EXHIBIT 5

Basic Econometrics

Fifth Edition

Damodar N. Gujarati

Professor Emeritus of Economics,
United States Military Academy, West Point

Dawn C. Porter

University of Southern California

McGraw Hill

Boston Burr Ridge, IL Dubuque, IA Madison, WI New York San Francisco St. Louis
Bangkok Bogotá Caracas Kuala Lumpur Lisbon London Madrid Mexico City
Milan Montreal New Delhi Santiago Seoul Singapore Sydney Taipei Toronto

The McGraw·Hill Companies

BASIC ECONOMETRICS
International Edition 2009

Published by McGraw-Hill/Irwin, a business unit of The McGraw-Hill Companies, Inc. 1221 Avenue of the Americas, New York, NY 10020.

10 09 08 07 06 05 04 03 02
20 09
CTF BJE

When ordering this title, use ISBN: 978-007-127625-2 or MHID: 007-127625-4

Printed in Singapore

www.mhhe.com

Chapter 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 subjects, 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 100 (pooled) observations on eggs produced and their prices.

Another example of panel data was given in Table 1.2, which gives data on investment, value of the firm, and capital stock for four companies for the period 1935–1954. The data for each company over the period 1935–1954 constitute *time series data*, with 20 observations; data, for all four companies for a given year is an example of *cross-section data*, with only four observations; and data for all the companies for all the years is an example of *panel data*, with a total of 80 observations.

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** (studying the movement over time of subjects through successive states or conditions), and **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 used increasingly 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 5,000 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 respondents are interviewed about their economic condition.
3. The **German Socio-Economic Panel (GESOEP)** studied 1,761 individuals every year between 1984 and 2002. Information on year of birth, gender, life satisfaction, marital status, individual labor earnings, and annual hours of work was collected for each individual for the period 1984 to 2002.

There are also many other surveys that are conducted by various governmental agencies, such as:

Household, Income and Labor Dynamics in Australia Survey (HILDA)
British Household Panel Survey (BHPS)
Korean Labor and Income Panel Study (KLIPS)

At the outset a warning is in order: The topic of panel data regressions is vast, and some of the mathematics and statistics involved are 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, PC-GIVE, 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 subject-specific variables, as we shall show shortly. We use the term **subject** 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 gives "more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency."
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

[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, Chapter 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, Chapter 11; W. H. Greene, *Econometric Analysis,* 6th ed., Prentice-Hall, Englewood Cliffs, NJ, 2008, Chapter 9; 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. For a detailed treatment of the subject with empirical applications, see Edward W. Frees, *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences,* Cambridge University Press, New York, 2004.

[2] Baltagi, op. cit., pp. 3–6.

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 some examples.

16.2 Panel Data: An Illustrative Example

To set the stage, let us consider a concrete example. Consider the data given as Table 16.1 on the textbook website, which were originally collected by Professor Moshe Kim and are reproduced from William Greene.[3] The data analyzes the costs of six airline firms for the period 1970–1984, for a total of 90 panel data observations.

The variables are defined as: I = airline id; T = year id; Q = output, in revenue passenger miles, an index number; C = total cost, in \$1,000; PF = fuel price; and LF = load factor, the average capacity utilization of the fleet.

Suppose we are interested in finding out how total cost (C) behaves in relation to output (Q), fuel price (PF), and load factor (LF). In short, we wish to estimate an airline cost function.

How do we go about estimating this function? Of course, we can estimate the cost function for each airline using the data for 1970–1984 (i.e., a time series regression). This can be accomplished with the usual ordinary least squares (OLS) procedure. We will have in all six cost functions, one for each airline. But then we neglect the information about the other airlines which operate in the same (regulatory) environment.

We can also estimate a cross-section cost function (i.e., a cross-section regression). We will have in all 15 cross-section regressions, one for each year. But this would not make much sense in the present context, for we have only six observations per year and there are three explanatory variables (plus the intercept term); we will have very few degrees of freedom to do a meaningful analysis. Also, we will not "exploit" the panel nature of our data.

Incidentally, the panel data in our example is called a **balanced panel;** a panel is said to be balanced if each subject (firm, individuals, etc.) has the same number of observations. If each entity has a different number of observations, then we have an **unbalanced panel.** For most of this chapter, we will deal with balanced panels. In the panel data literature you will also come across the terms **short panel** and **long panel.** In a short panel the number of cross-sectional subjects, N, is greater than the number of time periods, T. In a long panel, it is T that is greater than N. As we discuss later, the estimating techniques can depend on whether we have a short panel or a long one.

