Applied Microeconometrics, Assignment 2: Bounds

Sindri Engilbertsson (584872), Ilse van der Voort (584098)

September 2021

To answer the questions we have made extensive use of the lecture slides and recommended readings. Data analyses are done in R and Stata. The first 6 questions are based on Stata code and question 7 is attached in a different format with R code.

Note that for this assignment we assume that people who do not receive benefits are either in their search period or they have a job. There is no one unemployed who is neither in their search period nor receiving benefits.

Q1: 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.

Search period is a dummy variable that equals one when someone had a search period. There are 760 applicants with a search period and 905 applicants without a search period.
For applicants that had a search period the average of benefits_week10 is .5723684 and the average of benefits_week30 is .4144737. For applicants that did not have a search period the average of benefits_week10 is .7359116 and the average of benefits_week30 is .5403315. These are the average probabilities to receive benefits either 10 or 30 weeks after application. The numbers indicate that at first sight it seems that applicants with a search period have a lower probability of receiving benefits. A t-test reveals that the differences are indeed significant at the 1% level.

Q2: Make a balancing table in which you compare characteristics of applicants with and without a search period.

	n_0	mean_0	sd_0	n_1	mean_1	sd_1	Diff
sumincome_12monthsbefore	905	1.30	1.05	760	1.26	1.10	-0.037
sumincome_24monthsbefore	905	2.78	2.05	760	2.69	2.12	-0.096
age	904	39.93	9.03	760	37.26	8.66	-2.667***
female	904	0.40	0.49	760	0.37	0.48	-0.025
children	905	0.16	0.37	760	0.11	0.32	-0.049***
partner	905	0.13	0.33	760	0.11	0.31	-0.019
period1	905	0.26	0.44	760	0.22	0.42	-0.042**
period2	905	0.26	0.44	760	0.23	0.42	-0.023
period3	905	0.27	0.44	760	0.29	0.45	0.020
period4	905	0.21	0.41	760	0.26	0.44	0.045**
location1	905	0.18	0.38	760	0.11	0.32	-0.064***
location2	905	0.18	0.39	760	0.23	0.42	0.049**
location3	905	0.37	0.48	760	0.30	0.46	-0.073***
location4	905	0.10	0.30	760	0.22	0.42	0.122***
location5	905	0.17	0.37	760	0.13	0.34	-0.034*
educ_bachelormaster	905	0.26	0.44	760	0.27	0.44	0.003
educ_prepvocational	905	0.22	0.41	760	0.20	0.40	-0.018
educ_primaryorless	905	0.13	0.34	760	0.15	0.36	0.018
educ_unknown	905	0.01	0.12	760	0.05	0.22	0.036***
educ_vocational	905	0.37	0.48	760	0.33	0.47	-0.039*

The balancing table shows that there are systematic differences between the groups with and without a search period. There are significant differences at the 1% level in age and having children, where those with a search period are on average younger and have less children than those without a search period. There are also significant differences between the groups for period1 and period4, at the 5% level, and significant differences between locations. For education, there are significant differences for unknown education at the 1% level and for vocational education at the 10% level. This indicates that the group who gets a search period has on average less often vocational education and more often unknown education than the group without a search period.

Q3: 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.

For this exercise we run 6 regressions. First we run regressions of the dependent variables on search period only. Then we run regressions where we include all the available covariates and regressions with the covariates that were unbalanced according to the previous item. For this we use a 5 percent significance level.
The reason for the second set of regressions is that whether or not you receive benefits might depend on more variables than just having a search period. We add the balanced variables because these might independently affect the outcome variables. We furthermore argue that none of the covariates are intermediate variables and that we therefore do not have a bad control problem. In interpreting these two regressions it is important to note that we are working with dummy variables. The baseline group regards period1, location1, and unknown education.
The reason for the last set of regressions is that in the previous item we saw that whether or not you get a search period is not random. There is some sample selection. We include the unbalanced covariates in case the variables that are unbalanced also affect the dependent variables. We include the unbalanced variables to make sure that the impact they have on the dependent variables is controlled for. The baseline group here is different than in the previous graphs, namely period2 or period3 and location5.
Regression of benefits_week10 on searchperiod:

```
. reg benefits_week10 searchperiod

      Source |       SS       df       MS              Number of obs =    1665
-------------+------------------------------           F(  1,  1663) =   50.77
       Model |  11.0487416     1  11.0487416           Prob > F      =  0.0000
    Residual |   361.90261  1663  .217620331           R-squared     =  0.0296
-------------+------------------------------           Adj R-squared =  0.0290
       Total |  372.951351  1664  .224129418           Root MSE      =   .4665

------------------------------------------------------------------------------
benefits_~10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
searchperiod |  -.1635432   .0229523    -7.13   0.000    -.2085616   -.1185248
       _cons |   .7359116   .0155069    47.46   0.000     .7054965    .7663267
------------------------------------------------------------------------------
```

