Summary of Chapter 15

Panel data from NLS

Short and wide

T=5, N=716

TABLE 15.1 **Representative Observations from NLS Panel Data**

ID	YEAR	LWAGE	EDUC	SOUTH	BLACK	UNION	EXPER	TENURE
1	82	1.8083	12	0	1	1	7.6667	7.6667
1	83	1.8634	12	0	1	1	8.5833	8.5833
1	85	1.7894	12	0	1	1	10.1795	1.8333
1	87	1.8465	12	0	1	1	12.1795	3.7500
1	88	1.8564	12	0	1	1	13.6218	5.2500
2	82	1.2809	17	0	0	0	7.5769	2.4167
2	83	1.5159	17	0	0	0	8.3846	3.4167
2	85	1.9302	17	0	0	0	10.3846	5.4167
2	87	1.9190	17	0	0	1	12.0385	0.3333
2	88	2.2010	17	0	0	1	13.2115	1.7500
3	82	1.8148	12	0	0	0	11.4167	11.4167
3	83	1.9199	12	0	0	1	12.4167	12.4167
3	85	1.9584	12	0	0	0	14.4167	14.4167
3	87	2.0071	12	0	0	0	16.4167	16.4167
3	88	2.0899	12	0	0	0	17.8205	17.7500

A group of cross-sectional units who are observed over time

Example: Wage equation

EXAMPLE 15.1 | Revisited

For example, in Table 15.1, the outcome variable of interest is $y_{it} = LWAGE_{it} = \ln(WAGE_{it})$. Explanatory variables include $x_{2it} = EXPER_{it}$, $x_{3it} = TENURE_{it}$, $x_{4it} = SOUTH_{it}$, and $x_{5it} = UNION_{it}$. These explanatory variables vary across both individual and time. For the indicator variables *SOUTH* and *UNION*, it means that at least some individuals moved into or out of the *SOUTH* during the 1982–1988 period, and at least some workers joined or quit a *UNION* over those years. The variables $w_{1i} = EDUC_i$ and $w_{2i} = BLACK_i$ do not change for the 716 individuals in our sample over the years 1982–1988. Two unobserved **time-invariant variables** are $u_{1i} = ABILITY_i$ and $u_{2i} = PERSEVERANCE_i$. Unobserved time-specific variables might be $m_{1t} = UNEMPLOYMENT\ RATE_t$ or $m_{2t} = INFLATION\ RATE_t$. Note that it is possible to have observable variables that change over time but not across individuals, like an indicator variable $D82_t = 1$ if the year is 1982 and $D82_t = 0$ otherwise.

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it}) = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + v_{it}$$

The panel data regression function

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + (u_i + e_{it}) = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + v_{it}$$

Because the regression error in the equation has two components, one for the individual and one for the regression, it is often called an error components model.

Using panel data we must consider dynamic, time related effects and model assumptions should take them into account.

The model conditions on the unobservable time-invariant error u_i.

The panel data regression function

- When using panel data, it is important to separate out this component (unobservable time-invariant error u_i) of the random error term from other components if we can argue that the factors causing the individual differences are unchanging over time.

- We consider estimation procedures that employ a transformation to eliminate the individual heterogeneity from the estimation equation and thus solve the common **endogeneity** problem.

Estimation procedures

- The estimators we will consider are
  1. The difference estimator
  2. The within estimator
  3. The fixed effects estimator

The difference estimator: T=2

- Each individual in two different time periods, t = 1 and t = 2 are:
- $y_{i1} = \beta_1 + \beta_2 x_{2i1} + \alpha_1 w_{1i} + u_i + e_{i1}$ (15.7a)
- $y_{i2} = \beta_1 + \beta_2 x_{2i2} + \alpha_1 w_{1i} + u_i + e_{i2}$ (15.7b)
- Subtracting (15.7a) from (15.7b) creates a new equation
- $(y_{i2} - y_{i1}) = \beta_2 (x_{2i2} - x_{2i1}) + (e_{i2} - e_{i1})$
- ➔ $\Delta y_i = \beta_2 \Delta x_{i2} + \Delta e_i$

The difference estimator: T=2

- In basic panel data analysis, the difference estimator is usually not used. We introduce it to illustrate that we can *eliminate* the unobserved heterogeneity through a transformation. In practice, we usually use the equivalent, but more flexible, fixed effects estimator.

- Example 15.2 (p. 641) Using $T = 2$ Differenced Observations for a Production Function

- The difference estimator is consistent when unobserved heterogeneity is correlated with the explanatory variables, but the OLS estimator is not. Given the substantial difference in the estimates we might suspect that the OLS estimates are unreliable.

The within estimator: T=2

- The advantage of the within transformation is that it **generalizes** nicely to situations when we have more than T = 2 time observations on each individual.

- The time-average of the equations, (15.7a) and (15.7b):

$$\frac{1}{2}\sum_{t=1}^{T}(y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it})$$

- Instead of first-differenced variables, we have differences from the variable means. The time-invariant terms subtract out, including the unobservable heterogeneity term.

