Bios 601 H4 V1

Jiewen Liu 260825295

November 3, 2021

14.1.a

14.1.a.i

This paper uses data to investigate potential sex bias in UCB's admission. Someone gives the seemingly reasonable conclusion that discrimination exists because the admission rate for women is obviously lower than for men. This paper explains why this statement is tenuous and how the true relationship could even be reversed: women are favored in the admission. The culprit in the dark suggested by the author is "uneven applications to different departments" between male and female students, which confounds the thing. What is confounding in this context? Think about a more intuitive example. Suppose there is a kind of drug against heart attacks. The dose suggested for older people is higher, and the increased dose intake within a certain range for each age group help avoid heart attacks. If we do not take age into our consideration, what we simply observe is that people having more doses are more likely to experience heart attacks compared to people having fewer doses. This result is reversed compared to the true one and is meaningless since the observed people having more dose are more likely the older people, who are intrinsically weaker. The background of people compared are different, and it takes us nowhere. Age is the culprit.

Back to this example, the latent less conspicuous culprit is that more women apply to the departments with generally lower admission rates while more apply to the easy ones. To get the correct result, we should zoom in and check the department by department to make sure we are comparing the same group of people except for the only difference in gender.

14.1.a.ii

```
-----------------------------------
# Men admitted 1. Men rejected 0
men_1 = c(512,353,120,138,53,22)
men_0 = c(313,207,205,279,138,351)

# Women admitted 1. Women rejected 0
wom_1 = c(89,17,202,131,94,24)
wom_0 = c(19,8,391,244,299,317)

# Men ad rate (by faculty). Women ad rate (by faculty)
p1 = men_1 /(men_0 + men_1)
p2 = wom_1 /(wom_0 + wom_1)
-----------------------------------
# Get the weights
w = (men_1+men_0+wom_1+wom_0) / 4526
w -> 0.2061423 0.1292532 0.2028281 0.1749890 0.1290323 0.1577552
-----------------------------------
# Get the weighted admission rate
weighted_p1 = sum(w * p1)
weighted_p1 -> 0.3873186 (If all men, the weighted admission rate)
weighted_p2 = sum(w * p2)
weighted_p2 -> 0.4299554 (If all women, the weighted admission rate)
```

```
-----------------------------------
# How many should be admitted if all men or women for 4526 people?
4526 * weighted_p1 -> 1753.004 (for men)
4526 * weighted_p2 -> 1945.978 (for women)
-----------------------------------
# Calculate the variance for admission rate for each men and women respectively
var_p1 = weighted_p1*(1-weighted_p1) / (men_1+men_0)
var_p2 = weighted_p2*(1-weighted_p2) / (wom_1+wom_0)
-----------------------------------
# Weight them (assume following the proportion given in the question)
weighted_var_p1 = sum(w^2 * var_p1)
weighted_var_p1 -> 0.000103285 (for men)
weighted_var_p2 = sum(w^2 * var_p2)
weighted_var_p2 -> 0.0003255095 (for women)
-----------------------------------
Construct the interval
c(weighted_p1 - 1.96 * sqrt(weighted_var_p1), weighted_p1 + 1.96 * sqrt(weighted_var_p1))
-> 0.3673993 0.4072379 (for men)
c(weighted_p2 - 1.96 * sqrt(weighted_var_p2), weighted_p2 + 1.96 * sqrt(weighted_var_p2))
-> 0.3945933 0.4653175 (for women)
-----------------------------------
```

How many should be admitted? (4526)	Men	Women
Number	1753	1945
Proportion	0.387	0.430
95% C.I. for the estimated proportion	(0.367,0.407)	(0.395,0.465)

14.1.a.iii

It is equivalent to scaling the things, which is achieve by multiplying 0.3333 for the uniform case in C&H's example for standardization. But here we use the 933, 585, . . . 769 / 4526. Also, it can be viewed as a form of IPW, which eliminates the propensity brought by gender within the same faculty by dividing P(Gender | Department) so creating 1:1 ratio within each department, mentioned in the note.

