

## Chapter 14

# Introduction to panel data

## Overview

Increasingly, researchers are now using panel data where possible in preference to cross-sectional data. One major reason is that dynamics may be explored with panel data in a way that is seldom possible with cross-sectional data. Another is that panel data offer the possibility of a solution to the pervasive problem of omitted variable bias. A further reason is that panel data sets often contain very large numbers of observations and the quality of the data is high.

## Learning outcomes

After working through the corresponding chapter in the text, studying the corresponding slideshows, and doing the starred exercises in the text and the additional exercises in this guide, you should be able to:

- explain the differences between panel data, cross-sectional data, and time series data
- explain what the benefits that can be obtained using panel data
- explain the differences between OLS pooled regressions, fixed effects regressions, and random effects regressions
- explain the potential advantages of the fixed effects model over pooled OLS
- explain the differences between the within-groups, first differences, and least squares dummy variables variants of the fixed effects model
- explain the assumptions required for the use of the random effects model
- explain the advantages of the random effects model over the fixed effects model when the assumptions are valid
- explain how to use a Durbin–Wu–Hausman test to determine whether the random effects model may be used instead of the fixed effects model

## Additional exercises

A14.1 The *NLSY2000* data set contains the following data for a sample of 2,427 males and 2,392 females for the years 1980–2000: weight in pounds, years of schooling, age, marital status in the form of a dummy variable *MARRIED* defined to be 1 if the respondent was married, 0 if single, and height in inches. Hypothesizing that weight is influenced by schooling, age, marital status, and height, the following regressions were performed for males and females separately:

- (1) an ordinary least squares (OLS) regression pooling the observations
- (2) a within-groups fixed effects regression
- (3) a random effects regression

The results of these regressions are shown in the table. Standard errors are given in parentheses.

Dougherty: Introduction to Econometrics 3e  
Study Guide

	Males			Females		
	OLS	FE	RE	OLS	FE	RE
Years of schooling	-0.98 (0.09)	-0.02 (0.23)	-0.45 (0.16)	-1.95 (0.12)	-0.60 (0.27)	-1.25 (0.18)
Age	1.61 (0.04)	1.64 (0.02)	1.65 (0.02)	2.03 (0.05)	1.66 (0.03)	1.72 (0.03)
Married	3.70 (0.48)	2.92 (0.33)	3.00 (0.32)	-8.27 (0.59)	3.08 (0.46)	1.98 (0.44)
Height	5.07 (0.08)	dropped	4.95 (0.18)	3.48 (0.10)	dropped	3.38 (0.21)
constant	-209.52 (5.39)	dropped	-209.81 (12.88)	-105.90 (6.62)	dropped	-107.61 (13.43)
$R^2$	0.27	-	-	0.17	-	-
$n$	17,299	17,299	17,299	13,160	13,160	13,160
DWH $\chi^2(3)$			7.22			92.94

- Explain why height is excluded from the FE regression.
- Evaluate, for males and females separately, whether the fixed effects or random effects model should be preferred.
- For males and females separately, compare the estimates of the coefficients in the OLS and FE models and attempt to explain the differences.
- Explain in principle how one might test whether individual-specific fixed effects jointly have significant explanatory power, if the number of individuals is small. Explain why the test is not practical in this case.

A14.2 The *NLSY2000* data set contains the following data for a sample of 2,427 males and 2,392 females for the years 1980–2000: years of work experience, *EXP*, years of schooling, *S*, and age, *AGE*. A researcher investigating the impact of schooling on willingness to work regresses *EXP* on *S*, including potential work experience, *PWE*, as a control. *PWE* was defined as

$$PWE = AGE - S - 5$$

The following regressions were performed for males and females separately:

- an ordinary least squares (OLS) regression pooling the observations
- a within-groups fixed effects regression
- a random effects regression

The results of these regressions are shown in the table. Standard errors are given in parentheses.

