Econometrics II TA Session #8

Hiroki Kato

1 Empirical Application of Panel Data Model: Earnings Equation

1.1 Backgruond

A researcher wants to estimate the effect of full-time work experience on wages. He uses a *balanced* panel of 595 individuals from 1976 to 1982, taken from the Panel Study of Income Dynamics (PSID). The *balanced* panel data means that we can observe all individuals every year.

```r
dt <- read.csv("./data/wages.csv")
head(dt, 14)
```

```
##    exp wks bluecol ind south smsa married  sex union ed black   lwage id time
## 1    3  32      no   0   yes   no     yes male    no  9    no 5.56068  1    1
## 2    4  43      no   0   yes   no     yes male    no  9    no 5.72031  1    2
## 3    5  40      no   0   yes   no     yes male    no  9    no 5.99645  1    3
## 4    6  39      no   0   yes   no     yes male    no  9    no 5.99645  1    4
## 5    7  42      no   1   yes   no     yes male    no  9    no 6.06146  1    5
## 6    8  35      no   1   yes   no     yes male    no  9    no 6.17379  1    6
## 7    9  32      no   1   yes   no     yes male    no  9    no 6.24417  1    7
## 8   30  34     yes   0    no   no     yes male    no 11    no 6.16331  2    1
## 9   31  27     yes   0    no   no     yes male    no 11    no 6.21461  2    2
## 10  32  33     yes   1    no   no     yes male   yes 11    no 6.26340  2    3
## 11  33  30     yes   1    no   no     yes male    no 11    no 6.54391  2    4
## 12  34  30     yes   1    no   no     yes male    no 11    no 6.69703  2    5
## 13  35  37     yes   1    no   no     yes male    no 11    no 6.79122  2    6
## 14  36  30     yes   1    no   no     yes male    no 11    no 6.81564  2    7
```

The variable `id` and `time` indicate individual and time indexs. We use these two variables to apply panel data models. Additionally, we use the following variables:

- `exp`: years of full-time work experience
- `sqexp`: squared value of `exp`
- `sex`: a dummy variable taking 1 if an individual is female
- `ed`: years of education

- `lwage`: logarithm of wage

```
dt <- dt[,c("id", "time", "exp", "sex", "ed", "lwage")]
dt$sqexp <- dt$exp^2
dt$sex <- ifelse(as.character(dt$sex) == "female", 1, 0)
summary(dt)
```

```
##        id             time          exp             sex               ed
##  Min.   :  1   Min.   :1   Min.   : 1.00   Min.   :0.0000   Min.   : 4.00
##  1st Qu.:149   1st Qu.:2   1st Qu.:11.00   1st Qu.:0.0000   1st Qu.:12.00
##  Median :298   Median :4   Median :18.00   Median :0.0000   Median :12.00
##  Mean   :298   Mean   :4   Mean   :19.85   Mean   :0.1126   Mean   :12.85
##  3rd Qu.:447   3rd Qu.:6   3rd Qu.:29.00   3rd Qu.:0.0000   3rd Qu.:16.00
##  Max.   :595   Max.   :7   Max.   :51.00   Max.   :1.0000   Max.   :17.00
##      lwage           sqexp
##  Min.   :4.605   Min.   :   1.0
##  1st Qu.:6.395   1st Qu.: 121.0
##  Median :6.685   Median : 324.0
##  Mean   :6.676   Mean   : 514.4
##  3rd Qu.:6.953   3rd Qu.: 841.0
##  Max.   :8.537   Max.   :2601.0
```

1.2 Pooled OLS

Using the OLS method, we want to estimate the following linear panel data model:

$$\text{lwage}_{it} = \alpha + \beta_1 \cdot \text{exp}_{it} + \beta_2 \cdot \text{sqexp}_{it} + \beta_3 \cdot \text{sex}_{it} + \beta_4 \cdot \text{ed}_{it} + u_{it}.$$

We will discuss assumptions for applying the OLS method. Let $\mathbf{X}_{it}$ be a $1 \times K$ (stochastic) explanatory vector. This vector contains `exp`, `sqexp`, `sex` and `ed`. Let Y_{it} be a random variable of outcome, that is `lwage`. The balanced panel data is given by

	$i=1$	$i=2$	$\cdots$	$i=n$
$t=1$	$(Y_{11}, \mathbf{X}_{11})$	$(Y_{21}, \mathbf{X}_{21})$	$\cdots$	$(Y_{n1}, \mathbf{X}_{n1})$
$t=2$	$(Y_{12}, \mathbf{X}_{12})$	$(Y_{22}, \mathbf{X}_{22})$	$\cdots$	$(Y_{n2}, \mathbf{X}_{n2})$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$t=T$	$(Y_{1T}, \mathbf{X}_{1T})$	$(Y_{2T}, \mathbf{X}_{2T})$	$\cdots$	$(Y_{nT}, \mathbf{X}_{nT})$

Then, the linear panel data model can be rewritten as follows:

$$Y_{it} = \mathbf{X}_{it}\beta + u_{it}, \quad t = 1, \dots, T, \quad i = 1, \dots, n.$$

Using notations $\underline{\mathbf{X}}_i = (\mathbf{X}'_{i1}, \dots, \mathbf{X}'_{iT})'$ and $\underline{Y}_i = (Y_{i1}, \dots, Y_{iT})'$, and $\underline{u}_i = (u_{i1}, \dots, u_{iT})'$, we can reformulate this model as follows:

$$\underline{Y}_i = \underline{\mathbf{X}}_i\beta + \underline{u}_i, \quad \forall i.$$

Now, we assume

1. $E[\mathbf{X}_{it}'u_{it}] = 0$, $\forall i, t$. This assumption, called *(contempraneous) exogneity assumption*, implies that u_{it} and $\mathbf{X}_{it}$ are orthogonal in the conditional mean sence, $E[u_{it}|\mathbf{X}_{it}] = 0$. However, this assumption does not imply u_{it} is uncorrelated with the explanatory variables in all time periods (strictly exogeneity), that is, $E[u_{it}|\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT}] = 0$. This assumption palces no restriction on the relationship between $\mathbf{X}_{is}$ and u_{it} for $s \neq t$.
2. $E[\underline{\mathbf{X}}_i'\underline{\mathbf{X}}_i] \succ 0$.

