Assignment 3

David Gyaraki, Thao Le

Contents

```r
# load packages
if(!require(pacman)){install.packages("pacman")}

p_load(devtools,tidyverse,dplyr,ggplot2,latex2exp,
       sampleSelection, quantreg, plm, nlme, knitr,car)

#load data
#dfData = read.csv("assignment2a_2023.csv")
#attach(dfData)
```

1 Question 1

color	number of individual		average outcome	
	treated	control	treated	control
purple	100	100	9	7
blue	75	25	13	8
green	25	75	10	9

1.1 (i)

The treatment effect in theory is the difference between the outcomes if the individual is treated versus if the individual is not treated. Suppose that for individual i, the treatment effect is defined as:

$$TE_i = \Delta_i = y_{i,d=1} - y_{i,d=0}, \tag{1}$$

where the y marks the outcome and d is the dummy whether individual i was treated ($d = 1$) or not ($d = 0$). However, this is rarely possible to be observed for each individual, as treatment is generally considered mutually exclusive, so an individual is either treated or not. Therefore, in this example we cannot calculate TE measures for individuals within color groups, but we can pretend as if one color corresponds to one observation, and calculate the three treatment effects across the color groups.

Then, the treatment effect for "observation" purple is $9 - 7 = 2$, the treatment effect for "observation" blue is $13 - 8 = 5$ and the treatment effect for "observation" green is $10 - 9 = 1$.

1.2 (ii)

Using the assumption of Section 1.1, we can infer here that we are looking for the average treatment effect of the full population including all colors. The average treatment effect is generally given by:

$$ATE = E[\Delta] = E[y_{d=1} - y_{d=0}], \tag{2}$$

where we take the average across the observation differences. Our data in the exercise does not contain information again on the individual outcomes, however we can use the average outcomes of treatment and control in each color group to calculate the total differences between $y_{d=1}$ and $y_{d=0}$ on the full population. Therefore, we can calculate the ATE as:

```
E_diff_purple = 9 - 7
E_diff_blue = 13 - 8
E_diff_green = 10 - 9

ATE = (E_diff_purple * 200 + E_diff_blue * 100 + E_diff_green * 100) /
↪  400
ATE
```

```
[1] 2.5
```

1.3 (iii)

Compared to the solution in Section 1.2, now we consider the average treatment effect among treated (ATET). For this case, we need to modify the previous definition to:

$$ATET = E[\Delta|d=1] = E[y_{d=1} - y_{d=0}|d=1] = E[y_{d=1}] - E[y_{d=0}] = \frac{1}{N_T}\Sigma_{i=1}^{N_T} y_{i,d=1} - \frac{1}{N_{NT}}\Sigma_{i=1}^{N_{NT}} \tag{3}$$

where N_T is the number of individuals in the treatment group while N_{NT} is the number of individuals in the control group. Therefore, the ATET for the whole population can be calculated as:

```
E_treatment_atet <- (9*100 + 13*75 + 10*25) / (100+75+25)
E_control_atet <- (7*100 + 8*25 + 9*75) / (100+75+25)
```

```
ATET = E_treatment_atet - E_control_atet
ATET
```

```
[1] 2.75
```

1.4 (iv)

The ATE measure generally describes the expected gain in y achieved by treating a random member i from the population, i.e. how much one benefits from being selected for the treatment compared to people who were not. On the other hand, ATET describes the average gain achieved by the treatment for the treated group, i.e. the comparison is not to the population but peer-to-peer, what is the expected benefit for those who are selected. The ATET is more helpful if we are not mainly interested in the potentially positive effect of the treatment for those who are treated versus those who are not, but rather the magnitude of these positive effects. Suppose that we have an experiment where the government announces a new plan to introduce an additional level of health insurance, where the own risk cost would be cut in half, in order to investigate the effects of these on household savings. Arguably, cutting the own risk cost in half without changing the insurance monthly premiums would most likely have a positive effect on the wealth for those who are involved in the initial study, but if the government is rather interested in measuring the average savings surplus this would create for households, we would be more interested in the ATET measure as this policy will ideally be introduced for everyone later on and we are solely interested in the average savings surplus this would create for everyone involved.