What, then, are the options? There are four possibilities:

1. **Pooled OLS model.** We simply pool all 90 observations and estimate a "grand" regression, neglecting the cross-section and time series nature of our data.
2. The **fixed effects least squares dummy variable (LSDV) model.** Here we pool all 90 observations, but allow each cross-section unit (i.e., airline in our example) to have its own (intercept) dummy variable.

[3] William H. Greene, *Econometric Analysis,* 6th ed., 2008. Data are located at http://pages.stern.nyu.edu/~wgreen/Text/econometricanalysis.htm.

3. The **fixed effects within-group model.** Here also we pool all 90 observations, but for each airline we express each variable as a deviation from its mean value and then estimate an OLS regression on such *mean-corrected* or "de-meaned" values.
4. The **random effects model (REM).** Unlike the LSDV model, in which we allow each airline to have its own (fixed) intercept value, we assume that the intercept values are a random drawing from a much bigger population of airlines.

We now discuss each of these methods using the data given in Table 16.1. (See textbook website.)

16.3 Pooled OLS Regression or Constant Coefficients Model

Consider the following model:

$$C_{it} = \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it} \qquad \textbf{(16.3.1)}$$

$$i = 1, 2, \ldots, 6$$

$$t = 1, 2, \ldots, 15$$

where i is ith subject and t is the time period for the variables we defined previously. We have chosen the linear cost function for illustrative purposes, but in Exercise 16.10 you are asked to estimate a log linear, or double-log function, in which case the slope coefficients will give the elasticity estimates.

Notice that we have pooled together all 90 observations, but note that we are assuming the regression coefficients are the same for all the airlines. That is, there is no distinction between the airlines—one airline is as good as the other, an assumption that may be difficult to maintain.

It is assumed that the explanatory variables are nonstochastic. If they are stochastic, they are uncorrelated with the error term. Sometimes it is assumed that the explanatory variables are **strictly exogenous.** *A variable is said to be strictly exogenous if it does not depend on current, past, and future values of the error term* u_{it}.

It is also assumed that the error term is $u_{it} \sim iid(0, \sigma_u^2)$, that is, it is independently and identically distributed with zero mean and constant variance. For the purpose of hypothesis testing, it may be assumed that the error term is also normally distributed. Notice the double-subscripted notation in Eq. (16.3.1), which should be self-explanatory.

Let us first present the results of the estimated equation (16.3.1) and then discuss some of the problems with this model. The regression results based on *EViews*, Version 6 are presented in Table 16.2.

If you examine the results of the **pooled regression** and apply the conventional criteria, you will see that all the regression coefficients are not only highly statistically significant but are also in accord with prior expectations and that the R^2 value is very high. The only "fly in the ointment" is that the estimated Durbin–Watson statistic is quite low, suggesting that perhaps there is autocorrelation and/or spatial correlation in the data. Of course, as we know, a low Durbin–Watson could also be due to specification errors.

The major problem with this model is that it does not distinguish between the various airlines nor does it tell us whether the response of total cost to the explanatory variables over time is the same for all the airlines. In other words, by lumping together different airlines at different times we *camouflage* the **heterogeneity** (individuality or uniqueness) that may exist among the airlines. Another way of stating this is that the individuality of each subject is subsumed in the disturbance term u_{it}. As a consequence, it is quite possible that the error term may be correlated with some of the regressors included in the model. If that is the case, the estimated coefficients in Eq. (16.3.1) may be biased as well as inconsistent.

TABLE 16.2

```
Dependent Variable: C
Method: Least Squares
Included observations: 90

                     Coefficient   Std. Error   t Statistic    Prob.

C (intercept)        1158559.      360592.7      3.212930     0.0018
      Q              2026114.      61806.95      32.78134     0.0000
      PF             1.225348      0.103722      11.81380     0.0000
      LF            -3065753.      696327.3     -4.402747     0.0000

R-squared            0.946093    Mean dependent var.   1122524.
Adjusted R-squared   0.944213    S.D. dependent var.   1192075.
S.E. of regression   281559.5    F-statistic           503.1176
Sum squared resid.   6.82E+12    Prob. (F-statistic)   0.000000
                                 Durbin-Watson         0.434162
```

Recall that one of the important assumptions of the classical linear regression model is that there is no correlation between the regressors and the disturbance or error term.

To see how the error term may be correlated with the regressors, let us consider the following revision of model (16.3.1):

$$C_{it} = \beta_1 + \beta_2 PF_{it} + \beta_3 LF_{it} + \beta_4 M_{it} + u_{it} \qquad (16.3.2)$$

where the additional variable M = management philosophy or management quality. Of the variables included in Eq. (16.3.2), only the variable M is **time-invariant** (or **time-constant**) because it varies among subjects but is constant over time for a given subject (airline).