Regression of benefits_week30 on searchperiod:

```
. reg benefits_week30 searchperiod

      Source |       SS       df       MS              Number of obs =    1665
-------------+------------------------------           F(  1,  1663) =   26.59
       Model |  6.54347214     1  6.54347214           Prob > F      =  0.0000
    Residual |   409.21869  1663  .246072574           R-squared     =  0.0157
-------------+------------------------------           Adj R-squared =  0.0151
       Total |  415.762162  1664  .249857069           Root MSE      =  .49606

------------------------------------------------------------------------------
benefits_~30 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
searchperiod |  -.1258578   .0244066    -5.16   0.000    -.1737287   -.0779869
       _cons |   .5403315   .0164895    32.77   0.000     .5079891    .5726738
------------------------------------------------------------------------------
```

Regression of benefits_week10 on searchperiod and all covariates:

```
. reg benefits_week10 searchperiod s~12months~e s~24months~e age female children ///
> partner period2 period3 period4 location2 location3 location4 location5 ///
> educ_bache~r educ_prepv~l educ_prima~s educ_vocat~l

      Source |       SS       df       MS              Number of obs =    1663
-------------+------------------------------           F( 18,  1644) =    6.57
       Model |  24.9517209    18  1.38620672           Prob > F      =  0.0000
    Residual |  347.124046  1644  .211146013           R-squared     =  0.0671
-------------+------------------------------           Adj R-squared =  0.0568
       Total |  372.075767  1662  .223872302           Root MSE      =  .45951

------------------------------------------------------------------------------------------
         benefits_week10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------+----------------------------------------------------------------
            searchperiod |  -.1431602   .0235784    -6.07   0.000     -.189407   -.0969134
suminconme_12monthsbefore |   .0004347   .0265421     0.02   0.987    -.0516252    .0524946
suminconme_24monthsbefore |  -.0086113    .013645    -0.63   0.528    -.0353748    .0181521
                     age |   .0005504   .0012869     0.43   0.669    -.0019737    .0030746
                  female |  -.0099765   .0243116    -0.41   0.682    -.0576616    .0377085
                children |  -.0373422   .0374337    -1.00   0.319     -.110765    .0360806
                 partner |    .055736   .0404568     1.38   0.168    -.0236163    .1350882
                 period2 |   .0080654   .0325099     0.25   0.804    -.0556998    .0718305
                 period3 |   .0490961   .0315178     1.56   0.119    -.0127232    .1109154
                 period4 |  -.0494219   .0330051    -1.50   0.134    -.1141584    .0153146
               location2 |   .0256358   .0398984     0.64   0.521    -.0526212    .1038928
               location3 |   -.005591   .0362149    -0.15   0.877    -.0766232    .0654412
               location4 |  -.0653579   .0424718    -1.54   0.124    -.1486624    .0179466
               location5 |  -.0000488   .0425373    -0.00   0.999    -.0834818    .0833842
     educ_bachelormaster |    .288448   .0688544     4.19   0.000     .1533964    .4234996
     educ_prepvocational |     .39409   .0699307     5.64   0.000     .2569274    .5312526
      educ_primaryorless |   .3466469   .0722185     4.80   0.000      .204997    .4882968
         educ_vocational |   .3807055   .0678122     5.61   0.000     .2476981    .5137128
                   _cons |   .3922324   .0915833     4.28   0.000     .2126002    .5718647
------------------------------------------------------------------------------------------
```

Regression of benefits_week30 on searchperiod and all covariates:

```
. reg benefits_week30 searchperiod s~12months~e s~24months~e age female children ///
> partner period2 period3 period4 location2 location3 location4 location5 ///
> educ_bache~r educ_prepv~l educ_prima~s educ_vocat~l
```

Source	SS	df	MS
Model	26.812735	18	1.48959639
Residual	388.482515	1644	.236303233
Total	415.29525	1662	.249876805

