Assignment 4 Linear Panel Data

Nond Prueksiri

April 8, 2019

Exercise 1 Data

Data Preparation

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Warning: package 'plm' was built under R version 3.5.3

## Loading required package: Formula

##
## Attaching package: 'plm'

## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead

##
## Attaching package: 'data.table'

## The following object is masked from 'package:plm':
##
##     between

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## dummies-1.5.6 provided by Decision Patterns

## Warning: package 'httr' was built under R version 3.5.3
```

Panel dimension of wages for 5 randomly selected individuals

```
sample <- sample(unique(panel$PERSONID), 5)
sample_df <- subset(panel, panel$PERSONID %in% sample)
as.matrix(sample_df$LOGWAGE)
```

```
##         1    2    3    4    5    6    7    8    9   10   11   12   13   14
## 135    NA   NA 1.75   NA 1.85 1.67 1.92 2.46 2.75 2.64 2.41 2.63 2.55 2.39
## 846    NA   NA   NA 2.01   NA 1.86 1.64 1.50 2.19 2.04 1.81   NA   NA 1.74
## 1442 1.64 2.42 2.41 2.69 2.76 3.04 3.16 3.16 3.25 3.19 3.03 3.13 3.30 3.40
## 1452 2.51   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA 1.49
```

```
## 2113   NA   NA   NA   NA   NA 1.17 1.85   NA   NA 2.31 2.28 0.61   NA 2.20
```

Exercise 2 Random Effects

To estimate random effect by hand, I do the two following steps; * Step 1: estimate lamda (1 - ratio of standard error of residual over standard error of residual plus standard error of intercept term times number of time periods) by pooled OLS

```
rm(sample, sample_df)

OLS <- lm(LOGWAGE ~ EDUC + POTEXPER, data = panel)
var_alpha <- (coef(summary(OLS))[1,2])^2
var_resid <- var(OLS$residuals)
lamda <- 1 - sqrt(var_resid / (var_resid + 15*var_alpha))
lamda
```

```
## [1] 0.02308281
```

- Step 2: use lamda estimated from step 1 to transform variable, that is; x_transform = x - lamda*mean_x

```
panel <- data.table(panel)
setkey(panel,PERSONID)
panel <- panel[,mean_LOGWAGE:=mean(LOGWAGE),by=PERSONID]
panel <- panel[,mean_POTEXPER:=mean(POTEXPER),by=PERSONID]
panel <- panel[,mean_EDUC:=mean(EDUC),by=PERSONID]

panel <- panel[,re_LOGWAGE:= LOGWAGE - lamda*mean_LOGWAGE]
panel <- panel[,re_POTEXPER:= POTEXPER - lamda*mean_POTEXPER]
panel <- panel[,re_EDUC:= EDUC - lamda*mean_EDUC]

RE <- lm(re_LOGWAGE ~ re_POTEXPER + re_EDUC, data = panel)
summary(RE)
```

```
##
## Call:
## lm(formula = re_LOGWAGE ~ re_POTEXPER + re_EDUC, data = panel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48422 -0.27411  0.02243  0.31317  2.09991
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.7737118  0.0268342   28.83   <2e-16 ***
## re_POTEXPER 0.0374853  0.0008922   42.02   <2e-16 ***
## re_EDUC     0.0939844  0.0019441   48.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4782 on 17916 degrees of freedom
## Multiple R-squared:  0.1592, Adjusted R-squared:  0.1591
## F-statistic:  1696 on 2 and 17916 DF,  p-value: < 2.2e-16
```

- Check if the estimated random effects closed to the estimation provided by the package plm

