Chapter 14
Introduction to panel data

14.1 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 crosssectional 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. This chapter describes fixed effects regression and random effects regression, alternative techniques that exploit the structure of panel data.

14.2 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 subject guide, you should be able to:

- explain the differences between panel data, cross-sectional data, and time series data
- explain 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.

14.3 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: 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:

(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 below. Standard errors are given in parentheses.

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

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

A14.2 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 below, with standard errors placed in parentheses.

	OLS	FE	RE
SH	0.026	0.005	0.016
	(0.002)	(0.007)	(0.004)
SC	0.063	0.073	0.067
	(0.001)	(0.004)	(0.002)
EXP	0.033	0.032	0.033
	(0.004)	(0.003)	(0.003)
ASVABC	0.012	—	0.011
	(0.003)		(0.001)
MALE	0.193		0.197
	(0.004)		(0.009)
ETHBLACK	−0.040	—	−0.030
	(0.007)		(0.015)
ETHHISP	0.047	—	0.033
	(0.008)		(0.018)
constant	5.639	—	5.751
	(0.028)		(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.

- Discuss the differences in the estimates of the coefficient of *SH*.
- Discuss the differences in the estimates of the coefficient of *SC*.

A14.3 A researcher has data on G, the average annual rate of growth of GDP 2001–2005, and S, the average years of schooling of the workforce in 2005, for 28 European Union countries. She believes that G depends on S and on E, the level of entrepreneurship in the country, and a disturbance term u:

$$G = \beta_1 + \beta_2 S + \beta_3 E + u. \qquad (1)$$

u may be assumed to satisfy the usual regression model assumptions. Unfortunately the researcher does not have data on E.

- Explain intuitively and mathematically the consequences of performing a simple regression of G on S. For this purpose S and E may be treated as nonstochastic variables.

  The researcher does some more research and obtains data on G^*, the average annual rate of growth of GDP 1996–2000, and S^*, the average years of schooling of the workforce in 2000, for the same countries. She thinks that she

can deal with the unobservable variable problem by regressing ΔG, the change in G, on ΔS, the change in S, where:

$$\begin{aligned} \Delta G &= G - G^* \\ \Delta S &= S - S^* \end{aligned}$$

assuming that E would be much the same for each country in the two periods. She fits the equation:

$$\Delta G = \delta_1 + \delta_2 \Delta S + w \qquad (2)$$

where w is a disturbance term that satisfies the usual regression model assumptions.

- Compare the properties of the estimators of the coefficient of S in (1) and of the coefficient of ΔS in (2).
- Explain why in principle you would expect the estimate of δ_1 in (2) not to be significant. Suppose that nevertheless the researcher finds that the coefficient is significant. Give two possible explanations.

Random effects regressions have potential advantages over fixed effect regressions.

- Could the researcher have used a random effects regression in the present case?

A14.4 A researcher has the following data for 3,763 respondents in the National Longitudinal Survey of Youth 1979– : hourly earnings in dollars in 1994 and 2000, years of schooling as recorded in 1994 and 2000, and years of work experience as recorded in 1994 and 2000. The respondents were aged 14–21 in 1979, so they were aged 29–36 in 1994 and 35–42 in 2000. 371 of the respondents had increased their formal schooling between 1994 and 2000, 210 by one year, 101 by two years, 47 by three years, and 13 by more than three years, mostly at college level in non-degree courses. The researcher performs the following regressions:

(1) the logarithm of hourly earnings in 1994 on schooling and work experience in 1994

(2) the logarithm of hourly earnings in 2000 on schooling and work experience in 2000

(3) the change in the logarithm of hourly earnings from 1994 to 2000 on the changes in schooling and work experience in that interval.