Number of obs =	1663
F(18, 1644) =	6.30
Prob > F =	0.0000
R-squared =	0.0646
Adj R-squared =	0.0543
Root MSE =	.48611

benefits_week30	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
searchperiod	-.0989639	.0249435	-3.97	0.000	-.1478883	-.0500396
sumincome_12monthsbefore	-.0220997	.0280788	-0.79	0.431	-.0771737	.0329742
sumincome_24monthsbefore	-.0052608	.014435	-0.36	0.716	-.0335738	.0230522
age	.0041273	.0013614	3.03	0.002	.001457	.0067976
female	-.0281157	.0257192	-1.09	0.274	-.0785615	.0223301
children	.002061	.039601	0.05	0.958	-.0756127	.0797348
partner	.0776513	.0427991	1.81	0.070	-.0062952	.1615978
period2	.0453173	.0343921	1.32	0.188	-.0221396	.1127742
period3	.0258131	.0333426	0.77	0.439	-.0395854	.0912116
period4	-.0700476	.034916	-2.01	0.045	-.1385322	-.0015631
location2	-.0071861	.0422084	-0.17	0.865	-.0899739	.0756018
location3	.0306186	.0383116	0.80	0.424	-.0445261	.1057634
location4	-.026533	.0449308	-0.59	0.555	-.1146606	.0615946
location5	-.0477859	.0450001	-1.06	0.288	-.1360494	.0404776
educ_bachelormaster	.1540551	.0728409	2.11	0.035	.0111844	.2969258
educ_prepvocational	.2917187	.0739795	3.94	0.000	.1466148	.4368226
educ_primaryorless	.3025644	.0763997	3.96	0.000	.1527135	.4524154
educ_vocational	.2699342	.0717383	3.76	0.000	.1292261	.4106423
_cons	.1738533	.0968857	1.79	0.073	-.0161791	.3638857

Regression of benefits_week10 on searchperiod and only the unbalanced covariates:

```
. reg benefits_week10 searchperiod age children period1 period4 location1 location2 ///
> location3 location4 educ_unknown

      Source |       SS       df       MS              Number of obs =    1664
-------------+------------------------------           F( 10,  1653) =    9.51
       Model |  20.2718501    10  2.02718501           Prob > F      =  0.0000
    Residual |  352.241972  1653  .213092542           R-squared     =  0.0544
-------------+------------------------------           Adj R-squared =  0.0487
       Total |  372.513822  1663  .224001096           Root MSE      =  .46162

------------------------------------------------------------------------------
benefits_~10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
searchperiod |  -.1442487   .0236306    -6.10   0.000    -.1905978   -.0978996
         age |    .000624    .001284     0.49   0.627    -.0018944    .0031424
    children |  -.0086599   .0329504    -0.26   0.793    -.0732888    .0559689
     period1 |  -.0246033   .0279529    -0.88   0.379    -.0794301    .0302236
     period4 |   -.073935   .0286337    -2.58   0.010    -.1300972   -.0177728
   location1 |   .0177329   .0418397     0.42   0.672    -.0643316    .0997974
   location2 |   .0506062   .0388819     1.30   0.193    -.0256568    .1268691
   location3 |  -.0032511   .0351508    -0.09   0.926    -.0721959    .0656937
   location4 |  -.0470163   .0413138    -1.14   0.255    -.1280492    .0340167
educ_unknown |  -.3501335   .0662608    -5.28   0.000    -.4800975   -.2201695
       _cons |   .7342045   .0597029    12.30   0.000     .6171031    .8513059
------------------------------------------------------------------------------
```

Regression of benefits_week30 on searchperiod and only the unbalanced covariates:

```
. reg benefits_week30 searchperiod age children period1 period4 location1 location2 ///
> location3 location4 educ_unknown
```

Source	SS	df	MS
Model	16.7265574	10	1.67265574
Residual	398.802289	1653	.241259703
Total	415.528846	1663	.249867015

