16

PANEL DATA REGRESSION MODELS

In Chapter 1 we discussed briefly the types of data that are generally available for empirical analysis, namely, **time series, cross section, and panel.** In time series data we observe the values of one or more variables over a period of time (e.g., GDP for several quarters or years). In cross-section data, values of one or more variables are collected for several sample units, or entities, at the same point in time (e.g., crime rates for 50 states in the United States for a given year). *In panel data the same cross-sectional unit (say a family or a firm or a state) is surveyed over time.* In short, panel data have *space as well as time dimensions.*

We have already seen an example of this in Table 1.1, which gives data on eggs produced and their prices for 50 states in the United States for years 1990 and 1991. For any given year, the data on eggs and their prices represent a cross-sectional sample. For any given state, there are two time series observations on eggs and their prices. Thus, we have in all $(50 \times 2) = 100$ (*pooled*) observations on eggs produced and their prices.

There are other names for panel data, such as **pooled data** (pooling of time series and cross-sectional observations), **combination of time series and cross-section data, micropanel data, longitudinal data** (a study over time of a variable or group of subjects), **event history analysis** (e.g., studying the movement over time of subjects through successive states or conditions), **cohort analysis** (e.g., following the career path of 1965 graduates of a business school). Although there are subtle variations, all these names *essentially connote movement over time of cross-sectional units*. We will therefore use the term panel data in a generic sense to include one or more of these terms. *And we will call regression models based on such data panel data regression models.*

Panel data are now being increasingly used in economic research. Some of the well-known panel data sets are:

1. The **Panel Study of Income Dynamics (PSID)** conducted by the Institute of Social Research at the University of Michigan. Started in 1968, each year the Institute collects data on some 5000 families about various socioeconomic and demographic variables.

2. The Bureau of the Census of the Department of Commerce conducts a survey similar to PSID, called the **Survey of Income and Program Participation (SIPP).** Four times a year, the respondents are interviewed about their economic condition.

There are also many other surveys that are conducted by various governmental agencies.

At the outset a warning is in order. The topic of panel data regressions is vast, and some of the mathematics and statistics involved is quite complicated. We only hope to touch on some of the essentials of the panel data regression models, leaving the details for the references.[1] But be forewarned that some of these references are highly technical. Fortunately, user-friendly software packages such as Limdep, PcGive, SAS, STATA, Shazam, and Eviews, among others, have made the task of actually implementing panel data regressions quite easy.

16.1 WHY PANEL DATA?

What are the advantages of panel data over cross-section or time series data? Baltagi lists the following advantages of panel data[2]:

1. Since panel data relate to individuals, firms, states, countries, etc., over time, there is bound to be heterogeneity in these units. The techniques of panel data estimation can take such heterogeneity explicitly into account by allowing for individual-specific variables, as we shall show shortly. We use the term *individual* in a generic sense to include microunits such as individuals, firms, states, and countries.

2. By combining time series of cross-section observations, panel data give "more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency."

[1] Some of the references are G. Chamberlain, "Panel Data," in *Handbook of Econometrics,* vol. II, Z. Griliches and M. D. Intriligator, eds., North-Holland Publishers, 1984, Chap. 22.; C. Hsiao, *Analysis of Panel Data,* Cambridge University Press, 1986; G. G. Judge, R. C. Hill, W. E. Griffiths, H. Lutkepohl, and T. C. Lee, *Introduction to the Theory and Practice of Econometrics,* 2d ed., John Wiley & Sons, New York, 1985, Chap. 11; W. H. Greene, *Econometric Analysis, 4th ed.,* Prentice-Hall, Englewood Cliffs, N.J., 2000, Chap. 14; Badi H. Baltagi, *Econometric Analysis of Panel Data,* John Wiley and Sons, New York, 1995; and J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data,* MIT Press, Cambridge, Mass., 1999.

[2] Baltagi, op. cit., pp. 3–6.

3. By studying the repeated cross section of observations, panel data are better suited to study the *dynamics of change.* Spells of unemployment, job turnover, and labor mobility are better studied with panel data.

4. Panel data can better detect and measure effects that simply cannot be observed in pure cross-section or pure time series data. For example, the effects of minimum wage laws on employment and earnings can be better studied if we include successive waves of minimum wage increases in the federal and/or state minimum wages.

5. Panel data enables us to study more complicated behavioral models. For example, phenomena such as economies of scale and technological change can be better handled by panel data than by pure cross-section or pure time series data.

6. By making data available for several thousand units, panel data can minimize the bias that might result if we aggregate individuals or firms into broad aggregates.

In short, panel data can enrich empirical analysis in ways that may not be possible if we use only cross-section or time series data. This is not to suggest that there are no problems with panel data modeling. We will discuss them after we cover some theory and discuss an example.