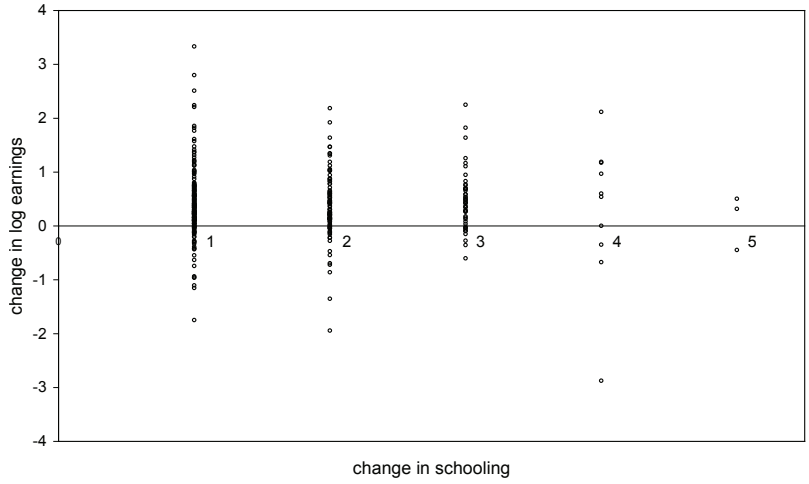
The results are shown in columns (1) – (3) in the table (t statistics in parentheses), and are presented at a seminar.

	(1)	(2)	(3)	(4)	(5)
Dependent variable	log earnings 1994	log earnings 2000	Change in log earnings 1994–2000	log earnings 2000	Change in log earnings 1994–2000
Schooling	0.114 (30.16)	0.116 (28.99)	—	0.108 (24.53)	—
Experience	0.052 (18.81)	0.038 (14.59)	—	0.037 (14.10)	— —
Cognitive ability score	—	—	—	0.004 (4.79)	—
Male	0.214 (12.03)	0.229 (11.77)	—	0.230 (11.88)	—
Black	−0.149 (−5.23)	−0.199 (−6.44)	—	−0.167 (−5.29)	—
Hispanic	0.039 (1.11)	0.053 (1.38)	—	0.071 (1.84)	—
Change in schooling	—	—	0.090 (5.00)	—	−0.006 (−0.16)
Change in experience	—	—	0.024 (2.75)	—	0.003 (0.15)
constant	4.899 (74.59)	5.023 (65.02)	0.102 (2.13)	4.966 (63.69)	0.389 (3.05)
R^2	0.265	0.243	0.007	0.248	0.0002
n	3,763	3,763	3,763	3,763	371

- The researcher is unable to explain why the coefficient of the change in schooling in regression (3) is so much lower than the schooling coefficients in (1) and (2). Someone says that it is because he has left out relevant variables such as cognitive ability, region of residence, etc, and the coefficients in (1) and (2) are therefore biased. Someone else says that cannot be the explanation because these variables are also omitted from regression (3). Explain what would be your view.
- He runs regressions (1) and (2) again, adding a measure of cognitive ability. The results for the 2000 regression are shown in column (4). The results for 1994 were very similar. Discuss possible reasons for the fact that the estimate of the schooling coefficient differs from those in (2) and (3).
- Someone says that the researcher should not have included a constant in regression (3). Explain why she made this remark and assess whether it is valid.
- Someone else at the seminar says that the reason for the relatively low coefficient of schooling in regression (3) is that it mostly represented non-degree schooling. Hence one would not expect to find the same relationship between schooling and earnings as for the regular pre-employment schooling of young people. Explain in general verbal terms what investigation the researcher should undertake in response to this suggestion.
- Another person suggests that the small minority of individuals who went back to school or college in their thirties might have characteristics different from

those of the individuals who did not, and that this could account for a different coefficient. Explain in general verbal terms what investigation the researcher should undertake in response to this suggestion.

- Finally, another person says that it might be a good idea to look at the relationship between earnings and schooling for the subsample who went back to school or college, restricting the analysis to these 371 individuals. The researcher responds by running the regression for that group alone. The result is shown in column (5) in the table. The researcher also plots a scatter diagram, reproduced below, showing the change in the logarithm of earnings and the change in schooling. For those with one extra year of schooling, the mean change in log earnings was 0.40. For those with two extra years, 0.37. For those with three extra years, 0.47. What conclusions might be drawn from the regression results?