```r
RE_check <- plm(LOGWAGE ~ POTEXPER + EDUC, data = panel_default, model = "random")
summary(RE_check)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = LOGWAGE ~ POTEXPER + EDUC, data = panel_default,
##     model = "random")
##
## Unbalanced Panel: n = 2178, T = 1-15, N = 17919
##
## Effects:
##                  var std.dev share
## idiosyncratic 0.1125  0.3355  0.48
## individual    0.1221  0.3494  0.52
## theta:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3074  0.6785  0.7095  0.6949  0.7427  0.7594
##
## Residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.33914 -0.16249  0.03227  0.00672  0.20687  2.24132
##
## Coefficients:
##              Estimate Std. Error z-value  Pr(>|z|)
## (Intercept) 0.5730949  0.0426738  13.430 < 2.2e-16 ***
## POTEXPER    0.0387453  0.0007203  53.791 < 2.2e-16 ***
## EDUC        0.1072682  0.0032864  32.640 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2708.3
## Residual Sum of Squares: 2040.7
## R-Squared:      0.24984
## Adj. R-Squared: 0.24976
## Chisq: 5861.23 on 2 DF, p-value: < 2.22e-16
```

Exercise 3 Fixed Effects

Between Estimator

```r
between_panel <- aggregate(panel, list(panel$PERSONID), mean)
Between <- lm(LOGWAGE ~ EDUC + POTEXPER, data = between_panel)
summary(Between)
```

```
##
## Call:
## lm(formula = LOGWAGE ~ EDUC + POTEXPER, data = between_panel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84875 -0.24624  0.02323  0.27538  1.45765
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.845569   0.077018  10.979  < 2e-16 ***
## EDUC        0.093100   0.004668  19.942  < 2e-16 ***
## POTEXPER    0.025999   0.003605   7.212 7.57e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3991 on 2175 degrees of freedom
## Multiple R-squared:  0.1553, Adjusted R-squared:  0.1546
## F-statistic:   200 on 2 and 2175 DF,  p-value: < 2.2e-16
```

within estimator

```
panel <- panel[,fe_LOGWAGE:= LOGWAGE - mean_LOGWAGE]
panel <- panel[,fe_POTEXPER:= POTEXPER - mean_POTEXPER]
panel <- panel[,fe_EDUC:= EDUC - mean_EDUC]
Within <- lm(fe_LOGWAGE ~ 0 + fe_EDUC + fe_POTEXPER, data = panel)
summary(Within)
```

```
##
## Call:
## lm(formula = fe_LOGWAGE ~ 0 + fe_EDUC + fe_POTEXPER, data = panel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44244 -0.14905  0.01822  0.17648  2.55015
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## fe_EDUC     0.1236620  0.0054003   22.90   <2e-16 ***
## fe_POTEXPER 0.0385611  0.0007109   54.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3144 on 17917 degrees of freedom
## Multiple R-squared:  0.1964, Adjusted R-squared:  0.1963
## F-statistic:  2190 on 2 and 17917 DF,  p-value: < 2.2e-16
```

First time difference

```
panel <- panel[,d_LOGWAGE:=c(NA,diff(LOGWAGE)),by=PERSONID]
panel <- panel[,d_EDUC:=c(NA,diff(EDUC)), by=PERSONID]
panel <- panel[,d_POTEXPER:=c(NA,diff(POTEXPER)), by=PERSONID]
panel <- panel[,d_TIMETRND:=c(NA,diff(TIMETRND)),by=PERSONID]
FD <- lm(d_LOGWAGE ~ 0 + d_EDUC + d_POTEXPER, data = panel[d_TIMETRND == 1])
summary(FD)
```

```
##
## Call:
## lm(formula = d_LOGWAGE ~ 0 + d_EDUC + d_POTEXPER, data = panel[d_TIMETRND ==
##     1])
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.78354 -0.13354 -0.01354  0.12646  2.51646
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## d_EDUC     0.043108   0.015179    2.84  0.00452 **
## d_POTEXPER 0.053537   0.002922   18.32  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3357 on 13682 degrees of freedom
## Multiple R-squared:  0.02451,    Adjusted R-squared:  0.02436
## F-statistic: 171.9 on 2 and 13682 DF,  p-value: < 2.2e-16
```

Comparison between the three models

```
FE <- coef(Between)
FE <- as.data.frame(rbind(FE, c(NA, coef(Within)) , c(NA,coef(FD))))
row.names(FE) <- c("Between", "Within", "First time difference")
FE
```

```
##                         (Intercept)       EDUC   POTEXPER
## Between                   0.8455688 0.09309987 0.02599874
## Within                           NA 0.12366202 0.03856107
## First time difference            NA 0.04310838 0.05353695
```

Exercise 4 Understanding Fixed Effects

- Write and optimize likelihood function to estimate "alpha" - individual fixed effects