16.2 PANEL DATA: AN ILLUSTRATIVE EXAMPLE

To set the stage, let us consider a concrete example. Consider the data given in Table 16.1, which are taken from a famous study of investment theory proposed by Y. Grunfeld.[3]

Grunfeld was interested in finding out how real gross investment (Y) depends on the real value of the firm (X_2) and real capital stock (X_3). Although the original study covered several companies, for illustrative purposes we have obtained data on four companies, General Electric (GE), General Motor (GM), U.S. Steel (US), and Westinghouse. Data for each company on the preceding three variables are available for the period 1935–1954. Thus, there are four cross-sectional units and 20 time periods. In all, therefore, we have 80 observations. A priori, Y is expected to be positively related to X_2 and X_3.

In principle, we could run four time series regressions, one for each company or we could run 20 cross-sectional regressions, one for each year, although in the latter case we will have to worry about the degrees of freedom.[4]

[3] Y. Grunfeld, “The Determinants of Corporate Investment,” unpublished Ph.D. thesis, Department of Economics, University of Chicago, 1958. The data are reproduced in several books. We have taken them from H. D. Vinod and Aman Ullha, *Recent Advances in Regression Methods,* Marcel Dekker, New York, 1981, pp. 259–261. The Grunfeld study has become a favorite of textbook writers as the data is manageable for illustration purposes.

[4] For each year, we have only four observations on the regressand and the regressors. If we also allow for the intercept, we will have to estimate three parameters, leaving only a single degree of freedom. Obviously, such a regression may not be meaningful.

TABLE 16.1 INVESTMENT DATA FOR FOUR COMPANIES, 1935–1954

Observation	I	F_{-1}	C_{-1}	Observation	I	F_{-1}	C_{-1}
GE				US			
1935	33.1	1170.6	97.8	1935	209.9	1362.4	53.8
1936	45.0	2015.8	104.4	1936	355.3	1807.1	50.5
1937	77.2	2803.3	118.0	1937	469.9	2673.3	118.1
1938	44.6	2039.7	156.2	1938	262.3	1801.9	260.2
1939	48.1	2256.2	172.6	1939	230.4	1957.3	312.7
1940	74.4	2132.2	186.6	1940	361.6	2202.9	254.2
1941	113.0	1834.1	220.9	1941	472.8	2380.5	261.4
1942	91.9	1588.0	287.8	1942	445.6	2168.6	298.7
1943	61.3	1749.4	319.9	1943	361.6	1985.1	301.8
1944	56.8	1687.2	321.3	1944	288.2	1813.9	279.1
1945	93.6	2007.7	319.6	1945	258.7	1850.2	213.8
1946	159.9	2208.3	346.0	1946	420.3	2067.7	232.6
1947	147.2	1656.7	456.4	1947	420.5	1796.7	264.8
1948	146.3	1604.4	543.4	1948	494.5	1625.8	306.9
1949	98.3	1431.8	618.3	1949	405.1	1667.0	351.1
1950	93.5	1610.5	647.4	1950	418.8	1677.4	357.8
1951	135.2	1819.4	671.3	1951	588.2	2289.5	341.1
1952	157.3	2079.7	726.1	1952	645.2	2159.4	444.2
1953	179.5	2371.6	800.3	1953	641.0	2031.3	623.6
1954	189.6	2759.9	888.9	1954	459.3	2115.5	669.7
GM				WEST			
1935	317.6	3078.5	2.8	1935	12.93	191.5	1.8
1936	391.8	4661.7	52.6	1936	25.90	516.0	0.8
1937	410.6	5387.1	156.9	1937	35.05	729.0	7.4
1938	257.7	2792.2	209.2	1938	22.89	560.4	18.1
1939	330.8	4313.2	203.4	1939	18.84	519.9	23.5
1940	461.2	4643.9	207.2	1940	28.57	628.5	26.5
1941	512.0	4551.2	255.2	1941	48.51	537.1	36.2
1942	448.0	3244.1	303.7	1942	43.34	561.2	60.8
1943	499.6	4053.7	264.1	1943	37.02	617.2	84.4
1944	547.5	4379.3	201.6	1944	37.81	626.7	91.2
1945	561.2	4840.9	265.0	1945	39.27	737.2	92.4
1946	688.1	4900.0	402.2	1946	53.46	760.5	86.0
1947	568.9	3526.5	761.5	1947	55.56	581.4	111.1
1948	529.2	3245.7	922.4	1948	49.56	662.3	130.6
1949	555.1	3700.2	1020.1	1949	32.04	583.8	141.8
1950	642.9	3755.6	1099.0	1950	32.24	635.2	136.7
1951	755.9	4833.0	1207.7	1951	54.38	732.8	129.7
1952	891.2	4924.9	1430.5	1952	71.78	864.1	145.5
1953	1304.4	6241.7	1777.3	1953	90.08	1193.5	174.8
1954	1486.7	5593.6	2226.3	1954	68.60	1188.9	213.5

Notes: $Y = I$ = gross investment = additions to plant and equipment plus maintenance and repairs, in millions of dollars deflated by P_1

$X_2 = F$ = value of the firm = price of common and preferred shares at Dec. 31 (or average price of Dec. 31 and Jan. 31 of the following year) times number of common and preferred shares outstanding plus total book value of debt at Dec. 31, in millions of dollars deflated by P_2

$X_3 = C$ = stock of plant and equipment = accumulated sum of net additions to plant and equipment deflated by P_1 minus depreciation allowance deflated by P_3 in these definitions

P_1 = implicit price deflator of producers' durable equipment (1947 = 100)

P_2 = implicit price deflator of GNP (1947 = 100)

P_3 = depreciation expense deflator = 10-year moving average of wholesale price index of metals and metal products (1947 = 100)

Source: Reproduced from H. D. Vinod and Aman Ullah, *Recent Advances in Regression Methods,* Marcel Dekker, New York, 1981, pp. 259–261.

