Week 3_Exercise

Fenqi Guo

2023-01-24

```
library("haven")
library("car")
library("ggplot2")
library("dplyr")
library(stargazer)

data <- read_dta("Teaching_Dataset.dta")
```

Question 1A

```
t.test(data$Teacher_Pay~data$Independent_School_Dummy)
```

```
##
##  Welch Two Sample t-test
##
## data:  data$Teacher_Pay by data$Independent_School_Dummy
## t = -0.93134, df = 57.356, p-value = 0.3556
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -5444.614  1987.491
## sample estimates:
## mean in group 0 mean in group 1
##        35417.83        37146.39
```

Question 1B

```
#Preparing data
data <- na.omit(data)
data$Degree_Sub <- NA
for (i in 1:nrow(data)){
  if (data$Degree_Sub_1[i]==1){data$Degree_Sub[i]="Arts"}
  if (data$Degree_Sub_2[i]==1){data$Degree_Sub[i]="Biology"}
  if (data$Degree_Sub_3[i]==1){data$Degree_Sub[i]="Business"}
  if (data$Degree_Sub_4[i]==1){data$Degree_Sub[i]="Combined"}
  if (data$Degree_Sub_5[i]==1){data$Degree_Sub[i]="Economics"}
  if (data$Degree_Sub_6[i]==1){data$Degree_Sub[i]="Education"}
```

```
  if (data$Degree_Sub_7[i]==1){data$Degree_Sub[i]="Engineering"}
  if (data$Degree_Sub_8[i]==1){data$Degree_Sub[i]="English"}
  if (data$Degree_Sub_9[i]==1){data$Degree_Sub[i]="Geography"}
  if (data$Degree_Sub_10[i]==1){data$Degree_Sub[i]="Law"}
  if (data$Degree_Sub_11[i]==1){data$Degree_Sub[i]="Linguistics"}
  if (data$Degree_Sub_12[i]==1){data$Degree_Sub[i]="Math"}
  if (data$Degree_Sub_13[i]==1){data$Degree_Sub[i]="Medicine"}
  if (data$Degree_Sub_14[i]==1){data$Degree_Sub[i]="Other"}
  if (data$Degree_Sub_15[i]==1){data$Degree_Sub[i]="Physics"}
  if (data$Degree_Sub_16[i]==1){data$Degree_Sub[i]="Politics"}
  if (data$Degree_Sub_17[i]==1){data$Degree_Sub[i]="Psychology"}
  if (data$Degree_Sub_18[i]==1){data$Degree_Sub[i]="Sociology"}
}
data <- data[, -9:-27]
data$Degree_Class <- NA
for (i in 1:nrow(data)){
  if (data$Degree_Class_1[i]==1){data$Degree_Class[i]="Class 1"}
  if (data$Degree_Class_2_1[i]==1){data$Degree_Class[i]="Class 2:1"}
  if (data$Degree_Class_2_2[i]==1){data$Degree_Class[i]="Class 2:2"}
  else {data$Degree_Class[i]="Other"}
}
data <- data[, -10:-13]
data$School_Taught <- NA
for (i in 1:nrow(data)){
  if (data$Schools_taught_1[i]==1){data$School_Taught[i]="1"}
  if (data$Schools_taught_2_3[i]==1){data$School_Taught[i]="2-3"}
  if (data$Schools_taught_4_5[i]==1){data$School_Taught[i]="4-5"}
  if (data$Schools_taught_6_7[i]==1){data$School_Taught[i]="6-7"}
  if (data$Schools_taught_8_9[i]==1){data$School_Taught[i]="8-9"}
  else {data$School_Taught[i]="10"}
}
data <- data[, -11:-16]
#Regression model
model1 <- lm(Teacher_Pay ~ Independent_School_Dummy + Age_Years + Male_Teacher +
               Ethnic_White + Russell_Group + ofsted_Good_Dum + Subject_Cat_STEM +
               Teach_Exp_dum_0_3 + Teach_Exp_dum_4_10 + PT +
               as.factor(Degree_Sub), data)
summary(model1)
```

```
##
## Call:
## lm(formula = Teacher_Pay ~ Independent_School_Dummy + Age_Years +
##     Male_Teacher + Ethnic_White + Russell_Group + ofsted_Good_Dum +
##     Subject_Cat_STEM + Teach_Exp_dum_0_3 + Teach_Exp_dum_4_10 +
##     PT + as.factor(Degree_Sub), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19394.9  -5417.0   -464.2   4855.7  26893.5
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      37209.83    4047.48   9.193  < 2e-16 ***
