CH 15

15.6 Using the NLS panel data on $N = 716$ young women, we consider only years 1987 and 1988. We are interested in the relationship between ln(*WAGE*) and experience, its square, and indicator variables for living in the south and union membership. Some estimation results are in Table 15.10.

TABLE 15.10 **Estimation Results for Exercise 15.6**

	(1) OLS 1987	(2) OLS 1988	(3) FE	(4) FE Robust	(5) RE
C	0.9348	0.8993	1.5468	1.5468	1.1497
	(0.2010)	(0.2407)	(0.2522)	(0.2688)	(0.1597)
EXPER	0.1270	0.1265	0.0575	0.0575	0.0986
	(0.0295)	(0.0323)	(0.0330)	(0.0328)	(0.0220)
$EXPER^2$	−0.0033	−0.0031	−0.0012	−0.0012	−0.0023
	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0007)
SOUTH	−0.2128	−0.2384	−0.3261	−0.3261	−0.2326
	(0.0338)	(0.0344)	(0.1258)	(0.2495)	(0.0317)
UNION	0.1445	0.1102	0.0822	0.0822	0.1027
	(0.0382)	(0.0387)	(0.0312)	(0.0367)	(0.0245)
N	716	716	1432	1432	1432

(standard errors in parentheses)

a. The OLS estimates of the ln(*WAGE*) model for each of the years 1987 and 1988 are reported in columns (1) and (2). How do the results compare? For these individual year estimations, what are you assuming about the regression parameter values across individuals (heterogeneity)?

b. The ln(*WAGE*) equation specified as a panel data regression model is

$$\ln(WAGE_{it}) = \beta_1 + \beta_2 EXPER_{it} + \beta_3 EXPER_{it}^2 + \beta_4 SOUTH_{it} + \beta_5 UNION_{it} + (u_i + e_{it}) \quad \text{(XR15.6)}$$

Explain any differences in assumptions between this model and the models in part (a).

a. The coefficient estimates for these two years are quite similar, indicating model stability. However, it assums that there's no unobserved heterogenity across individuals, which may lead to endogenous problems, causing biased estimate. Pooled OLS let individual heterogeneity all belong to error.

b. Panel data add time-invariant variables, allowing more consistent estimation.

c. Column (3) contains the estimated fixed effects model specified in part (b). Compare these estimates with the OLS estimates. Which coefficients, apart from the intercepts, show the most difference?

d. The F-statistic for the null hypothesis that there are no individual differences, equation (15.20), is 11.68. What are the degrees of freedom of the F-distribution if the null hypothesis (15.19) is true? What is the 1% level of significance critical value for the test? What do you conclude about the null hypothesis.

e. Column (4) contains the fixed effects estimates with cluster-robust standard errors. In the context of this sample, explain the different assumptions you are making when you estimate with and without cluster-robust standard errors. Compare the standard errors with those in column (3). Which ones are substantially different? Are the robust ones larger or smaller?

c.

Panel Data Model 是一個統稱，泛指結合「個體 (cross-section)」與「時間 (time series)」資料結構的模型。
在 Panel Data 模型中，我們可以根據對「個體異質性」的處理方式，進一步區分為：

類型	說明
Fixed Effects Model (FE)	假設個體效應 u_i 是固定參數，允許與解釋變數相關 → 適用於控制**內生性**
Random Effects Model (RE)	假設 u_i 是隨機變數，且與解釋變數無關 → 適用於更高效率，但前提假設較強

EXPER FE 95% C.I. = [-0.0085, 0.1235], OLS 的 estimate 不在裡面

可能是因為 EXPER 造成了個体差異.

d. $df_1 = N - 1 = 716 - 1 = 715$

$df_2 = NT - N - (K-1) = 716 \times 2 - 716 - (5-1) = 712$

$\because F = 11.68 > F(715, 712, 0.99) = 1.19$ (right-tailed test)

$\therefore$ reject H_0. → Individual difference exists.

e. Cluster-standard errors are larger since it allow the heterogeneity and "autocorrelation" in the data.

f. Column (5) contains the random effects estimates. Which coefficients, apart from the intercepts, show the most difference from the fixed effects estimates? Use the Hausman test statistic (15.36) to test whether there are significant differences between the random effects estimates and the fixed effects estimates in column (3) (Why that one?). Based on the test results, is random effects estimation in this model appropriate?

