HW0512

Yung-Jung Cheng

2025-05-25

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 $\ln(\text{WAGE})$ 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)	(0.0200)
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)

15.6(a)

The OLS estimates of the $\ln(\text{WAGE})$ 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)?

Ans

The OLS estimates for 1987 and 1988 are remarkably similar, suggesting stability in the wage-experience relationship across years. For example, the coefficient on $EXPER$ is 0.1270 in 1987 and 0.1265 in 1988, and the coefficient on $EXPER^2$ is -0.0033 in 1987 and -0.0031 in 1988. The effects of $SOUTH$ and $UNION$ are also consistent in direction and magnitude across the two years.

By estimating the model separately for each year, we are implicitly assuming that the regression parameters are homogeneous across individuals within each year and that there is no correlation between unobserved individual characteristics and the explanatory variables. This approach ignores unobserved heterogeneity and does not account for the panel nature of the data.

15.6(b)

The $\ln(\text{WAGE}_{it})$ equation specified as a panel data regression model is

$$\ln(\text{WAGE}_{it}) = \beta_1 + \beta_2\text{EXPER}_{it} + \beta_3\text{EXPER}_{it}^2 + \beta_4\text{SOUTH}_{it} + \beta_5\text{UNION}_{it} + (u_i + \varepsilon_{it})$$

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

Ans

The panel data regression model in part (b) introduces individual-specific effects u_i that are constant over time but vary across individuals. This model acknowledges the panel structure of the data and allows for unobserved heterogeneity across individuals to be controlled.

In contrast, the separate OLS regressions in part (a) assume that all individual differences are captured by the included regressors and that the error term is homoskedastic and uncorrelated with the regressors. There is no mechanism to control for unobserved, time-invariant individual characteristics in the OLS models.

Thus, the panel model assumes that differences across individuals can be accounted for by including fixed or random effects, while the OLS models assume that such differences either do not exist or are uncorrelated with the regressors.

15.6(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?

Ans

Comparing the fixed effects estimates in column (3) with the OLS estimates in columns (1) and (2), we observe that the most notable differences appear in the coefficients for $EXPER$ and $SOUTH$.

- The coefficient on $EXPER$ drops from about 0.127 in the OLS models to 0.0575 in the fixed effects model.
- The coefficient on $SOUTH$ becomes more negative, from roughly -0.21 to -0.33.

This suggests that when individual-specific, time-invariant characteristics are accounted for, the return to experience is substantially lower and the regional wage gap is even larger. The fixed effects model removes bias from omitted variables that are constant over time but vary across individuals, such as innate ability or long-term educational background.

15.6(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?

Ans

The null hypothesis is that there are no individual-specific effects, i.e., $u_i = 0$ for all i. The F-statistic for testing this is 11.68.

Degrees of freedom for the F-distribution: - Numerator (df1): $N - 1 = 716 - 1 = 715$ - Denominator (df2): $N(T - 1 - K)$, where $T = 2$ and $K = 4$ (number of regressors excluding the intercept). This simplifies to roughly 716, since the actual denominator degrees of freedom are estimated from the residuals.

The 1% critical value for the F-distribution with these degrees of freedom is approximately **2.33**.

Since the computed F-statistic (11.68) is much greater than the critical value, we **reject the null hypothesis**. This provides strong evidence that individual-specific effects are present and that the fixed effects model is preferred over pooled OLS.

15.6(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?

Ans

Cluster-robust standard errors allow for arbitrary correlation in the error terms within individuals over time, making them more reliable in panel settings. In contrast, the conventional standard errors in column (3) assume that errors are independent and identically distributed (i.i.d.).

Comparing columns (3) and (4), most standard errors are similar, but the standard error for $SOUTH$ increases substantially—from 0.1258 to 0.2495. This suggests that the original standard error underestimated variability due to within-individual correlation for that variable.

In general, cluster-robust standard errors tend to be **larger** when intra-individual correlation is present, reflecting more conservative inference. Thus, using cluster-robust SEs provides more accurate standard errors and confidence intervals in the presence of serial correlation or heteroskedasticity within individuals.

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.

15.20(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?

Ans

```
# (a) Estimate a simple OLS regression with no fixed or random effects
model_15_20a <- lm(readscore ~ small + aide + tchexper + boy + white_asian + freelunch, data
= star)
summary(model_15_20a)
```

```
##
## 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
```

Students in small classes perform significantly better in reading, with an average score increase of about 5.82 points ($p < 0.01$). The presence of a teacher's aide has a small positive effect (0.82 points) but is not statistically significant ($p = 0.391$). Teacher experience has a positive and highly significant effect: each additional year of experience increases reading scores by about 0.49 points. Male students perform worse than females, with an average deficit of 6.16 points ($p < 0.01$). White or Asian students score significantly higher (about 3.91 points) than other racial groups. Students receiving free lunch, a proxy for lower socioeconomic status, perform significantly worse—about 14.77 points lower on average.

15.20(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.]*

Ans

```r
library(plm)

# 將資料轉換為 panel 資料格式
pdata_star <- pdata.frame(star, index = c("schid", "id"))

# 使用固定效應模型（學校固定效應）
model_15_20b <- plm(readscore ~ small + aide + tchexper + boy + white_asian + freelunch,
                    data = pdata_star, model = "within")

summary(model_15_20b)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = readscore ~ small + aide + tchexper + boy + white_asian +
##     freelunch, data = pdata_star, 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
```

After controlling for school fixed effects, the effect of being in a small class remains significantly positive and even slightly increases to about 6.49 points ($p < 0.01$), confirming the benefit of smaller class sizes. The effect of having a teacher's aide remains statistically insignificant. Teacher experience continues to have a positive and significant impact on reading scores, though the magnitude is reduced to 0.29 points per year. Boys still perform significantly worse than girls, with a gap of about 5.46 points. The effect of being White or Asian becomes even larger and remains highly significant, while the negative effect of receiving free lunch remains large and highly significant (about -14.59 points). Overall, the key conclusions from the OLS model are robust to the inclusion of school fixed effects.

15.20(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?

Ans

```
# F test for the significance of school fixed effects
# Compare fixed effect model vs. pooled OLS (no fixed effect)
pooling_model <- plm(readscore ~ small + aide + tchexper + boy + white_asian + freelunch,
                     data = pdata_star, model = "pooling")

pFtest(model_15_20b, pooling_model)
```

```
##
##  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
```

The F-test for individual (school) fixed effects yields an F-statistic of 16.698 with a p-value < 2.2e-16. This provides strong evidence against the null hypothesis of no school effects, indicating that school fixed effects are jointly significant and should be included in the model.

We would expect the inclusion of significant fixed effects to have little influence on the coefficients of other explanatory variables if those variables are uncorrelated with the unobserved school-level heterogeneity. However, if these explanatory variables (such as class size or teacher experience) are correlated with school-specific characteristics, omitting fixed effects could bias the coefficient estimates.