Applied Microeconometrics - Assignment 2

Walter Verwer (589962) & Bas Machielsen (590049)

September 4, 2021

1. Compute the average probability to receive benefits 10 and 30 weeks after application for applicants that had a search period and applicants that did not have a search period.

```
dataset %>%
    group_by(searchperiod) %>%
    summarize(prob_10weeks = mean(benefits_week10), prob_30weeks = mean(benefits_week30)) %>%
    knitr::kable()
```

searchperiod	prob_10weeks	prob_30weeks
0	0.7359116	0.5403315
1	0.5723684	0.4144737

It seems that there is a large difference in unconditional means in the outcome variable among treated and controlled groups. Individuals exposed to the treatment (a search period) have much lower probabilities of ultimately receiving benefits, whether this is after 10 weeks, or after 30 weeks. This could be a potential indication of the presence of a treatment effect, but a more rigorous examination should ensue.

2. Make a balancing table in which you compare characteristics of applicants with and without a search period.

```
modelsummary::datasummary_balance(~ searchperiod,
                                  data = dataset %>%
                                      select(c(1,4:23)) %>%
                                      mutate(searchperiod = if_else(
                                          searchperiod == 1,
                                          "With Search",
                                          "Without Search")),
                                  output = "latex",
                                  fmt = "%.3f",
                                  dinm = TRUE,
                                  dinm_statistic = "p.value"
                                    ) %>%
    kableExtra::kable_styling(font_size = 10)
```

It seems that all covariates are rather balanced, indicated by the absence of significant differences in means among the treated and the control group. Of course, because we are dealing with a large number of joint null-hypotheses, we should only reject the null hypothesis according to a Bonferroni-corrected p-value. If our regular p-value criterion would be $p < 0.05$, in this case, we reject the null hypothesis when $p < \frac{0.05}{20} = 0.0025$. Even with this criterion, most of the location dummies are still significantly different in treatment and control groups, indicating that perhaps the treatment was administered in different regions, but was stratified according to all other observables. Adding region-specific fixed effects to the regression specifications should solve this problem.

	With Search (N=760)		Without Search (N=905)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
sumincome_12monthsbefore	1.259	1.099	1.296	1.052	0.037	0.485
sumincome_24monthsbefore	2.689	2.125	2.785	2.054	0.096	0.352
age	37.259	8.657	39.926	9.031	2.667	0.000
female	0.372	0.484	0.397	0.490	0.025	0.301
children	0.114	0.319	0.164	0.370	0.049	0.004
partner	0.107	0.309	0.126	0.332	0.019	0.218
period1	0.222	0.416	0.264	0.441	0.042	0.048
period2	0.233	0.423	0.256	0.437	0.023	0.267
period3	0.286	0.452	0.265	0.442	-0.020	0.356
period4	0.259	0.438	0.214	0.411	-0.045	0.033
location1	0.113	0.317	0.177	0.382	0.064	0.000
location2	0.232	0.422	0.182	0.386	-0.049	0.014
location3	0.300	0.459	0.373	0.484	0.073	0.002
location4	0.222	0.416	0.101	0.301	-0.122	0.000
location5	0.133	0.340	0.167	0.373	0.034	0.052
educ_bachelormaster	0.267	0.443	0.264	0.441	-0.003	0.890
educ_prepvocational	0.200	0.400	0.218	0.413	0.018	0.376
educ_primaryorless	0.149	0.356	0.130	0.337	-0.018	0.285
educ_unknown	0.050	0.218	0.014	0.119	-0.036	0.000
educ_vocational	0.334	0.472	0.373	0.484	0.039	0.095

