Code for QSS Chapter 2: Causality

Kosuke Imai

First Printing

Section 2.1: Racial Discrimination in the Labor Market

```
resume <- read.csv("resume.csv")

dim(resume)
```

```
## [1] 4870    4
```

```
head(resume)
```

```
##   firstname    sex  race call
## 1   Allison female white    0
## 2   Kristen female white    0
## 3   Lakisha female black    0
## 4   Latonya female black    0
## 5    Carrie female white    0
## 6       Jay   male white    0
```

```
summary(resume)
```

```
##    firstname        sex           race           call
##  Tamika : 256   female:3746   black:2435   Min.   :0.00000
##  Anne   : 242   male  :1124   white:2435   1st Qu.:0.00000
##  Allison: 232                              Median :0.00000
##  Latonya: 230                              Mean   :0.08049
##  Emily  : 227                              3rd Qu.:0.00000
##  Latoya : 226                              Max.   :1.00000
##  (Other):3457
```

```
race.call.tab <- table(race = resume$race, call = resume$call)
race.call.tab
```

```
##        call
## race       0    1
##   black 2278  157
##   white 2200  235
```

```
addmargins(race.call.tab)
```

```
##        call
## race       0    1  Sum
##   black 2278  157 2435
##   white 2200  235 2435
##   Sum   4478  392 4870
```

```
## overall callback rate: total callbacks divided by the sample size
sum(race.call.tab[, 2]) / nrow(resume)
```

```
## [1] 0.08049281
```

```
## callback rates for each race
race.call.tab[1, 2] / sum(race.call.tab[1, ]) # black
```

```
## [1] 0.06447639
```

```
race.call.tab[2, 2] / sum(race.call.tab[2, ]) # white
```

```
## [1] 0.09650924
```

```
race.call.tab[1, ]  # the first row
```

```
##    0    1
## 2278  157
```

```
race.call.tab[, 2]  # the second column
```

```
## black white
##   157   235
```

```
mean(resume$call)
```

```
## [1] 0.08049281
```

Section 2.2: Subsetting the Data in R

```
class(TRUE)
```

```
## [1] "logical"
```

```
as.integer(TRUE)
```

```
## [1] 1
```

```
as.integer(FALSE)
```

```
## [1] 0
```

```
x <- c(TRUE, FALSE, TRUE) # a vector with logical values

mean(x) # proportion of TRUEs
```

```
## [1] 0.6666667
```

```
sum(x) # number of TRUEs
```

```
## [1] 2
```

```
FALSE & TRUE
```

```
## [1] FALSE
```

```
TRUE & TRUE
```

```
## [1] TRUE
```

```
TRUE | FALSE
```

```
## [1] TRUE
```

```
FALSE | FALSE
```

```
## [1] FALSE
```

```r
TRUE & FALSE & TRUE
```

```
## [1] FALSE
```

```r
(TRUE | FALSE) & FALSE # the parentheses evaluate to TRUE
```

```
## [1] FALSE
```

```r
TRUE | (FALSE & FALSE) # the parentheses evaluate to FALSE
```

```
## [1] TRUE
```

```r
TF1 <- c(TRUE, FALSE, FALSE)
TF2 <- c(TRUE, FALSE, TRUE)
TF1 | TF2
```

```
## [1]  TRUE FALSE  TRUE
```

```r
TF1 & TF2
```

```
## [1]  TRUE FALSE FALSE
```