```
sample100 = sample(unique(panel$PERSONID), 100)
panel100 <- subset(panel_default, panel_default$PERSONID %in% sample100)
D <- dummy("PERSONID", data = panel100, sep="")
X <- panel100[,2]
X <- cbind(X, panel100[,4], D)
Y <- panel100[,3]

mle <- function(coeff) {
  beta <- coeff[1:102]
  sigma <- exp(coeff[103])
  p <- (Y - X%*%beta)/sigma
  ll <- sum(log(dnorm(p)/sigma))
  return(-ll)
}

coeff <- matrix(0, nrow = 103)

est <- nlm(mle,coeff)
```

- Regression of fixed effects

```
panel_alpha <- as.matrix(est$estimate)[3:102,]
panel_alpha <- cbind(panel_alpha, sort(sample100))
FE_character <- unique(merge(panel_alpha, panel100, by.x ="V2", by.y="PERSONID")[,-(3:6)])
```

```
lm_Character <- lm(panel_alpha ~ ABILITY + MOTHERED +FATHERED + BRKNHOME + SIBLINGS, data=FE_character)
summary(lm_Character)
```

```
##
## Call:
## lm(formula = panel_alpha ~ ABILITY + MOTHERED + FATHERED + BRKNHOME +
##     SIBLINGS, data = FE_character)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0564 -0.2150  0.0620  0.2319  0.8850
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.100761   0.223438   0.451    0.653
## ABILITY     -0.027219   0.052480  -0.519    0.605
## MOTHERED    -0.014468   0.017459  -0.829    0.409
## FATHERED     0.009831   0.014971   0.657    0.513
## BRKNHOME    -0.091310   0.107311  -0.851    0.397
## SIBLINGS    -0.003254   0.020752  -0.157    0.876
##
## Residual standard error: 0.3908 on 94 degrees of freedom
## Multiple R-squared:  0.01565,    Adjusted R-squared:  -0.03671
## F-statistic: 0.2989 on 5 and 94 DF,  p-value: 0.9124
```

- The standard errors The standard error here may not be correct due to the fact that we random only 100 observation from the whole sample of more than 2,000 individual. The possible method to improve this is to calculate them by "bootstrap" in order to increase the precision of standard errors given there are more simulated datasets

```
boot_coef <- matrix(0, nrow = 1, ncol=6)

for (i in 1:49) {
sample100 = sample(unique(panel$PERSONID), 100)
panel100 <- subset(panel_default, panel_default$PERSONID %in% sample100)
D <- dummy("PERSONID", data = panel100, sep="")
X <- panel100[,2]
X <- cbind(X, panel100[,4], D)
Y <- panel100[,3]
coeff <- matrix(0, nrow = 103)

est_boot <- nlm(mle,coeff)
panel_alpha <- as.matrix(est_boot$estimate)[3:102,]
panel_alpha <- cbind(panel_alpha, sort(sample100))
FE_character <- unique(merge(panel_alpha, panel100, by.x ="V2", by.y="PERSONID")[,-(3:6)])
lm_Character <- lm(panel_alpha ~ ABILITY + MOTHERED +FATHERED + BRKNHOME + SIBLINGS, data=FE_character)
coeff <- as.matrix(coef(lm_Character))
boot_coef <- rbind(boot_coef, t(coeff))
}

boot_sd <- (c(sd(boot_coef[,1]), sd(boot_coef[,2]), sd(boot_coef[,3]), sd(boot_coef[,4]), sd(boot_coef[
boot_sd <- as.data.frame(boot_sd)
rownames(boot_sd) <- c("intercept", "ABILITY" ,"MOTHERED" ,"FATHERED" ,"BRKNHOME" ,"SIBLINGS")
boot_sd
```

```
##              boot_sd
## intercept 0.24321976
## ABILITY   0.05515494
## MOTHERED  0.02131728
## FATHERED  0.01668546
## BRKNHOME  0.11832996
## SIBLINGS  0.02151379
```

The above shows standard errors from bootstrap method using 49 replications.