

Stat 021 Homework 5

Kayonna L Tindle

Due: Friday, Nov. 1, 12:00pm

Instructions: A **pdf** version of your homework must be submitted to Gradescope by **noon** on the due date. The course passcode is **MPKJ4Z**. If you are having trouble getting your *.Rmd* file to compile, you need to get help with this **before** the due date.

You are allowed to hand in **only one** late homework assignment throughout the semester. If you need to hand in this particular assignment late, you must let me know via email by noon on the due date.

You are encouraged to study with your peers to help complete the homework assignments but no copying is allowed. If I see that two or more homework assignments are copied, all students involved will receive a grade of 0 on that assignment and will forfeit (perhaps retroactively) the opportunity to hand in a late homework.

Q1) Sketch (by hand) residual plots (with \hat{y}_i , predicted response values, on the horizontal axis) that show each of the following: (5 points) 1. constant variance and linearity; 1. non-constant variance and linearity; 1. constant variance and non-linearity; 1. non-constant variance and non-linearity.

Q2) Suppose we have two random variables X and Y . What are the differences among the following assumptions regarding X and Y :

- X and Y are uncorrelated,
- X and Y are independent,
- X and Y have the same variance, and
- X and Y have the same distribution? (5 points)

Q3) Read the Wikipedia page for Simpson's Paradox: https://en.wikipedia.org/wiki/Simpson%27s_paradox. Then, import the "Stand your ground" data set uploaded on Moodle. This data (from 2015) is related to the Stand Your Ground law in Florida. Each observational unit consists of a case where the Stand Your Ground law was a part of the defense strategy, the defendant's race (white or non-white), the victim's race (white or non-white), and a binary variable indicating whether or not the defendant was convicted. With this categorical data we are not going to fit a regression model but we are going to examine this data and look out for Simpson's paradox. (10 points)

a) Create and print the following tables to summarize the data:

1. Defendant's race vs convicted for all observational units;
2. Defendant's race vs convicted for cases with minority victims only;
3. Defendant's race vs convicted for cases with white victims only;
4. The table created by adding Tables 2 and 3 together.

b) What are the overall conviction rates for minority and white defendants, respectively? What are the conviction rates for minority and white defendants among the cases with minority victims? What are the conviction rates for minority and white defendants among the cases with white victims?

c) Explain what is going on here in terms of Simpson's paradox and interpret what this means with respect to racial bias in the criminal justice system.

Q2 Suppose we have two random variables X and Y . What are the differences among the following assumptions regarding X and Y :