2 Question 2

```
# Make sure to delete this before handin, just have it as setting
↪ working directory
#setwd("~/GitHub/bds_block3/Econometrics_II/Assignment 3")

dfData = read.csv("bonus.csv")
attach(dfData)

dfData <- na.omit(dfData)
```

2.1 (i)

```r
df_noR <- dfData[dfData$bonus0 == 1,]
df_lowR <- dfData[dfData$bonus500 == 1,]
df_highR <- dfData[dfData$bonus1500 == 1,]

sum_noR <- sapply(df_noR, mean, na.rm=TRUE)
sum_lowR <- sapply(df_lowR, mean, na.rm=TRUE)
sum_highR <- sapply(df_highR, mean, na.rm=TRUE)

summary_table <- cbind(sum_noR, sum_lowR, sum_highR)

kable(summary_table, caption="The means of the predictor and dependent
↪ variables across the three groups", col.names = c("no reward", "low
↪ reward", "high reward"), digits = 3)
```

Table 1: The means of the predictor and dependent variables across the three groups

	no reward	low reward	high reward
p0	0.558	0.525	0.576
job	0.760	0.840	0.811
stp2001	34.129	31.998	33.311
stp2004	89.183	82.268	86.311
dropout	0.360	0.346	0.324
myeduc	12.253	12.111	12.662
fyeduc	13.467	13.420	13.581
bonus0	1.000	0.000	0.000
bonus500	0.000	1.000	0.000
bonus1500	0.000	0.000	1.000
effort	19.549	18.273	18.483
pass	0.200	0.198	0.243
math	5.440	5.395	5.405

As we can see from the summary table above, the no reward group saw a 19.5% pass rate, the low reward group saw a 20.2% pass rate while the high reward group saw a 24.1% pass rate. At the same time, we can also look at the other numeric variables to investigate their means and check how balanced the other predictors are. Naturally since we grouped the students based on the incentives, the three *bonus* variables will not be balanced. However, the other ones are quite similar among the three groups implying that the background characteristics are relatively balanced. In particular, parental educational background (*myeduc* and *fyeduc*) and high-school math scores (*math*) are quite similar to one another among groups.

2.2 (ii)

```
prob_lm1 <- lm(pass ~ bonus500 + bonus1500, data = dfData)
```

Looking at the summary statistics for the first model, we can see that the effects of the treatments are not statistically significant. Thus, we cannot say whether the treatment effect has an influence on the students' study achievements or not.

```
prob_lm2 <- lm(pass ~ bonus500 + bonus1500 + fyeduc + p0 + math, data =
  ↪  dfData)
summary(prob_lm2)
```

```

Call:
lm(formula = pass ~ bonus500 + bonus1500 + fyeduc + p0 + math,
    data = dfData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61424 -0.23811 -0.09870  0.08326  0.92864

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.601623   0.131274  -4.583 7.62e-06 ***
bonus500     0.010991   0.058597   0.188   0.8514
bonus1500    0.043247   0.059871   0.722   0.4708
fyeduc      -0.000789   0.007340  -0.107   0.9145
p0           0.237672   0.097726   2.432   0.0158 *
math         0.124943   0.018710   6.678 1.89e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3651 on 224 degrees of freedom
Multiple R-squared:  0.2255,    Adjusted R-squared:  0.2082
F-statistic: 13.04 on 5 and 224 DF,  p-value: 3.783e-11
```

In the second model, we add 3 additional control variables, namely, father's education *fyeduc*, subjective self-assessment of the pass probability *p0*, and highschool math score *math*.

However, the significance of the treatment variables *bonus500* and *bonus1500* do not change. Thus, from both of this model, we say that the financial incenntive does not influence student's study performance in the first year.