Pooling, or combining, all the 80 observations, we can write the Grunfeld investment function as:

$$
\begin{aligned}
Y_{it} &= \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \\
& i = 1, 2, 3, 4 \\
& t = 1, 2, \ldots, 20
\end{aligned}
\tag{16.2.1}
$$

where i stands for the ith cross-sectional unit and t for the tth time period. As a matter of convention, we will let i denote the cross-section identifier and t the time identifier. It is assumed that there are a maximum of N cross-sectional units or observations and a maximum of T time periods. If each cross-sectional unit has the same number of time series observations, then such a panel (data) is called a **balanced panel.** In the present example we have a balanced panel, as each company in the sample has 20 observations. If the number of observations differs among panel members, we call such a panel an **unbalanced panel.** In this chapter we will largely be concerned with a balanced panel.

Initially, we assume that the X's are nonstochastic and that the error term follows the classical assumptions, namely, $E(u_{it}) \sim N(0, \sigma^2)$.

Notice carefully the double and triple subscripted notation, which should be self-explanatory.

How do we estimate (16.2.1)? The answer follows.

16.3 ESTIMATION OF PANEL DATA REGRESSION MODELS: THE FIXED EFFECTS APPROACH

Estimation of (16.2.1) depends on the assumptions we make about the intercept, the slope coefficients, and the error term, u_{it}. There are several possibilities[5]:

1. Assume that the intercept and slope coefficients are constant across time and space and the error term captures differences over time and individuals.

2. The slope coefficients are constant but the intercept varies over individuals.

3. The slope coefficients are constant but the intercept varies over individuals and time.

4. All coefficients (the intercept as well as slope coefficients) vary over individuals.

5. The intercept as well as slope coefficients vary over individuals and time.

[5] This discussion is influenced by Judge et al., op. cit., and Hsiao, op. cit., pp. 9–10.

As you can see, each of these cases introduces increasing complexity (and perhaps more reality) in estimating panel data regression models, such as (16.2.1). Of course, the complexity will increase if we add more regressors to the model because of the possibility of collinearity among the regressors.

To cover each of the preceding categories in depth will require a separate book, and there are already several ones on the market.[6] In what follows, we will cover some of the main features of the various possibilities, especially the first four. Our discussion is nontechnical.

1. All Coefficients Constant across Time and Individuals

The simplest, and possibly naive, approach is to disregard the space and time dimensions of the pooled data and just estimate the usual OLS regression. That is, stack the 20 observations for each company one on top of the other, thus giving in all 80 observations for each of the variables in the model. The OLS results are as follows

$$\begin{aligned}
\hat{Y} &= -63.3041 \ + \ 0.1101X_2 + \ 0.3034X_3 \\
\text{se} &= \ (29.6124) \quad (0.0137) \quad\quad (0.0493) \\
t &= \ (-2.1376) \quad (8.0188) \quad\quad (6.1545) \qquad\qquad \textbf{(16.3.1)} \\
& R^2 = 0.7565 \qquad \text{Durbin–Watson} = 0.2187 \\
& \qquad\qquad\qquad n = 80 \qquad \text{df} = 77
\end{aligned}$$

If you examine the results of the **pooled regression,** and applying the conventional criteria, you will see that all the coefficients are individually statistically significant, the slope coefficients have the expected positive signs and the R^2 value is reasonably high. As expected, Y is positively related to X_2 and X_3. The "only" fly in the ointment is that the estimated Durbin–Watson statistic is quite low, suggesting that perhaps there is autocorrelation in the data. Of course, as we know, a low Durbin–Watson value could be due to specification errors also. For instance, the estimated model assumes that the intercept value of GE, GM, US, and Westinghouse are the same. It also assumes that the slope coefficients of the two X variables are all identical for all the four firms. Obviously, these are highly restricted assumptions. Therefore, despite its simplicity, the pooled regression (16.2.1) may distort the true picture of the relationship between Y and the X's across the four companies. What we need to do is find some way to take into account the specific nature of the four companies. How this can be done is explained next.

[6] Besides the books mentioned in footnote 1, see Terry E. Dielman, *Pooled Cross-sectional and Time Series Data Analysis,* Marcel Dekker, New York, 1989, and Lois W. Sayrs, *Pooled Time Series Analysis,* Sage Publications, Newbury Park, California, 1989.