14.1.a.iv

```
-----------------------------------
# Difference in the proportions (female to male)
weighted_p_diff = weighted_p2 - weighted_p1
weighted_var_p_diff = weighted_var_p1 + weighted_var_p2
weighted_p_diff -> 0.0426368
c(weighted_p_diff - 1.96*sqrt(weighted_var_p_diff), weighted_p_diff + 1.96*sqrt(weighted_var_p_diff))
-> 0.002050386 0.083223209
-----------------------------------
# Ratio of the proportions (admitted)  (female to male)
weighted_ratio =  weighted_p2 / weighted_p1
var_log_ratio =  1/men_1 - 1/(men_1+men_0) + 1/wom_1 - 1/(wom_1+wom_0)
weighted_var_log_ratio = sum(w^2 * var_log_ratio)
weighted_ratio -> 1.110082
c(weighted_ratio /* exp(1.96*sqrt(weighted_var_log_ratio)))
-> 0.9887297 1.2463285
-----------------------------------
# Ratio of the proportions (rejected)  (female to male)
weighted_ratio =  (1-weighted_p2) / (1-weighted_p1)
weighted_ratio -> 0.9304095
c(weighted_ratio /( exp(1.96*sqrt(weighted_var_log_ratio)))
-> 0.8286987 1.0446038
```

```
----------------------------------
# Ratio of the odds (admitted) (female to male). Using 1753/4526 and 1945/4526
weighted_or = (weighted_p2 / (1-weighted_p2)) / (weighted_p1 / (1-weighted_p1))
weighted_var_log_or =  1/1946 + 1/(4526-1946) + 1/1753 + 1/(4526-1753)
weighted_or -> 1.193111
c(weighted_or /* exp(1.96*sqrt(weighted_var_log_or)))
-> 1.097089 1.297538
```

What weighted things we are interested in? (female to male)	Value	95% C.I.
Difference in the proportions	0.043	(0.002,0.083)
Ratio of the proportions (admitted)	1.11	(0.989,1.246)
Ratio of the proportions (rejected)	0.930	(0.829,1.045)
Ratio of the odds (admitted). **The calculation is based on this weighted pseudo-popliulation – 1753/4526 and 1946/4526.**	1.193	(1.097,1.298)

I like OR. We could flip (reciprocal) it to get the reverse quickly.

14.1.c

Hi Jim. I had done this in my style before you shared your codes. But I compare the (identity glm) results generated by codes. They are the same. Hope my codes are easy to read and save your time.

14.1.c – Missing Variable – Department

First, we could verify that if we do not put in the variable department and it confounds the things, we should observe the plus + or minus - for the coefficient of the gender term is reversed. Here we've figured out the gender should give a positive impact on the admission rate. Thus, it will be reversed to negative if the variable department is missing. I omit the case of log and logit. The identity case is given here.

```
-------------------------------------
# Put the data in.
admitted = c(men_1,wom_1)
rejected = c(men_0,wom_0)
dept = rep(c('A','B','C','D','E','F'),times=2)
gender = rep(c('M','W'),each=6)
-------------------------------------
# Identity link and the variable department not considered.
m_identity <-glm(cbind(admitted,rejected)~gender,family=binomial(link='identity'))
summary(m_identity)
------
Call:
glm(formula = cbind(admitted, rejected) ~ gender, family = binomial(link = "identity"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-16.7915   -4.7613   -0.4365    5.1025   11.2022

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.44519    0.00958  46.468   <2e-16 ***
genderW     -0.14164    0.01439  -9.845   <2e-16 ***

-------------------------------------
```