Number of obs	=	1664
F(10, 1653)	=	6.93
Prob > F	=	0.0000
R-squared	=	0.0403
Adj R-squared	=	0.0344
Root MSE	=	.49118

benefits_~30	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
searchperiod	-.1003018	.0251439	-3.99	0.000	-.1496191	-.0509845
age	.0044789	.0013662	3.28	0.001	.0017992	.0071586
children	.0359117	.0350605	1.02	0.306	-.032856	.1046795
period1	-.0274732	.029743	-0.92	0.356	-.0858112	.0308648
period4	-.0984597	.0304674	-3.23	0.001	-.1582186	-.0387009
location1	.0807131	.0445192	1.81	0.070	-.0066068	.1680331
location2	.0801569	.0413719	1.94	0.053	-.00099	.1613038
location3	.0855637	.0374019	2.29	0.022	.0122036	.1589237
location4	.0590534	.0439596	1.34	0.179	-.0271689	.1452758
educ_unknown	-.2510632	.0705042	-3.56	0.000	-.3893502	-.1127762
_cons	.3214715	.0635264	5.06	0.000	.1968709	.4460722

The first two graphs show a significant negative effect of having a search period on receiving benefits at the 1% level. Having benefits is a bad thing, so this negative estimate is good. The outcome variable is a dummy variable, so we can interpret the estimates as percentages. Having a search period leads to a 16 or 12 percentage point lower chance of receiving benefits. Without a search period, the percentage of people receiving benefits is 74 and 54 in the first and second regression respectively. When including control variables, the search period still has a significant negative effect at the 1% level in all regressions. The estimates differ between the first two regressions and the regressions with control variables, where we argue that the latter are more reliable. There is rarely any difference between the second and third sets of regressions, indicating that the balanced variables are not needed as controls.

Q4: Compute the no-assumption bounds for the treatment effects.

The lower no-assumption bound is the lower bound of the potentially treated minus the upper bound of the potentially untreated. The upper no-assumption bound is the upper bound of the potentially treated minus the lower bound of the potentially untreated. This gives the following:

$$
\begin{gathered}
(E[Y|D=1]-y_{max})Pr(D=1)-(E[Y|D=0]-y_{min})Pr(D=0)= \\
E[Y|D=1]Pr(D=1)-E[Y|D=0]Pr(D=0)+(y_{min}+y_{max})Pr(D=0)-y_{max} \\
\leq E[Y_1^*]-E[Y_0^*]\leq \\
E[Y|D=1]Pr(D=1)-E[Y|D=0]Pr(D=0)+(y_{min}+y_{max})Pr(D=0)-y_{min} \\
=(E[Y|D=1]-y_{min})Pr(D=1)-(E[Y|D=0]-y_{max})Pr(D=0)
\end{gathered}
$$

Here, Y is benefits_week10 or benefits_week30. $D=1$ indicates that someone got a search period. $y_{min}=0$ and $y_{max}=1$. The probability of having a search period is $Pr(D=1)=\frac{760}{1665}$ and the probability of not having a search period is $Pr(D=0)=1-Pr(D=0)=\frac{905}{1665}$.
For benefits_week10 this gives:

$$
-0.595\approx 0.5723684\times\frac{760}{1665}-0.7359116\times\frac{905}{1665}+\frac{905}{1665}-1\leq E[Y_1^*]-E[Y_0^*]\leq
$$
$$
0.5723684\times\frac{760}{1665}-0.7359116\times\frac{905}{1665}+\frac{905}{1665}\approx 0.405
$$

For benefits_week30 this gives:

$$-0.561 \approx 0.4144737 \times \frac{760}{1665} - 0.5403315 \times \frac{905}{1665} + \frac{905}{1665} - 1 \leq E[Y_1^*] - E[Y_0^*] \leq$$
$$0.4144737 \times \frac{760}{1665} - 0.5403315 \times \frac{905}{1665} + \frac{905}{1665} \approx 0.439$$

The bounds are tighter than the possible bounds without data, but are still very wide and not very helpful in determining the treatment effect and possibly giving policy advice.

Q5: Assume that caseworkers only apply search periods to applicants who benefit from it. How does this affects the bounds.