Under these two assumptions, the true parameter can be identified by

$$\beta = E[\underline{\mathbf{X}}_i'\underline{\mathbf{X}}_i]^{-1}E[\underline{\mathbf{X}}_i'\underline{Y}_i].$$

Hence, the OLSE (pooled OLSE) is given by

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n \underline{\mathbf{X}}_i'\underline{\mathbf{X}}_i\right)\left(\frac{1}{n}\sum_{i=1}^n \underline{\mathbf{X}}_i'\underline{Y}_i\right) = \left(\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T \mathbf{X}_{it}'\mathbf{X}_{it}\right)\left(\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T \mathbf{X}_{it}'Y_{it}\right).$$

The pooled OLS estimator is consistent and asymptotically normally distributed.

$$\sqrt{n}(\hat{\beta} - \beta) \sim N(0, A^{-1}BA^{-1}),$$

where $A = E[\underline{\mathbf{X}}_i'\underline{\mathbf{X}}_i]$ and $B = E[\underline{\mathbf{X}}_i'\underline{u}_i\underline{u}_i'\underline{\mathbf{X}}_i]$. The consistent estimator of the asymptotic variance covariance matrix is given by

$$\hat{A}^{-1}\hat{B}\hat{A}^{-1} = \left(\frac{1}{n}\sum_{i=1}^n \underline{\mathbf{X}}_i'\underline{\mathbf{X}}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \underline{\mathbf{X}}_i'\underline{u}_i\underline{u}_i'\underline{\mathbf{X}}_i\right)\left(\frac{1}{n}\sum_{i=1}^n \underline{\mathbf{X}}_i'\underline{\mathbf{X}}_i\right)^{-1}$$

The standard errors caclulated by this matrix is called *robust standard errors clustered by individuals*.

In `R`, the pooled OLSE can be obtained by `lm` function. However, the `lm` function does not return the cluster-robust standard errors. Thus, you need to caclulate them by yourself. Here is a sample code.

```r
# OLSE
pool <- lm(lwage ~ exp + sqexp + sex + ed, data = dt)

# Clustered SE
X <- model.matrix(pool); uhat <- pool$residuals
uhatset <- matrix(0, nrow = nrow(X), ncol = nrow(X))

i_from <- 1; j_from <- 1
```

```r
for (i in 1:max(dt$id)) {
  x <- as.numeric(rownames(dt))[dt$id == i]
  usq <- uhat[x] %*% t(uhat[x])
  i_to <- i_from + nrow(usq) - 1
  j_to <- j_from + ncol(usq) - 1
  uhatset[i_from:i_to, j_from:j_to] <- usq
  i_from <- i_to + 1; j_from <- j_to + 1
}

Ahat <- t(X) %*% X
Bhat <- t(X) %*% uhatset %*% X
clust_vcov <- solve(Ahat) %*% Bhat %*% solve(Ahat)
clust_se <- sqrt(diag(clust_vcov))

print("Pooled OLSE"); coef(pool)
```

```
## [1] "Pooled OLSE"

##   (Intercept)           exp         sqexp           sex            ed
##  5.2759322858  0.0427794655 -0.0007022517 -0.4305537695  0.0747976640
```

```r
print("SE of pooled OLSE"); clust_se
```

```
## [1] "SE of pooled OLSE"

##  (Intercept)          exp        sqexp          sex           ed
## 0.0846130179 0.0047667140 0.0001103059 0.0339740598 0.0048139358
```

Alternatively, using the `plm` function (the package `plm`) and the `coeftest` function (the package `lmtest`), you can obtain the asymptotic variance covariance matrix of pooled OLSE easily. The `plm` function provides the panel data model. When you want to estimate pooled OLS, you need to specify `model = "pooling"`. Moreover, you should specify individual and time index using `index` augment. This augment passes `index = c("individual index", "time index")`. After estimating the pooled OLS by the `plm` function, you must use the `coeftest` function to obtain the cluster-robust standard errors. To caclulate the clustered standard errors, you should use the `vcovHC` function in the `vcov` augment.

```r
library(plm)
library(lmtest)
library(sandwich)
test <- plm(
  lwage ~ exp + sqexp + sex + ed,
  data = dt, model = "pooling", index = c("id", "time"))
coeftest(test, vcov = vcovHC(test, type = "HC0", cluster = "group"))
```

```
##
## t test of coefficients:
```

```
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  5.27593229  0.08461302  62.3537 < 2.2e-16 ***
## exp          0.04277947  0.00476671   8.9746 < 2.2e-16 ***
## sqexp       -0.00070225  0.00011031  -6.3664 2.145e-10 ***
## sex         -0.43055377  0.03397406 -12.6730 < 2.2e-16 ***
## ed           0.07479766  0.00481394  15.5377 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```