We see the beta for genderW is -0.14164! How about the results when log, logit and identity links are used respectively and the variable department is considered?

```
--------------------------------------
# Logit link (canonical one!)
m_logit <-glm(cbind(admitted,rejected)~dept+gender,family=binomial(link='logit'))
summary(m_logit)
------
Call:
glm(formula = cbind(admitted, rejected) ~ dept + gender, family = binomial(link = "logit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.58205    0.06899   8.436   <2e-16 ***
deptB       -0.04340    0.10984  -0.395    0.693
deptC       -1.26260    0.10663 -11.841   <2e-16 ***
deptD       -1.29461    0.10582 -12.234   <2e-16 ***
deptE       -1.73931    0.12611 -13.792   <2e-16 ***
deptF       -3.30648    0.16998 -19.452   <2e-16 ***
genderW      0.09987    0.08085   1.235    0.217
--------------------------------------
# Log link
m_log <- glm(cbind(admitted,rejected)~dept+gender,family=binomial(link='log'))
summary(m_log)
------
Call:
glm(formula = cbind(admitted, rejected) ~ dept + gender, family = binomial(link = "log"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.44863    0.02497 -17.969  < 2e-16 ***
deptB       -0.01519    0.03983  -0.381  0.70287
deptC       -0.67820    0.05571 -12.173  < 2e-16 ***
deptD       -0.68878    0.05714 -12.054  < 2e-16 ***
deptE       -1.01260    0.07889 -12.836  < 2e-16 ***
deptF       -2.35060    0.14537 -16.170  < 2e-16 ***
genderW      0.11605    0.04269   2.718  0.00656 **
--------------------------------------
# Identity link
m_identity <- glm(cbind(admitted,rejected)~dept+gender,family=binomial(link='identity'))
summary(m_identity)
------
Call:
glm(formula = cbind(admitted, rejected) ~ dept + gender, family = binomial(link = "identity"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.64283    0.01574  40.831   <2e-16 ***
deptB       -0.01096    0.02537  -0.432    0.666
deptC       -0.30141    0.02323 -12.974   <2e-16 ***
deptD       -0.31019    0.02344 -13.236   <2e-16 ***
deptE       -0.40059    0.02487 -16.108   <2e-16 ***
deptF       -0.58528    0.01857 -31.510   <2e-16 ***
genderW      0.01476    0.01297   1.139    0.255
--------------------------------------
```

14.1.c – Confidence Intervals

How about the C.I. mentioned in the question? We could do something like what we've done in the last question. Since we do not include the interaction term in our GLM, which means the same universal effect by gender across different departments, we could build C.I. out of them and compare with our previous results.

- **Identity Link**: The relationship between x and the proportion here is linear.
- **Log Link**: The relationship between x and the log ratio of the proportions here is linear.
- **Logit Link**: The relationship between x and the log odds ratio here is linear.

	Value	95% C.I.
Difference in the proportions (Identity)	0.03582	(0.020,0.052)
Log ratio of the proportions (admitted) (Log Link)	0.11605	(0.032,0.200)
Ratio of the proportions (admitted) (Log Link)	1.12305	(1.033,1.221)
Log ratio of the odds (admitted) (Logit Link).	0.09987	(-0.059,0.258)
Ratio of the odds (admitted) (Logit Link).	1.10503	(0.943,1.295)

I find somewhat different results estimated using our pseudo-population weights but the trend (+ or -) is kept. If my calculations are correct, I guess the reason is that the glm just finds the optimal estimate by MLE (Fisher Scoring) with the explicitly assumed distribution and link function, and the weights of estimating equations are still based on the data observed. Even if I put the pseudo-population data in, I could not obtain the results using simple standardization. Emmmmmmm..... Besides, how would glm with inverse weights applied look like? Not enough time.

14.2.a

14.2.a.i

The experiment introduced in this paper aims at studying how the stereotype accounts for the women's under-achievement in math. The experiment pipeline designed by the author is that: 1. Female test-takers are separated into 4 groups; 2. The math test ONE is given to measure their math performance; 3. A verbal test is in the middle which contains different reading articles exploring the different reasons that result in women's math under-achievement; 4. The math test TWO is given to re-measure how female test-takers math performance is influenced by the different ideas perceived about such a stereotype perceived. In the statistical analysis part, the authors find significantly different math performance before the ideas of stereotype indoctrinated in the verbal part. They make some adjustments to make sure a fair comparison will be given. The final result reveals that genetic and a normal (standard) stereotype about women in the math field make female test-takers more likely to be worse performance in the test.

