, Adjusted R-squared:  0.9519
## F-statistic: 347.3 on 2 and 33 DF,  p-value: < 2.2e-16
```

(e) Explain the discrepancy between these estimates.

Patients that get the new treatment are more severe, given their age. Being more severe increases their length of stay. This masks the beneficial effect of the new treatment.