We assume that the caseworkers apply search periods to those who benefit from it, we also assume that they only don't apply search periods to applicants that won't benefit from it. For every applicant case workers can choose between assigning a search period or not, they always assign the option that gives the applicant a lower expected value of Y. This gives $E[Y_0^*|D=0] \leq E[Y_1^*|D=0]$ and $E[Y_1^*|D=1] \leq E[Y_0^*|D=1]$.
These new assumptions give us a higher y_{min} than for the no-assumption bounds, making the bounds tighter. In calculating the bounds we now use $y_{min} = E[Y|D=1]$ and $y_{min} = E[Y|D=0]$ for $E[Y_1^*]$ and $E[Y_0^*]$ respectively, whereas we used $y_{min} = 0$ for the no-assumption bounds. Our new bounds for $E[Y_1^*]$ are now:

$$E[Y|D=1]Pr(D=1) + y_{min}Pr(D=0) \leq E[Y_1^*] \leq E[Y|D=1]Pr(D=1) + y_{max}Pr(D=0)$$
$$E[Y|D=1]Pr(D=1) + E[Y|D=0]Pr(D=0) \leq E[Y_1^*] \leq E[Y|D=1]Pr(D=1) + y_{max}Pr(D=0)$$

Our new bounds for $E[Y_0^*]$ are:

$$y_{min}Pr(D=1) + E[Y|D=0]Pr(D=0) \leq E[Y_0^*] \leq y_{max}Pr(D=1) + E[Y|D=0]Pr(D=0)$$
$$E[Y|D=1]Pr(D=1) + E[Y|D=0]Pr(D=0) \leq E[Y_0^*] \leq y_{max}Pr(D=1) + E[Y|D=0]Pr(D=0)$$

Our new lower bound for $E[Y_1^*] - E[Y_0^*]$ is again the lower bound of the potentially treated minus the upper bound of the potentially untreated. The new upper bound is the upper bound of the potentially treated minus the lower bound of the potentially untreated:

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

Filling out the numbers for benefits_week10 this gives:

$$-0.195 \approx (0.5723684 - 1)\frac{760}{1665} \leq E[Y_1^*] - E[Y_0^*] \leq (1 - 0.7359116)\frac{905}{1665} \approx 0.146$$

Filling out the numbers for benefits_week30 this gives:

$$-0.267 \approx (0.414473 - 1)\frac{760}{1665} \leq E[Y_1^*] - E[Y_0^*] \leq (1 - 0.540331)\frac{905}{1665} \approx 0.250$$

The new bounds are tighter than the no-assumption bounds but still include 0.

Q6: Next, imposed the monotone treatment response and the monotone treatment selection assumption separately and also jointly.

The Monotone Treatment Selection (MTS) assumption assumes that individuals assigned to treatment have a better expected value from the treatment than those not assigned treatment. Furthermore, with MTS people who get treatment overall have better Y values. We then get that the following statements hold:

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

$E[Y_1^*]$ is now bounded from below ($E[Y|D=1] \leq E[Y_1^*]$) and $E[Y_0^*]$ is now bounded from above ($E[Y_0^*] \leq E[Y|D=0]$). This changes the lower bound of $E[Y_1^*] - E[Y_0^*]$ and we get:

$$E[Y|D=1] - E[Y|D=0]$$
$$\leq E[Y_1^*] - E[Y_0^*] \leq$$
$$E[Y|D=1]P(D=1) - E[Y|D=0]P(D=0) + (y_{min} + y_{max})P(D=0) - y_{min}$$

In numbers this gives the following bounds for benefits_week10:

$$-0.164 \leq E[Y_1^*] - E[Y_0^*] \leq 0.405$$

In numbers this gives the following bounds for benefits_week30:

$$-0.126 \leq E[Y_1^*] - E[Y_0^*] \leq 0.439$$

The Monotone Treatment Response (MTR) assumption states that treatment can only improve the outcomes, i.e. that $y_{min} \leq Y_1^* \leq Y_0^* \leq y_{max}$. $E[Y_1^*]$ is now bounded from above ($E[Y_1^*] \leq E[Y|D=1]Pr(D=1) + E[Y|D=0]Pr(D=0)$) and $E[Y_0^*]$ is now bounded from below ($E[Y_0^*] \geq E[Y|D=1]Pr(D=1) + E[Y|D=0]Pr(D=0)$). Combining this shows that the lower bound does not change, but the upper bound equals zero. This gives:

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

In numbers this gives the following bounds for benefits_week10:

$$-0.595 \leq E[Y_1^*] - E[Y_0^*] \leq 0$$

In numbers this gives the following bounds for benefits_week30:

$$-0.561 \leq E[Y_1^*] - E[Y_0^*] \leq 0$$

Combining the MTS and MTR gives us the bounds:

$$E[Y|D=1] - E[Y|D=0] \leq E[Y_1^*] - E[Y_0^*] \leq 0$$

These bounds show that treatment effects are always expected to be negative (which is beneficial). In numbers this gives the following bounds for benefits_week10:

$$-0.164 \leq E[Y_1^*] - E[Y_0^*] \leq 0$$

In numbers this gives the following bounds for benefits_week30:

$$-0.126 \leq E[Y_1^*] - E[Y_0^*] \leq 0$$

Code to load dataset and compute regressions

Commenting style follows http://adv-r.had.co.nz/Style.html‘

```
rm(list = ls())

library(foreign)
library(cobalt)
library(xtable)
library(stargazer)
library(plm)
library(dplyr)
```