14.2.a.ii

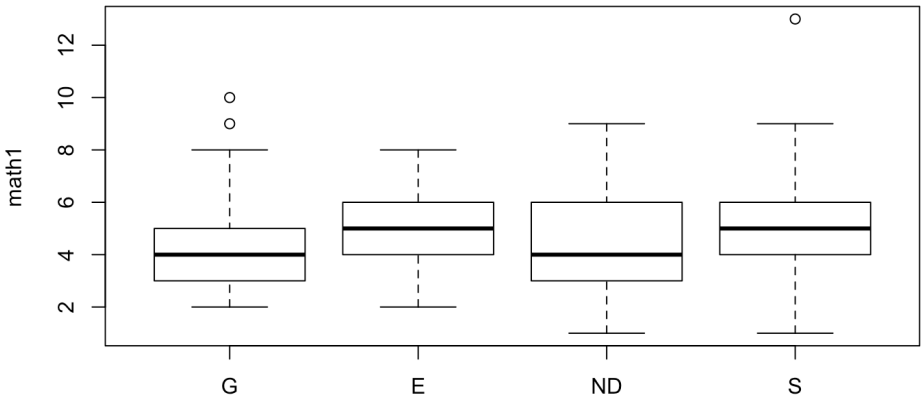
I follow what has been done in your R file and keep the same style. Some codes are omitted.

The similar question **(14.2b.ii)** that is not limited to these two groups is asked later. I give the results for all these four groups in this question.

The standardized mean difference (SMD is also a good benchmark especially when we need to adjust them based on many variables, using IPW or OW) from our data is:

	G	E	ND	S	SMD
n	28	27	27	29	
math1 (mean (SD))	4.46 (2.15)	5.22 (1.72)	4.52 (2.03)	5.41 (2.35)	0.285
math2 (mean (SD))	3.57 (2.06)	5.22 (1.83)	4.89 (2.39)	4.14 (2.07)	0.460

Let's have a look of the boxplot:



We see there exists much difference in the math ability before the exposure to the verbal section. We should adjust it before comparing them in math2.

Here are codes for the above stuff:

```
df = data.frame(c,math1,math2)
df$c = as.factor(df$c)
levels(df$c) = c('G','E','ND','S')

library(tableone)
library(knitr)
boxplot(math1~c,data = df)
tab = CreateTableOne(strata = "c", vars = c('math1','math2'),data = df, test = FALSE, smd = TRUE)
k = print(tab,smd = TRUE)
kable(k, format = "latex")
```

14.2.a.iii

Let me follow what has been done in Jim ur file.

```
----------------------
# Restrict to these two groups!
df = df[df$c=='ND'|df$c=='S',]
# Do the cutting and get weights
df$bin = cut(df$math1,breaks=c(0,3,4,6,13))
f = table(df$bin)
w=as.numeric(f)
w=w/sum(w)
----------------------
ybar =matrix(aggregate(df$math2,by = list(bin=df$bin,category=df$c), mean)$x ,4,2)
ybar
         [,1]     [,2]
[1,] 3.333333 3.250000
[2,] 4.000000 3.142857
[3,] 5.875000 4.166667
[4,] 7.750000 5.833333


# SS is the SST in each cell. We can ust it to  calculate the pooled variance.
n = table(df$bin,df$c)[,c(3,4)]
SS = (n-1) * matrix(aggregate(df$math2,by =list(bin=df$bin,category=df$c), var)$x ,4,2)
sigma.sq.hat = sum(SS)/( sum(n-1))
sigma.sq.hat -> 3.61942
```

```
----------------------
# The weighted mean
( w.ybar1 = sum ( ybar[,1] * w ) )
-> 5.184524 (if all ND)
( w.ybar2 = sum ( ybar[,2] * w ) )
-> 4.013818  (if all S)
( V.1 = sum ( (sigma.sq.hat/n[,1] ) * w^2 ) )
-> 0.1407424
( V.2 = sum ( (sigma.sq.hat/n[,2] ) * w^2 ) )
-> 0.134335
c(w.ybar1 - 1.96 * V.1, w.ybar1 + 1.96 * V.1)
-> 4.908669 5.460379
c(w.ybar2 - 1.96 * V.2, w.ybar2 + 1.96 * V.2)
-> 3.750521 4.277115
----------------------
```