2. Slope Coefficients Constant but the Intercept Varies across Individuals: The Fixed Effects or Least-Squares Dummy Variable (LSDV) Regression Model

One way to take into account the "individuality" of each company or each cross-sectional unit is to let the intercept vary for each company but still assume that the slope coefficients are constant across firms. To see this, we write model (16.2.1) as:

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \qquad \textbf{(16.3.2)}$$

Notice that we have put the subscript i on the intercept term to suggest that the intercepts of the four firms may be different; the differences may be due to special features of each company, such as managerial style or managerial philosophy.

In the literature, model (16.3.2) is known as the **fixed effects** (regression) model (**FEM**). The term "fixed effects" is due to the fact that, although the intercept may differ across individuals (here the four companies), each individual's intercept does not vary over time; that is, it is *time invariant*. Notice that if we were to write the intercept as β_{1it}, it will suggest that the intercept of each company or individual is *time variant*. It may be noted that the FEM given in (16.3.2) assumes that the (slope) coefficients of the regressors do not vary across individuals or over time.

How do we actually allow for the (fixed effect) intercept to vary between companies? We can easily do that by the dummy variable technique that we learned in Chapter 9, particularly, the **differential intercept dummies.** Therefore, we write (16.3.2) as:

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \qquad \textbf{(16.3.3)}$$

where $D_{2i} = 1$ if the observation belongs to GM, 0 otherwise; $D_{3i} = 1$ if the observation belongs to US, 0 otherwise; and $D_{4i} = 1$ if the observation belongs to WEST, 0 otherwise. Since we have four companies, we have used only three dummies to avoid falling into the **dummy-variable trap** (i.e., the situation of perfect collinearity). Here there is no dummy for GE. In other words, α_1 represents the intercept of GE and α_2, α_3, and α_4, the *differential intercept* coefficients, tell by how much the intercepts of GM, US, and WEST differ from the intercept of GE. In short, GE becomes the comparison company. Of course, you are free to choose any company as the comparison company.

Incidentally, if you want explicit intercept values for each company, you can introduce four dummy variables provided you run your regression through the origin, that is, drop the common intercept in (16.3.3); if you do not do this, you will fall into the dummy variable trap.

Since we are using dummies to estimate the fixed effects, in the literature the model (16.3.3) is also known as the **least-squares dummy variable (LSDV) model.** So, the terms fixed effects and LSDV can be used interchangeably. In passing, note that the LSDV model (16.3.3) is also known as the **covariance model** and X_2 and X_3 are known as *covariates*.

The results based on (16.3.3) are as follows:

$$\begin{array}{rccccccc}
\hat{Y}_{it} = & -245.7924 & +\,161.5722D_{2i} & +\,339.6328D_{3i} & +\,186.5666D_{3i} & +\,0.1079X_{2i} & +\,0.3461X_{3i} \\
\text{se} = & (35.8112) & (46.4563) & (23.9863) & (31.5068) & (0.0175) & (0.0266) \\
t = & (-6.8635) & (3.4779) & (14.1594) & (5.9214) & (6.1653) & (12.9821)
\end{array}$$

$$R^2 = 0.9345 \qquad d = 1.1076 \qquad \text{df} = 74 \tag{16.3.4}$$

Compare this regression with (16.3.1). In (16.3.4) all the estimated coefficients are individually highly significant, as the *p values* of the estimated *t* coefficients are extremely small. The intercept values of the four companies are statistically different; being −245.7924 for GE, −84.220 (= −245.7924 + 161.5722) for GM, 93.8774 (= −245.7924 + 339.6328) for US, and −59.2258 (= −245.7924 + 186.5666) for WEST. These differences in the intercepts may be due to unique features of each company, such as differences in management style or managerial talent.

Which model is better—(16.3.1) or (16.3.4)? The answer should be obvious, judged by the statistical significance of the estimated coefficients, and the fact that the R^2 value has increased substantially and the fact that the Durbin–Watson *d* value is much higher, suggesting that model (16.3.1) was mis-specified. The increased R^2 value, however, should not be surprising as we have more variables in model (16.3.4).

We can also provide a formal test of the two models. In relation to (16.3.4), model (16.3.1) is a restricted model in that it imposes a common intercept on all the companies. Therefore, we can use the **restricted *F* test** discussed in Chapter 8. Using formula (8.7.10), the reader can easily check that in the present instance the *F* value is:

$$F\,\frac{\left(R^2_{\text{UR}} - R^2_{R}\right)/3}{\left(1 - R^2_{\text{UR}}\right)/74} = \frac{(0.9345 - 0.7565)/3}{(1 - 0.9345)/74} = 66.9980 \tag{16.3.5}$$

where the restricted R^2 value is from (16.3.1) and the unrestricted R^2 is from (16.3.4) and where the number of restrictions is 3, since model (16.3.1) assumes that the intercepts of the GE, GM, US, and WEST are the same.

Clearly, the *F* value of 66.9980 (for 3 numerator df and 74 denominator df) is highly significant and, therefore, the restricted regression (16.3.1) seems to be invalid.

