Assignment 4

Bas Machielsen & Walter Verwer

January 27, 2021

Problem 1: Judges and Prison Sentences

(i) Use the Wald estimator to compute the causal effect of a prison sentence on the probability of being arrested later.

The Wald estimator is defined as follows:

$$\frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]}$$

In our case, Y is the future arrest, $Z=1$ if the judge is Jones, and it is equal to zero if it is Smith, and $D=1$ if the individual went to prison, and zero otherwise. Filling in the formula with the numbers given, results in following:

$$\frac{(70\% \cdot 40\% + 30\% \cdot 60\%) - (40\% \cdot 20\% + 60\% \cdot 50\%)}{70\% - 40\%} = 0.27$$

(ii) What is the interpretation of the estimated effect? And for which fraction of the population does this causal effect hold?

The interpretation of this is that sending an individual to prison results in a 27% higher probability that the individual has to go to prison again. The fraction for which this causal effect hold is the people that comply. That is, the fraction of the people that go to prison if they are assigned to Jones and do not go to prison if they are assigned to Smith plus the fraction of people that do not go to prison under Jones, but go to prison under Smith. This fraction is equal to $0.7 \cdot 0.6 + 0.4 \cdot 0.3 = 0.54$.

(iii) Explain what an always taker is in this setting and which fraction of the population are always takers? An always taker is someone who always takes up treatment. In this situation an always taker is someone who will always go to prison. This would be someone who comitted a very bad crime, such that both judges sentence the individual to prison, think about someone who comitted a murder. This fraction equals $0.7 \cdot 0.4 = 0.28$.

Problem 2: Eating and Drinking

(i) Perform a power calculation for the number of students that the teacher should include in the field experiment.

Compute the power as follows.

$$MDE = (t_{1-\alpha/2} - t_{1-q})\sqrt{\frac{1}{p(1-p)}}\sqrt{\frac{\sigma^2}{n}}$$

Filling in the values given, results in the following.

$$0.1 = (1.96 + 0.524)\sqrt{\frac{1}{0.25}}\sqrt{\frac{0.25}{n}} \implies n \approx 617$$

In this calculation, the variance follows from a standard Bernoulli variance calculation. In our case $p = 0.5$, which implies that the variance is 0.25.

(ii)The teacher assumes that 20% of the students randomized in the treatment group will actually have breakfast. How does this change the number of students required to participate in the field experiment?

We take the formula from the previous question and change it accordingly. This results the following formula.

$$MDE = (t_{1-\alpha/2} - t_{1-q})\sqrt{\frac{1}{p(1-p)}}\sqrt{\frac{\sigma^2}{n}}\frac{1}{r_t - t_c}$$

Filling in the values again gives us the following.

$$0.1 = (1.96 + 0.524)\sqrt{\frac{1}{0.25}}\sqrt{\frac{0.25}{n}}\frac{1}{0.8-0} \implies n \approx 964$$

Concluding, we observe that the partial compliance increases the number of observations needed.

Problem 3: Flu shots for young children

(i) Compute for the children assigned to the control group the variance in u incidence. If the researcher aims at reducing flu incidence by 0.05, how many children should participate in the randomized experiment.

First, we calculate the variance in the population (without the treatment) of getting the flu:

```
var <- flu %>%
    filter(TreatGroup == 0 ) %>%
    summarize(var = var(Flu)) %>%
    pull()

var
```

```
## [1] 0.2355284
```

Then, we calculate the power based on the following specification, with MDE 0.045, $t_{1-\alpha/2} = 1.96$, $t_{1-q} = -0.52$, and the proportion of treated subjects $p = 0.80$:

$$MDE = (t_{1-\alpha/2} - t_{1-q})\sqrt{\frac{1}{p(1-p)}}\sqrt{\frac{\sigma^2}{n}}$$

```r
# Proportion of treatment
p <- flu %>%
    summarize(prop_treated = mean(TreatGroup)) %>%
    pull()

#Effect size
mde <- 0.05

# Alpha, and Q: alpha = 5%, alpha/2 = 2.5%, power  = 0.7
t_1_min_alpha_div_2 <- qnorm(0.975)
t_1_min_power <- qnorm(0.3)

# Compute the required sample size
n = var * ((t_1_min_alpha_div_2 - t_1_min_power)^2) / ((mde*sqrt(p*(1-p)))^2)
n <- round(n, 2)
```

Hence, n should be greater than approximately 3658.96.

(ii) Compute which fraction of the children in the treatment group actually received a flu shot. What is the implication for the power analysis of the experiment?

```r
fraction <- flu %>%
    filter(TreatGroup == 1) %>%
    summarize(fraction = mean(Treatment)) %>%
    pull() %>%
    round(2)
```

Only 0.67 percent of the individuals in the treatment group actually received the treatment. The previously effectuated power analysis therefore underestimates the sample size needed to discover the effect at the required α level with the required power.

(iii) Make a table with summary statistics for (1) the control group, (2) the treated treatment group, and (3) the untreated treatment group. What do you conclude?

```r
#gender of the child, age of the mother at birth, years of education of the mother, whether or not the

#control group
ctrl <- flu %>%
    filter(TreatGroup == 0) %>%
    summarize(across(c(GenderChild, AgeMother, EducationMother,
                       Married, Nationality, Hhincome), mean)) %>%
    pivot_longer(everything(), names_to = "var", values_to = "mean_control")

#treated treatment group
tt <- flu %>%
    filter(TreatGroup == 1, Treatment == 1) %>%
    summarize(across(c(GenderChild, AgeMother, EducationMother,
                       Married, Nationality, Hhincome), mean)) %>%
    pivot_longer(everything(), names_to = "var", values_to = "mean_tt")

#untreated treatment group
utt <- flu %>%
    filter(TreatGroup == 1, Treatment == 0) %>%
    summarize(across(c(GenderChild, AgeMother, EducationMother,
```

```
                         Married, Nationality, Hhincome), mean)) %>%
    pivot_longer(everything(), names_to = "var", values_to = "mean_utt")

merge(ctrl, tt) %>%
    merge(utt) %>%
    kable(digits = 3)
```

var	mean_control	mean_tt	mean_utt
AgeMother	26.093	26.594	24.881
EducationMother	12.340	12.524	11.834
GenderChild	0.508	0.503	0.501
Hhincome	2269.884	2373.871	2110.712
Married	0.957	0.977	0.939
Nationality	0.278	0.239	0.341

The researcher first focuses on only those children randomized in the treatment group. The researcher specifies the linear regression model

$$Flu_i = \alpha + \delta FluShot_i + U_i$$

(iv) Estimate this model using OLS. Next, include subsequently the individual characteristics. What do you learn from these regressions?