f. (1) EXPER^2 係数差 $\frac{0.0023}{0.0012} = 1.92$ 倍

(2) $H_0: \beta_{FE} = \beta_{RE}$, no endogenity u_i v.s. $H_1: \beta_{FE} \neq \beta_{RE}$, endogenity exists

Hausman test: $t = \frac{b_{FE,k} - b_{RE,k}}{\sqrt{\hat{var}(b_{FE,k}) - \hat{var}(b_{RE,k})}}$

$t_{EXPER} = \frac{0.0575 - 0.0986}{\sqrt{0.033^2 - 0.022^2}} = -1.67$

$t_{EXPER^2} = \frac{-0.0012 - (-0.0023)}{\sqrt{0.0011^2 - 0.0007^2}} = 1.296$

$t_{SOUTH} = \frac{-0.3261 - (-0.2326)}{\sqrt{0.1258^2 - 0.0317^2}} = -0.77$

$t_{UNION} = \frac{0.0822 - 0.1027}{\sqrt{0.0312^2 - 0.0245^2}} = -1.06$

$\longrightarrow$ $RR = \{t \mid t < -1.96 \text{ or } t > 1.96\}$

We all fail to reject H_0.

$\longrightarrow$ use β_{RE}.

15.17 The data file *liquor* contains observations on annual expenditure on liquor (*LIQUOR*) and annual income (*INCOME*) (both in thousands of dollars) for 40 randomly selected households for three consecutive years.

a. Create the first-differenced observations on *LIQUOR* and *INCOME*. Call these new variables *LIQUORD* and *INCOMED*. Using OLS regress *LIQUORD* on *INCOMED* without a constant term. Construct a 95% interval estimate of the coefficient.

a.

```
Call:
lm(formula = liquord ~ incomed - 1, data = liquor5_diff)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6852 -0.9196 -0.0323  0.9027  3.3620

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
incomed  0.02975    0.02922   1.018    0.312

Residual standard error: 1.417 on 79 degrees of freedom
Multiple R-squared:  0.01295,   Adjusted R-squared:  0.0004544
F-statistic: 1.036 on 1 and 79 DF,  p-value: 0.3118
```

$\longrightarrow$ 95% C.I. for $\beta_{incomed} = 0.02975 \pm 1.96 \times 0.02922 = [-0.027212, 0.087012]$

b. Estimate the model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + u_i + e_{it}$ using random effects. Construct a 95% interval estimate of the coefficient on *INCOME*. How does it compare to the interval in part (a)?

c. Test for the presence of random effects using the LM statistic in equation (15.35). Use the 5% level of significance.

b.

```
Call:
plm(formula = liquor ~ income, data = pdat, model = "random")

Balanced Panel: n = 40, T = 3, N = 120

Effects:
                 var std.dev share
idiosyncratic 0.9640  0.9819 0.571
individual    0.7251  0.8515 0.429
theta: 0.4459

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-2.263634 -0.697383  0.078697  0.552680  2.225798

Coefficients:
             Estimate Std. Error z-value  Pr(>|z|)
(Intercept) 0.9690324  0.5210052  1.8599 0.0628957 .
income      0.0265755  0.0070126  3.7897 0.0001508 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    126.61
Residual Sum of Squares: 112.88
R-Squared:      0.1085
Adj. R-Squared: 0.10095
Chisq: 14.3618 on 1 DF, p-value: 0.00015083
```

→ $\hat{LIQUOR}_{it} = 0.969\,0324 + 0.026\,5755\ INCOME_{it}$

→ 95% C.I. for $\beta_{income} = 0.026\,5755 \pm 1.96 \times 0.007\,0126 = [0.012\,8311,\ 0.040\,31983]$

→ The C.I. in part (a) contains zero,

indicating income having no statistically significant effect on liquor.

However, part (b) does not contain zero.

→ The estimate from first difference will be unbiased but larger SE.

