3036274164\_q1

PENG, Xin

2024-04-15

knitr::opts\_chunk$set(echo = TRUE)  
setwd("/Users/amypxx/Library/CloudStorage/OneDrive-TheUniversityofHongKong-Connect/adv stat/final exam")  
library(openxlsx)  
library(ggplot2)  
library(survival)

# Q1 A

q1 <- read.csv('exam\_question1.csv')  
head(q1)

## id male age heart comorbid covid anxiety  
## 1 1 1 35 0 0 1 0  
## 2 2 0 28 1 1 1 0  
## 3 3 1 35 0 0 1 0  
## 4 4 0 19 0 0 1 0  
## 5 5 0 22 0 0 1 0  
## 6 6 1 22 0 0 1 0

summary(q1)

## id male age heart   
## Min. : 1.0 Min. :0.000 Min. :18.00 Min. :0.0000   
## 1st Qu.: 750.8 1st Qu.:0.000 1st Qu.:39.00 1st Qu.:0.0000   
## Median :1500.5 Median :1.000 Median :52.00 Median :0.0000   
## Mean :1500.5 Mean :0.541 Mean :51.38 Mean :0.2017   
## 3rd Qu.:2250.2 3rd Qu.:1.000 3rd Qu.:64.00 3rd Qu.:0.0000   
## Max. :3000.0 Max. :1.000 Max. :95.00 Max. :1.0000   
## comorbid covid anxiety   
## Min. :0.000 Min. :0.0 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.:0.0 1st Qu.:0.0000   
## Median :0.000 Median :0.5 Median :0.0000   
## Mean :0.402 Mean :0.5 Mean :0.1333   
## 3rd Qu.:1.000 3rd Qu.:1.0 3rd Qu.:0.0000   
## Max. :2.000 Max. :1.0 Max. :1.0000

lm1 <- glm(anxiety ~ covid + male + age + heart + comorbid, family = "binomial", data = q1)  
  
ci <- exp(confint(lm1))

or <- exp(coef(lm1))  
  
results <- cbind(  
 Variable = rownames(ci),  
 OR = round(or,2),  
 round(ci,2)  
)  
  
knitr::kable(results, caption = "OR and 95% CI", digits = 2)

OR and 95% CI

|  | Variable | OR | 2.5 % | 97.5 % |
| --- | --- | --- | --- | --- |
| (Intercept) | (Intercept) | 0.01 | 0.01 | 0.02 |
| covid | covid | 1.6 | 1.28 | 2 |
| male | male | 1.17 | 0.94 | 1.46 |
| age | age | 1.04 | 1.03 | 1.05 |
| heart | heart | 1.31 | 1.02 | 1.68 |
| comorbid | comorbid | 1.07 | 0.92 | 1.23 |

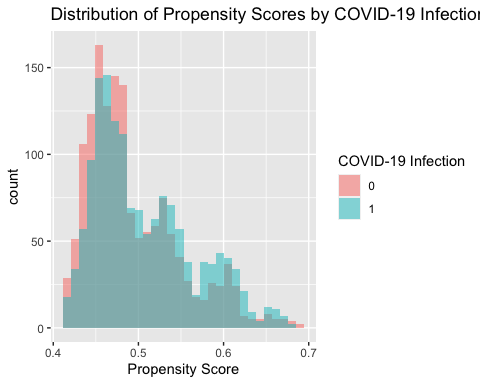
COVID-19 infections significantly increase the risk of anxiety disorder adjusting for various confounders.

