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.

```
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`
- `lwage`: logarithm of wage

```r
dt <- dt[,c("id", "time", "exp", "lwage")]
dt$sqexp <- dt$exp^2
summary(dt)
```

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

To examine the effect of labor experience on wages, we want to estimate the following linear panel data model:

$$\text{lwage}_{it} = \beta_1 \cdot \text{exp}_{it} + \beta_2 \cdot \text{sqexp}_{it} + u_{it}.$$

We can define the regression equation as the `formula` object in `R`. To exclude the intercept, we must specify `-1` in the rhs of regression equation. Thus, in `R`, we define the linear panel data model as follows:

```r
model <- lwage ~ -1 + exp + sqexp
```

1.2 Pooled OLS

We want to estimate the above regression equation by the OLS method. We will discuss assumptions for implementation. Let $\mathbf{X}_{it}$ be a $1 \times K$ (stochastic) explanatory vector. This vector contains `exp`, `sqexp`. Let Y_{it} be a random variable of outcome, that is `lwage`. 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 is given 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

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

Using the full matrix notation, the OLS estimator is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'Y),$$

where $\mathbf{X} = (\underline{\mathbf{X}}_1, \dots, \underline{\mathbf{X}}_n)'$ and $Y = (\underline{Y}_1, \dots, \underline{Y}_n)'$.

In `R` programming, the `lm` function provides the pooled OLSE in the context of panel data model. Another way is the `plm` function in the package `plm`. When you want to estimate pooled OLS by the `plm` function, 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")`.

```
bols1 <- lm(model, data = dt)
bols2 <- plm(model, data = dt, model = "pooling", index = c("id", "time"))
```

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

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

Thus, estimator of asymptotic variance of the pooled OLSE is

$$A\hat{v}ar(\hat{\beta}) = \left(\sum_{i=1}^{n}\underline{\mathbf{X}}_i'\underline{\mathbf{X}}_i\right)^{-1}\left(\sum_{i=1}^{n}\underline{\mathbf{X}}_i'\underline{u}_i\underline{u}_i'\underline{\mathbf{X}}_i\right)\left(\sum_{i=1}^{n}\underline{\mathbf{X}}_i'\underline{\mathbf{X}}_i\right)^{-1}.$$

Using the full matrix notations, we can reformulate

$$A\hat{v}ar(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1},$$

where

$$\Omega = \begin{pmatrix} \underline{u}_1\underline{u}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \underline{u}_2\underline{u}_2' & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \underline{u}_n\underline{u}_n' \end{pmatrix}.$$

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

In R, the `lm` and `plm` function provide the standard errors based on $\widehat{Avar}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$, where $\hat{\sigma}^2 = \hat{u}\hat{u}'/(nT - K)$ and $\hat{u} = Y - X\hat{\beta}$. There are two ways to obtain cluster robust standard errors. The first way is to caclulate by yourself. The second way is to use the `coeftest` function in the package `lmtest`. When you use this function, we should use the `plm` function to estimate the pooled OLSE, and the `vcovHC` function (the package `sandwich`) in the `vcov` augment of `coeftest` fucntion.

```r
# Setup
N <- length(unique(dt$id)); T <- length(unique(dt$time))
X <- model.matrix(bols1); k <- ncol(X)

# Inference
uhat <- bols1$residuals
uhatset <- matrix(0, nrow = nrow(X), ncol = nrow(X))

i_from <- 1; j_from <- 1
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
vcovols <- solve(Ahat) %*% Bhat %*% solve(Ahat)
seols <- sqrt(diag(vcovols))

# Easy way
library(lmtest)
library(sandwich)
easy_cluster <- coeftest(
  bols2, vcov = vcovHC(bols2, type = "HC0", cluster = "group"))
```

The result is shown in the first column of **??**. The partial effect of experience represents the percent change of wages. Thus,

$$(\text{\% Change of Wage}) = 64.6 - 2 \cdot 1.3 \cdot \text{exp}.$$

For example, wages increase by 12.99% at a mathematical mean of labor experience (`exp`). Moreover, this result implies diminishing marginal returns of labor experience.