Stat 021 Homework 5

Ankur Malik

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; 2. non-constant variance and linearity; 3. constant variance and non-linearity; 4. 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)

The definitions of covariance and correlation for two random variables X and Y are pertinent to our analysis of the first two statements. They are as follows:

$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$

$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

If X and Y are independent, we know that: $E[XY] = E[X]E[Y]$

$=> Cov(X,Y) = 0$

$=> Corr(X,Y) = \frac{0}{\sqrt{Var(X)Var(Y)}} = 0$

Thus, the correlation of X and Y is 0 - in other words, they are uncorrelated. If X and Y are independent, it is always true that X and Y are uncorrelated.

However, if X and Y are uncorrelated, it does not necessarily follow that X and Y are indepedent. Let us look at an example where X and Y are uncorrelated but not independent.

Let $X \sim N(0,1), Y = X^3$. Clearly, since Y is based on the value of X, X and Y are dependent (ie. they are not independent). From the distribution of X, we know that: $E[X] = 0, E[X^4] = 0$

$Cov(X,Y) = E[XY] - E[X]E[Y] = E[X(X^3)] - E[X]E[Y] = E[X^4] - E[X]E[Y] = 0 - 0 * E[Y] = 0$

$=> Corr(X,Y) = \frac{0}{\sqrt{Var(X)Var(Y)}} = 0$

Thus, X and Y are uncorrelated. However, they are not independent.

If X and Y have the same distribution, X and Y have the same variance and the same mean. This is obviously a more exclusive statement than simply saying that X and Y have the same variance.

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.

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.

Part a):

```
stand <- read_csv("~/Documents/STAT 21/HW5/stand_your_ground.csv")
```

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

```
race_con <- stand %>% select(Convicted, Accused)
race_con_min <- stand %>% filter(MinVictim == 1)
race_con_min <- race_con_min %>% select(Convicted, Accused)
race_con_white <- stand %>% filter(WhiteVictim == 1)
race_con_white <- race_con_white %>% select(Convicted, Accused)

white <- race_con %>% filter(Accused == "White") %>% count(Convicted)
min <- race_con %>% filter(Accused == "Minority") %>% count(Convicted)

white_min <- race_con_min %>% filter(Accused == "White") %>% count(Convicted)
min_min <- race_con_min %>% filter(Accused == "Minority") %>% count(Convicted)

white_white <- race_con_white %>% filter(Accused == "White") %>% count(Convicted)
min_white <- race_con_white %>% filter(Accused == "Minority") %>% count(Convicted)

#table for all observational units
part1 <- tibble("WhiteConvictions" = 0, "MinorityConvictions" = 0)
part1[1] = white[2,2]
part1[2] = min[2,2]
```

```
#table for cases with minority victims only
part2 <- tibble("WhiteConvictions" = 0, "MinorityConvictions" = 0)
part2[1] = white_min[2,2]
part2[2] = min_min[2,2]

#table for cases with white victims only
part3 <- tibble("WhiteConvictions" = 0, "MinorityConvictions" = 0)
part3[1] = white_white[2,2]
part3[2] = min_white[2,2]

#table created by adding second and third tables together
bottom_row <- tibble("WhiteConvictions" = 0, "MinorityConvictions" = 0)
bottom_row[1] = part1[1]
bottom_row[2] = part1[2]
part4 <- rbind(part2, part3, bottom_row)

print(part1)
```

```
## # A tibble: 1 x 2
##   WhiteConvictions MinorityConvictions
##              <int>               <int>
## 1               45                  29
```

```
print(part2)
```

```
## # A tibble: 1 x 2
##   WhiteConvictions MinorityConvictions
##              <int>               <int>
## 1                5                  19
```

```
print(part3)
```

```
## # A tibble: 1 x 2
##   WhiteConvictions MinorityConvictions
##              <int>               <int>
## 1               40                  10
```

```
print(part4)
```

```
## # A tibble: 3 x 2
##   WhiteConvictions MinorityConvictions
##              <int>               <int>
## 1                5                  19
## 2               40                  10
## 3               45                  29
```

Part b):

```
white_rate = white[2,2]/(white[2,2]+white[1,2])
white_rate <- white_rate %>% specify_decimal(4)
min_rate = min[2,2]/(min[2,2]+min[1,2])
min_rate <- min_rate %>% specify_decimal(4)

white_min_rate = white_min[2,2]/(white_min[2,2]+white_min[1,2])
white_min_rate <- white_min_rate %>% specify_decimal(4)
min_min_rate = min_min[2,2]/(min_min[2,2]+min_min[1,2])
min_min_rate <- min_min_rate %>% specify_decimal(4)
```

```
white_white_rate = white_white[2,2]/(white_white[2,2]+white_white[1,2])
white_white_rate <- white_white_rate %>% specify_decimal(4)
min_white_rate = min_white[2,2]/(min_white[2,2]+min_white[1,2])
min_white_rate <- min_white_rate %>% specify_decimal(4)
```

Overall conviction rate for minority defendants: 0.3258

Overall conviction rate for white defendants: 0.3435

Conviction rate for minority defendants among cases with minority victims: 0.2969

Conviction rate for white defendants among cases with minority victims: 0.2083

Conviction rate for minority defendants among cases with white victims: 0.4000

Conviction rate for white defendants among cases with white victims: 0.3738

Part c):

Let us examine what is going on here. For white defendants, the conviction rate is much higher among cases with white victims, than it is among cases with minority victims (0.37 vs 0.2). This large discrepancy suggests that there is a racial bias against minority victims in cases where the defendant is white - there may be underconviction of white defendants due to the victim being a minority. However, the number of cases where the defendant is white and the victim is also white is much larger than the number of cases where the defendant is white and the victim is a minority (107 vs 24). Since the higher conviction rate for white defendants applies to a much larger number of the cases against them, the overall conviction rate for white defendants is driven up towards 0.37 (it is 0.34).

