PWRD Aggregation Weights Demo

2022-05-31

In this R Markdown, we present a demonstration of how to calculate PWRD aggregation weights and use them in analysis. We begin with examining synthetic data that mirrors the design we observed in BURST. We then create an R function that will calculate weights for each cohort year. We obtain the weights used in our analysis and then conduct the analysis itself.

Synthetic Data

We begin by examining our data.

```
## # A tibble: 6 x 12
##   blocks Sch   EID    cohort_yr treatment Grade   Yrs Race_White Gend_Fem
##   <chr>  <chr> <chr>  <chr>         <dbl> <dbl> <dbl>      <dbl>    <dbl>
## 1 A      A0    C1A106 11                0     0     1          1        0
## 2 A      A0    C1A106 12                0     1     2          1        0
## 3 A      A0    C1A106 13                0     2     3          1        0
## 4 A      A0    C1A106 14                0     3     4          1        0
## 5 A      A0    C1A170 11                0     0     1          1        0
## 6 A      A0    C1A170 12                0     1     2          1        0
## # ... with 3 more variables: Free_Lunch <dbl>, Y <dbl>, Eligible <dbl>
```

```
##  [1] "blocks"     "Sch"        "EID"        "cohort_yr"  "treatment"
##  [6] "Grade"      "Yrs"        "Race_White" "Gend_Fem"   "Free_Lunch"
## [11] "Y"          "Eligible"
```

This data set consists of the following variables:

- `blocks`: the matched set to which the student belongs.
- `Sch`: a variable denoting the school to which the student belongs.
- `EID`: the individual student's identification code.
- `cohort_yr`: a variable denoting the student's cohort-year.
- `treatment`: an indicator denoting whether the student belongs to a treatment or control school.
- `Grade`: the grade of the student.
- `Yrs`: the number of years the student has participated in the study.
- `Race_White`: an indicator denoting the student is white.
- `Gend_Fem`: an indicator denoting the student is female.
- `Free_Lunch`: an indicator denoting the student is eligible for free or reduced price lunch.
- `Y`: the student's end of year outcome.
- `Eligible`: a variable denoting whether the student has ever become eligible for supplemental instruction.

The cohorts are as follows:

```
##
##   11   12   13   14   21   22   23   31   32   41
## 8000 6000 4000 2000 2000 2000 2000 2000 2000 2000
```

The first character denotes the cohort number (i.e. the study year during which they entered) and the second character denotes the year of follow up.

We now examine one treatment student below.

```
## # A tibble: 4 x 7
##   EID    cohort_yr Grade   Yrs     Y Eligible treatment
##   <chr>  <chr>     <dbl> <dbl> <dbl>    <dbl>     <dbl>
## 1 T1D205 11            0     1  104.        0         1
## 2 T1D205 12            1     2  140.        0         1
## 3 T1D205 13            2     3  121.        1         1
## 4 T1D205 14            3     4  157.        1         1
```

Note that in the first two years of follow-up, the student was not eligible for supplemental instruction. After their performance stalled in the third year of follow-up, the student was eligible and remained eligible for the remainder of the study.

The proportion of control students who are eligible for supplemental instruction in each cohort-year is as follows:

```
##    Group.1       x
## 1       11 0.29375
## 2       12 0.45500
## 3       13 0.57400
## 4       14 0.65800
## 5       21 0.47000
## 6       22 0.58300
## 7       23 0.65900
## 8       31 0.49100
## 9       32 0.61300
## 10      41 0.46700
```

Function to create PWRD aggregation weights

We now create a function that inputs data and creates PWRD aggregation weights for each cohort-year.

```
pwrd_cohort_weights = function(control){

  # We begin by calculating the proportion of students who have ever
  # become eligible for the intervention for each cohort year.
  # This serves as p0_hat.
  p0_hat <- aggregate(control$Eligible, by = list(control$cohort_yr),
                    mean, na.rm = TRUE)
  p0_hat <- as.matrix(p0_hat[,2])

  # We now use a grouping of control-group residuals to estimate sigma.

  # We begin by fitting a regression that models the outcome as a function of
  # a set of covariates. In this simple case, we just use the intercept, which
  # returns the mean outcome. We could also incorporate covariates which would
  # then return a covariate-adjusted predicted value for each student.
  mod_cont  <- lm(Y ~ 1, data = control)
  fits_cont <- predict(mod_cont, control)

  # We now create a design matrix denoting the cohort-year of each observation.
  dummies <- model.matrix(~factor(control$cohort_yr)-1)
  d_names <- c('CY11','CY12','CY13','CY14', 'CY21', 'CY22', 'CY23', 'CY31',
              'CY32', 'CY41')
  colnames(dummies) <- d_names

  # We now model the residual (i.e. the outcome Y minus the predicted mean
  # from fits_cont) as a function of the cohort-year.
  mod_c <- lm(Y ~ dummies - 1, offset = fits_cont, data = control)

  # We now examine the relative precision and mutual correlations between the
  # different cohort-years by calculating a cluster robust covariance matrix
  # for the cohort-years.
  sigma <- vcovCR(mod_c, cluster = control$blocks, type = 'CR2')

  # Now that we have sigma and p0, we calculate w = p0 * Sigma^-1
  weights = t(p0_hat)%*%solve(sigma)

  # We now incorporate our non-negativity constraint.
  weights = pmax(weights, 0)

  # We now normalize so our weights sum to 1.
  weights = weights/sum(weights)

  return(weights)
}
```

We now calculate the weights for PWRD aggregation using the control data.

```
weights <- pwrd_cohort_weights(df[(df$treatment == 0),])
weights
```

```
##      dummiesCY11 dummiesCY12 dummiesCY13 dummiesCY14 dummiesCY21 dummiesCY22
## [1,]           0           0   0.1231395   0.2865784           0   0.1560793
##      dummiesCY23 dummiesCY31 dummiesCY32 dummiesCY41
## [1,]  0.08427931    0.242324   0.1075995           0
```

Analysis

Now that we have our weights, we can fit our outcome model.

```
# We fit our model, interacting treatment with cohort-year. This provides
# our ATE estimates for each cohort-year. We could further adjust for the
# matched set among other factors but choose not to for this simple analysis.
mod_pwrd <- lm(Y ~ treatment*cohort_yr - treatment + Race_White +
                 Gend_Fem + Free_Lunch, data = df)

# We extract the 10 ATE estimates, one for each cohort-year.
ests <- coef(mod_pwrd)[14:23]

# We obtain a cluster-robust matrix with clusters at the school level.
test_cov <- vcovCR(mod_pwrd, cluster = df$Sch, type = "CR2")

# Calculate the aggregated effect with appropriate test statistic and p-value
K <- rbind(c(rep(0, 13), weights))
maxt <- glht(mod_pwrd, K, alternative = c('greater'), vcov = test_cov)

# Results

tstat <- summary(maxt)[[9]]$coefficients
pval <- summary(maxt)[[9]]$pvalues[1]
```

```
tstat
```

```
##        1
## 4.786509
```

```
pval
```

```
## [1] 0.000185496
```