	Males			Females		
	OLS	FE	RE	OLS	FE	RE
<i>S</i>	0.78 (0.01)	0.65 (0.01)	0.72 (0.01)	0.89 (0.01)	0.71 (0.02)	0.85 (0.01)
<i>PWE</i>	0.83 (0.003)	0.94 (0.001)	0.94 (0.001)	0.74 (0.004)	0.88 (0.002)	0.87 (0.002)
constant	-10.16 (0.09)	dropped	-10.56 (0.14)	-11.11 (0.12)	dropped	-12.39 (0.19)
$R^2$	0.79	—	—	0.71	—	—
<i>n</i>	24,057	24,057	24,057	18,758	18,758	18,758
DWH $\chi^2(2)$			10.76			1.43

- (a) Explain why the researcher included *PWE* as a control.
- (b) Evaluate the results of the Durbin–Wu–Hausman tests
- (c) For males and females separately, explain the differences in the coefficients of *S* in the OLS and FE regressions.
- (d) For males and females separately, explain the differences in the coefficients of *PWE* in the OLS and FE regressions.

A14.3 Using the *NLSY2000* data set, a researcher fits OLS and fixed effects regressions of the logarithm of hourly wages on schooling, years of work experience, *EXP*, *ASVABC* score, and dummies *MALE*, *ETHBLACK*, and *ETHHISP* for being male, black, or hispanic. Schooling was split into years of high school, *SH*, and years of college, *SC*. The results are shown in the table, with standard errors placed in parentheses.

	OLS	FE	RE
<i>SH</i>	0.026 (0.002)	0.005 (0.007)	0.016 (0.004)
<i>SC</i>	0.063 (0.001)	0.073 (0.004)	0.067 (0.002)
<i>EXP</i>	0.033 (0.0004)	0.032 (0.0003)	0.033 (0.0003)
<i>ASVABC</i>	0.012 (0.0003)	—	0.011 (0.001)
<i>MALE</i>	0.193 (0.004)	—	0.197 (0.009)
<i>ETHBLACK</i>	-0.040 (0.007)	—	-0.030 (0.015)
<i>ETHHISP</i>	0.047 (0.008)	—	0.033 (0.018)
constant	5.639 (0.028)	—	5.751 (0.051)
$R^2$	0.0367	—	—
DWH $\chi^2(3)$	—	—	9.31

If an individual reported being in high school or college, the observation for that individual for that year was deleted from the sample. As a consequence, the observations for most individuals in the sample begin when the formal education of that individual has been completed. However a small minority of individuals, having apparently completed their formal education and having taken employment, subsequently resumed their formal education, either to complete high school with a general educational development (GED) degree equivalent to the high school diploma, or to complete one or more years of college.

- (a) Discuss the differences in the estimates of the coefficient of  $SH$ .
- (b) Discuss the differences in the estimates of the coefficient of  $SC$ .

## Answers to the additional exercises

A14.1 (a) Height is constant over observations. Hence, for each individual,  $HEIGHT_{it} - \overline{HEIGHT}_i = 0$  for all  $t$ , where  $\overline{HEIGHT}_i$  is the mean height for individual  $i$  for the observations for that individual. Hence height has to be dropped from the regression model.

(b) The critical value of chi-squared, with three degrees of freedom, is 7.82 at the 5 percent level and 16.27 at the 0.1 percent level. Hence there is a possibility that the random effects model may be appropriate for males, but it is definitely not appropriate for females.

(c) *Males*

The OLS regression suggests that schooling has a small (one pound less per year of schooling) but highly significant negative effect on weight. The fixed effects regression eliminates the effect, indicating that an unobserved effect is responsible: males with unobserved qualities that favour more schooling, controlling for other measured variables, have lower weight as a consequence of the same unobserved qualities. We cannot compare estimates of the effect of height since it is dropped from the FE regression. The effect of age is the same in the two regressions. There is a small but highly significant positive effect of being married, the OLS estimate possibly being inflated by an unobserved effect.

*Females*

The main difference is in the marriage coefficient. The OLS regression suggests that marriage reduces weight by eight pounds, a remarkable amount. The FE regression suggests the opposite, that marriage leads to an *increase* in weight that is similar to that for males. The clear implication is that women who weigh less are relatively successful in the marriage market, but even they put on weight once they are married.

For schooling the story is much the same as for males, except that the OLS coefficient is much larger and the coefficient remains significant at the 5 percent level in the FE regression. The effect of age appears to be exaggerated in the OLS regression, for reasons that are not obvious.

(d) One would perform a LSDV regression, eliminating the intercept in the model and adding a dummy variable for each individual. One would compare  $RSS$  for this regression with that for the regression without the dummy variables, using a standard  $F$  test. In the present case it is not a practical proposition because there are more than 17,000 males and 13,000 females.