# Q1 B

psq1 <- glm(covid ~ male + age + heart + comorbid, family = "binomial", data = q1)  
  
summary(psq1)

##   
## Call:  
## glm(formula = covid ~ male + age + heart + comorbid, family = "binomial",   
## data = q1)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.431475 0.128471 -3.359 0.000784 \*\*\*  
## male 0.052912 0.073794 0.717 0.473358   
## age 0.005158 0.002351 2.194 0.028213 \*   
## heart 0.200203 0.094209 2.125 0.033579 \*   
## comorbid 0.244814 0.053941 4.539 5.66e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4158.9 on 2999 degrees of freedom  
## Residual deviance: 4119.3 on 2995 degrees of freedom  
## AIC: 4129.3  
##   
## Number of Fisher Scoring iterations: 4

q1$ps <- predict(psq1, type = "response")  
  
p1 <- ggplot(q1, aes(x = ps, fill = as.factor(covid))) +  
 geom\_histogram(alpha = 0.5, position = "identity", bins = 30) +  
 labs(title = "Distribution of Propensity Scores by COVID-19 Infection",  
 x = "Propensity Score",  
 fill = "COVID-19 Infection")  
  
p1



# Exponentiate coefficients to get odds ratios  
exp(coef(psq1))

## (Intercept) male age heart comorbid   
## 0.6495501 1.0543371 1.0051716 1.2216504 1.2773839

**Female, older people, people with heart disease and people with more comorbidities are more likely to get COVID-19 infection.**

**Common support in the range of 0.4 to 0.7.**

# Q1 C

## stratification by propensity score quintiles  
  
ps.boundary <- quantile(q1$ps, 0:5/5)   
ps.boundary

## 0% 20% 40% 60% 80% 100%   
## 0.4161432 0.4519327 0.4711594 0.4997593 0.5460530 0.6902536

q1$psq <- cut(q1$ps, ps.boundary, right=F, include.lowest=T, label=1:5)  
head(q1$psq)

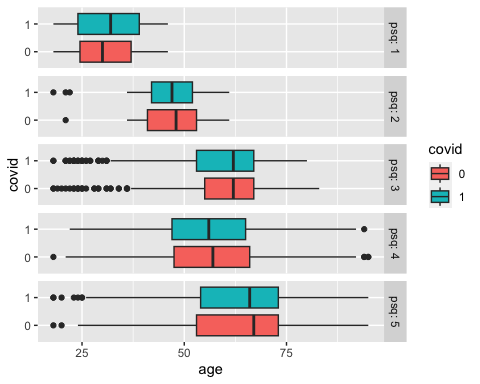
## [1] 1 4 1 1 1 1  
## Levels: 1 2 3 4 5

head(q1)

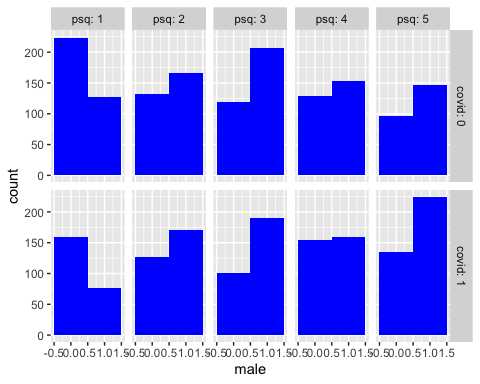
## id male age heart comorbid covid anxiety ps psq  
## 1 1 1 35 0 0 1 0 0.4506554 1  
## 2 2 0 28 1 1 1 0 0.5394115 4  
## 3 3 1 35 0 0 1 0 0.4506554 1  
## 4 4 0 19 0 0 1 0 0.4173971 1  
## 5 5 0 22 0 0 1 0 0.4211649 1  
## 6 6 1 22 0 0 1 0 0.4341151 1

q1$covid<-as.factor(q1$covid)

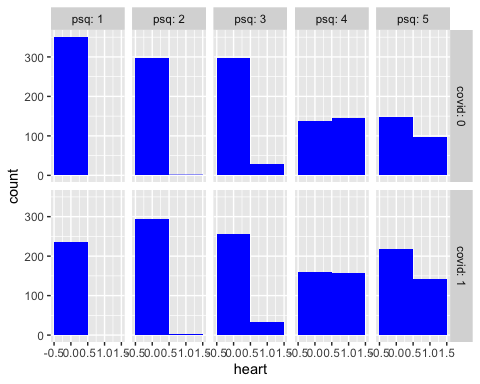
### age by PS strata  
ggplot(q1, aes(x=covid, y=age)) +   
 geom\_boxplot(aes(fill=covid)) + facet\_grid(psq ~ ., labeller=label\_both) +   
 coord\_flip()