The Time Effect. Just as we used the dummy variables to account for individual (company) effect, we can allow for *time effect* in the sense that the Grunfeld investment function shifts over time because of factors such as technological changes, changes in government regulatory and/or tax policies, and external effects such as wars or other conflicts. Such time effects can be easily accounted for if we introduce time dummies, one for each year. Since we have data for 20 years, from 1935 to 1954, we can introduce 19 time dummies (why?), and write the model (16.3.3) as:

$$Y_{it} = \lambda_0 + \lambda_1 \text{Dum35} + \lambda_2 \text{Dum36} + \cdots + \lambda_{19} \text{Dum53} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \tag{16.3.6}$$

where Dum35 takes a value of 1 for observation in year 1935 and 0 otherwise, etc. We are treating the year 1954 as the base year, whose intercept value is given by λ_0 (why?)

We are not presenting the regression results based on (16.3.6), for none of the individual time dummies were individually statistically significant. The R^2 value of (16.3.6) was 0.7697, whereas that of (16.3.1) was 0.7565, an increment of only 0.0132. It is left as an exercise for the reader to show that, on the basis of the restricted F test, this increment is not significant, which probably suggests that the year or time effect is not significant. This might suggest that perhaps the investment function has not changed much over time.

We have already seen that the individual company effects were statistically significant, but the individual year effects were not. Could it be that our model is mis-specified in that we have not taken into account both individual and time effects together? Let us consider this possibility.

3. Slope Coefficients Constant but the Intercept Varies over Individuals As Well As Time

To consider this possibility, we can combine (16.3.4) and (16.3.6), as follows:

$$\begin{aligned} Y_{it} = \alpha_1 + \alpha_2 D_{\text{GM}_i} + \alpha_3 D_{\text{US}_i} + \alpha_4 D_{\text{WEST}_i} + \lambda_0 + \lambda_1 \text{Dum35} + \cdots \\ + \lambda_{19} \text{Dum53} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_{it} \end{aligned} \tag{16.3.7}$$

When we run this regression, we find the company dummies as well as the coefficients of the X are individually statistically significant, but none of the time dummies are. Essentially, we are back to (16.3.4).

The overall conclusion that emerges is that perhaps there is pronounced individual company effect but no time effect. In other words, the investment functions for the four companies are the same except for their intercepts. In all the cases we have considered, the X variables had a strong impact on Y.

4. All Coefficients Vary across Individuals

Here we assume that the intercepts and the slope coefficients are different for all individual, or cross-section, units. This is to say that the investment functions of GE, GM, US, and WEST are all different. We can easily extend our LSDV model to take care of this situation. Reconsider (16.3.4). There we introduced the individual dummies in an *additive* manner. But in Chapter 9 on dummy variables, we showed how *interactive,* or *differential, slope dummies,* can account for differences in slope coefficients. To do this in the context of the Grunfeld investment function, what we have to do is multiply each of the company dummies by each of the X variables [this will add six more variables to (16.3.4)]. That is, we estimate the following model:

$$\begin{aligned} Y_{it} = {} & \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \gamma_1 (D_{2i} X_{2it}) + \gamma_2 (D_{2i} X_{3it}) \\ & + \gamma_3 (D_{3i} X_{2it}) + \gamma_4 (D_{3i} X_{3it}) + \gamma_5 (D_{4i} X_{2it}) + \gamma_6 (D_{4i} X_{3it}) + u_{it} \qquad \textbf{(16.3.8)} \end{aligned}$$

You will notice that the γ's are the *differential slope coefficients,* just as α_2, α_3, and α_4 are the *differential intercepts*. If one or more of the γ coefficients are statistically significant, it will tell us that one or more slope coefficients are different from the base group. For example, say β_2 and γ_1 are statistically significant. In this case $(\beta_2 + \gamma_1)$ will give the value of the slope coefficient of X_2 for General Motors, suggesting that the GM slope coefficient of X_2 is different from that of General Electric, which is our comparison company.

If all the differential intercept and all the differential slope coefficients are statistically significant, we can conclude that the investment functions of General Motors, United States Steel, and Westinghouse are different from that of General Electric. If this is in fact the case, there may be little point in estimating the pooled regression (16.3.1).

Let us examine the regression results based on (16.3.8). For ease of reading, the regression results of (16.3.8) are given in tabular form in Table 16.2.

As these results reveal, Y is significantly related to X_2 and X_3. However, several differential slope coefficients are statistically significant. For instance, the slope coefficient of X_2 is 0.0902 for GE, but 0.1828 (0.0902 + 0.092) for GM. Interestingly, none of the differential intercepts are statistically significant.