3. Regress the outcome variables first only on whether or not a search period was applied (which should give the difference-in-means estimate) and next include other covariates in the regression.

```
model1 <- lm(data = dataset, formula = benefits_week10 ~ searchperiod)
model2 <- lm(data = dataset, formula = benefits_week30 ~ searchperiod)
model3 <- update(model1, . ~ . + period1 + period2 + period3 + period4 +
                     location1 + location2 + location3 + location4)
model4 <- update(model2, . ~ . + period1 + period2 + period3 + period4 +
                     location1 + location2 + location3 + location4)
model5 <- update(model3, . ~ . + sumincome_12monthsbefore +
                     sumincome_24monthsbefore + age + female + children +
                     partner + educ_bachelormaster + educ_prepvocational +
                     educ_primaryorless + educ_unknown + educ_vocational)
model6 <- update(model4, . ~ . + sumincome_12monthsbefore +
                     sumincome_24monthsbefore + age + female + children +
                     partner + educ_bachelormaster + educ_prepvocational +
                     educ_primaryorless + educ_unknown + educ_vocational)

models <- list(model1, model2, model3, model4, model5, model6)
```

```
stargazer(models, title = "Estimations of the Effect of Search on P(Benefits)",
          label = "tab:reg", header=FALSE, model.names = FALSE,
          column.sep.width="0pt",
          df=F,
          dep.var.labels = c(rep("Benefits",6)),
          column.labels= c(rep(c("10 Weeks", "30 Weeks"),3)),
          omit = c("period1", "period2", "period3", "period4","location"),
          add.lines = list(c("Period Dummies", rep("No", 2), rep("Yes", 4)),
                           c("Region Dummies", rep("No", 2), rep("Yes", 4))),
          omit.stat = c("ll", "ser", "rsq"))
```

The results imply that the treatment is effective in reducing by 10-percentage points the probability of receiving benefits on the long-term (30 weeks), and slightly higher (15 percentage points) on the short-term (10-weeks). If there is no selection on unobservables, these estimates give a good estimate of the ATE. But to what extent can these estimates be trusted?

Table 2: Estimations of the Effect of Search on P(Benefits)

	Dependent variable:					
	Benefits	Benefits	Benefits	Benefits	Benefits	Benefits
	10 Weeks	30 Weeks	10 Weeks	30 Weeks	10 Weeks	30 Weeks
	(1)	(2)	(3)	(4)	(5)	(6)
searchperiod	−0.164***	−0.126***	−0.157***	−0.121***	−0.143***	−0.099***
	(0.023)	(0.024)	(0.023)	(0.025)	(0.024)	(0.025)
sumincome_12monthsbefore					0.0004	−0.022
					(0.027)	(0.028)
sumincome_24monthsbefore					−0.009	−0.005
					(0.014)	(0.014)
age					0.001	0.004***
					(0.001)	(0.001)
female					−0.010	−0.028
					(0.024)	(0.026)
children					−0.037	0.002
					(0.037)	(0.040)
partner					0.056	0.078*
					(0.040)	(0.043)
educ_bachelormaster					−0.092***	−0.116***
					(0.029)	(0.031)
educ_prepvocational					0.013	0.022
					(0.032)	(0.033)
educ_primaryorless					−0.034	0.033
					(0.037)	(0.039)
educ_unknown					−0.381***	−0.270***
					(0.068)	(0.072)
educ_vocational						
Constant	0.736***	0.540***	0.682***	0.404***	0.723***	0.326***
	(0.016)	(0.016)	(0.038)	(0.040)	(0.068)	(0.072)
Period Dummies	No	No	Yes	Yes	Yes	Yes
Region Dummies	No	No	Yes	Yes	Yes	Yes
Observations	1,665	1,665	1,665	1,665	1,663	1,663
Adjusted R^2	0.029	0.015	0.034	0.020	0.057	0.054
F Statistic	50.771***	26.592***	8.301***	5.298***	6.565***	6.304***

Note: *p<0.1; **p<0.05; ***p<0.01

4. Compute the no-assumption bounds for the treatment effects.

```r
# Implement the no assumption bounds
no_assumption_bounds <- function(dataset, y_min, y_max, treatmentvar, depvar){
  depvar <- dplyr::enquo(depvar)
  treatmentvar <- dplyr::enquo(treatmentvar)

  pr_treated <- dataset %>%
    summarize(mean = mean(UQ(treatmentvar), na.rm = TRUE)) %>%
    pull()

  pr_untreated <- 1-pr_treated

  expected_y_given_deq1 <- dataset %>%
    dplyr::filter(UQ(treatmentvar) == 1) %>%
            summarize(mean = mean(UQ(depvar), na.rm = TRUE)) %>%
            pull()

  expected_y_given_deq0 <- dataset%>%
    dplyr::filter(UQ(treatmentvar) == 0) %>%
            summarize(mean = mean(UQ(depvar), na.rm = TRUE)) %>%
            pull()

  # bounds on y^*_1:
  lower_bound_y1 <- expected_y_given_deq1 * pr_treated + y_min * pr_untreated
  upper_bound_y1 <- expected_y_given_deq1 * pr_treated + y_max * pr_untreated

  # bounds on y^*_0:
  lower_bound_y0 <- expected_y_given_deq0 * pr_untreated + y_min * pr_treated
  upper_bound_y0 <- expected_y_given_deq0 * pr_untreated + y_max * pr_treated

  # bounds on the ATE:
  lower_bound_ate <- expected_y_given_deq1*pr_treated - expected_y_given_deq0*pr_untreated +
    (y_min + y_max)*pr_untreated - y_max
  upper_bound_ate <- expected_y_given_deq1*pr_treated - expected_y_given_deq0*pr_untreated +
    (y_min + y_max)*pr_untreated - y_min

  out <- tribble(~"lower_bound_y1", ~"upper_bound_y1", ~"lower_bound_y0",
                 ~"upper_bound_y0", ~"lower_bound_ate", ~"upper_bound_ate",
          lower_bound_y1, upper_bound_y1, lower_bound_y0, upper_bound_y0, lower_bound_ate, upper_bound_ate)

  return(out)
}

no_assumption_bounds(dataset, 0,1,searchperiod, benefits_week10) %>%
  knitr::kable(booktabs=T) %>%
    kableExtra::kable_styling(font_size = 7, latex_options = "hold_position")
```