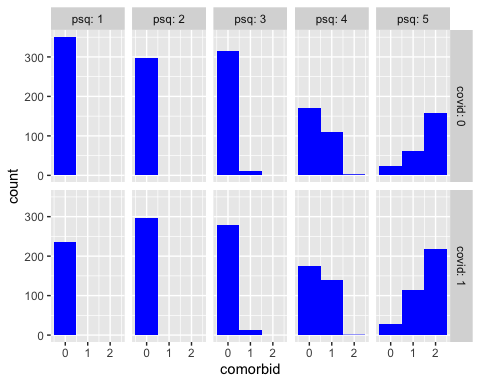
# sex by PS strata  
ggplot(q1, aes(x=male)) +   
 geom\_histogram(binwidth=1, fill='blue') +   
 facet\_grid(covid ~ psq, labeller=label\_both)



# heart disease by PS strata  
ggplot(q1, aes(x=heart)) +   
 geom\_histogram(binwidth=1, fill='blue') +   
 facet\_grid(covid ~ psq, labeller=label\_both)



# comorbidity by PS strata  
ggplot(q1, aes(x=comorbid)) +   
 geom\_histogram(binwidth=1, fill='blue') +   
 facet\_grid(covid ~ psq, labeller=label\_both)



**Patient characteristics are more balanced within each propensity score strata.**

# Q1 D

Conditional Independence: This assumption presumes that all known confounders are included in the model. However, the existence of unobserved confounders cannot be completely excluded.

Common Support: The overlap observed in the propensity score distribution between the treatment and control groups suggests that this assumption is reasonably met.

# Q1 E

mean.cp <- aggregate(anxiety~covid+psq, data=q1,   
 FUN=mean) # infection rates   
count.cp <- aggregate(anxiety~covid+psq, data=q1,   
 FUN=length) # number of infections  
cbind(mean.cp, count.cp$anxiety)

## covid psq anxiety count.cp$anxiety  
## 1 0 1 0.02564103 351  
## 2 1 1 0.08016878 237  
## 3 0 2 0.09731544 298  
## 4 1 2 0.10774411 297  
## 5 0 3 0.11692308 325  
## 6 1 3 0.21649485 291  
## 7 0 4 0.10954064 283  
## 8 1 4 0.16825397 315  
## 9 0 5 0.19341564 243  
## 10 1 5 0.21944444 360

n.psq <- as.numeric(table(q1$psq))  
n.psq

## [1] 588 595 616 598 603

covid.strata <- mean.cp$anxiety[mean.cp$covid==1]-  
 mean.cp$anxiety[mean.cp$covid==0]   
  
covid.est <- sum(covid.strata\*n.psq)/sum(n.psq)   
  
## Estimated overall effect  
  
round(covid.est,2) # Weighted Mean of the Stratum-specific effect

## [1] 0.05

var.cp <- aggregate(anxiety~covid+psq, data=q1, FUN=var)  
cbind(var.cp, count.cp$covid)

## covid psq anxiety count.cp$covid  
## 1 0 1 0.02505495 0  
## 2 1 1 0.07405421 1  
## 3 0 2 0.08814092 0  
## 4 1 2 0.09646010 1  
## 5 0 3 0.10357075 0  
## 6 1 3 0.17020974 1  
## 7 0 4 0.09788738 0  
## 8 1 4 0.14039025 1  
## 9 0 5 0.15665068 0  
## 10 1 5 0.17176571 1

covid.var <- sum((var.cp$anxiety/count.cp$anxiety)\*rep(n.psq, each=2)^2)/sum(n.psq)^2   
  
## Variance of estimated effect  
covid.var

## [1] 0.0001540265

## 95% CI  
cl<-covid.est-1.96\*sqrt(covid.var)  
cu<-covid.est+1.96\*sqrt(covid.var)  
  
print(paste('95%CI:',round(cl,2),",",round(cu,2)))

## [1] "**95%CI: 0.03 , 0.07**"