2.3 (iii)

```
prob_lm3 <- lm(pass ~ bonus500 + bonus1500 + fyeduc + p0 + math +job
 ↪  +effort, data = dfData)
summary(prob_lm3)
```

```
Call:
lm(formula = pass ~ bonus500 + bonus1500 + fyeduc + p0 + math +
    job + effort, data = dfData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66148 -0.23088 -0.09068  0.13310  0.85853

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6782570  0.1438032  -4.717 4.24e-06 ***
bonus500     0.0234916  0.0576245   0.408  0.68391
bonus1500    0.0556653  0.0587878   0.947  0.34473
fyeduc       0.0003153  0.0072150   0.044  0.96519
p0           0.1703763  0.0978230   1.742  0.08295 .
math         0.1243602  0.0183332   6.783 1.05e-10 ***
job         -0.0621005  0.0597798  -1.039  0.30002
effort       0.0076559  0.0023862   3.208  0.00153 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3576 on 222 degrees of freedom
Multiple R-squared:  0.2637,    Adjusted R-squared:  0.2405
F-statistic: 11.36 on 7 and 222 DF,  p-value: 2.699e-12
```

In the third model, we include the variable indicating whether a student has a job *job* and the average number of study hours *effort*.

To comment on this approach, we do not think these two control variables are good addition to the model as it might be correlated with the treatment variables and with each other. For example, the amount of study effort one puts in might correlate with the type of reward they are assign (e.g., a higher reward means more study efforts). On the other hand, if the reward of studying is low, a student might be more incentivized to take up part-time job, and the fact that a student has a job might lower the average number of hours they put into the study (which is measures in the variable *effort*).

2.4 (iv)

For this question, we use a linear probabi

```
drop_lm <- lm(dropout ~ bonus500 + bonus1500 + math + fyeduc + p0 +
 ↪  effort + job, data=dfData)
credsYear1_lm <- lm(stp2001 ~ bonus500 + bonus1500 + math + fyeduc + p0
 ↪  + effort + job, data=dfData)
credsYear3_ln <- lm(stp2004 ~ bonus500 + bonus1500 + math + fyeduc + p0
 ↪  + effort + job, data=dfData)

stargazer::stargazer(drop_lm,credsYear1_lm,credsYear3_ln, type = "text")
```

```
==============================================================
                                   Dependent variable:
                              --------------------------------
                                dropout    stp2001    stp2004
                                  (1)        (2)        (3)
--------------------------------------------------------------
bonus500                        -0.048     -0.144     -2.388
                                (0.068)    (2.690)    (7.999)

bonus1500                       -0.057      0.227      0.294
                                (0.070)    (2.744)    (8.160)

math                          -0.077***   6.565***  15.561***
                                (0.022)    (0.856)    (2.545)

fyeduc                           0.009     -0.309     -1.231
                                (0.009)    (0.337)    (1.002)

p0                              -0.150    13.582***   18.312
                                (0.116)    (4.567)   (13.579)

effort                        -0.018***   0.928***   2.920***
```

```
                              (0.003)   (0.111)    (0.331)

job                            0.034     -0.904     5.628
                              (0.071)   (2.791)    (8.298)

Constant                      1.071***  -22.455*** -50.458**
                              (0.170)   (6.713)    (19.962)

-------------------------------------------------------------
Observations                    230       230        230
R2                             0.232     0.445      0.388
Adjusted R2                    0.207     0.427      0.369
Residual Std. Error (df = 222) 0.424     16.695     49.642
F Statistic (df = 7; 222)     9.563***  25.388***  20.112***
=============================================================
Note:                            *p<0.1; **p<0.05; ***p<0.01
```

For this purpose, we build 3 linear probability models that include the treatment variables and all exogenous variables as the predictors. The summary shows that the effect of the financial incentive is not significant.