lower_bound_y1	upper_bound_y1	lower_bound_y0	upper_bound_y0	lower_bound_ate	upper_bound_ate
0.2612613	0.8048048	0.4	0.8564565	-0.5951952	0.4048048

```r
no_assumption_bounds(dataset, 0,1,searchperiod, benefits_week30) %>%
  knitr::kable(booktabs=T) %>%
    kableExtra::kable_styling(font_size = 7, latex_options = "hold_position")
```

lower_bound_y1	upper_bound_y1	lower_bound_y0	upper_bound_y0	lower_bound_ate	upper_bound_ate
0.1891892	0.7327327	0.2936937	0.7501502	-0.560961	0.439039

5. Assume that caseworkers only apply search periods to applicants who benefit from it. How does this affects the bounds.

If people only select into the treatment if it works (meaning, decreasing the probability of benefits), we have:

$$\mathbb{E}[Y_1^*|D=1] \leq \mathbb{E}[Y_0^*|D=1] \text{ and } \mathbb{E}[Y_0^*|D=0] \leq \mathbb{E}[Y_1^*|D=0]$$

Since the case is the opposite of the case that is worked out on the lecture slides, we cannot blindly apply the formulate, but realizing that:

$$y_{min} \leq \mathbb{E}[Y_1^*|D=1] \leq \mathbb{E}[Y_0^*|D=1] \leq y_{max} \text{ and} y_{min} \leq \mathbb{E}[Y_0^*|D=0] \leq \mathbb{E}[Y_1^*|D=0] \leq y_{max}$$

We can evaluate $\mathbb{E}[Y_1^*]$, and we get:

$$\mathbb{E}[Y_1^*|D=1] * \Pr[D=1] + \Pr[D=0] * \mathbb{E}[Y_0^*|D=0] \leq \mathbb{E}[Y_1^*] \leq \mathbb{E}[Y_1^*|D=1] * \Pr[D=1] + y_{max} * \Pr[D=0]$$

And for $\mathbb{E}[Y_0^*]$, we get:

$$\mathbb{E}[Y_0^*|D=0] * \Pr[D=0] + \Pr[D=1] * \mathbb{E}[Y_1^*|D=1] \leq \mathbb{E}[Y_0^*] \leq \mathbb{E}[Y_0^*|D=0] * \Pr[D=0] + \Pr[D=1] * y_{max}$$

Then, realizing that the lower bound of $\mathbb{E}[Y_1^* - Y_0^*]$ is the lower bound of $\mathbb{E}[Y_1^*]$ minus the upper bound of $\mathbb{E}[Y_0^*]$, and *mutatis mutandis* for the upper bound of $\mathbb{E}[Y_1^* - Y_0^*]$, after rewriting, we find:

$$-\Pr(D=1) \cdot (y_{max} - \mathbb{E}[Y_1^*|D=1]) \leq \mathbb{E}[Y_1^* - Y_0^*] \leq \Pr(D=0) \cdot (y_{max} - \mathbb{E}[Y_0^*|D=0])$$

Which corresponds to the same properties as found in the lecture slides (i.e. narrower bounds, but without ever excluding zero).

6. Next, imposed the monotone treatment response and the monotone treatment selection assumption separately and also jointly.

7. Usually higher educated workers have more favorable labor market outcomes. Use education as monotone instrumental variable and compute the bounds.