Stat 021 Homework 5

Suzanne Thornton

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 horizantal 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)

X and Y are uncorrelated doesn't necessarily mean they are independent of each other, but X and Y are independent means X and Y are definitely uncorrelated. Cor(X,Y)=0 therefore Cov(X,Y)=0. i.e. X~N(0,1) and Y = x^2, X and Y are uncorrelated but Y is dependent on X.

X and Y have the same variance means they have the same spread or range of data. X and Y have same distribution means they have the same shape of distribution but the variance is not necessarily the same.

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 defendent's race (white or non-white), the victim's race (white or non-white), and a binary variable indicating whether or not the defendent 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;

```
stand_your_ground <- read_csv("stand your ground.csv")
```

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

```
overall_tbl <- table(stand_your_ground$Accused,stand_your_ground$Convicted)
overall_tbl
```

```
##
##            No Yes
##   Minority 60  29
##   White    86  45
```

1. Defendant's race vs convicted for cases with minority victims only;

```
minV <- stand_your_ground %>% filter(MinVictim == 1)
minV_tbl <- table(minV$Accused,minV$Convicted)
minV_tbl
```

```
##
##            No Yes
##   Minority 45  19
##   White    19   5
```

1. Defendant's race vs convicted for cases with white victims only;

```
whiteV <- stand_your_ground %>% filter(WhiteVictim ==1)
whiteV_tbl <- table(whiteV$Accused, whiteV$Convicted)
whiteV_tbl
```

```
##
##            No Yes
##   Minority 15  10
##   White    67  40
```

1. The table created by adding Tables 2 and 3 together.

```
added_tbl <- minV_tbl+whiteV_tbl
added_tbl
```

```
##
##            No Yes
##   Minority 60  29
##   White    86  45
```

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?

```
#overall conviction rates
minC_rate <- 29/(60+29)
minC_rate
```

```
## [1] 0.3258427
```

```
whiteC_rate <- 45/(45+86)
whiteC_rate
```

```
## [1] 0.3435115
```

```
#minority victims
minC_minV_rate <- 19/(19+45)
minC_minV_rate
```

```
## [1] 0.296875
```

```
whiteC_minV_rate <- 5/(19+5)
whiteC_minV_rate
```

```
## [1] 0.2083333
```

```
#WHITE victims
minC_whiteV_rate <-10/(15+10)
minC_whiteV_rate
```

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

```
whiteC_whiteV_rate <- 40/(40+67)
whiteC_whiteV_rate
```

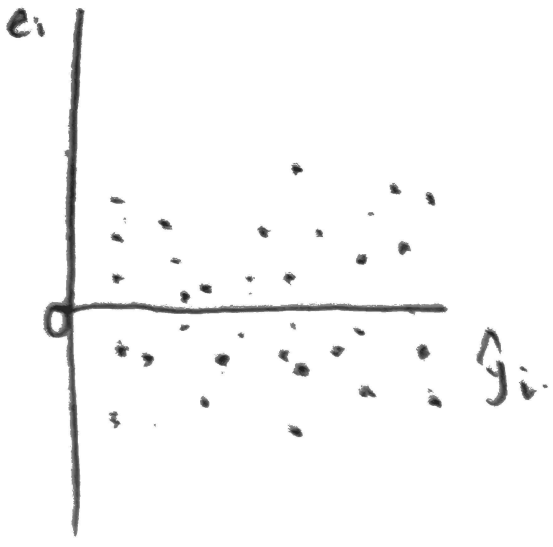
```
## [1] 0.3738318
```

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.

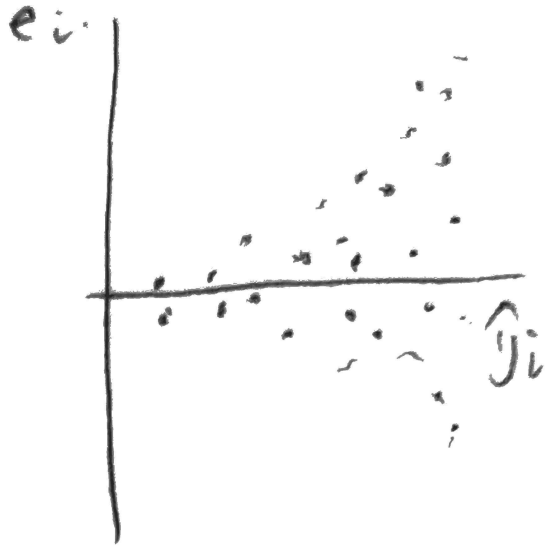
Overall conviction rate of white defendants is higher than minority's (0.34>0.32). But among specifc cases with either minority victims or white victims, conviction rates of minority defendants are all higher than white defendants'(0.29>0.20, 0.4>0.37).

This means that racial bias against the minority exists in crimincal justice system. Though overall conviction rate of white defendants is higher, the Simpsons's paradox indicates there are latent factors influencing this. For example,averagepoverty rates in minority's residency maybe higher and police activity in miniority's neighborhood maybe more frequent that in the white's.

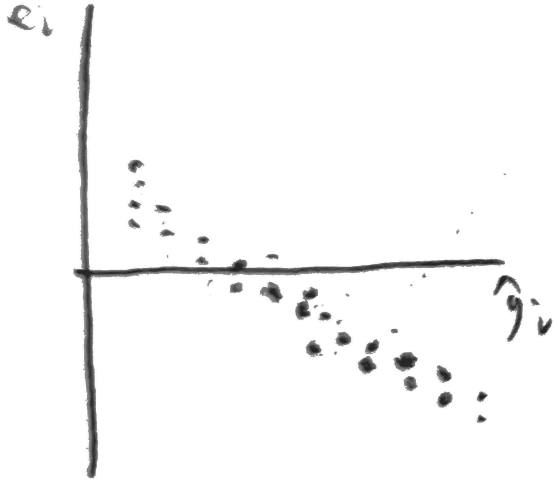
Q.1)



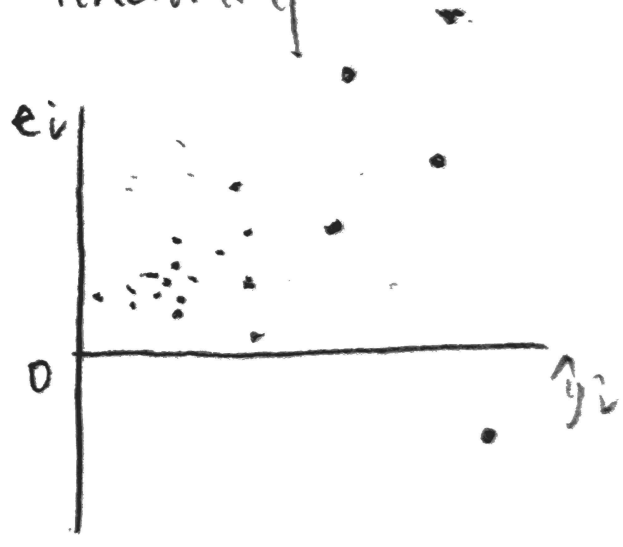
constant variance
linearity.



non-constant variance
linearity



constant variance
non linearity.



non constant variance
non linearity.