```

```
## Independent_School_Dummy           3501.57    1750.12   2.001 0.046398 *
## Age_Years                            14.88      59.65   0.249 0.803163
## Male_Teacher                       3908.73    1108.22   3.527 0.000492 ***
## Ethnic_White                       2905.58    2255.33   1.288 0.198713
## Russell_Group                     -1572.93    1202.63  -1.308 0.191993
## ofsted_Good_Dum                     556.17    1254.30   0.443 0.657817
## Subject_Cat_STEM                  -4210.04    5860.23  -0.718 0.473113
## Teach_Exp_dum_0_3                -12870.96    1600.23  -8.043 2.60e-14 ***
## Teach_Exp_dum_4_10                -3021.43    1267.77  -2.383 0.017837 *
## PT                                -9570.86    1113.19  -8.598 6.18e-16 ***
## as.factor(Degree_Sub)Biology       2386.43    2421.65   0.985 0.325263
## as.factor(Degree_Sub)Business      1126.43    3528.99   0.319 0.749821
## as.factor(Degree_Sub)Combined      -603.62    2551.46  -0.237 0.813160
## as.factor(Degree_Sub)Economics    -5674.22    3783.28  -1.500 0.134805
## as.factor(Degree_Sub)Education    -2671.86    2237.42  -1.194 0.233436
## as.factor(Degree_Sub)Engineering   3618.87    7136.45   0.507 0.612491
## as.factor(Degree_Sub)English       1576.96    2410.45   0.654 0.513517
## as.factor(Degree_Sub)Geography     1586.61    2516.15   0.631 0.528842
## as.factor(Degree_Sub)Law            489.50    4424.40   0.111 0.911986
## as.factor(Degree_Sub)Linguistics   -849.32    2708.93  -0.314 0.754119
## as.factor(Degree_Sub)Math          1588.67    6454.23   0.246 0.805753
## as.factor(Degree_Sub)Medicine      -669.98   10483.59  -0.064 0.949090
## as.factor(Degree_Sub)Other         -200.10    1957.73  -0.102 0.918663
## as.factor(Degree_Sub)Physics       2265.92    6347.56   0.357 0.721383
## as.factor(Degree_Sub)Politics      4254.76    4446.44   0.957 0.339460
## as.factor(Degree_Sub)Psychology   -5065.52    2837.60  -1.785 0.075336 .
## as.factor(Degree_Sub)Sociology     -412.24    3344.41  -0.123 0.901989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8107 on 276 degrees of freedom
## Multiple R-squared:  0.4617, Adjusted R-squared:  0.4091
## F-statistic: 8.769 on 27 and 276 DF,  p-value: < 2.2e-16
```

Question 2A

```
t.test(Pr_1~Ability_Own_Good, data)
```

```
##
##  Welch Two Sample t-test
##
## data:  Pr_1 by Ability_Own_Good
## t = 0.76215, df = 118.39, p-value = 0.4475
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -4.316981  9.719332
## sample estimates:
## mean in group 0 mean in group 1
##        16.28378        13.58261
```

```
t.test(Pr_2~Ability_Own_Good, data)
```

```
##
##  Welch Two Sample t-test
##
## data:  Pr_2 by Ability_Own_Good
## t = 1.2876, df = 120.62, p-value = 0.2003
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.735357 12.911386
## sample estimates:
## mean in group 0 mean in group 1
##        25.40541        20.31739
```