The estimate from random effects will have smaller SE but sometimes biased.

c.

```
        Lagrange Multiplier Test - (Breusch-Pagan)

data:  liquor ~ income
chisq = 20.68, df = 1, p-value = 5.429e-06
alternative hypothesis: significant effects
```

H_0 : no random effects v.s. H_1 : random effects exist

∵ $p-value < 0.05$ (right — tailed test)

∴ reject H_0.

d. For each individual, compute the time averages for the variable *INCOME*. Call this variable *INCOMEM*. Estimate the model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + \gamma INCOMEM_i + c_i + e_{it}$ using the random effects estimator. Test the significance of the coefficient γ at the 5% level. Based on this test, what can we conclude about the correlation between the random effect u_i and *INCOME*? Is it OK to use the random effects estimator for the model in (b)?

d.

```
Call:
plm(formula = liquor ~ income + INCOMEM, data = pdat2, model = "random")

Balanced Panel: n = 40, T = 3, N = 120

Effects:
                 var std.dev share
idiosyncratic 0.9640  0.9819 0.571
individual    0.7251  0.8515 0.429
theta: 0.4459

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-2.300955 -0.703840  0.054992  0.560255  2.257325

Coefficients:
             Estimate Std. Error z-value Pr(>|z|)
(Intercept) 0.9163337  0.5524439  1.6587  0.09718 .
income      0.0207421  0.0209083  0.9921  0.32117
INCOMEM     0.0065792  0.0222048  0.2963  0.76700
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    126.61
Residual Sum of Squares: 112.79
R-Squared:      0.10917
Adj. R-Squared: 0.093945
Chisq: 14.3386 on 2 DF, p-value: 0.00076987
```

$\widehat{LIQUOR}_{it} = 0.9163337 + 0.0207421\, INCOME + 0.0065792\, INCOMEM$

$H_0: \gamma = 0,\ cov(u_i, X_{it}) = 0$ v.s. $H_1: \gamma \neq 0,\ cov(u_i, X_{it}) \neq 0$

$\because$ p-value of $\gamma > 0.05$

$\therefore$ fail to reject H_0. There's no correlation between u_i and INCOME

$\rightarrow$ Random effects model is valid.

15.20 This exercise uses data from the STAR experiment introduced to illustrate fixed and random effects for grouped data. In the STAR experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star*.

a. Estimate a regression equation (with no fixed or random effects) where *READSCORE* is related to *SMALL*, *AIDE*, *TCHEXPER*, *BOY*, *WHITE_ASIAN*, and *FREELUNCH*. Discuss the results. Do students perform better in reading when they are in small classes? Does a teacher's aide improve scores? Do the students of more experienced teachers score higher on reading tests? Does the student's sex or race make a difference?

b. Reestimate the model in part (a) with school fixed effects. Compare the results with those in part (a). Have any of your conclusions changed? [*Hint*: specify *SCHID* as the cross-section identifier and *ID* as the "time" identifier.]

a.

```
Call:
lm(formula = readscore ~ small + aide + tchexper + boy + white_asian +
    freelunch, data = star)

Residuals:
     Min       1Q   Median       3Q      Max
-107.220  -20.214   -3.935   14.339  185.956

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 437.76425    1.34622 325.180  < 2e-16 ***
small         5.82282    0.98933   5.886 4.19e-09 ***
aide          0.81784    0.95299   0.858    0.391
tchexper      0.49247    0.06956   7.080 1.61e-12 ***
boy          -6.15642    0.79613  -7.733 1.23e-14 ***
white_asian   3.90581    0.95361   4.096 4.26e-05 ***
freelunch   -14.77134    0.89025 -16.592  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.19 on 5759 degrees of freedom
  (因為不存在，20 個觀察量被刪除了)
Multiple R-squared:  0.09685,   Adjusted R-squared:  0.09591
F-statistic: 102.9 on 6 and 5759 DF,  p-value: < 2.2e-16
```

Small class, teacher experience both help students' performance but aide does NOT.

Girl better than boy, white_asian also become better.

However, students with free lunch have poor performance.

b.

```
Call:
plm(formula = readscore ~ small + aide + tchexper + boy + white_asian +
    freelunch, data = pdata, effect = "individual", model = "within")

Unbalanced Panel: n = 79, T = 34-137, N = 5766

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-102.6381  -16.7834   -2.8473   12.7591  198.4169

Coefficients:
             Estimate Std. Error  t-value  Pr(>|t|)
small        6.490231   0.912962   7.1090 1.313e-12 ***
aide         0.996087   0.881693   1.1297    0.2586
tchexper     0.285567   0.070845   4.0309 5.629e-05 ***
boy         -5.455941   0.727589  -7.4987 7.440e-14 ***
white_asian  8.028019   1.535656   5.2277 1.777e-07 ***
freelunch  -14.593572   0.880006 -16.5835 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    4628000
Residual Sum of Squares: 4268900
R-Squared:      0.077592
Adj. R-Squared: 0.063954
F-statistic: 79.6471 on 6 and 5681 DF, p-value: < 2.22e-16
```