What the Math2 should be if all 56? (Following 13, 13, 20, 10 distribution)	ND	S
Weighted Mean	5.18	4.01
Variance	0.387	0.430
95% C.I. for the Weighted Mean	(4.91,5.46)	(3.75,4.28)

14.2.a.iv

```
w.ybar1 - w.ybar2
-> 1.170706
V.1+V.2
-> 0.2750774
c(w.ybar1 - w.ybar2 - 1.96*sqrt(V.1+V.2),w.ybar1 - w.ybar2 + 1.96*sqrt(V.1+V.2))
-> 0.1427285 2.1986831
```

Weighted Difference (ND - S)	Variance	95% C.I.
1.17	0.275	(0.143,2.199)

14.2.b

14.2.b.i

```
----------------------
# We now care about all 4 groups.
df = data.frame(c,math1,math2)
df$c = as.factor(df$c)
levels(df$c) = c('G','E','ND','S')
summary(lm(math2 ~ c, data = df))
Call:
lm(formula = math2 ~ c, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1379 -1.5714 -0.2222  1.1111  5.4286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.5714     0.3959   9.021 8.36e-15 ***
cE            1.6508     0.5651   2.921  0.00425 **
cND           1.3175     0.5651   2.331  0.02160 *
cS            0.5665     0.5551   1.021  0.30975
```

```
----------------------
```

G is set as the reference group.

14.2.b.ii

For the plot and SMD, see 14.2.a.ii. We should notice the obvious difference across these four groups in math performance at the beginning. The comparison's fairness is impaired.

14.2.b.iii

```
----------------------
df$math1c = df$math1 - mean(df$math1)
summary(lm(math2 ~ c+math1c, data = df))
Call:
lm(formula = math2 ~ c + math1c, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6073 -1.4088  0.1348  0.9459  4.4481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.81582    0.33558  11.371  < 2e-16 ***
cE           1.23513    0.48014   2.572  0.01148 *
cND          1.28772    0.47612   2.705  0.00797 **
cS           0.04578    0.47410   0.097  0.92326
math1c       0.54841    0.08200   6.688 1.09e-09 ***
----------------------
```

According to the regression's result, as compared to G's mean, the (adjusted) estimates of the between group differences in Math2:

E - G	ND - G	S - G
1.235	1.288	0.046

.

14.2.b.iv

Such action is equivalent to adding the prior knowledge that the slope of math1 should be 1. I prefer not to do so.

In my opinion, it really depends. And in the general case, I'd like to be more conservative and say keep it on the right hand side. There could be some benefits by doing so. If Math1 and Math2 are different measurements, doing the subtraction could be problematic – what does the difference between Math1 and Math2 mean here?

Nowadays, new GRE introduces the notion of adaptive testing for increasing the repeatability, i.e. making people test results less fickle, Math1 is used to determine the difficulty index for the second test, and thus, the subtraction is not so interpretable here.

In addition, what if we cast doubt on some factors that influence people's performance? Even if they are good measurements on the same scale, people's performance on Math2 could be affected by a variety of factors. They may just feel tired, hungry or terrible if their good moods are ruined by hard questions in previous Math1 or Verbal section.