Section 2.2.2: Relational Operators

```r
4 > 3
```

```
## [1] TRUE
```

```r
"Hello" == "hello"  # R is case-sensitive
```

```
## [1] FALSE
```

```r
"Hello" != "hello"
```

```
## [1] TRUE
```

```r
x <- c(3, 2, 1, -2, -1)
x >= 2
```

```
## [1]  TRUE  TRUE FALSE FALSE FALSE
```

```r
x != 1
```

```
## [1]  TRUE  TRUE FALSE  TRUE  TRUE
```

```r
## logical conjunction of two vectors with logical values
(x > 0) & (x <= 2)
```

```
## [1] FALSE  TRUE  TRUE FALSE FALSE
```

```r
## logical disjunction of two vectors with logical values
(x > 2) | (x <= -1)
```

```
## [1]  TRUE FALSE FALSE  TRUE  TRUE
```

```r
x.int <- (x > 0) & (x <= 2) # logical vector
x.int
```

```
## [1] FALSE  TRUE  TRUE FALSE FALSE
```

```r
mean(x.int) # proportion of TRUEs
```

```
## [1] 0.4
```

```r
sum(x.int)  # number of TRUEs
```

```
## [1] 2
```

Section 2.2.3: Subsetting

```r
## callback rate for black-sounding names
mean(resume$call[resume$race == "black"])
```

```
## [1] 0.06447639
```

```r
## race of first 5 observations
resume$race[1:5]
```

```
## [1] white white black black white
## Levels: black white
```

```r
## comparison of first 5 observations
(resume$race == "black")[1:5]
```

```
## [1] FALSE FALSE  TRUE  TRUE FALSE
```

```r
dim(resume) # dimension of original data frame
```

```
## [1] 4870    4
```

```r
## subset blacks only
resumeB <- resume[resume$race == "black", ]
dim(resumeB) # this data.frame has fewer rows than the original data.frame
```

```
## [1] 2435    4
```

```r
mean(resumeB$call) # callback rate for blacks
```

```
## [1] 0.06447639
```

```r
## keep "call" and "firstname" variables
## also keep observations with black female-sounding names
resumeBf <- subset(resume, select = c("call", "firstname"),
                   subset = (race == "black" & sex == "female"))
head(resumeBf)
```

```
##    call firstname
## 3     0   Lakisha
## 4     0   Latonya
## 8     0     Kenya
## 9     0   Latonya
## 11    0     Aisha
## 13    0     Aisha
```

```r
## ## an alternative syntax with the same results
## resumeBf <- resume[resume$race == "black" & resume$sex == "female",
##                    c("call", "firstname")]
## black male
resumeBm <- subset(resume, subset = (race == "black") & (sex == "male"))
## white female
resumeWf <- subset(resume, subset = (race == "white") & (sex == "female"))
## white male
```

```r
resumeWm <- subset(resume, subset = (race == "white") & (sex == "male"))
## racial gaps
mean(resumeWf$call) - mean(resumeBf$call) # among females
```

```
## [1] 0.03264689
```

```r
mean(resumeWm$call) - mean(resumeBm$call) # among males
```

```
## [1] 0.03040786
```

Section 2.2.4: Simple Conditional Statements

```r
resume$BlackFemale <- ifelse(resume$race == "black" &
                                 resume$sex == "female", 1, 0)
table(race = resume$race, sex = resume$sex,
      BlackFemale = resume$BlackFemale)
```

```
## , , BlackFemale = 0
##
##        sex
## race    female male
##   black      0  549
##   white   1860  575
##
## , , BlackFemale = 1
##
##        sex
## race    female male
##   black   1886    0
##   white      0    0
```