A14.5 In the discussion of the DWH test, it was stated that the test compares the coefficients of those variables not dropped in the FE regression. Explain why the constant is not included in the comparison.

14.4 Answer to the starred exercise in the textbook

14.9 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.

	Males			Females		
	OLS	FE	RE	OLS	FE	RE
Year 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.

Answer:

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.

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.

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 have a positive effect on educational attainment, 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, and very striking, 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 once they are married they put on weight.

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.

One might test whether individual-specific fixed effects jointly have significant explanatory power by performing 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.

14.5 Answers to the additional exercises

A14.1 *Explain why the researcher included PWE as a control.*

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.

Evaluate the results of the Durbin–Wu–Hausman tests

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.

For males and females separately, explain the differences in the coefficients of S in the OLS and FE regressions.

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.

For males and females separately, explain the differences in the coefficients of PWE in the OLS and FE regressions.

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.

A14.2 First, note that the DWH statistic is significant at the 5 per cent level (critical value 7.82) but not at the 1 per cent 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.

Discuss the differences in the estimates of the coefficient of SH.

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 courses 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 per cent, with the coefficient being highly significant, whereas the FE coefficient indicates that the effect is only 0.5 per cent and not significant.

Discuss the differences in the estimates of the coefficient of SC.

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 to 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 are 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 this 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 the exercise. This illustrates a basic point in econometrics: understanding the context of the data is often just as important as being proficient at technical analysis.

A14.3 *Explain intuitively and mathematically the consequences of performing a simple regression of G on S. For this purpose S and E may be treated as nonstochastic variables.*

If one fits the regression:

$$\widehat{G} = \widehat{\beta}_1 + \widehat{\beta}_2 S$$

then

$$\begin{aligned}
\widehat{\beta}_2 &= \frac{\sum \left(S_i - \overline{S}\right)\left(G_i - \overline{G}\right)}{\sum \left(S_i - \overline{S}\right)^2} \\
&= \frac{\sum \left(S_i - \overline{S}\right)\left((\beta_1 + \beta_2 S_i + \beta_3 E_i + u_i) - \left(\beta_1 + \beta_2 \overline{S} + \beta_2 \overline{E} + \overline{u}\right)\right)}{\sum \left(S_i - \overline{S}\right)^2} \\
&= \beta_2 + \beta_3 \frac{\sum \left(S_i \overline{S}\right)\left(E_i - \overline{E}\right)}{\sum \left(S_i - \overline{S}\right)^2} + \frac{\sum \left(S_i - \overline{S}\right)(u_i - \overline{u})}{\sum \left(S_i - \overline{S}\right)^2}.
\end{aligned}$$

Taking expectations, and making use of the invitation to treat S and E as nonstochastic:

$$\begin{aligned}
E(\widehat{\beta}_2) &= \beta_2 + \beta_3 \frac{\sum \left(S_i \overline{S}\right)\left(E_i - \overline{E}\right)}{\sum \left(S_i - \overline{S}\right)^2} + \frac{\sum \left(S_i - \overline{S}\right) E\,(u_i - \overline{u})}{\sum \left(S_i - \overline{S}\right)^2} \\
&= \beta_2 + \beta_3 \frac{\sum \left(S_i - \overline{S}\right)\left(E_i - \overline{E}\right)}{\sum \left(S_i - \overline{S}\right)^2}.
\end{aligned}$$

Hence the estimator is biased unless S and E happen to be uncorrelated in the sample. As a consequence, the standard errors will be invalid.

Compare the properties of the estimators of the coefficient of S in (1) and of the coefficient of ΔS in (2).

Given (1), the differenced model should have been:

$$\Delta G = \delta_2 \Delta S + w$$

where $w = u - u^*$.