TABLE 16.2 RESULTS OF REGRESSION (16.3.8)

Variable	Coefficient	Std. error	*t* value	*p* value
Intercept	-9.9563	76.3518	-0.1304	0.8966
D_{2i}	-139.5104	109.2808	-1.2766	0.2061
D_{3i}	-40.1217	129.2343	-0.3104	0.7572
D_{4i}	9.3759	93.1172	0.1006	0.9201
X_{2i}	0.0926	0.0424	2.1844	0.0324
X_{3i}	0.1516	0.0625	2.4250	0.0180
$D_{2i}X_{2i}$	0.0926	0.0424	2.1844	0.0324
$D_{2i}X_{3i}$	0.2198	0.0682	3.2190	0.0020
$D_{3i}X_{2i}$	0.1448	0.0646	2.2409	0.0283
$D_{3i}X_{3i}$	0.2570	0.1204	2.1333	0.0365
$D_{4i}X_{2i}$	0.0265	0.1114	0.2384	0.8122
$D_{4i}X_{3i}$	-0.0600	0.3785	-0.1584	0.8745

$R^2 = 0.9511 \quad d = 1.0896$

All in all, it seems that the investment functions of the four companies are different. This might suggest that the data of the four companies are not "poolable," in which case one can estimate the investment functions for each company separately. (See exercise 16.13.) This is a reminder that panel data regression models may not be appropriate in each situation, despite the availability of both time series and cross-sectional data.

A Caution on the Use of the Fixed Effects, or LSDV, Model. Although easy to use, the LSDV model has some problems that need to be borne in mind.

First, if you introduce too many dummy variables, as in the case of model (16.3.7), you will run up against the degrees of freedom problem. In the case of (16.3.7), we have 80 observations, but only 55 degrees of freedom—we lose 3 df for the three company dummies, 19 df for the 19 year dummies, 2 for the two slope coefficients, and 1 for the common intercept.

Second, with so many variables in the model, there is always the possibility of multicollinearity, which might make precise estimation of one or more parameters difficult.

Third, suppose in the FEM (16.3.1) we also include variables such as sex, color, and ethnicity, which are time invariant too because an individual's sex color, or ethnicity does not change over time. Hence, the LSDV approach may not be able to identify the impact of such time-invariant variables.

Fourth, we have to think carefully about the error term u_{it}. All the results we have presented so far are based on the assumption that the error term follows the classical assumptions, namely, $u_{it} \sim N(0, \sigma^2)$. Since the i index refers to cross-sectional observations and t to time series observations, the classical assumption for u_{it} may have to be modified. There are several possibilities.

1. We can assume that the error variance is the same for all cross-section units or we can assume that the error variance is heteroscedastic.

2. For each individual we can assume that there is no autocorrelation over time. Thus, for example, we can assume that the error term of the investment function for General Motors is nonautocorrelated. Or we could assume that it is autocorrelated, say, of the AR(1) type.

3. For a given time, it is possible that the error term for General Motors is correlated with the error term for, say, U.S. Steel or both U.S. Steel and Westinghouse.[7] Or, we could assume that there is no such correlation.

4. We can think of other permutations and combinations of the error term. As you will quickly realize, allowing for one or more of these possibilities will make the analysis that much more complicated. Space and mathematical demands preclude us from considering all the possibilities. A somewhat accessible discussion of the various possibilities can be found in

[7] This leads to the so-called **seemingly unrelated regression (SURE) modeling,** originally proposed by Arnold Zellner. For a discussion of this model, see Terry E. Dielman, op. cit.

Dielman, Sayrs, and Kmenta.[8] However, some of the problems *may* be alleviated if we resort to the so-called **random effects model,** which we discuss next.

16.4 ESTIMATION OF PANEL DATA REGRESSION MODELS: THE RANDOM EFFECTS APPROACH

Although straightforward to apply, fixed effects, or LSDV, modeling can be expensive in terms of degrees of freedom if we have several cross-sectional units. Besides, as Kmenta notes:

> An obvious question in connection with the covariance [i.e., LSDV] model is whether the inclusion of the dummy variables—and the consequent loss of the number of degrees of freedom—is really necessary. The reasoning underlying the covariance model is that in specifying the regression model we have failed to include relevant explanatory variables that do not change over time (and possibly others that do change over time but have the same value for all cross-sectional units), and that the inclusion of dummy variables is a *cover up of our ignorance* [emphasis added].[9]

If the dummy variables do in fact represent a lack of knowledge about the (true) model, why not express this ignorance through the disturbance term u_{it}? This is precisely the approach suggested by the proponents of the so-called **error components model (ECM) or random effects model (REM).**

The basic idea is to start with (16.3.2):

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \tag{16.4.1}$$

Instead of treating β_{1i} as fixed, we assume that it is a random variable with a mean value of β_1 (no subscript i here). And the intercept value for an individual company can be expressed as

$$\beta_{1i} = \beta_1 + \varepsilon_i \qquad i = 1, 2, \ldots, N \tag{16.4.2}$$

where ε_i is a random error term with a mean value of zero and variance of σ_ε^2.

What we are essentially saying is that the four firms included in our sample are a drawing from a much larger universe of such companies and that they have a common mean value for the intercept ($= \beta_1$) and the individual differences in the intercept values of each company are reflected in the error term ε_i.

