Assignment 3

Walter Verwer & Bas Machielsen

January 20, 2021

Question 1

1.1

```r
data <- tribble(
    ~ color, ~ no_treated, ~ no_contr, ~ avg_treated, ~avg_control,
    "purple", 100, 100, 9, 7,
    "blue", 75, 25, 13, 8,
    "green", 25, 75, 10, 9
)

data %>%
    group_by(color) %>%
    summarize(treatment_effect = avg_treated - avg_control) %>%
    kable(caption = "Treatment effect per color")
```

Table 1: Treatment effect per color

color	treatment_effect
blue	5
green	1
purple	2

1.2

The ATE is defined as $\mathbb{E}[\delta] = \mathbb{E}[Y_1^*] - \mathbb{E}[Y_0^*]$ which are the expectations of the potential outcomes. In general, these two variables are not observed. Under the *random assignment* assumption, we assume that $\mathbb{E}[Y_1^*] = \mathbb{E}[Y_1^*|D = 1]$ and $\mathbb{E}[Y_0^*] = \mathbb{E}[Y_0^*|D = 0]$, which can be estimated by their sample by their sample equivalents:

```r
data %>%
    summarize(
        e_y1_d_is_1 = (9*100 + 13 * 75 + 10 * 25) / sum(no_treated),
        e_y0_d_is_0 = (7 * 100 + 8 * 25 + 9 * 75) / sum(no_contr)) %>%
    summarize(ate = e_y1_d_is_1 - e_y0_d_is_0) %>%
    kable(caption = "ATE")
```

Table 2: ATE

ate
2.75

1.3

The ATT is defined as $\mathbb{E}[\delta|D = 1] = \mathbb{E}[Y^1|D = 1] - \mathbb{E}[Y^0|D = 1]$. The first term is readily observable. The second term is estimated by us as $\hat{\mathbb{E}}[Y^0|D = 1] = \mathbb{E}[Y^0|D = 0]$. Hence:

```
data_ate <- data %>%
    mutate(n = no_treated + no_contr) %>%
    summarize(e_y1_d_is_1 = (9 * 100 + 13 * 75 + 10 * 25) / sum(no_treated),
              e_y0_d_is_0 = (7 * 100 + 8 * 25 + 9 * 75) / sum(no_contr))

data_ate
```

```
## # A tibble: 1 x 2
##   e_y1_d_is_1 e_y0_d_is_0
##         <dbl>       <dbl>
## 1        10.6        7.88
```

```
data_ate %>%
    summarize(att = e_y1_d_is_1 - e_y0_d_is_0) %>%
    kable(caption = "ATT")
```

Table 3: ATT

att
2.75

So the $ATE = ATT$ (because of randomization).

Question 2

2.1

Compute the fraction of students in all three groups (control, low-reward and high-reward) that complete all first-year courses before the start of the second academic year. Show within a table that background characteristics are balanced over the treatment groups.

```
bonus_clean <- bonus %>%
    pivot_longer(cols = c(bonus0, bonus500, bonus1500),
                 names_to = "kind_treatment",
                 values_to = "treatment") %>%
    filter(treatment == 1)
```

```
bonus_clean %>%
    group_by(kind_treatment) %>%
    summarize(fraction_pass = sum(pass)/n()) %>%
    kable(caption = "Fraction passed per treatment", digits = 3)
```

Table 4: Fraction passed per treatment

kind_treatment	fraction_pass
bonus0	0.195
bonus1500	0.241
bonus500	0.202

```
#Drop the outcome variables
bonus_clean %>%
    group_by(kind_treatment) %>%
    select(-c(pass, stp2001, stp2004, dropout)) %>%
    summarize(across(p0:math,
                  list(mean = ~ mean(., na.rm = TRUE),
                       sd = ~ sd(., na.rm = TRUE)))) %>%
    pivot_longer(-kind_treatment, names_to = "variable") %>%
    separate(variable, into = c("var", "statistic"), sep = "_") %>%
    pivot_wider(names_from = c(kind_treatment, statistic)) %>%
    kable(caption = "Means and SDs according to treatment", digits = 3)
```

Table 5: Means and SDs according to treatment

var	bonus0_mean	bonus0_sd	bonus1500_mean	bonus1500_sd	bonus500_mean	bonus500_sd
p0	0.553	0.265	0.573	0.248	0.530	0.251
job	0.760	0.430	0.805	0.399	0.829	0.379
myeduc	12.293	3.041	12.590	2.992	12.119	3.316
fyeduc	13.378	3.416	13.422	3.596	13.524	3.273
effort	19.549	9.460	18.303	10.592	18.477	10.475
math	5.476	1.468	5.388	1.258	5.386	1.360

2.2

Use the linear probability model to regress the dummy variable for completing all courses on the assignment of the three treatment groups. Interpret the treatment effects. Next include as additional regressors father's education, high-school math score and the subjective assessment about the pass probability.