The within estimator: T=2

- The time-averaged model for i = 1,…,N is

$$\bar{y}_{i.} = \beta_1 + \beta_2 \bar{x}_{2i.} + \alpha_1 w_{1i} + u_i + \bar{e}_{i.}$$

- The within transformation subtracts (15.10) from the original observations to obtain

$$y_{it} - \bar{y}_{i.} = \beta_2 (x_{2it} - \bar{x}_{2i.}) + (e_{it} - \bar{e}_{i.})$$

- The within-transformed model is

$$\tilde{y}_{it.} = \beta_2 \tilde{x}_{2it} + \tilde{e}_{it}$$

The within estimator: T=2

- **EXAMPLE 15.4** Using the Within Transformation with $T = 2$ Observations for a Production Function

- Notice that the within estimates are exactly the same as the first-difference estimates in Example 15.2.

The within estimator: T>2

- Suppose that we have T observations on each individual
- $y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it},\ i = 1, \dots, N, \quad t = 1, \dots, T$
- Averaging over all time observations:
- $\frac{1}{T}\sum_{t=1}^{T}(y_{it} = \beta_1 + \beta_2 x_{2it} + \alpha_1 w_{1i} + u_i + e_{it})$
- ➔

$$\bar{y}_{i.} = \beta_1 + \beta_2 \bar{x}_{2i.} + \alpha_1 w_{1i} + u_i + \bar{e}_{i.}$$

$$y_{it} - \bar{y}_{i.} = \beta_2 (x_{2it} - \bar{x}_{2i.}) + (e_{it} - \bar{e}_{i.})$$

$$\tilde{y}_{it.} = \beta_2 \tilde{x}_{2it} + \tilde{e}_{it}$$

The least squares dummy variable model

- It turns out that the within estimator is numerically equivalent to another estimator that has long been used in empirical work and that is logically appealing.
- Unobserved heterogeneity is controlled for by including in the panel data regression

$$D_{1i} = \begin{cases} 1 & i = 1 \\ 0 & \text{otherwise} \end{cases} \quad D_{2i} = \begin{cases} 1 & i = 2 \\ 0 & \text{otherwise} \end{cases} \quad D_{3i} = \begin{cases} 1 & i = 3 \\ 0 & \text{otherwise} \end{cases}$$

- $y_{it} = \beta_{11} D_{1i} + \beta_{12} D_{2i} + \cdots + \beta_{1N} D_{Ni} + \beta_2 x_{2it} + \cdots + \beta_K x_{Kit} + e_{it}$

- This is called the fixed effects model, or sometimes the least squares dummy variable model.

The least squares dummy variable model

- Testing for individual differences in the fixed effects model is a test of the joint hypothesis

$$H_0: \beta_{11} = \beta_{12} = \beta_{13} = \cdots = \beta_{1,N-1} = \beta_{1N} \text{ and}$$

$$H_1: \textit{the } \beta_{1i} \textit{ are not all equal}$$

Remarks

- To summarize, the within estimator, the fixed effects estimator and the least squares dummy variable estimator are all names for the same estimators of β2,…, β*K* in (15.17). In practice, no choice is required. Use the computer software option for "fixed effects" estimation.

- Consider now the fixed effects estimation procedure that employs the "within" transformation
  - The within transformation removes the unobserved heterogeneity so that only the idiosyncratic error e_{it} remains
  - It is possible that within the cluster of observations defining each individual cross-sectional unit there remains serial correlation and/or heteroscedasticity
  - EXAMPLE 15.8 Using Fixed Effects and Cluster-Robust Standard Errors for a Production Function (p. 651)

The random effects estimator

- Panel data applications fall into one of two types:

  1. The first type of application is when the unobserved heterogeneity term u_i is correlated with one or more of the explanatory variables ➔ fixed effects model or OLS estimators (need to test the joint significance of fixed effects)

  2. The second type of application is when the unobserved heterogeneity term u_i is not correlated with any of the explanatory variables ➔ random effects model or OLS estimators (need to test the presence of random effects)

- The panel data regression model with unobserved heterogeneity (included in the error term; and u_i are random and called the **random effects**) is sometimes called the random effects model.

The random effects estimator

- There are two random errors in the panel data model we have been using
  - u_i , accounts for time invariant unobserved heterogeneity across individuals
  - e_{it}, is the "usual" regression error that varies across individuals and time

- Combining the two homoskedasticity assumptions and the statistical independence of u_i and e_{it}, we have: (15.26)

$$\text{var}(v_{it}) = \text{E}\left(v_{it}^2\right) = \sigma_v^2 = \sigma_u^2 + \sigma_e^2$$

The random effects estimator

- The minimum variance, efficient, estimator for the random effects model is a GLS estimator.
- The FGLS estimator is called the random effects estimator.

➢ A key feature of the random effects model is that time-invariant variables are not eliminated (different from the fixed effects model where time-invariant variables are eliminated).