Substituting (16.4.2) into (16.4.1), we obtain:

$$\begin{aligned} Y_{it} &= \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_i + u_{it} \\ &= \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + w_{it} \end{aligned} \tag{16.4.3}$$

[8] Dielman, op. cit., Sayrs, op. cit., Jan Kmenta, *Elements of Econometrics,* 2d ed., Macmillan, New York, 1986, Chap. 12.

[9] Kmenta, op. cit., p. 633.

where

$$w_{it} = \varepsilon_i + u_{it} \tag{16.4.4}$$

The composite error term w_{it} consists of two components, ε_i, which is the cross-section, or individual-specific, error component, and u_{it}, which is the combined time series and cross-section error component. The term *error components model* derives its name because the composite error term w_{it} consists of two (or more) error components.

The usual assumptions made by ECM are that

$$\begin{gathered}\varepsilon_i \sim N(0, \sigma_\varepsilon^2)\\ u_{it} \sim N(0, \sigma_u^2)\\ E(\varepsilon_i u_{it}) = 0 \qquad E(\varepsilon_i \varepsilon_j) = 0 \quad (i \neq j)\\ E(u_{it}u_{is}) = E(u_{it}u_{jt}) = E(u_{it}u_{js}) = 0 \quad (i \neq j; t \neq s).\end{gathered} \tag{16.4.5}$$

that is, the individual error components are not correlated with each other and are not autocorrelated across both cross-section and time series units.

Notice carefully the difference between FEM and ECM. In FEM each cross-sectional unit has its own (fixed) intercept value, in all N such values for N cross-sectional units. In ECM, on the other hand, the intercept β_1 represents the mean value of all the (cross-sectional) intercepts and the error component ε_i represents the (random) deviation of individual intercept from this mean value. However, keep in mind that ε_i is not directly observable; it is what is known as an **unobservable,** or **latent, variable.**

As a result of the assumptions stated in (16.4.5), it follows that

$$E(w_{it}) = 0 \tag{16.4.6}$$

$$\operatorname{var}(w_{it}) = \sigma_\varepsilon^2 + \sigma_u^2 \tag{16.4.7}$$

Now if $\sigma_\varepsilon^2 = 0$, there is no difference between models (16.2.1) and (16.4.3), in which case we can simply pool all the (cross-sectional and time series) observations and just run the pooled regression, as we did in (16.3.1).

As (16.4.7) shows, the error term w_{it} is homoscedastic. However, it can be shown that w_{it} and w_{is} $(t \neq s)$ are correlated; that is, the error terms of a given cross-sectional unit at two different points in time are correlated. The correlation coefficient, corr (w_{it}, w_{is}), is as follows:

$$\operatorname{corr}(w_{it}, w_{is}) = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_u^2} \tag{16.4.8}$$

Notice two special features of the preceding correlation coefficient. *First,* for any given cross-sectional unit, the value of the correlation between error terms at two different times remains the same no matter how far apart the two time periods are, as is clear from (16.4.8). This is in strong contrast to the first-order [AR(1)] scheme that we discussed in Chapter 12, where we found that the correlation between time periods declines over time. *Second,* the correlation structure given in (16.4.8) remains the same for all cross-sectional units; that is, it is identical for all individuals.

If we do not take this correlation structure into account, and estimate (16.4.3) by OLS, the resulting estimators will be inefficient. The most appropriate method here is the method of *generalized least squares* (GLS).

We will not discuss the mathematics of GLS in the present context because of its complexity.[10] Since most modern statistical software packages now have routines to estimate ECM (as well as FEM), we will only present the results for our investment example. But before we do that, it may be noted that we can easily extend (16.4.4) to allow for a random error component to take into account variation over time (see exercise 16.6).

The results of ECM estimation of the Grunfeld investment function are presented in Table 16.3. Several aspects of this regression should be noted. *First,* if you sum the random effect values given for the four companies, it will be zero, as it should (why?). *Second,* the mean value of the random error component, ε_i, is the common intercept value of −73.0353. The random effect value of GE of −169.9282 tells us by how much the random error component of GE differs from the common intercept value. Similar interpretation applies to the other three values of the random effects. *Third,* the R^2 value is obtained from the transformed GLS regression.

If you compare the results of the ECM model given in Table 16.3 with those obtained from FEM, you will see that generally the coefficient values of the two X variables do not seem to differ much, except for those given in Table 16.2, where we allowed the slope coefficients of the two variables to differ across cross-sectional units.

TABLE 16.3 ECM ESTIMATION OF THE GRUNFELD INVESTMENT FUNCTION

Variable	Coefficient	Std. error	*t* statistic	*p* value
Intercept	-73.0353	83.9495	-0.8699	0.3870
X_2	0.1076	0.0168	6.4016	0.0000
X_3	0.3457	0.0168	13.0235	0.0000
Random effect:				
GE	-169.9282			
GM	-9.5078			
USS	165.5613			
Westinghouse	13.87475			
	R^2 = 0.9323 (GLS)			

[10] The interested reader may refer to Kmenta, op. cit., pp. 625–630 for an accessible discussion.

