Assignment 4

Bas Machielsen & Walter Verwer

January 29, 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
## For mean:
#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")

## For sd:
#control group
ctrl_sd <- flu %>%
    filter(TreatGroup == 0) %>%
    summarize(across(c(GenderChild, AgeMother, EducationMother,
                       Married, Nationality, Hhincome), sd)) %>%
    pivot_longer(everything(), names_to = "var", values_to = "sd_control")

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

#untreated treatment group
utt_sd <- flu %>%
    filter(TreatGroup == 1, Treatment == 0) %>%
    summarize(across(c(GenderChild, AgeMother, EducationMother,
                       Married, Nationality, Hhincome), sd)) %>%
    pivot_longer(everything(), names_to = "var", values_to = "sd_utt")

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

var	mean_control	sd_control	mean_tt	d_tt	mean_utt	sd_utt
AgeMother	26.093	3.022	26.594	2.942	24.881	2.917
EducationMother	12.340	1.732	12.524	1.698	11.834	1.695
GenderChild	0.508	0.500	0.503	0.500	0.501	0.500
Hhincome	2269.884	1007.743	2373.871	1059.784	2110.712	905.214
Married	0.957	0.202	0.977	0.151	0.939	0.240
Nationality	0.278	0.448	0.239	0.426	0.341	0.474

We conclude that the covariates are still fairly balanced among the three groups: we can observe that, compared to the treated treatment group, the untreated treatment group (the individuals who chose not to opt for the treatment despite being assigned) had slightly younger mothers, slightly less education and income, and were slightly less likely to be married. In addition, they were more likely to be (Nationality?). The differences, however, are very small. We also observe that if we compare the standard deviations of the mean estimates, we see no significant differences in means. In other words, all means are well within two standard deviations from each other.

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?

```
modele_uno <- lm(data = flu,
    formula = Flu ~ TreatGroup)

modele_duo <- lm(data = flu,
    formula = Flu ~ TreatGroup + AgeMother + EducationMother +
        GenderChild + Hhincome + Married + Nationality)

stargazer(modele_uno, modele_duo,
          header = FALSE,
          column.labels = c("Without Controls", "With Controls")
          )
```

Table 2:

	Dependent variable:	
	Flu	
	Without Controls	With Controls
	(1)	(2)
TreatGroup	−0.129***	−0.133***
	(0.011)	(0.010)
AgeMother		−0.052***
		(0.002)
EducationMother		−0.030***
		(0.003)
GenderChild		0.013
		(0.008)
Hhincome		0.00000
		(0.00000)
Married		−0.042*
		(0.022)
Nationality		0.106***
		(0.009)
Constant	0.621***	2.336***
	(0.010)	(0.041)
Observations	12,583	12,583
R^2	0.011	0.168
Adjusted R^2	0.010	0.167
Residual Std. Error	0.497 (df = 12581)	0.456 (df = 12575)
F Statistic	134.184*** (df = 1; 12581)	361.627*** (df = 7; 12575)
Note:		*p<0.1; **p<0.05; ***p<0.01

We conclude that the treatment effect of a flu vaccination has a significant and negative effect on flu incidence. This result is robust to certain factors that are correlated with taking the treatment, such as Parental Education and Household Income. However, there is only partial compliance with the treatment assignment. Intuitively, the OLS estimator incorporates the effect of the treatment on the treated subjects, but also the effect of not taking the treatment among the subjects that were assigned the treatment. Concretely, assuming the effect of treatment is negative, the OLS estimator would underestimate the effect of the treatment, because it is (falsely) implied that the subjects assigned to take the treatment also take it.

(v) Use 2SLS to estimate δ and check the robustness with respect to adding individual characteristics.