List of working directories

Default working directory

```
cd <- 'C:/Users/sindr/Desktop/Tinbergen/2nd-year/Block 1/Applied Microeconometrics/Assignments/code'
```

Folder containing data and folder for outputs

```
dt <- 'C:/Users/sindr/Desktop/Tinbergen/2nd-year/Block 1/Applied Microeconometrics/Assignments/data'
tb <- 'C:/Users/sindr/Desktop/Tinbergen/2nd-year/Block 1/Applied Microeconometrics/Assignments/Final'
```

Reading in data

```
setwd(dt)
df <- as.data.frame(read.dta('searchperiod.dta'))
setwd(tb)
```

Assignment

Naive estimation

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

Solution

In our data set we have 5 education-related dummy-variables. Let us get a feeling for those variables by considering some simple values for all of them:

```
# How common is each value?
Nz0 <- sum(df[["educ_primaryorless"]]==1)
Nz0
```

```
## [1] 231
```

```r
Nz1 <- sum(df[["educ_prepvocational"]]==1)
Nz1
```

```
## [1] 349
```

```r
Nz2 <- sum(df[["educ_vocational"]]==1)
Nz2
```

```
## [1] 592
```

```r
Nz3 <- sum(df[["educ_bachelormaster"]]==1)
Nz3
```

```
## [1] 442
```

```r
NzU <- sum(df[["educ_unknown"]]==1)
NzU
```

```
## [1] 51
```

We see that a fairly even amount of applicants have prep-vocational schooling, vocational schooling or a bachelor or master degree. However fewer have only a primary education or less, and the education level is unknown for very few applicants.

We can then consider the mean values for applicants from each of those categories:

```r
# 10 weeks mean for different education levels
E0 <- mean(df[["benefits_week10"]][df[["educ_primaryorless"]]==1])
E0
```

```
## [1] 0.6709957
```

```r
E1 <- mean(df[["benefits_week10"]][df[["educ_prepvocational"]]==1])
E1
```

```
## [1] 0.7191977
```

```r
E2 <- mean(df[["benefits_week10"]][df[["educ_vocational"]]==1])
E2
```

```
## [1] 0.6993243
```

```r
E3 <- mean(df[["benefits_week10"]][df[["educ_bachelormaster"]]==1])
E3
```

```
## [1] 0.6040724
```

```r
unk <- mean(df[["benefits_week10"]][df[["educ_unknown"]]==1])
unk
```

```
## [1] 0.2745098
```

```r
# 30 weeks mean for different education levels
E03 <- mean(df[["benefits_week30"]][df[["educ_primaryorless"]]==1])
E03
```

```
## [1] 0.5670996
```

```r
E13 <- mean(df[["benefits_week30"]][df[["educ_prepvocational"]]==1])
E13
```

```
## [1] 0.5444126
```

```r
E23 <- mean(df[["benefits_week30"]][df[["educ_vocational"]]==1])
E23
```

```
## [1] 0.5084459
```

```r
E33 <- mean(df[["benefits_week30"]][df[["educ_bachelormaster"]]==1])
E33
```

```
## [1] 0.3891403
```

```r
unk3 <- mean(df[["benefits_week30"]][df[["educ_unknown"]]==1])
unk3
```

```
## [1] 0.1960784
```

Again we see that the applicants whose educational level is unkown stick out like a sore thumb, and so we remove those individuals from our sample:

```r
df <- df[df$educ_unknown != 1,]
```