The estimator of the coefficient of ΔS in (2) should be unbiased, while that of S in (1) will be subject to omitted variable bias. However:

- it is possible that the bias in (1) may be small. This would be the case if E were a relatively unimportant determinant of G or if its correlation with S were low.
- it is possible that the variance in ΔS is smaller than that of S. This would be the case if S were changing slowly in each country, or if the rate of change of S were similar in each country.

Thus there may be a trade-off between bias and variance and it is possible that the estimator of β_2 using specification (1) could actually be superior according to some criterion such as the mean square error. It should be noted that the inclusion of δ_1 in (2) will make the estimation of δ_2 even less efficient.

Explain why in principle you would expect the estimate of δ_1 in (2) not to be significant. Suppose that nevertheless the researcher finds that the coefficient is significant. Give two possible explanations.

If specification (1) is correct, there should be no intercept in (2) and for this reason the estimate of the intercept should not be significantly different from zero. If it is significant, this could have occurred as a matter of Type I error. Alternatively, it might indicate a shift in the relationship between the two time periods. Suppose that (1) should have included a dummy variable set equal to 0 in the first time period and 1 in the second. $\widehat{\delta}_1$ would then be an estimate of its coefficient.

Could the researcher have used a random effects regression in the present case?

Random effects requires the sample to be drawn randomly from a population and for unobserved effects to be uncorrelated with the regressors. The first condition is not satisfied here, so random effects would be inappropriate.

A14.4 *The researcher is unable to explain why the coefficient of the change in schooling in regression (3) is so much lower than the schooling coefficients in (1) and (2). Someone says that it is because he has left out relevant variables such as cognitive ability, region of residence, etc, and the coefficients in (1) and (2) are therefore biased. Someone else says that cannot be the explanation because these variables are also omitted from regression (3). Explain what would be your view.*

Suppose that the true model is:

$$\begin{aligned} LGEARN &= \beta_1 + \beta_2 S + \beta_3 EXP + \beta_4 ASVABC + \beta_5 MALE \\ &\quad + \beta_6 ETHBLACK + \beta_7 ETHHISP + \beta_8 X_8 + u \end{aligned}$$

where X_8 is some further fixed characteristic of the respondent. $ASVABC$ and X_8 are absent from regressions (1) and (2) and so those regressions will be subject to omitted variable bias. In particular, since $ASVABC$ is likely to be positively correlated with S, and to have a positive coefficient, its omission will tend to bias the coefficient of S upwards.

However, if the specification is valid for both 1994 and 2000 and unchanged, one can eliminate the omitted variable bias by taking first differences as in regression (3):

$$\Delta LGEARN = \beta_2 \Delta S + \beta_3 \Delta EXP + \Delta u.$$

By fitting this specification one should obtain unbiased estimates of the coefficients of schooling and experience, and the former should therefore be smaller than in (1)

and (2). Note that all the fixed characteristics have been washed out. The suggestion that *ASVABC* should have been included in (3) is therefore incorrect.

Note that (3) should not have included an intercept. This is discussed later in the question.

He runs regressions (1) and (2) again, adding a measure of cognitive ability. The results for the 2000 regression are shown in column (4). The results for 1994 were very similar. Discuss possible reasons for the fact that the estimate of the schooling coefficient differs from those in (2) and (3).

The estimate of the coefficient of S differs from that in (2) because the omitted variable bias attributable to the omission of *ASVABC* in that specification has now been corrected. However it is still biased if X_8 (representing other omitted characteristics) is a determinant of earnings and is correlated with S. This partial rectification of the omitted variable problem accounts for the fact that the coefficient of S in (4) lies between those in (2) and (3).

Someone says that the researcher should not have included a constant in regression (3). Explain why she made this remark and assess whether it is valid.