Section 2.2.5: Factor Variables

```r
resume$type <- NA
resume$type[resume$race == "black" & resume$sex == "female"] <- "BlackFemale"
resume$type[resume$race == "black" & resume$sex == "male"] <- "BlackMale"
resume$type[resume$race == "white" & resume$sex == "female"] <- "WhiteFemale"
resume$type[resume$race == "white" & resume$sex == "male"] <- "WhiteMale"

## check object class
class(resume$type)
```

```
## [1] "character"
```

```r
## coerce new character variable into a factor variable
resume$type <- as.factor(resume$type)
## list all levels of a factor variable
levels(resume$type)
```

```
## [1] "BlackFemale" "BlackMale"   "WhiteFemale" "WhiteMale"
```

```r
## obtain the number of observations for each level
table(resume$type)
```

```
##
## BlackFemale   BlackMale WhiteFemale   WhiteMale
##        1886         549        1860         575
```

```
tapply(resume$call, resume$type, mean)
```

```
## BlackFemale   BlackMale WhiteFemale   WhiteMale
##  0.06627784  0.05828780  0.09892473  0.08869565
```

```
## turn first name into a factor variable
resume$firstname <- as.factor(resume$firstname)
## compute callback rate for each first name
callback.name <- tapply(resume$call, resume$firstname, mean)
## sort the result in the increasing order
sort(callback.name)
```

```
##      Aisha    Rasheed     Keisha   Tremayne     Kareem    Darnell
## 0.02222222 0.02985075 0.03825137 0.04347826 0.04687500 0.04761905
##     Tyrone      Hakim     Tamika    Lakisha    Tanisha       Todd
## 0.05333333 0.05454545 0.05468750 0.05500000 0.05797101 0.05882353
##      Jamal       Neil      Brett   Geoffrey    Brendan       Greg
## 0.06557377 0.06578947 0.06779661 0.06779661 0.07692308 0.07843137
##      Emily       Anne       Jill     Latoya      Kenya    Matthew
## 0.07929515 0.08264463 0.08374384 0.08407080 0.08673469 0.08955224
##    Latonya      Leroy    Allison      Ebony   Jermaine     Laurie
## 0.09130435 0.09375000 0.09482759 0.09615385 0.09615385 0.09743590
##      Sarah   Meredith     Carrie    Kristen        Jay       Brad
## 0.09844560 0.10160428 0.13095238 0.13145540 0.13432836 0.15873016
```

Section 2.3: Causal Effects and the Counterfactual

```
resume[1, ]
```

```
##   firstname    sex  race call BlackFemale        type
## 1   Allison female white    0           0 WhiteFemale
```

Section 2.4: Randomized Controlled Trials

Section 2.4.1: The Role of Randomization

Section 2.4.2: Social Pressure and Voter Turnout

```
social <- read.csv("social.csv") # load the data

summary(social) # summarize the data
```

```
##      sex          yearofbirth    primary2004          messages
##  female:152702   Min.   :1900   Min.   :0.0000   Civic Duty: 38218
##  male  :153164   1st Qu.:1947   1st Qu.:0.0000   Control   :191243
##                  Median :1956   Median :0.0000   Hawthorne : 38204
##                  Mean   :1956   Mean   :0.4014   Neighbors : 38201
##                  3rd Qu.:1965   3rd Qu.:1.0000
```

```
##                    Max.   :1986   Max.   :1.0000
##   primary2006         hhsize
##  Min.   :0.0000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.0000   Median :2.000
##  Mean   :0.3122   Mean   :2.184
##  3rd Qu.:1.0000   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :8.000
```

```
## turnout for each group
tapply(social$primary2006, social$messages, mean)
```

```
## Civic Duty    Control  Hawthorne  Neighbors
##  0.3145377  0.2966383  0.3223746  0.3779482
```

```
## turnout for control group
mean(social$primary2006[social$messages == "Control"])
```

```
## [1] 0.2966383
```

```
## subtract control group turnout from each group
tapply(social$primary2006, social$messages, mean) -
    mean(social$primary2006[social$messages == "Control"])
```

```
## Civic Duty    Control  Hawthorne  Neighbors
## 0.01789934 0.00000000 0.02573631 0.08130991
```

```
social$age <- 2006 - social$yearofbirth # create age variable
tapply(social$age, social$messages, mean)
```

```
## Civic Duty    Control  Hawthorne  Neighbors
##   49.65904   49.81355   49.70480   49.85294
```

```
tapply(social$primary2004, social$messages, mean)
```

```
## Civic Duty    Control  Hawthorne  Neighbors
##  0.3994453  0.4003388  0.4032300  0.4066647
```

```
tapply(social$hhsize, social$messages, mean)
```

```
## Civic Duty    Control  Hawthorne  Neighbors
##   2.189126   2.183667   2.180138   2.187770
```