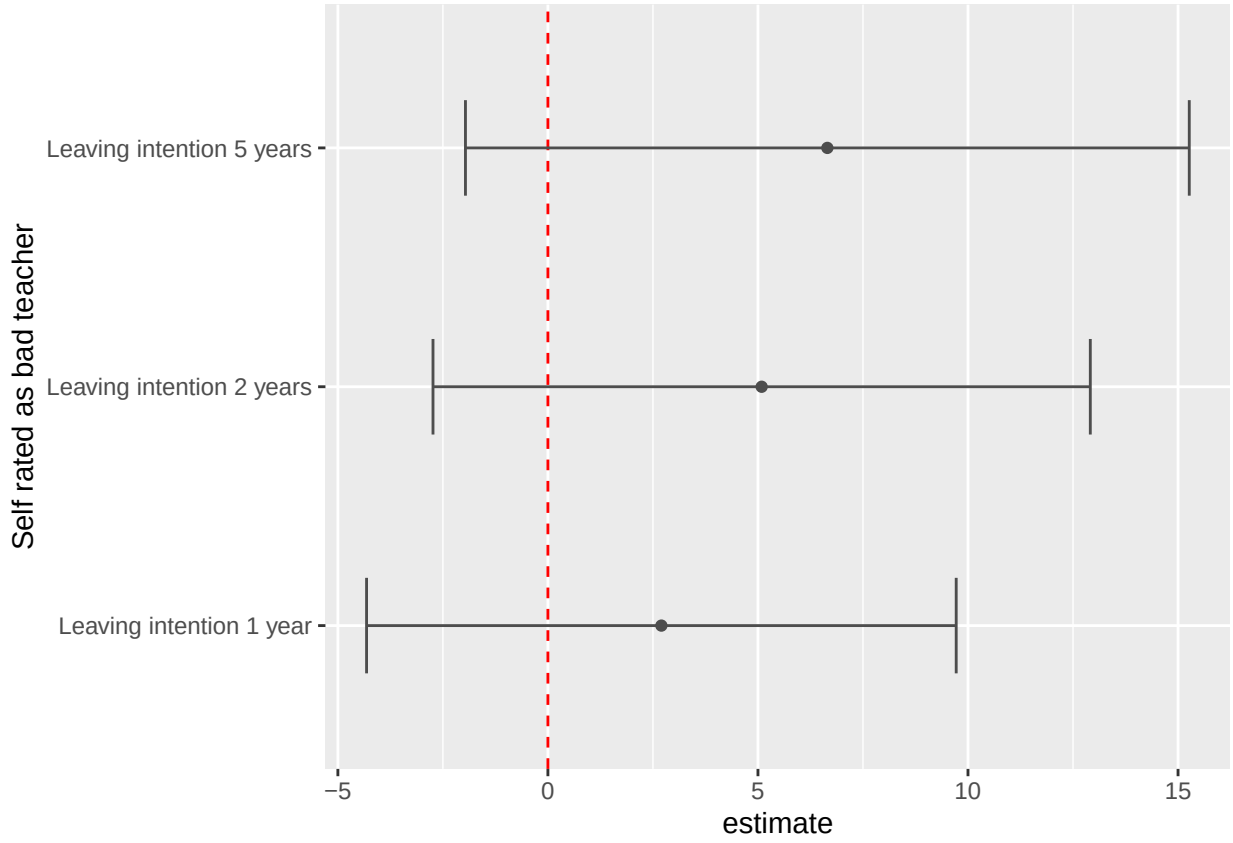
```
t.test(Pr_3~Ability_Own_Good, data)
```

```
##
##  Welch Two Sample t-test
##
## data:  Pr_3 by Ability_Own_Good
## t = 1.5272, df = 133.55, p-value = 0.1291
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.963056 15.266934
## sample estimates:
## mean in group 0 mean in group 1
##        42.24324        35.59130
```