16.5 FIXED EFFECTS (LSDV) VERSUS RANDOM EFFECTS MODEL

The challenge facing a researcher is: Which model is better, FEM or ECM? The answer to this question hinges around the assumption one makes about the likely correlation between the individual, or cross-section specific, error component ε_i and the X regressors.

If it is assumed that ε_i and the X's are *uncorrelated*, ECM may be appropriate, whereas if ε_i and the X's are *correlated,* FEM may be appropriate.

Why would one expect correlation between the individual error component ε_i and one or more regressors? Consider an example. Suppose we have a random sample of a large number of individuals and we want to model their wage, or earnings, function. Suppose earnings are a function of education, work experience, etc. Now if we let ε_i stand for innate ability, family background, etc., then when we model the earnings function including ε_i it is very likely to be correlated with education, for innate ability and family background are often crucial determinants of education. As Wooldridge contends, "In many applications, the whole reason for using panel data is to allow the unobserved effect [i.e., ε_i] to be correlated with the explanatory variables."[11]

The assumptions underlying ECM is that the ε_i are a random drawing from a much larger population. But sometimes this may not be so. For example, suppose we want to study the crime rate across the 50 states in the United States. Obviously, in this case, the assumption that the 50 states are a random sample is not tenable.

Keeping this fundamental difference in the two approaches in mind, what more can we say about the choice between FEM and ECM? Here the observations made by Judge et al. may be helpful[12]:

1. If T (the number of time series data) is large and N (the number of cross-sectional units) is small, there is likely to be little difference in the values of the parameters estimated by FEM and ECM. Hence the choice here is based on computational convenience. On this score, FEM may be preferable.

2. When N is large and T is small, the estimates obtained by the two methods can differ significantly. Recall that in ECM $\beta_{1i} = \beta_1 + \varepsilon_i$, where ε_i is the cross-sectional random component, whereas in FEM we treat β_{1i} as fixed and not random. In the latter case, statistical inference is conditional on the observed cross-sectional units in the sample. This is appropriate if we strongly believe that the individual, or cross-sectional, units in our sample are not random drawings from a larger sample. In that case, FEM is appropriate. However, if the cross-sectional units in the sample are regarded as random drawings, then ECM is appropriate, for in that case statistical inference is unconditional.

3. If the individual error component ε_i and one or more regressors are correlated, then the ECM estimators are biased, whereas those obtained from FEM are unbiased.

[11] Wooldridge, op. cit., p. 450.

[12] Judge et al., op. cit., pp. 489–491.

4. If N is large and T is small, and if the assumptions underlying ECM hold, ECM estimators are more efficient than FEM estimators.[13]

Is there a formal test that will help us to choose between FEM and ECM? Yes, a test was developed by Hausman in 1978.[14] We will not discuss the details of this test, for they are beyond the scope of this book.[15] The null hypothesis underlying the Hausman test is that the FEM and ECM estimators do not differ substantially. The test statistic developed by Hausman has an asymptotic χ^2 distribution. If the null hypothesis is rejected, the conclusion is that ECM is not appropriate and that we may be better off using FEM, in which case statistical inferences will be conditional on the ε_i in the sample.

Despite the Hausman test, it is important to keep in mind the warning sounded by Johnston and DiNardo. In deciding between fixed effects or random effects models, they argue that, " . . . there is no simple rule to help the researcher navigate past the Scylla of fixed effects and the Charybdis of measurement error and dynamic selection. Although they are an improvement over cross-section data, panel data do not provide a cure-all for all of an econometrician's problems."[16]

16.6 PANEL DATA REGRESSIONS: SOME CONCLUDING COMMENTS

As noted at the outset, the topic of panel data modeling is vast and complex. We have barely scratched the surface. Among the topics that we have not discussed, the following may be mentioned.

1. Hypothesis testing with panel data.
2. Heteroscedasticity and autocorrelation in ECM.
3. Unbalanced panel data.
4. Dynamic panel data models in which the lagged value(s) of the regressand (Y_{it}) appears as an explanatory variable.
5. Simultaneous equations involving panel data.
6. Qualitative dependent variables and panel data.

One or more of these topics can be found in the references cited in this chapter, and the reader is urged to consult them to learn more about this topic. These references also cite several empirical studies in various areas of business and economics that have used panel data regression models. The beginner is well advised to read some of these applications to get a feel about how researchers have actually implemented such models.

[13] Taylor has shown that for $T \geq 3$ and $(N - K) \geq 9$, where K is the number of regressors, the statement holds. See W. E. Taylor, "Small Sample Considerations in Estimation from Panel Data," *Journal of Econometrics,* vol. 13, 1980, pp. 203–223.

[14] J. A. Hausman, "Specification Tests in Econometrics," *Econometrica,* vol. 46, 1978, pp. 1251–1271.

[15] For the details, see Baltagi, op. cit., pp. 68–73.

[16] Jack Johnson and John DiNardo, *Econometric Methods,* 4th ed., McGraw-Hill, 1997, p. 403.