14.4

Age		P-Y	P	D	D/100,000PY†	VE
16-39						
	Unvaccinated	8669.5	35449	17	0.20	
	Vaccinated*	567.3	3040	0	0.00	vv.v
		9236.8				
40-59						
	Unvaccinated	1230.3	4803	33	2.68	
	Vaccinated*	629.8	3318	2	0.32	vv.v
		1860.1				
≥ 60						
	Unvaccinated	81.4	380	24	29.49	
	Vaccinated*	351.0	1952	24	6.84	vv.v
		432.4				
===		===	===	==	===	
ALL						
	Unvaccinated	9981.2	40632	74	xx.xx	
	Vaccinated*	1548.1	8310	26	yy.yy	VV.V
		11529.3				

14.4.i

Let's do a quick check. $17/8669.5 * 100 \approx 0.20$. It should be D/100PY.

14.4.ii

$$xx.xx = 74/9981.2 * 100 \approx 0.74$$
$$yy.yy = 26/1548.1 * 100 \approx 1.68$$
$$\hat{VE} = 100 * (1 - 1.68/0.74) = -127.03$$
$$\hat{SE} = \sqrt{1/74 + 1/26} \approx 0.228$$
$$95\%\ for\ VE \rightarrow 100 * (1 - 1.68/0.74 * /exp(1.96)) \approx (-255.93, -45.21)$$

14.4.iii

The VE is meaningless here. The VE is a benchmark for quantifying the effectiveness of the vaccine. Some factors need to be controlled. Why? An unfair comparison makes no sense. You cannot say this vaccine is useless because you observed a lower death rate in the unvaccinated younger group of people and a higher death rate in the vaccinated older group of people. Older people are at higher risks. Now without any control, the average age of the vaccinated group is higher than one of the unvaccinated group, which means the vaccinated group's overall health conditions are intrinsically worse, and you should anticipate the higher death rate of this group of people even if they are not vaccinated. If you are on the wrong side, what you are studying is nothing, and C.I. takes you nowhere.

14.4.iv

```
----------------------
a = c(9236.8,1860.1,432.4)
w = a / sum(a)
----------------------
# Unvaccinated
d1 = c(17,33,24)
N1 = c(8669.5,1230.3,81.4)
p1 = d1/N1
weighted_death_rate1 = sum(p1*a) / sum(a) * 1000
var_death_rate1 = d1/N1^2 * 1000^2
weighted_var_death_rate1 = sum(w^2 * var_death_rate1)

sum(p1*a)
```

```
-> 195.4943 (Number)
weighted_death_rate1
-> 16.9563 (Rate per 1000 PY)
weighted_var_death_rate1
-> 5.807477 (Variance)
c(weighted_death_rate1 -+ 1.96*sqrt(weighted_var_death_rate1))
-> 12.23296 26.21406 (C.I)
----------------------
# Vaccinated
d2 = c(0,2,24)
N2 = c(567.3,629.8,351.0)

# Repeat the codes. Get the following
-> 35.47277 (Number)
-> 3.076749 (Rate per 1000 PY)
-> 0.4052551 (Variance)
-> 1.829020 5.522299 (C.I)

# However, it should be problematic. The number of death is fine.
# But the variance and thus the C.I. should be problematic due to 0 death observed in Age 16-39 group.
# A simple fix is to plug in a simple and seemingly reasonable value. How about 1?
d2 = c(1,2,24)
N2 = c(567.3,629.8,351.0)
-> 51.7548 (Number)
-> 4.488981 (Rate per 1000 PY)
-> 2.399652 (Variance)
-> 1.452782 10.439931 (C.I)
----------------------
```

How many deaths are expected ? (11529.3 PY)	Unvaccinated	Vaccinated (* Simple Fix)
Number	195	35 (*52)
Rate (per 1,000 PY)	16.96	3.08 (*4.49)
Variance for rates (per 1,000 PY)	5.81	0.41 (*2.40)
95% C.I. for the estimated rate	(12.23,26.21)	(1.83,5.52) (*1.45,10.44)

For the vaccinated group, the zero numerator is a big issue, especially for the variance. The normal approximation won't be a reasonable choice. How about having an exact interval and picking a value from it?): My simple fix is not good also. Need more time to think about this.