2.3

Next also include as regressors in your model whether a student has a job and the amount of study effort. Comment on this approach. Do you consider this an improvement over (ii)?

```
attach(bonus)

# We need to regress pass on the treatment group assignment. This is
# a simple linear probability model. Model is denoted by prob_1.
# Note, we need to omit bonus0, cause of a dummy variable trap.
prob_1 <- lm(pass ~ bonus500 + bonus1500, data=bonus)
# Same as before, but now with some extra variables
prob_2 <- lm(pass ~ bonus500 + bonus1500 + fyeduc + math + p0, data=bonus)
# for q2.3:
prob_3 <- lm(pass ~ bonus500 + bonus1500 + fyeduc + math + p0 + job + effort, data=bonus)

# # q2.4; I think we should take the largest model. BAS???:
# prob_4 <- lm(pass ~ bonus500 + bonus1500 + fyeduc + math + p0 + job + effort + dropout + stp2001 + st

# Implement HC-robust standard errors (standard in linear prob model)
#library(sandwich)
#cov <- vcovHC(reg.model, type = "HC")
#robust.se <- sqrt(diag(cov))

# Create table:
stargazer(prob_1, prob_2, prob_3, header=FALSE, style='aer')
```

In the first model, we find that students who receive the 500 bonus are 0.007 percentage points more likely to pass the first year, and students who receive the f 1,500 bonus are 0.046 percentage points more likely to pass the first year, both relative to the group that receives no bonus. This effect is, however, not statistically significant, indicating that there is a high variance in passing within treatment groups, or, alternatively, that the sample size is too sample to statistically detect a relatively small effect size.

The point estimates change only slightly when including a vector of control variables, indicating that the treatment conditional on these (potential) confounders does not significantly increase the probability of passing.

2.4

Use your preferred model specifcation to estimate the effects of the financial incentives on some other outcomes: dropping out and credit points collected (in the first year and after three years).

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Jan 20, 2021 - 12:33:12

2.5

Given the sample size and the estimates you have obtained above, what would be the minimum detectable effect size of this experiment? In the next model, we observe that the point estimates for the treatment effects increase a bit, indicating that effort was a confounder. Hence, we do consider this to be an improvement, although the treatment effects are still not significantly different from zero.

```
# For now the constants:
n <- length(p0)
df <- n - length(prob_3$coefficients)
alpha <- 0.05
t_95 <- qt(1-alpha/2, df)
```

Table 6:

| | pass | | |
	(1)	(2)	(3)
bonus500	0.007	0.015	0.023
	(0.064)	(0.057)	(0.058)
bonus1500	0.046	0.048	0.056
	(0.064)	(0.058)	(0.059)
fyeduc		−0.001	0.000
		(0.007)	(0.007)
math		0.119***	0.124***
		(0.018)	(0.018)
p0		0.249***	0.170*
		(0.095)	(0.098)
job			−0.062
			(0.060)
effort			0.008***
			(0.002)
Constant	0.195***	−0.576***	−0.678***
	(0.045)	(0.128)	(0.144)
Observations	249	245	230
R^2	0.002	0.209	0.264
Adjusted R^2	−0.006	0.192	0.240
Residual Std. Error	0.411 (df = 246)	0.366 (df = 239)	0.358 (df = 222)
F Statistic	0.297 (df = 2; 246)	12.629*** (df = 5; 239)	11.359*** (df = 7; 222)

Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

	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)
fyeduc	0.009	−0.309	−1.231
	(0.009)	(0.337)	(1.002)
math	−0.077***	6.565***	15.561***
	(0.022)	(0.856)	(2.545)
p0	−0.150	13.582***	18.312
	(0.116)	(4.567)	(13.579)
job	0.034	−0.904	5.628
	(0.071)	(2.791)	(8.298)
effort	−0.018***	0.928***	2.920***
	(0.003)	(0.111)	(0.331)
Constant	1.071***	−22.455***	−50.458**
	(0.170)	(6.713)	(19.962)
Observations	230	230	230
R^2	0.232	0.445	0.388
Adjusted R^2	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***

Table 7:

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

```
sigma2 <- var(prob_4$residuals)

# Ik heb deze maar pre specified. We hebben twee verschillende treatments,
# zouden we dan niet twee power calculaties moeten doen?
p<-0.5

# How much power do we want? Range of powers:
q <- seq(from=0.6, to=0.9, by=0.1)

# Init empty vector of t values for the power:
t_q = rep(0, length(q))

# Fill vector iteratively and compute MDE for bonus500:
mde <- cbind(rep(0, length(q)),rep(0, length(q)))
counter <- 1
for (i in q){
    t_q[counter] <- qt(1-i, df)   # A check: from slides, t_0.7 = -0.525...

    mde[counter,1] <- (t_95 - t_q[counter]) * sqrt( (1/(p*(1-p))) * (sigma2 / n) )

    counter <- counter + 1
}

# Treatment bonus1500:
counter <- 1
for (i in q){
    t_q[counter] <- qt(1-i, df)   # A check: from slides, t_0.7 = -0.525...

    mde[counter,2] <- (t_95 - t_q[counter]) * sqrt( (1/(p*(1-p))) * (sigma2 / n) )

    counter <- counter + 1
}


kable(mde)
```

```
## Warning in kable_pipe(x = structure(c("0.1175597", "0.1319127", "0.1487270", :
## The table should have a header (column names)
```

0.1175597	0.1175597
0.1319127	0.1319127
0.1487270	0.1487270
0.1720942	0.1720942

2.6

Initially, the researchers were aiming at an increases in the pass rate of 10% points. How large should the sample size of the experiment have been in that case?

The proportion of treated subjects $p = \frac{83+84}{82+83+84} = 0.67$.

```
bonus_clean %>%
    group_by(kind_treatment) %>%
    summarize(test = n()) %>%
    kable()
```

`summarise()` ungrouping output (override with `.groups` argument)

kind_treatment	test
bonus0	82
bonus1500	83
bonus500	84

The estimated variance of the error term in the full model is

We know that the minimum effect size as a function of p, power q, level α,