What we then note, is that the difference between the outcomes for applicants in the categories "primaryorless", "prepvocational", and "vocational", is far smaller and less pronounced than the difference between the outcomes for those applications and those from "bachelormaster". Simple t-tests reveal that only "bachelormaster" is always significantly different from all other categories with regards to outcomes.

```r
# 10 weeks: prepvocational - vocational
t.test(df[["benefits_week10"]][df[["educ_prepvocational"]]==1]-df[["benefits_week10"]][df[["educ_vocatio
```

```
## Warning in df[["benefits_week10"]][df[["educ_prepvocational"]] == 1] -
## df[["benefits_week10"]][df[["educ_vocational"]] == : longer object length is not
## a multiple of shorter object length
```

```
##
##  One Sample t-test
##
## data:  df[["benefits_week10"]][df[["educ_prepvocational"]] == 1] - df[["benefits_week10"]][df[["educ_
## t = 1.0907, df = 591, p-value = 0.2758
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.02299089  0.08042333
## sample estimates:
##  mean of x
## 0.02871622
```

```
# 30 weeks: primaryorless - prepvocational
t.test(df[["benefits_week30"]][df[["educ_primaryorless"]]==1]-df[["benefits_week30"]][df[["educ_prepvoca
```

```
## Warning in df[["benefits_week30"]][df[["educ_primaryorless"]] == 1] -
## df[["benefits_week30"]][df[["educ_prepvocational"]] == : longer object length is
## not a multiple of shorter object length
```

```
##
##  One Sample t-test
##
## data:  df[["benefits_week30"]][df[["educ_primaryorless"]] == 1] - df[["benefits_week30"]][df[["educ_p
## t = -0.076363, df = 348, p-value = 0.9392
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.07666479  0.07093414
## sample estimates:
##   mean of x
## -0.00286533
```

It thus seems logical to use the variable "educ_bachelormaster" as our monotone instrumental variable against all other forms of education. In this way, we also avoid splitting our fairly limited dataset of only 1614 observations into too many cells. So, we set $Z = 1$ if "educ_bachelormaster"$= 1$.

The Monotone Instrumental Variable assumption we make is that $E[Y_d^*|Z = 0] \geq E[Y_d^*|Z = 1], \quad d = 0, 1$. Keep in mind that a lower value of Y is preferred by the case worker.

To calculate the bounds, we take the following as a starting point:

$$E[Y_1^*] = Pr(Z = 0)E[Y_1^*|Z = 0] + Pr(Z = 1)E[Y_1^*|Z = 1]$$

The probabilities are easily calculated:

```
Pz1 <- sum(df[["educ_bachelormaster"]]==1)/nrow(df)
Pz1
```

```
## [1] 0.2738538
```

```
Pz0 <- 1-Pz1
Pz0
```

```
## [1] 0.7261462
```

The best lower bound for $E[Y_1^*|Z=0]$ is the maximum value of the bounds $LB(d=1,Z=0)$ and $LB(d=1,Z=1)$. We calculate these lower bounds in the following way:

$$E[Y_d^*|Z=0] \geq LB(d,Z=0) = E[Y_d|z=0]Pr(Z=0) + y_{min}Pr(Z=1), \quad d \in 0,1$$

For $Z=0$, and:

$$E[Y_d^*|Z=1] \geq LB(d,Z=1) = E[Y_d|Z=1]Pr(Z=1) + y_{min}Pr(Z=0), \quad d \in 0,1$$

Since we have that $y_min = 0$, the above equations simplify to just $E[Y_1^*|Z=i] = E[Y_1|z=i], \quad i \in 1,2$. We will call the better/higher lower bound $LB_{max}(1,z)$, and the worse lower bound $LB_{min}(1,z)$ for $z \in 0,1$. Let us calculate those values:

```
# 10 weeks
LB10d1z0 <- mean(df[df$searchperiod == 1 & df$educ_bachelormaster == 0,]$benefits_week10)*Pz0
LB10d1z1 <- mean(df[df$searchperiod == 1 & df$educ_bachelormaster == 1,]$benefits_week10)*Pz1

LB10d0z0 <- mean(df[df$searchperiod == 0 & df$educ_bachelormaster == 0,]$benefits_week10)*Pz0
LB10d0z1 <- mean(df[df$searchperiod == 0 & df$educ_bachelormaster == 1,]$benefits_week10)*Pz1

# 30 weeks
LB30d1z0 <- mean(df[df$searchperiod == 1 & df$educ_bachelormaster == 0,]$benefits_week30)*Pz0
LB30d1z1 <- mean(df[df$searchperiod == 1 & df$educ_bachelormaster == 1,]$benefits_week30)*Pz1

LB30d0z0 <- mean(df[df$searchperiod == 0 & df$educ_bachelormaster == 0,]$benefits_week30)*Pz0
LB30d0z1 <- mean(df[df$searchperiod == 0 & df$educ_bachelormaster == 1,]$benefits_week30)*Pz1
```