14.4.v

```
weighted_death_diff = (weighted_death_rate1 - weighted_death_rate2)
weighted_death_diff
-> 13.87955
weighted_var_death_diff = weighted_var_death_rate1 + weighted_var_death_rate2
weighted_var_death_diff
-> 2.804907
c(weighted_death_diff-+1.96*sqrt(weighted_var_death_diff))
-> 10.59697 17.16213 (C.I.)
```

Weighted Difference (Unvac - Vac) Rate per 1,000 PY	Variance (SD)	95% C.I.
13.88	2.80 (1.67)	(10.60,17.16)

Let's have a look of glms.

```
----------------------
# crude
```

```
summary(glm(D~ -1+PY+V.PY, family=poisson(link="identity") ) )
Call:
glm(formula = D ~ -1 + PY + V.PY, family = poisson(link = "identity"))

Deviance Residuals:
     1       2       3       4       5       6
-7.024  -4.365   6.092  -3.239  11.402   5.584

Coefficients:
      Estimate Std. Error z value Pr(>|z|)
PY   0.0074139  0.0008619   8.602  < 2e-16 ***
V.PY 0.0093808  0.0034046   2.755  0.00586 **
---------------------
```

Crude One: If we focus merely on the data collected from these two group sand make no adjustment, the vaccination seems to increase the death rate.

```
---------------------
summary(glm(D~ -1+ S.1.PY + S.2.PY + S.3.PY , family=poisson(link="identity") ) )
Call:
glm(formula = D ~ -1 + S.1.PY + S.2.PY + S.3.PY, family = poisson(link = "identity"))

Deviance Residuals:
      1        2        3        4        5        6
 0.2586  -1.4451   1.9231  -3.5474   4.1182  -2.5822

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
S.1.PY 0.0018405  0.0004464   4.123 3.74e-05 ***
S.2.PY 0.0188162  0.0031805   5.916 3.30e-09 ***
S.3.PY 0.1110083  0.0160227   6.928 4.26e-12 ***
---------------------
```

Separate Age Group: I think the main takeaway from this glm is that we should notice the intrinsic different death rates across these study groups categorized by ages. Suspect the age play the role of the confounder that renders the effect of the vaccine seemingly harmful?

```
---------------------
summary(glm(D~ -1+ S.1.PY + S.2.PY + S.3.PY + V.PY ,family=poisson(link="identity")))
Error: no valid set of coefficients has been found: please supply starting values.
summary(glm(D~ -1+ S.1.PY + S.2.PY + S.3.PY + V.PY , start = c(0.01,0.02,0.03,0.01),family=poisson(link="i
Call:
glm(formula = D ~ -1 + S.1.PY + S.2.PY + S.3.PY + V.PY, family = poisson(link = "identity"),
    start = c(0.01, 0.02, 0.03, 0.01))

Deviance Residuals:
      1        2        3        4        5        6
-0.3378  -0.0001   1.8883  -3.2429   4.0847  -2.5248

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
S.1.PY  0.002126   0.000494   4.304 1.68e-05 ***
S.2.PY  0.018944   0.003130   6.052 1.43e-09 ***
S.3.PY  0.112037   0.015976   7.013 2.33e-12 ***
V.PY   -0.002126   0.000494  -4.304 1.68e-05 ***
---------------------
```

I try to make this run. Now the effect of vaccine is positive, reducing the rate of infection (-0.002126 the negative beta). What I get is that the reduction in rate is 13.88 per 1,000 PY. Here it is 21.26.

My thoughts here are the same as what I give in 14.1.c. Yet, the good thing is that the direction of changes of vaccine is consistent as revealed in both GLM and pseudo-population.