Section 2.5: Observational Studies

Section 2.5.1: Minimum Wage and Unemployment

```
minwage <- read.csv("minwage.csv") # load the data

dim(minwage) # dimension of data
```

```
## [1] 358   8
```

```
summary(minwage) # summary of data
```

```
##         chain          location     wageBefore      wageAfter
##  burgerking:149   centralNJ: 45   Min.   :4.250   Min.   :4.250
```

```
##   kfc       : 75   northNJ  :146   1st Qu.:4.250   1st Qu.:5.050
##   roys      : 88   PA       : 67   Median :4.500   Median :5.050
##   wendys    : 46   shoreNJ  : 33   Mean   :4.618   Mean   :4.994
##                    southNJ  : 67   3rd Qu.:4.987   3rd Qu.:5.050
##                                    Max.   :5.750   Max.   :6.250
##     fullBefore       fullAfter        partBefore      partAfter
##   Min.   : 0.000   Min.   : 0.000   Min.   : 0.00   Min.   : 0.00
##   1st Qu.: 2.125   1st Qu.: 2.000   1st Qu.:11.00   1st Qu.:11.00
##   Median : 6.000   Median : 6.000   Median :16.25   Median :17.00
##   Mean   : 8.475   Mean   : 8.362   Mean   :18.75   Mean   :18.69
##   3rd Qu.:12.000   3rd Qu.:12.000   3rd Qu.:25.00   3rd Qu.:25.00
##   Max.   :60.000   Max.   :40.000   Max.   :60.00   Max.   :60.00
```

```r
## subsetting the data into two states
minwageNJ <- subset(minwage, subset = (location != "PA"))
minwagePA <- subset(minwage, subset = (location == "PA"))

## proportion of restaurants whose wage is less than $5.05
mean(minwageNJ$wageBefore < 5.05) # NJ before
```

```
## [1] 0.9106529
```

```r
mean(minwageNJ$wageAfter < 5.05)  # NJ after
```

```
## [1] 0.003436426
```

```r
mean(minwagePA$wageBefore < 5.05) # PA before
```

```
## [1] 0.9402985
```

```r
mean(minwagePA$wageAfter < 5.05)  # PA after
```

```
## [1] 0.9552239
```

```r
## create a variable for proportion of full-time employees in NJ and PA
minwageNJ$fullPropAfter <- minwageNJ$fullAfter /
    (minwageNJ$fullAfter + minwageNJ$partAfter)
minwagePA$fullPropAfter <- minwagePA$fullAfter /
    (minwagePA$fullAfter + minwagePA$partAfter)

## compute the difference in means
mean(minwageNJ$fullPropAfter) - mean(minwagePA$fullPropAfter)
```

```
## [1] 0.04811886
```

Section 2.5.2: Confounding Bias

```r
prop.table(table(minwageNJ$chain))
```

```
##
## burgerking        kfc       roys     wendys
##   0.4054983  0.2233677  0.2508591  0.1202749
```

```r
prop.table(table(minwagePA$chain))
```

```
##
## burgerking        kfc       roys     wendys
```

```
##  0.4626866  0.1492537  0.2238806  0.1641791
```

```
## subset Burger King only
minwageNJ.bk <- subset(minwageNJ, subset = (chain == "burgerking"))
minwagePA.bk <- subset(minwagePA, subset = (chain == "burgerking"))

## comparison of full-time employment rates
mean(minwageNJ.bk$fullPropAfter) - mean(minwagePA.bk$fullPropAfter)
```

```
## [1] 0.03643934
```

```
minwageNJ.bk.subset <-
    subset(minwageNJ.bk, subset = ((location != "shoreNJ") &
                                     (location != "centralNJ")))

mean(minwageNJ.bk.subset$fullPropAfter) - mean(minwagePA.bk$fullPropAfter)
```

```
## [1] 0.03149853
```