The opposite trend applies to cases against minority defendants. For minority defendants, the conviction rate is much higher among cases with white victims, than it is among cases with minority victims (0.4 vs 0.29). This large discrepancy suggests that there is a racial bias against minority defendants in cases where the victim is white - there may be overconviction of minority defendants due to the victim being white. However, the number of cases where the defendant is a minority and the victim is white is much smaller than the number of cases where the defendant is a minority and the victim is also a minority (25 vs 69). Since the higher conviction rate for minority defendants applies to a much smaller number of the cases against them, the overall conviction rate for minority defendants is driven down towards 0.29 (it is 0.32).

Therefore, while the conviction rate for minority defendants is higher than that of white defendants in both individual groups (minority victims vs white victims), the overall conviction rate for minority defendants is lower than that of white defendants. The overall conviction rates give us the misleading impression that there is no racial bias in the criminal justice system, when there does in fact appear to be racial bias in the system based on this data. Racial bias lowers conviction rates in the favor of white defendants, and raises conviction rates to the detriment of minority defendants. Overall conviction rates are misleading because of the sample sizes involved - the number of racially biased cases is much lower than the number of racially unbiased cases. Thus, the overall conviction rate for white defendants is higher than it might otherwise be, and the overall conviction rate for minority defendants is lower than it might otherwise be. In this dataset, this causes the overall conviction rate for minority defendants to be lower than that of white defendants, misleading us.

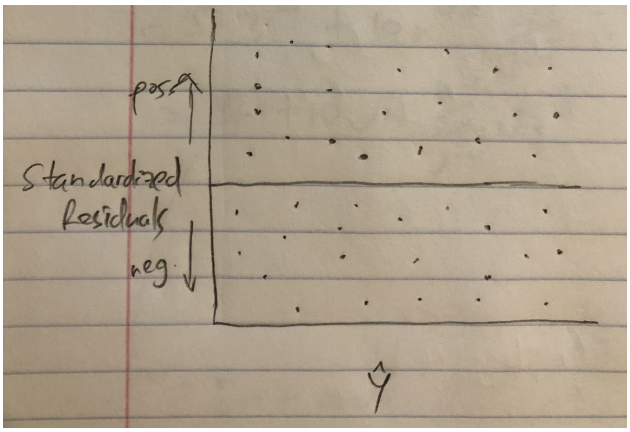

Figure 1: Constant variance, linearity

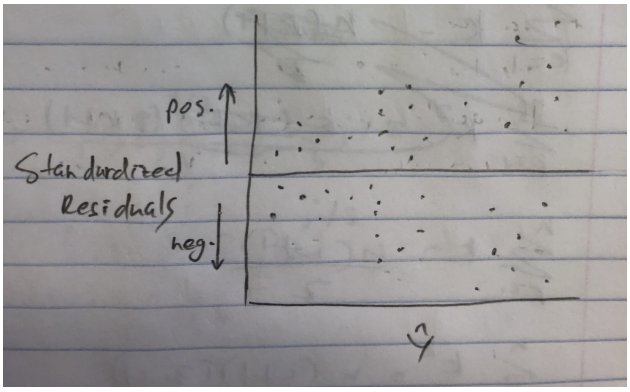

Figure 2: Non-constant variance, linearity

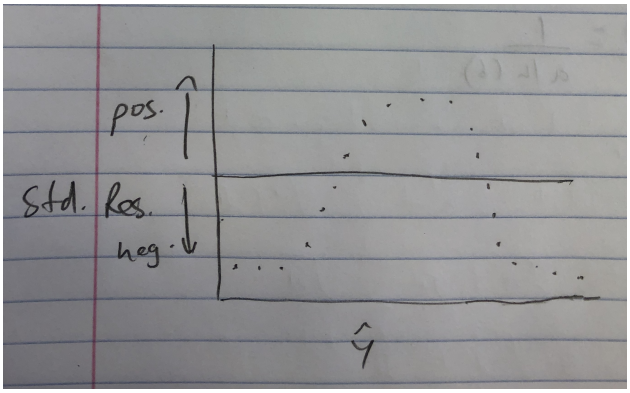

Figure 3: Constant variance, non-linearity

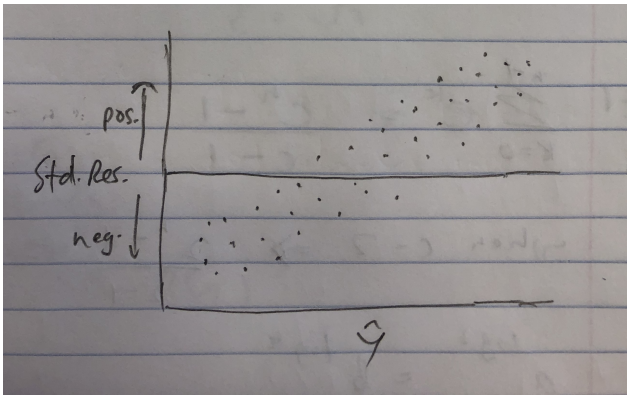

Figure 4: Non-constant variance, non-linearity