Adding the control variables (individual characteristics), increases the adjusted r squared. Also, it appears that most of the added variables are highly significant. These findings indicate a robust finding because the coefficient of treatment changes by only a fraction of the standard error. Thus, if we would test whether there is a significant difference between the two coefficients, we would not be able to reject the hypothesis that the two are significantly different at common stated levels.

```
ivmodel1 <- ivreg(data = flu, formula = Flu ~ Treatment | TreatGroup)

ivmodel2 <- ivreg(data = flu, formula = Flu ~ Treatment + AgeMother +
                      EducationMother + GenderChild + Hhincome
                  + Married + Nationality |
                      TreatGroup + AgeMother + EducationMother +
        GenderChild + Hhincome +
            Married + Nationality)

stargazer(ivmodel1, ivmodel2, header = FALSE)
```

(vi) Estimate the first-stage regression using OLS. Are you afraid of a weak instruments problem?

```
firststage_1 <- lm(data = flu,
                   formula = Treatment ~ TreatGroup
                 )

firststage_2 <- lm(data = flu,
                   formula = Treatment ~ TreatGroup +
                       AgeMother + EducationMother +
        GenderChild + Hhincome + Married + Nationality)

stargazer(firststage_1, firststage_2, header = FALSE,
            title = "First stage regressions")
```

No, the F-statistics are very high in both models.

(vii) Explain why in this case the local average treatment effect is the same as the average treatment effect on the treated.

If we have no defiers (by assumption), and we have no always takers (you cannot take the vaccine if you have not been assigned to the treatment), the actually treated population consists only of never takers and compliers. Let p be the proportion of compliers in the assigned treatment group, q be the proportion of compliers in the control group. Then, we know:

$$\mathbb{E}[Y|Z=1] = (1-p)Y^*_{0,NT} + pY^*_{1,C}$$

$$\mathbb{E}[Y|Z=0] = (1-q)Y^*_{0,NT} + qY^*_{0,C}$$

The IV (Wald) estimator is defined as:

$$\delta_W = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{Pr(D=1|Z=1) - Pr(D=1|Z=0)}$$

Which, after realizing that because of randomization, $p=q$, that $Pr(D=1|Z=0)=0$, and substitution of the two above equations simplifies to:

Table 3:

	Dependent variable:	
	Flu	
	(1)	(2)
Treatment	−0.193***	−0.198***
	(0.016)	(0.015)
AgeMother		−0.046***
		(0.002)
EducationMother		−0.027***
		(0.003)
GenderChild		0.013*
		(0.008)
Hhincome		0.00000
		(0.00000)
Married		−0.028
		(0.022)
Nationality		0.090***
		(0.009)
Constant	0.621***	2.153***
	(0.010)	(0.040)
Observations	12,583	12,583
R^2	0.059	0.185
Adjusted R^2	0.059	0.184
Residual Std. Error	0.485 (df = 12581)	0.451 (df = 12575)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 4: First stage regressions

	Dependent variable:	
	Treatment	
	(1)	(2)
TreatGroup	0.668***	0.669***
	(0.009)	(0.009)
AgeMother		0.027***
		(0.001)
EducationMother		0.012***
		(0.003)
GenderChild		0.001
		(0.007)
Hhincome		0.00001***
		(0.00000)
Married		0.068***
		(0.020)
Nationality		−0.080***
		(0.008)
Constant	0.000	−0.925***
	(0.008)	(0.037)
Observations	12,583	12,583
R^2	0.285	0.334
Adjusted R^2	0.285	0.334
Residual Std. Error	0.422 (df = 12581)	0.407 (df = 12575)
F Statistic	5,016.237*** (df = 1; 12581)	901.599*** (df = 7; 12575)
Note:		*p<0.1; **p<0.05; ***p<0.01

$$\delta_W = \frac{p(Y^*_{1,C} - Y^*_{0,C} + (1-p)(Y^*_{0,NT} - Y^*_{0,NT}))}{p}$$

Which finally simplifies to:

$$\delta_W = Y^*_{1,c} - Y^*_{0,c} = ATET$$

where the $ATET$ is the definition of the average treatment effect on the treated in this setting, because the only subjects that are treated are compliers (i.e. always takers do not exist).