2.5 (v)

The minimum detectable effect of this experiment is measure using the formula:

(insert latex formula)

We need to calculate the MDE for both low-reward versus control group and high-reward versus control group

```
MinDE <- function(dfData, input_model, sTreatment, dAlpha, dPower){
  num_coef <- length(input_model$coefficients)

  # get the number of observations
  n_obs <- nrow(dfData)
  # Observations in treated group
  n_treatment <- sum(dfData[,sTreatment])
  # Proportion of treatment observations
  proportion <- n_treatment/n_obs

  # T_statistics of alpha and power levels
  t_stats_alpha <- qt(1-dAlpha/2, n_obs - num_coef)
```

```
  t_power <- qt(1-dPower, n_obs - num_coef)

  # get variance of residuals
  dVariance <- var(input_model$residuals)

  # get the MDE
  MDE <- (t_stats_alpha - t_power) * sqrt(1/(proportion*(1-proportion)))
↪   * sqrt(dVariance/n_obs)
  return(MDE)
}

# Initialize desired alpha and power
dAlpha = 0.05
dPower = 0.75

# Prepare dataframe
df_low_treatment = dfData[dfData$bonus1500 != 1,]
df_high_treatment = dfData[dfData$bonus500 != 1,]

#Estimate linear model with low reward treatment and control group

Low_treatment = lm(pass ~ bonus500 + math + fyeduc + p0 + effort + job,
↪   data=df_low_treatment)

#Get MDE of low-reward versus control group
MDE_low = MinDE(df_low_treatment,Low_treatment,"bonus500",dAlpha,
↪   dPower)

#Estimate linear model with high reward treatment and control group
High_treatment = lm(pass ~ bonus1500 + math + fyeduc + p0 + effort +
↪   job, data=df_high_treatment)

#Get MDE of high-reward versus control group
MDE_high = MinDE(df_high_treatment, High_treatment,"bonus1500",dAlpha,
↪   dPower)

cat("The MDE of the low reward treatment group is: ", MDE_low, "\n",
↪   ",and the MDE of the high reward treatment group is: ", MDE_high)
```

```
The MDE of the low reward treatment group is:  0.1483831
 ,and the MDE of the high reward treatment group is:  0.1579427
```

2.6 (vi)

An increases in the pass rate of 10% points correspond with the Minimum Detectable Effect size of 10%. Given this, we can calculate the minimum number of observations needed using:

(insert Latex formula from lecture)

Assumptions: since we do not know n and the degree of freedom to calculate the t-stats, we assume n is big enough for a normal distribution and calculate the statictics using a standard normal distribution.

```r
get_n <- function(dfData, input_model, sTreatment, dAlpha, dPower,
↪  MinDE=0.1, dSigma_sq){
  # T_statistics of alpha and power levels
  t_stats_alpha <- qnorm(1-dAlpha/2, 0,dSigma_sq)
  t_power <- qnorm(1-dPower, 0, dSigma_sq)

  # get the number of observations
  n_obs <- nrow(dfData)
  # Observations in treated group
  n_treatment <- sum(dfData[,sTreatment])
  # Proportion of treatment observations
  proportion <- n_treatment/n_obs

  # get the desired sample size
  sample_size =
↪  ((t_stats_alpha-t_power)/MinDE)^2*(dSigma_sq/(proportion/(1-proportion)))
  return(round(sample_size))
}

# Initialize, set sigma and alpha fixed
dSigma_sq = 1
dAlpha= 0.5
dPower = 0.8

# required sample sizez needed in each treatment group
sample_low_required = get_n(df_low_treatment, input_model =
↪  Low_treatment,"bonus500", dAlpha = dAlpha, dPower = dPower,
↪  dSigma_sq = dSigma_sq)
sample_low_required
```

```
[1] 213
```

```
sample_high_required = get_n(df_high_treatment, input_model =
 ↪  High_treatment,"bonus1500", dAlpha = dAlpha, dPower = dPower,
 ↪  dSigma_sq = dSigma_sq)
sample_high_required
```

```
[1] 233
```

2.7 (vii)

Looking at the formula for MDE (page 35 of lecture slide), fewer students in the experiment mean higher MDE. If we want to achieve the same MDE with fewer students, we need to choose the proportion of control versus treatment group such that the term $p(1-p)$ is minimized. Taking the derivative of that term with respect to p, we find the optimal value of $p = 0.5$

in which: p is the proportion of treated students. To get $p = 0.5$, we indeed need to increase number of students in control group.

Do you agree? now we need to relate this with our dataset, do some count functions, then we're done!