Question 2B

[?]I didn't find a proper substitution for the STATA *foreach* function in R.

```
res_list = list()
res_list[[1]] <- t.test(Pr_1~Ability_Own_Good, data)
res_list[[2]] <- t.test(Pr_2~Ability_Own_Good, data)
res_list[[3]] <- t.test(Pr_3~Ability_Own_Good, data)
dat1 = data.frame(id=c("Leaving intention 1 year",
                       "Leaving intention 2 years",
                       "Leaving intention 5 years"),
                  estimate=sapply(res_list,
                                  function(x) x$estimate[1]-x$estimate[2]),
                  conf_int_lower=sapply(res_list, function(x) x$conf.int[1]),
                  conf_int_upper=sapply(res_list, function(x) x$conf.int[2]))
ggplot(data=dat1, aes(x=estimate, y=id)) +
    geom_vline(xintercept=0, color="red", linetype=2) +
    geom_point(color="grey30") +
    geom_errorbarh(aes(xmin=conf_int_lower, xmax=conf_int_upper),
                   color="grey30", height=0.4) +
    ylab("Self rated as bad teacher")
```



Question 3

```
coef_list <- list()
A <- lm(Pr_3~Ability_Own_Good, data)
coef_list[[1]] <- summary(A)
#Teacher characteristics
B <- update(A, . ~ . + Age_Years + Male_Teacher + Ethnic_White +
              Dep_Child_Dum + Partner_Earn_More + as.factor(School_Taught))
coef_list[[2]] <- summary(B)
#Teacher contract
C <- update(B, . ~ . + Teacher_Pay + Teacher_Hours_Actual + PT)
coef_list[[3]] <- summary(C)
#Teacher education
D <- update(C, . ~ . + as.factor(Degree_Sub) + as.factor(Degree_Class) +
              Russell_Group)
coef_list[[4]] <- summary(D)
#School characteristics
E <- update(D, . ~ . + Independent_School_Dummy + Early_Primary_Dum +
              FSM_Eligible + Class_Size + ofsted_Good_Dum)
coef_list[[5]] <- summary(E)
#Model presentation
stargazer(A, B, C, D, E, title="Results", type="text", align=TRUE,
          dep.var.labels="Probability to leave in 5 years",
          keep="Ability_Own_Good", omit.stat=c("ser","f"), no.space=TRUE)
```

```
##
## Results
## ================================================================#
#                                Dependent variable:
##                    ------------------------------------------#
#                        Probability to leave in 5 years
##                       (1)      (2)      (3)      (4)      (5)
## ----------------------------------------------------------------#
# Ability_Own_Good -6.652  -9.170** -6.452  -5.723  -4.848
##                   (4.558) (4.385)  (4.443) (4.505) (4.491)#
# ----------------------------------------------------------------#
# Observations       304      304      304      304      304
## R2                0.007    0.115    0.145    0.227    0.256
## Adjusted R2       0.004    0.094    0.116    0.145    0.162
## ================================================================#
# Note:                            *p<0.1; **p<0.05; ***p<0.01
```

```
dat2 = data.frame(id=c("A", "B", "C", "D", "E"),
                  estimate=sapply(coef_list, function(x) x$coefficients[2,1]),
                  conf_int_lower=sapply(coef_list,
                  function(x) x$coefficients[2,1]-1.96*x$coefficients[2,2]),
                  conf_int_upper=sapply(coef_list,
                  function(x) x$coefficients[2,1]+1.96*x$coefficients[2,2]))
ggplot(data=dat2, aes(x=estimate, y=id)) +
    geom_vline(xintercept=0, color="red", linetype=2) +
    geom_point(color="grey30") +
    geom_errorbarh(aes(xmin=conf_int_lower, xmax=conf_int_upper),
                   color="grey30", height=0.4) +
    ylab("Estimated effect of teacher payment")
```