- X and Y are uncorrelated, this idea relates back to covariance wherein if a two variables are uncorrelated, we cannot fit a linear relationship to the two variables. We note that there the covariance between the two random variables equals zero.
- X and Y are independent, this idea relates back more strongly to probability wherein we could state that the probability of X and Y combined is the same (in other words, does “not” change). More plainly, the outcome of one variable, X , does not affect the results of variable Y . When two variables are independent, they will be inherently uncorrelated as well.
- X and Y have the same variance, and X and Y have the same distribution? Observing equal variance relates to not having heteroscedasticity, wherein the spread of the data points are about equal across the data set. However, in relation to random variables X and Y , we will observe similar spread between the data sets wherein variance of both random variables may equal 1. Nonetheless, when we observe distribution in the data set, we are displaying the values of a data set and how frequently we will observe those values. Since variance is not the only factor that goes into a distribution and its shape (i.e we also look at the mean). we can not say that two random variables with the same variance will have the same distribution. (5 points)

```
#setwd("~/Stat 21 HW")
setwd("~/Google Drive Swat/Swat docs/Stat 21/Homework")
#read_table2("stand_you_ground.txt")
#dtable1 <- read_table2("stand_you_ground.txt") ##typo! "your" not "you"
dtable1 <- read_csv("stand_your_ground.csv")

## Parsed with column specification:
## cols(
##   Convicted = col_character(),
##   Accused = col_character(),
##   WhiteVictim = col_double(),
##   MinVictim = col_double()
## )

dtable1 <- read_csv("~/Google Drive Swat/Swat docs/Stat 21/Homework/stand_your_ground.csv") ##sometimes

## Parsed with column specification:
## cols(
##   Convicted = col_character(),
##   Accused = col_character(),
##   WhiteVictim = col_double(),
##   MinVictim = col_double()
## )

##also, note that we are using the "read_csv()" function since this data is a .csv file

print(dtable1)

## # A tibble: 220 x 4
##   Convicted Accused WhiteVictim MinVictim
##   <chr>      <chr>      <dbl>      <dbl>
## 1 Yes      White          1          0
## 2 Yes      White          1          0
## 3 Yes      White          1          0
## 4 Yes      White          1          0
## 5 Yes      White          1          0
## 6 Yes      White          1          0
## 7 Yes      White          1          0
```

```
## 8 Yes      White      1      0
## 9 Yes      White      1      0
## 10 Yes     White      1      0
## # ... with 210 more rows
```

```
View(dtable1)
dtable2 <- dtable1
dtable2 <- dtable1 %>% select(Convicted, Accused, MinVictim) %>% filter(MinVictim==1)
print(dtable2)
```

```
## # A tibble: 88 x 3
##   Convicted Accused MinVictim
##   <chr>      <chr>      <dbl>
## 1 Yes      White      1
## 2 Yes      White      1
## 3 Yes      White      1
## 4 Yes      White      1
## 5 Yes      White      1
## 6 No       White      1
## 7 No       White      1
## 8 No       White      1
## 9 No       White      1
## 10 No      White      1
## # ... with 78 more rows
```

```
View(dtable2)
dtable3 <- dtable1
dtable3 <- dtable1 %>% select(Convicted, Accused, WhiteVictim) %>% filter(WhiteVictim==1)
print(dtable3)
```

```
## # A tibble: 132 x 3
##   Convicted Accused WhiteVictim
##   <chr>      <chr>      <dbl>
## 1 Yes      White      1
## 2 Yes      White      1
## 3 Yes      White      1
## 4 Yes      White      1
## 5 Yes      White      1
## 6 Yes      White      1
## 7 Yes      White      1
## 8 Yes      White      1
## 9 Yes      White      1
## 10 Yes     White      1
## # ... with 122 more rows
```

```
View(dtable3)
```

```
# Conviction rate defined: "conviction rate of a prosecutor or government is the number of convictions
# overall conviction rates for minority and white defendants, respectively
# percent of cases brought of white defendants: divided by convictions is 34 percent
q3b1 <- dtable1 %>% filter(Accused=="White", Convicted=="Yes")
q3b2 <- dtable1 %>% filter(Accused=="White")
count(q3b1)/count(q3b2)
```

```
##           n
## 1 0.3435115
```

```
# percent of cases brought of minority defendants: divided by convictions is 33 percent
q3b3 <- dtable1 %>% filter(Accused=="Minority", Convicted=="Yes")
q3b4 <- dtable1 %>% filter(Accused=="Minority")
count(q3b3)/count(q3b4)
```

```
##          n
## 1 0.3258427
```

```
# percent the conviction rates for minority and white defendants among the cases with minority victims
q3b5 <- dtable1 %>% filter(MinVictim=="1", Convicted=="Yes")
q3b6 <- dtable1 %>% filter(MinVictim=="1")
count(q3b5)/count(q3b6)
```

```
##          n
## 1 0.2727273
```

```
# percent the conviction rates for minority and white defendants among the cases with white victims is .
q3b7 <- dtable1 %>% filter(WhiteVictim=="1", Convicted=="Yes")
q3b8 <- dtable1 %>% filter(WhiteVictim=="1")
count(q3b7)/count(q3b8)
```

```
##          n
## 1 0.3787879
```

```
# part c. these findings apply to the Simpson's paradox by showing how if we were to observe the data s
```