→ Controlling schools' heteroskedasticity makes the conclusion still remain same.

c. Test for the significance of the school fixed effects. Under what conditions would we expect the inclusion of significant fixed effects to have little influence on the coefficient estimates of the remaining variables?

d. Reestimate the model in part (a) with school random effects. Compare the results with those from parts (a) and (b). Are there any variables in the equation that might be correlated with the school effects? Use the LM test for the presence of random effects.

c.

```
        F test for individual effects

data:  readscore ~ small + aide + tchexper + boy + white_asian + freelunch
F = 16.698, df1 = 78, df2 = 5681, p-value < 2.2e-16
alternative hypothesis: significant effects
```

H_0 : no individual effects (pooled OLS) v.s. H_1 : individual effects exist (FE)

$\because F = 16.698 > F = 10$

$\therefore$ reject H_0

→ If fixed effect only give each school time variant change instead of making each school different, then the estimates of variables do NOT change.

d.

```
Call:
plm(formula = readscore ~ small + aide + tchexper + boy + white_asian +
    freelunch, data = pdata, model = "random")

Unbalanced Panel: n = 79, T = 34-137, N = 5766

Effects:
                 var std.dev share
idiosyncratic 751.43   27.41 0.829
individual    155.31   12.46 0.171
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6470  0.7225  0.7523  0.7541  0.7831  0.8153

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-97.483 -17.236  -3.282   0.037  12.803 192.346

Coefficients:
             Estimate Std. Error  z-value  Pr(>|z|)
(Intercept) 436.126774   2.064782 211.2217 < 2.2e-16 ***
small         6.458722   0.912548   7.0777 1.466e-12 ***
aide          0.992146   0.881159   1.1260    0.2602
tchexper      0.302679   0.070292   4.3060 1.662e-05 ***
boy          -5.512081   0.727639  -7.5753 3.583e-14 ***
white_asian   7.350477   1.431376   5.1353 2.818e-07 ***
freelunch   -14.584332   0.874676 -16.6740 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    6158000
Residual Sum of Squares: 4332100
R-Squared:      0.29655
Adj. R-Squared: 0.29582
Chisq: 493.205 on 6 DF, p-value: < 2.22e-16
```

```
       Lagrange Multiplier Test - (Breusch-Pagan)

data:  readscore ~ small + aide + tchexper + boy + white_asian + freelunch
chisq = 6677.4, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects
```

The coefficients are similar with those from pooled OLS and fixed effects.

This shows that regressors have little correlation with unobserved school-level heterogeneity.

However, by LM test, we reject H_0. It has school-level heterogeneity. (panel data)

e. Using the *t*-test statistic in equation (15.36) and a 5% significance level, test whether there are any significant differences between the fixed effects and random effects estimates of the coefficients on *SMALL*, *AIDE*, *TCHEXPER*, *WHITE_ASIAN*, and *FREELUNCH*. What are the implications of the test outcomes? What happens if we apply the test to the fixed and random effects estimates of the coefficient on *BOY*?

f. Create school-averages of the variables and carry out the Mundlak test for correlation between them and the unobserved heterogeneity.

e.

```
small       :  t =  1.15,  p = 0.252
aide        :  t =  0.13,  p = 0.898
tchexper    :  t = -1.94,  p = 0.053
white_asian :  t =  1.22,  p = 0.223
freelunch   :  t = -0.10,  p = 0.924
```

```
        Hausman Test

data:  readscore ~ small + aide + tchexper + boy + white_asian + freelunch
chisq = 13.809, df = 6, p-value = 0.03184
alternative hypothesis: one model is inconsistent
```

Hausman Test ——

$H_0: \beta_{FE} = \beta_{RE}$ v.s. $H_0: \beta_{FE} \neq \beta_{RE}$

∵ all p-value > 0.5

∴ fail to reject H_0. RE is appropriate.

However, Hausman Test for all shows RE is not appropriate.

→ 另外，BOY 无法計算

f.

```
Total Sum of Squares:    6007200
Residual Sum of Squares: 4281300
R-Squared:      0.28737
Adj. R-Squared: 0.28586
Chisq: 500.306 on 12 DF, p-value: < 2.22e-16
```

Mundlak Test

$H_0: \gamma = 0$, no endogeneity v.s. $H_1: \gamma \neq 0$, endogeneity exists.

∵ P-value < 0.05

∴ reject H_0. There's correlation and thus RE is not appropriate.