Section 2.5.3: Before-and-After and Difference-in-Differences Designs

```
## full-time employment proportion in the previous period for NJ
minwageNJ$fullPropBefore <- minwageNJ$fullBefore /
    (minwageNJ$fullBefore + minwageNJ$partBefore)

## mean difference between before and after the minimum wage increase
NJdiff <- mean(minwageNJ$fullPropAfter) - mean(minwageNJ$fullPropBefore)
NJdiff
```

```
## [1] 0.02387474
```

```
## full-time employment proportion in the previous period for PA
minwagePA$fullPropBefore <- minwagePA$fullBefore /
    (minwagePA$fullBefore + minwagePA$partBefore)
## mean difference between before and after for PA
PAdiff <- mean(minwagePA$fullPropAfter) - mean(minwagePA$fullPropBefore)
## difference-in-differences
NJdiff - PAdiff
```

```
## [1] 0.06155831
```

```
## full-time employment proportion in the previous period for PA
minwagePA$fullPropBefore <- minwagePA$fullBefore /
    (minwagePA$fullBefore + minwagePA$partBefore)
## mean difference between before and after for PA
PAdiff <- mean(minwagePA$fullPropAfter) - mean(minwagePA$fullPropBefore)
## difference-in-differences
NJdiff - PAdiff
```

```
## [1] 0.06155831
```

Section 2.6: Descriptive Statistics for a Single Variable

Section 2.6.1: Quantiles

```
## cross-section comparison between NJ and PA
median(minwageNJ$fullPropAfter) - median(minwagePA$fullPropAfter)
```

```
## [1] 0.07291667
```

```
## before and after comparison
NJdiff.med <- median(minwageNJ$fullPropAfter) -
    median(minwageNJ$fullPropBefore)
NJdiff.med
```

```
## [1] 0.025
```

```
## median difference-in-differences
PAdiff.med <- median(minwagePA$fullPropAfter) -
    median(minwagePA$fullPropBefore)
NJdiff.med - PAdiff.med
```

```
## [1] 0.03701923
```

```
## summary shows quartiles as well as minimum, maximum, and mean
summary(minwageNJ$wageBefore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.25    4.25    4.50    4.61    4.87    5.75
```

```
summary(minwageNJ$wageAfter)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   5.050   5.050   5.081   5.050   5.750
```

```
## interquartile range
IQR(minwageNJ$wageBefore)
```

```
## [1] 0.62
```

```
IQR(minwageNJ$wageAfter)
```

```
## [1] 0
```

```
## deciles (10 groups)
quantile(minwageNJ$wageBefore, probs = seq(from = 0, to = 1, by = 0.1))
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
## 4.25 4.25 4.25 4.25 4.50 4.50 4.65 4.75 5.00 5.00 5.75
```

```
quantile(minwageNJ$wageAfter, probs = seq(from = 0, to = 1, by = 0.1))
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
## 5.00 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.15 5.75
```

2.6.2: Standard Deviation

```
sqrt(mean((minwageNJ$fullPropAfter - minwageNJ$fullPropBefore)^2))
```

```
## [1] 0.3014669
```

```r
mean(minwageNJ$fullPropAfter - minwageNJ$fullPropBefore)
```

```
## [1] 0.02387474
```

```r
## standard deviation
sd(minwageNJ$fullPropBefore)
```

```
## [1] 0.2304592
```

```r
sd(minwageNJ$fullPropAfter)
```

```
## [1] 0.2510016
```

```r
## variance
var(minwageNJ$fullPropBefore)
```

```
## [1] 0.05311145
```

```r
var(minwageNJ$fullPropAfter)
```

```
## [1] 0.0630018
```