14.4.vi

To avoid NaN in calculating the ratio, I the simple fix one, using 1 instead of 0 here for the observed number of death.

```
----------------------
weighted_death_ratio = (weighted_death_rate1 / weighted_death_rate2)
weighted_death_ratio
-> 3.777317
var_log_death_rate1 = 1/var_death_rate1 * p1 * 1000 (scale p1 to 1,000PY)
var_log_death_rate2 = 1/var_death_rate2 * p2 * 1000 (scale p1 to 1,000PY)
var_log_death_ratio = var_log_death_rate1 + var_log_death_rate2
weighted_var_log_death_ratio = sum(w^2 * var_log_death_ratio)
c(weighted_death_ratio/*1.96*sqrt(weighted_var_log_death_ratio))
-> 1.855192 7.126906
----------------------
```

Weighted Ratio (Unvac / Vac)	95% C.I.
3.78	(1.86,7.13)

14.4.vii

```
----------------------
# crude
summary(fit)
Call:
glm(formula = D ~ Vaccinated + offset(log(PY)), family = poisson)

Deviance Residuals:
     1       2       3       4       5       6
-7.024  -4.365   6.092  -3.239  11.402   5.584

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.9044     0.1162 -42.192  < 2e-16 ***
Vaccinated    0.8177     0.2280   3.587 0.000335 ***
----------------------
```

The story is similar. It is confounding. But the interesting thing here is that the effect of vaccination is quantified almost exactly in these two crude models. Since they all ignore the age and thus have a single baseline, it is easy. Now turn the multiplicative thing to additive one:

$$\textbf{Effect of vaccination for data in hand} = exp(0.8177) * exp(-4.9044) - exp(-4.9044) = 0.009380674$$

$$\textbf{Crude glm with identity link shown in 14.4.v gives the coefficient for V.PY :}\ \hat{\beta} = 0.0093808$$

```
----------------------
# incl. age
summary(fit)
Call:
glm(formula = D ~ as.factor(Stratum) + Vaccinated + offset(log(PY)),
    family = poisson)

Deviance Residuals:
       1         2         3         4         5         6
 0.05457  -0.66931   0.23114  -0.77772  -0.30670   0.32007
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.2476     0.2426 -25.757   <2e-16 ***
as.factor(Stratum)2   2.5886     0.2964   8.735   <2e-16 ***
as.factor(Stratum)3   5.0883     0.3054  16.658   <2e-16 ***
Vaccinated           -1.5894     0.2568  -6.190    6e-10 ***
----------------------
```

After addressing the confounder age, we see the vaccination very significantly reduce the death rates! If we want to go back to additive model, it is much more messier here.

14.4.viii

I've seen some of them before. But, if it is possible, I hope I could figure out the pseudo-population behind it that gives the exact match with the glm coefficients.

14.4.ix

Everything is good.

Some thoughts on 14.2:

The core of this paper is about stereotype/impression related stuff, and this paper's statements contain two things: 1."different availability of aptitude at the high end". 2."What President Summers perhaps intended to be a provocative call for more empirical research on biological bases of achievement may inadvertently exacerbate ***." I am not against the experiments and the conclusion based on this result suggested by the author. But, I don't see the necessary direct/indirect relationship between it and Summers' one for two reasons.

1. A clear point stressed in Summers's paper is variability. In a rarefied high-end level, males outcompete females. The paper is not talking exactly the same group of people in Summers's paper. Will the fiercely intelligent one-in-a-million females' performance be hampered by such a stereotype? I guess they are already smart enough and know themselves well. More researches should be done if Summers's view is questioned. GRE is not a test for differentiating genius from normal people but tests like the Olympics test does, which means **standardization** is not possible here.

2. I think it is social media and some other people's duty to better understand, interpret and broadcast their ideas. I find much similar criticism online. For example, Swarthmore College's Magazine says: *"those who completely missed Summers's point about variability, including Los Angeles Times, David Gelernter, a computer scientist at Yale and occasional conservative commentator, wrote: "[Summers] suggested that, on average, maybe women are less good than men at science. . . ." Well, no, he didn't. But in the public debate, that is how his statement was interpreted."*

So I think Summers's work should not be that blameworthy or suffer from any implicative criticism in this way.