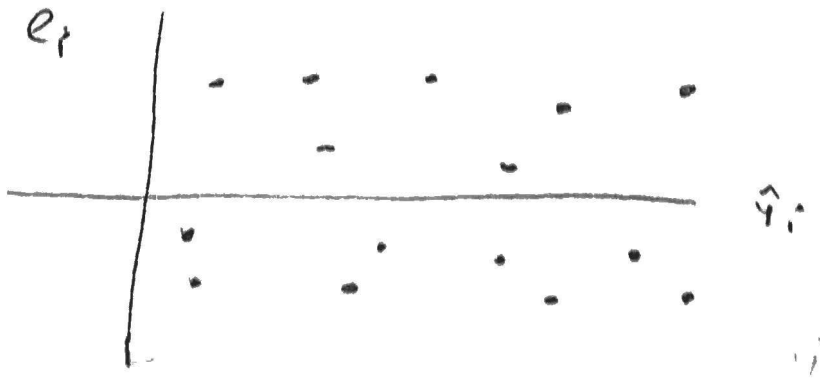
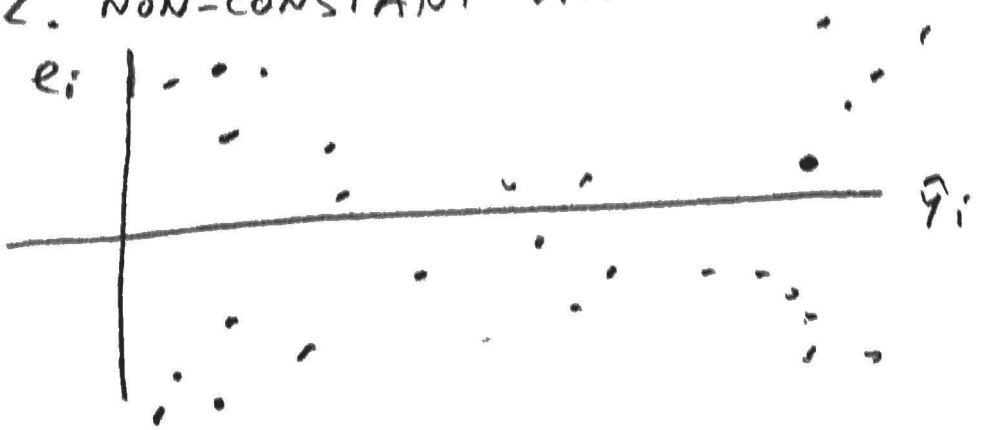
Wen Seng
STAT 021 HW 5

Q 1)