Given the specification in (1) and (2), there should have been no intercept in the first differences specification (3). One would therefore expect the estimate of the intercept to be somewhere near zero in the sense of not being significantly different from it. Nevertheless, it is significantly different at the 5 percent level. However, suppose that the relationship shifted between 1994 and 2000, and that the shift could be represented by a dummy variable D equal to zero in 1994 and 1 in 2000, with coefficient δ. Then (3) should have an intercept δ. Its estimate, 0.102, suggests that earnings grew by 10 percent from 1994 to 2000, holding other factors constant. This seems entirely reasonable, perhaps even a little low.

Alternatively, the apparently significant t statistic might have arisen as a matter of Type I error.

Someone else at the seminar says that the reason for the relatively low coefficient of schooling in regression (3) is that it mostly represented non-degree schooling. Hence one would not expect to find the same relationship between schooling and earnings as for the regular preemployment schooling of young people. Explain in general verbal terms what investigation the researcher should undertake in response to this suggestion.

Divide S into two variables, schooling as of 1994 and extra schooling as of 2000, with separate coefficients. Then use a standard F test (or t test) of a restriction to test whether the coefficients are significantly different.

Another person suggests that the small minority of individuals who went back to school or college in their thirties might have characteristics different from those of the individuals who did not, and that this could account for a different coefficient. Explain in general verbal terms what investigation the researcher should undertake in response to this suggestion.

The issue is sample selection bias and an appropriate procedure would be that proposed by Heckman. One would use probit analysis with an appropriate set of determinants to model the decision to return to school between 1994 and 2000, and a regression model to explain variations in the logarithm of earnings of those

respondents who do return to school, linking the two models by allowing their disturbance terms to be correlated. One would test whether the estimate of this correlation is significantly different from zero.

Finally, another person says that it might be a good idea to look at the relationship between earnings and schooling for the subsample who went back to school or college, restricting the analysis to these 371 individuals. The researcher responds by running the regression for that group alone. The result is shown in column (5) in the table. The researcher also plots a scatter diagram, reproduced below, showing the change in the logarithm of earnings and the change in schooling. For those with one extra year of schooling, the mean change in log earnings was 0.40. For those with two extra years, 0.37. For those with three extra years, 0.47. What conclusions might be drawn from the regression results?

The schooling coefficient is effectively zero! [These are real data, incidentally.] The scatter diagram shows why. Irrespective of whether the respondent had one, two, or three years of extra schooling, the gain is about the same, on average. (These are the only categories with large numbers of observations, given the information at the beginning of the question, confirmed by the scatter diagram.) So the results indicate that the fact of going back to school, rather than the duration of the schooling, is the relevant determinant of the change in earnings. The intercept indicates that this subsample on average increased their earnings between 1994 and 2000 by 38.9 percent. (As a first approximation. The actual proportion would be better estimated as $e^{0.389} - 1 = 0.476$.) This figure is confirmed by the diagram, and it would appear to be much greater than the effect of regular schooling. One explanation could be sample selection bias, as already discussed. A more likely possibility is that the respondents were presented with opportunities to increase their earnings substantially if they undertook certain types of formal course, and they took advantage of these opportunities.

A14.5 In a random effects regression, the interpretation of an intercept is not affected by the estimation technique. In a fixed effects regression, the intercept is washed out. Hence there is no basis for a comparison. In general, the model is fitted without an intercept. The only case where an intercept should be included is in first-differences fixed effects estimation of a model containing a deterministic trend. For example, suppose one is fitting the model:

$$Y_{it} = \beta_1 + \beta_2 X_{it} + \delta t + u_{it}.$$

For individual i in the previous time period, one has:

$$Y_{i,t-1} = \beta_1 + \beta_2 X_{i,t-1} + \delta(t-1) + u_{i,t-1}.$$

Subtracting, one obtains:

$$Y_{it} - Y_{i,t-1} = \beta_2(X_{it} - X_{i,t-1}) + \delta + u_{it} - u_{i,t-1}.$$

The model now does have an intercept, but its meaning is different from that in the original specification. It now provides an estimate of δ, not β_1.