In the same way we acquire the upper bounds for $E[Y_1^*|Z=0]$ is the minimum value of the bounds $UB(d=1,Z=0)$ and $UB(d=1,Z=1)$. Which we calculate in the following way:

$$E[Y_1^*|Z=0] \leq UB(d=1,Z=0) = E[Y_1|z=0]Pr(Z=0) + y_{max}Pr(Z=1)$$

For $Z=0$, and:

$$E[Y_1^*|Z=1] \leq UB(d=1,Z=1) = E[Y_1|Z=1]Pr(Z=1) + y_{max}Pr(Z=0).$$

We can calculate the upper bounds:

```
# 10 weeks
UB10d1z0 <- (LB10d1z0+Pz1)
UB10d1z1 <- (LB10d1z1+Pz0)

UB10d0z0 <- (LB10d0z0+Pz1)
UB10d0z1 <- (LB10d0z1+Pz0)

# 30 weeks
UB30d1z0 <- (LB30d1z0+Pz1)
UB30d1z1 <- (LB30d1z1+Pz0)

UB30d0z0 <- (LB30d0z0+Pz1)
UB30d0z1 <- (LB30d0z1+Pz0)
```

Using this to make some substitutions, we get:

$$E[Y_1^*] \geq Pr(Z=0)LB_{max}(1,0) + Pr(Z=1)LB_{min}(1,1)$$

Which gives us the complete lower bound:

```r
# 10 weeks
LB10_d1 <- Pz0*max(LB10d1z1, LB10d1z0)+Pz1*min(LB10d1z1, LB10d1z0)
LB10_d0 <- Pz0*max(LB10d0z1, LB10d0z0)+Pz1*min(LB10d0z1, LB10d0z0)

# 30 weeks
LB30_d1 <- Pz0*max(LB30d1z1, LB30d1z0)+Pz1*min(LB30d1z1, LB30d1z0)
LB30_d0 <- Pz0*max(LB30d0z1, LB30d0z0)+Pz1*min(LB30d0z1, LB30d0z0)
```

Similarly for the upper bound we get:

$$E[Y_1^*] \leq Pr(Z=0)UB_{max}(1,0) + Pr(Z=1)UB_{min}(1,1)$$

Which gives us the complete upper bound:

```r
# 10 weeks
UB10_d1 <- Pz0*max(UB10d1z1, UB10d1z0)+Pz1*min(UB10d1z1, UB10d1z0)
UB10_d0 <- Pz0*max(UB10d0z1, UB10d0z0)+Pz1*min(UB10d0z1, UB10d0z0)

# 30 weeks
UB30_d1 <- Pz0*max(UB30d1z1, UB30d1z0)+Pz1*min(UB30d1z1, UB30d1z0)
UB30_d0 <- Pz0*max(UB30d0z1, UB30d0z0)+Pz1*min(UB30d0z1, UB30d0z0)
```

Putting it all together, we see that the bounds for $E[Y_0^*]$ are:

$$E[Y_d^*] \geq Pr(Z=0)LB_{max}(0,0)+Pr(Z=1)LB_{min}(0,1)E[Y_d^*] \leq Pr(Z=0)UB_{max}(0,0)+Pr(Z=1)UB_{min}(0,1)$$

This gives us the final, complete MIV bounds as:

```r
LB10_MIV <- LB10_d1 - UB10_d0
UB10_MIV <- UB10_d1 - LB10_d0
```

So we see that using education as a monotone instrumental variable gives us the bounds: -0.524 $\leq E[Y_{10weeks}^*] \leq 0.374$.

```r
LB30_MIV <- LB30_d1 - UB30_d0
LB30_MIV
```

```
## [1] -0.5350021
```

```r
UB30_MIV <- UB30_d1 - LB30_d0
UB30_MIV
```

```
## [1] 0.4187329
```

So we see that using education as a monotone instrumental variable gives us the bounds: -0.535 $\leq E[Y_{30weeks}^*] \leq 0.419$.

As we can see the bounds are slightly tighter than for the no-assumption bounds, but the difference is very small. This leads us to believe that using education as a monotone variable as constructed by us is not very helpful. The results might differ if we use an education variable where we use four categories.