Problem Set 3

psid <- read.csv("psid.csv", header = TRUE)  
attach(psid)  
library(lmtest)

## Warning: package 'lmtest' was built under R version 4.0.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(sandwich)

## Warning: package 'sandwich' was built under R version 4.0.3

#### 1.

psid$u74 <- as.numeric(re74 == 0)  
psid$u75 <- as.numeric(re75 == 0)

#### 2.

mean\_stat <- aggregate(psid, by = list(treat), mean)  
sd\_stat <- aggregate(psid, by = list(treat), sd)  
stat <- matrix(rep(0, 60), 10, 6)  
stat[,1] <- t(mean\_stat[2, -c(1, 2, 3)])   
stat[,3] <- t(mean\_stat[1, -c(1, 2, 3)])   
stat[,2] <- t(sd\_stat[2, -c(1, 2, 3)])   
stat[,4] <- t(sd\_stat[1, -c(1, 2, 3)])   
stat[,5] <- stat[,1] - stat[,3]  
stat[,6] <- stat[,5]/sqrt((stat[,2]^2 + stat[,4]^2)/2)  
stat

## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] 25.81621622 7.1550193 34.85060241 10.4407572 -9.03438619 -1.0094327  
## [2,] 10.34594595 2.0106503 12.11686747 3.0824346 -1.77092152 -0.6805170  
## [3,] 0.18918919 0.3927217 0.86626506 0.3404357 -0.67707587 -1.8423347  
## [4,] 0.84324324 0.3645579 0.25060241 0.4334470 0.59264083 1.4798025  
## [5,] 0.05945946 0.2371244 0.03253012 0.1774389 0.02692934 0.1285908  
## [6,] 0.70810811 0.4558666 0.30522088 0.4605934 0.40288722 0.8792132  
## [7,] 1.12260108 2.2525173 4.69380530 3.1069321 -3.57120422 -1.3160567  
## [8,] 1.20773243 2.0587875 4.45289530 3.0566155 -3.24516287 -1.2453102  
## [9,] 0.70810811 0.4558666 0.08634538 0.2809298 0.62176273 1.6420964  
## [10,] 0.60000000 0.4912274 0.10000000 0.3000603 0.50000000 1.2284224

#### 3.

reg1 <- lm(re78~treat, data = psid)  
reg2 <- lm(re78~treat + re74 + re75 + u74 + u75, data = psid)  
reg3 <- lm(re78~treat + re74 + re75 + u74 + u75 + education + married + black + hispanic + age + nodegree, data = psid)  
  
coeftest(reg1, vcov = sandwich)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.55392 0.31167 69.157 < 2.2e-16 \*\*\*  
## treat -15.20478 0.65567 -23.190 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(reg2, vcov = sandwich)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.13496 0.81480 25.9388 < 2.2e-16 \*\*\*  
## treat -6.25361 1.02253 -6.1158 1.101e-09 \*\*\*  
## re74 0.20626 0.10709 1.9261 0.0542 .   
## re75 0.19477 0.10814 1.8011 0.0718 .   
## u74 -2.63154 1.83473 -1.4343 0.1516   
## u75 -11.89263 1.67499 -7.1001 1.592e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coeftest(reg3, vcov = sandwich)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.212302 2.459859 -4.5581 5.395e-06 \*\*\*  
## treat 3.784736 1.233011 3.0695 0.002166 \*\*   
## re74 0.220364 0.094966 2.3205 0.020391 \*   
## re75 0.161437 0.096145 1.6791 0.093251 .   
## u74 -3.929826 1.808688 -2.1727 0.029887 \*   
## u75 -11.451390 1.646183 -6.9563 4.379e-12 \*\*\*  
## education 1.762652 0.143223 12.3070 < 2.2e-16 \*\*\*  
## married 3.521446 0.669962 5.2562 1.587e-07 \*\*\*  
## black -3.215248 0.546280 -5.8857 4.460e-09 \*\*\*  
## hispanic 1.060721 1.583099 0.6700 0.502898   
## age 0.251149 0.028425 8.8354 < 2.2e-16 \*\*\*  
## nodegree 0.352852 0.806019 0.4378 0.661588   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The estimate is not robust to the specification, estimated treatment effect is not consistent across specifications.

#### 4.

**i)**

Treatment group mean: (24 + 61)/2 = 42.5

Control group mean: (42 + 33 + 46 + 37)/4 = 39.5

Observed statistic value = 42.5 - 39.5 = 3

**ii)**

Treatment group mean: (24 + 42)/2 = 33

Control group mean: (61 + 33 + 46 + 37)/4 = 44.25

Test statistic = 33 - 44.25 = -11.25

**iii)**

= 15 different ways of assigning treatment and control

**iv)**

{1, 1, 0, 0, 0, 0} = 3

{1, 0, 1, 0, 0, 0} = -11.25

{1, 0, 0, 1, 0, 0} = -18

{1, 0, 0, 0, 1, 0} = -8.25

{1, 0, 0, 0, 0, 1} = -15

{0, 1, 1, 0, 0, 0} = 16.5

{0, 1, 0, 1, 0, 0} = 9.75

{0, 1, 0, 0, 1, 0} = 19.5

{0, 1, 0, 0, 0, 1} = 12.75

{0, 0, 1, 1, 0, 0} = -4.5

{0, 0, 1, 0, 1, 0} = 5.25

{0, 0, 1, 0, 0, 1} = -1.5

{0, 0, 0, 1, 1, 0} = -1.5

{0, 0, 0, 1, 0, 1} = -8.25

{0, 0, 0, 0, 1, 1} = 1.5

distribution <- as.numeric(list(3, -11.25, -18, -8.25, -15, 16.5, 9.75, 19.5, 12.75, -4.5, 5.25, -1.5, -1.5, -8.25, 1.5))  
hist(distribution, breaks = 30)

**v)**

Proportion = 6/15 = 0.4

**vi)**

2-sided p-value = 0.4\*2 = 0.8

#### 5.

You cannot estimate the treatment effect correctly by comparing the mean difference since the treatment is not randomly assigned. This makes the presence of confounders and selection bias a serious issue, which would lead to an inaccurate ATE. Some possible confounders are:

* Income/socioeconomic status: Individuals that have a low income might be more likely to smoke. They are unable to afford high quality healthcare, which could lead to a higher disease rate.
* Age: Older individuals might be more likely to smoke. They have a weaker immune system, which could lead to a higher disease rate.
* Occupation: Individuals that work in high-risk jobs (i.e. construction) might be more likely to smoke. A higher rate of injury while working could lead to a higher disease rate.

#### 6.

**i)**

where equals the selection bias

**ii)**

Selection bias can go away under the unconfoundedness assumption.

#### 7.

**i)**

Old: ATE = 30 - 19 = 11

Mid-aged: ATE = 26 - 21 = 5

Young: ATE = 22 - 18 = 4

**ii)**

Old: 25 + 140 = 165

Mid-aged: 90 + 170 = 260

Young: 235 + 155 = 390

Total: 165 + 260 + 390 = 815

ATE = (165/815)\*11 + (260/815)\*5 + (390/815)\*4 = 5.736196