A14.2 (a) Clearly actual work experience is positively influenced by  $PWE$ . Omitting it would cause the coefficient of  $S$  to be biased downwards since  $PWE$  and  $S$  are negatively correlated.

(b) With two degrees of freedom, the critical value of chi-squared is 5.99 at the 5 percent level and 9.21 at the 1 percent level. Thus the random effects model is rejected for males but seemingly not for females.

- (c) For both sexes the OLS estimate is greater than the FE estimate. One possible reason is that some unobserved characteristics, for example drive, are positively correlated with both acquiring schooling, and seeking and gaining employment.
- (d) Since  $S$  and  $PWE$  are negatively correlated, these same unobserved characteristics would cause the OLS estimate of the coefficient of  $PWE$  to be biased downwards.

A14.3 First, note that the DWH statistic is significant at the 5 percent level (critical value 7.82) but not at the 1 percent level (critical value 11.35).

The coefficients of  $SH$  and  $SC$  in the OLS regression is an estimate of the impact of variations in years of high school and years of college among all the individuals in the sample. Most individuals in fact completed high school and so had  $SH = 12$ . However, a small minority did not and this variation made possible the estimation of the  $SH$  coefficient. The majority of the remainder did not complete any years of college and therefore had  $SC = 0$ , but a substantial minority did have a partial or complete college education, some even pursuing postgraduate studies, and this variation made possible the estimation of the  $SC$  coefficient.

Most individuals completed their formal education before entering employment. For them,  $SH_{it} = \overline{SH}_i$  for all  $t$  and hence  $SH_{it} - \overline{SH}_i = 0$  for all  $t$ . As a consequence, the observations for such individuals provide no variation in the  $SH$  variable. Likewise they provide no variation in the  $SC$  variable. If all observations pertained to such individuals, schooling would be washed out in the FE regression along with other unchanging characteristics such as sex, ethnicity, and  $ASVABC$  score. The schooling coefficients in the FE regression therefore relate to those individuals who returned to formal education after a break in which they found employment.

The fact that these individuals account for a relatively small proportion of the observations in the data set has an adverse effect on the precision of the FE estimates of the coefficients of  $SH$  and  $SC$ . This is reflected in standard errors that are much larger than those obtained in the OLS pooled regression.

- (a) Most of the variation in  $SH$  in the FE regressions come from individuals earning the GED degree. This degree provides an opportunity for high school drop-outs to make good their shortfall by taking courses and passing the examinations required for this diploma. These course may be civilian or military adult education classes, but very often they are programmes offered to those in jail. In principle the GED should be equivalent to the high school diploma, but there is some evidence that standards are sometimes lower. The results in the table appear to corroborate this view. The OLS regression indicates that a year of high school raises earnings by 2.6 percent, with the coefficient being highly significant, whereas the FE coefficient indicates that the effect is only 0.5 percent and not significant.
- (b) Some of the variation in  $SC$  in the FE regressions comes from individuals entering employment for a year or two after finishing high school and then going to college, resuming their formal education. However most comes from individuals returning college for a year or two after having been employment for a number of years. A typical example is a high school graduate who has settled down in an occupation and who has then decided to upgrade his or her professional skills by taking a two-year associate of arts degree. Similarly one encounters college graduates who upgrade to masters level after having worked for some time. One would expect such students to be especially well motivated—they are often undertaking studies that a relevant to an established career, and they are often bearing high opportunity costs from loss of earnings while studying—and accordingly one might expect the payoff in terms of increased earnings to be relatively high. This seems to be borne out in a comparison of the OLS and FE estimates of the coefficient of  $SC$ , though the difference is not dramatic.

On the surface, this exercise appeared to be about how one might use FE to eliminate the bias in OLS pooled regression caused by unobserved effects. Has the analysis been successful in is respect?

Absolutely not. In particular, the apparent conclusion that high school education has virtually no effect

on earnings should not be taken at face value. The reason is that the issue of biases attributable to unobserved effects has been overtaken by the much more important issue of the difference in the interpretation of the *SH* and *SC* coefficients discussed in (a) and (b). This illustrates a basic point in econometrics: understanding the context of the data is often just as important as being proficient at technical analysis.