➢ **EXAMPLE 15.9** Random Effects Estimation of a Production Function

➢ ➔ The cluster-robust standard errors for the random effects estimates are slightly larger than the conventional FGLS standard errors, suggesting that there may be serial correlation and/or heteroskedasticity in the overall error component e_{it}.

The random effects estimator

TABLE 15.5 **Example 15.10: Fixed and Random Effects Estimates of a Wage Equation**

	Fixed Effects			Random Effects		
Variable	**Coefficient**	**Std. Error***	***t*-Value**	**Coefficient**	**Std. Error***	***t*-Value**
C	1.4500	0.0401	36.12	0.5339	0.0799	6.68
EDUC				0.0733	0.0053	13.74
EXPER	0.0411	0.0066	6.21	0.0436	0.0064	6.86
$EXPER^2$	−0.0004	0.0003	−1.50	−0.0006	0.0003	−2.14
TENURE	0.0139	0.0033	4.24	0.0142	0.0032	4.47
$TENURE^2$	−0.0009	0.0002	−4.35	−0.0008	0.0002	−3.88
BLACK				−0.1167	0.0302	−3.86
SOUTH	−0.0163	0.0361	−0.45	−0.0818	0.0224	−3.65
UNION	0.0637	0.0143	4.47	0.0802	0.0132	6.07

*Conventional standard errors.

Testing for random effects

- We can test for the presence of heterogeneity by testing the null hypothesis $H_0: \sigma^2_u = 0$ against the alternative hypothesis $H_1: \sigma^2_u > 0$

- If the null hypothesis is rejected we conclude that there are random individual differences among sample members and that the **random effects model** might be appropriate

- if we fail to reject the null hypothesis, then we have no evidence to conclude that random effects are present

Testing for random effects

- The Lagrange multiplier (LM) principle for test construction is very convenient in this case

- If the null hypothesis is true, then $u_i = 0$ and the random effects model reduces to the usual linear regression model

- The test statistic is based on the OLS residuals

$$\hat{e}_{it} = y_{it} - b_1 - b_2 x_{2it} - a_1 w_{1i}$$

Testing for random effects

- We reject H_0 at significance level α and accept the alternative H_1: $\sigma^2_u > 0$:
- if **LM > $z_{(1-\alpha)}$**, where $z_{(1-\alpha)}$ is the 100(1–α) percentile of the standard normal distribution
- This critical value is 1.645 if α = 0.05 and 2.326 if α = 0.01

EXAMPLE 15.11 | Testing for Random Effects in a Production Function

Using the $N = 1000$ Chinese chemical firms data from *chemical3*, the value of the test statistic in (15.35) is $LM = 44.0637$. This is far greater than the $\alpha = 0.01$ critical value 2.326, so we reject the null hypothesis $H_0 : \sigma_u^2 = 0$ and conclude that $\sigma_u^2 > 0$; there is evidence of unobserved heterogeneity, or random effects, in the data.

A Hausman test for endogeneity in the random effects model

- The problem of endogenous regressors is common in random effects models because the individual-specific error component u_i may well be **correlated** with some of the explanatory variables
  - Such a correlation will cause the random effects estimator to be inconsistent
- To check for any correlation between the error component u_i and the regressors in a random effects model, we can use a **Hausman test**
- The test compares the coefficient estimates from the random effects model to those from the fixed effects model

Hausman test

EXAMPLE 15.13 | Testing for Endogenous Random Effects in a Wage Equation

Using the Hausman contrast test to compare the fixed and random effects estimates of the wage equation in Table 15.5 is limited to the six common coefficients. Using the individual coefficient t-tests you will find significant differences at the 5% level for the coefficients of $TENURE^2$, *SOUTH*, and *UNION*. The joint test for the equality of the common coefficients yields a χ^2-statistic value of 20.73 while $\chi^2_{(0.95,6)} = 12.592$. Thus both approaches lead us to conclude that there is correlation between the individual heterogeneity term and one or more of the explanatory variables and therefore the random effects estimator should not be used.

The Hausman-Taylor estimator

- The outcome from our comparison of the fixed and random effects estimates of the wage equation in Example 15.10 poses a dilemma. Correlation between the explanatory variables and the random effects means the random effects estimator will be inconsistent. We can overcome the inconsistency problem by using the fixed effects estimator, but doing so means we can no longer estimate the effects of the time-invariant variables *EDUC* and *BLACK*. The wage return for an extra year of education, and whether or not there is wage discrimination on the basis of race, might be two important questions that we would like to answer.

- The Hausman–Taylor estimator is an instrumental variables estimator applied to the random effects model, to overcome the problem of inconsistency caused by correlation between the random effects (the individual-specific error component u_i) and some of the explanatory variables.

Steps of analysis

- Test for any correlation between the error component u_i and the regressors in the regression model using **Hausman test**

- Rejecting the null ➔ choose fixed effects model ➔ test for joint significance of fixed effects (F test)

- Fail to reject the null ➔ choose random effects model ➔ test for the presence of random effects (LM test)