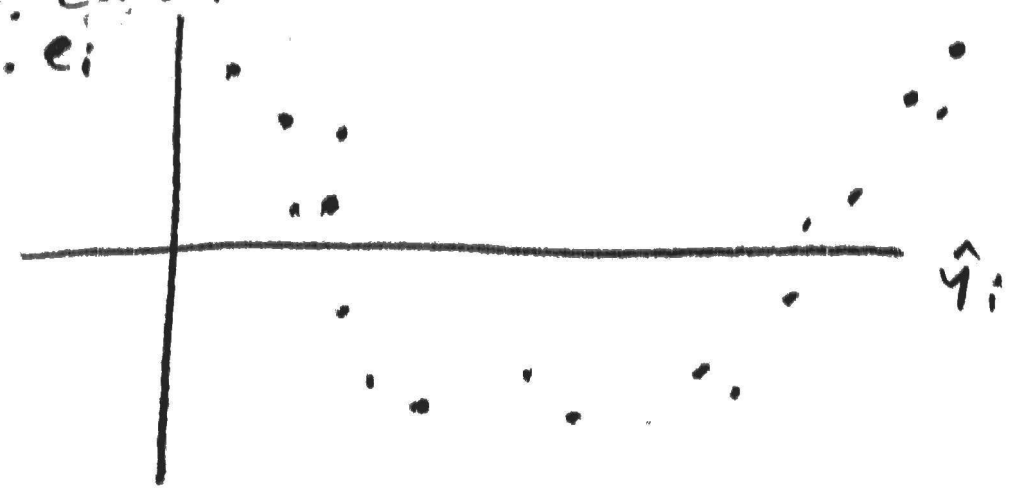
1. CONSTANT VARIANCE & LINEARITY



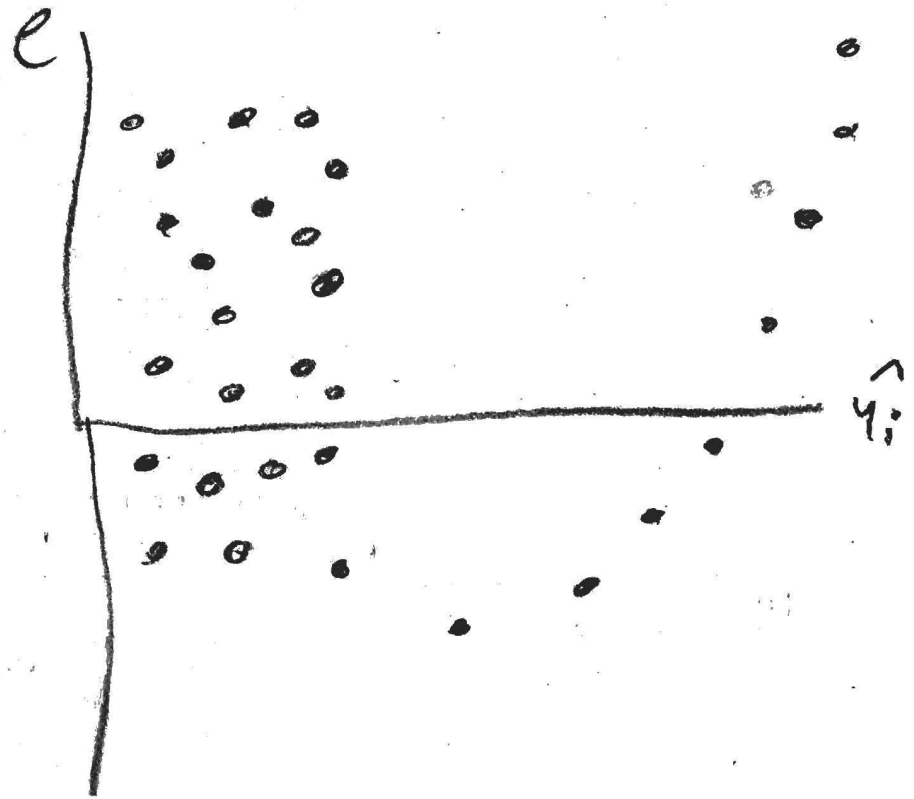
2. NON-CONSTANT VARIANCE & LINEARITY



3. CONSTANT VARIANCE AND NON-LINEARITY



NON-CONSTANT VARIANCE AND NON-LINEARITY



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)

Answer: When X and Y are independent, then the Cov(X,Y)=0, so we can conclude that X and Y are uncorrelated. Generally, independence means that P(X=x and Y=y)=P(X=x)*P(Y=y). This means that knowing the value of X tells us nothing about the distribution of Y. On the other hand, when two random variables X and Y are uncorrelated, then the correlation coefficient is zero. However, it is still possible that X and Y can be dependent or independent.

Random variables X and Y have the same variance if they have the same spread over the relevant range of data values. X and Y have the same distribution when both X and Y are characterized by the same type of distribution. For example, if X is a uniform distribution from [0,1] and Y is a uniform distribution from [0,2], X and Y have the same distribution, but their variances are different because they have different spread.

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;
   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.

```
library(tidyverse)
stand_ground_dat<-read_table2("stand_you_ground.txt",col_types=list(col_character(),
                              col_character(),col_double(),col_double()))

#creating table for all observational units
race_convicted<-table(stand_ground_dat$Accused,stand_ground_dat$Convicted)
race_convicted
```

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

```
29/89   #32.6% of minority defendants are convicted
```

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

```
45/131  #34.4% of white defendants are convicted
```

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

```
#minority victims only
stand_ground_min_vic<-stand_ground_dat%>%filter(MinVictim==1)
race_convicted2<-table(stand_ground_min_vic$Accused,stand_ground_min_vic$Convicted)
race_convicted2
```

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

```
19/64   #given that victims are minority, 29.7% of minority defendants get convicted
```

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

```
5/24    #given that victims are minority, 20.8% of white defendants get convicted
```

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

```
#white victims only
stand_ground_white<-stand_ground_dat%>%filter(WhiteVictim==1)
race_convicted3<-table(stand_ground_white$Accused,stand_ground_white$Convicted)
race_convicted3
```

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

```
10/25   #given that victims are white, 40% of minority defendants get convicted
```

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

```
40/117  #given that victims are white, 37.4% of white defendants get convicted
```

```
## [1] 0.3418803
```

```
#let's add the tables together
stand_ground_agg<-race_convicted2+race_convicted3
```

```
stand_ground_agg
```

```
##
##            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?

Answer: The overall conviction rate for minority defendants is 29/89=32.6% and the overall conviction rate for white defendants is 45/131=34.4%. So the aggregated data indicate that white defendants are more likely to be be convicted.

Now let's observe the unaggregated data: When the victims are minorities, minority defendants have a 29.7% conviction rate (19/64) while white defendants have a 20.8% conviction rate (5/24). So according to this data, minority defendants are more likely to be convicted for murders given that there are minority victims.

When the victims are white, minority defendants have a 40% conviction rate (10/25) while white defendants have a 37.4% conviction rate (40/107). According to this data, minority defendants seem to be more likely to be convicted for murders of 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.

In Simpson's paradox, the conclusion for unaggregated data, which is often separated into two tables by a confounding variable, is the reverse of the conclusion for aggregated data. In part (b), we see from the tables that white defendants have a higher conviction rate than minority defendants, so there is no bias against minority defendants here.

However, when we split the data into the victims' races, the opposite conclusion can be made: minority defendants have a higher conviction rate for the murders of both victim groups. Now the data show a potential bias against minority defendants. One possible reason is because of the great number of white defendants being convicted (40 in this case) in crimes on white victims, which would impact the result for aggregated data.