Although it is time-invariant, the variable M is not directly observable and therefore we cannot measure its contribution to the cost function. We can, however, do this indirectly if we write Eq. (16.3.2) as

$$C_{it} = \beta_1 + \beta_2 PF_{it} + \beta_3 LF_{it} + \alpha_i + u_{it} \qquad (16.3.3)$$

where α_i, called the **unobserved,** or **heterogeneity, effect,** reflects the impact of M on cost. Note that for simplicity we have shown only the unobserved effect of M on cost, but in reality there may be more such unobserved effects, for example, the nature of ownership (privately owned or publicly owned), whether it is a minority-owned company, whether the CEO is a man or a woman, etc. Although such variables may differ among the subjects (airlines), they will probably remain the same for any given subject over the sample period.

Since α_i is not directly observable, why not consider it random and include it in the error term u_{it}, and thereby consider the composite error term $v_{it} = \alpha_i + u_{it}$? We now write Eq. (16.3.3) as:

$$C_{it} = \beta_1 + \beta_2 PF_{it} + \beta_3 LF_{it} + v_{it} \qquad (16.3.4)$$

But if the α_i term included in the error term v_{it} is correlated with any of the regressors in Eq. (16.3.4), we have a violation of one of the key assumptions of the classical linear regression model—namely, that the error term is not correlated with the regressors. As we know in this situation, the OLS estimates are not only biased but they are also inconsistent.

There is a real possibility that the unobservable α_i is correlated with one or more of the regressors. For example, the management of one airline may be astute enough to buy future contracts of the fuel price to avoid severe price fluctuations. This will have the effect of lowering the cost of airline services. As a result of this correlation, it can be shown that $\operatorname{cov}(v_{it}, v_{is}) = \sigma_u^2$; $t \neq s$, which is non-zero, and therefore, the (unobserved) heterogeneity induces *autocorrelation* and we will have to pay attention to it. We will show later how this problem can be handled.

The question, therefore, is how we account for the unobservable, or heterogeneity, effect(s) so that we can obtain consistent and/or efficient estimates of the parameters of the variables of prime interest, which are output, fuel price, and load factor in our case. Our prime interest may not be in obtaining the impact of the unobservable variables because they remain the same for a given subject. That is why such unobservable, or heterogeneity, effects are called **nuisance parameters.** How then do we proceed? It is to this question we now turn.

16.4 The Fixed Effect Least-Squares Dummy Variable (LSDV) Model

The least-squares dummy variable (LSDV) model allows for heterogeneity among subjects by allowing each entity to have its own intercept value, as shown in model (16.4.1). Again, we continue with our airlines example.

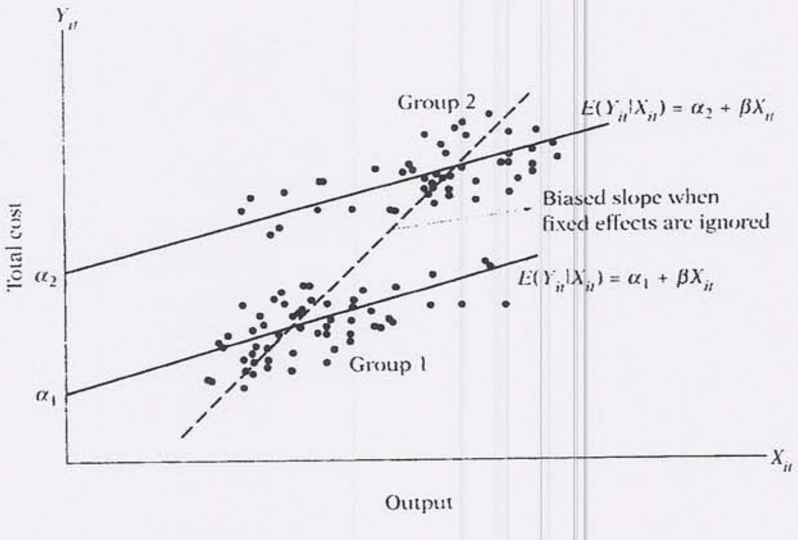
$$C_{it} = \beta_{1i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it} \qquad \textbf{(16.4.1)}$$
$$i = 1, 2, \ldots, 6$$
$$t = 1, 2, \ldots, 15$$

Notice that we have put the subscript i on the intercept term to suggest that the intercepts of the six airlines may be different. The difference may be due to special features of each airline, such as managerial style, managerial philosophy, or the type of market each airline is serving.

In the literature, model (16.4.1) 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 subjects (here the six airlines), each entity'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 would suggest that the intercept of each entity or individual is **time-variant.** It may be noted that the FEM given in Eq. (16.4.1) assumes that the (slope) coefficients of the regressors do not vary across individuals or over time.

Before proceeding further, it may be useful to visualize the difference between the pooled regression model and the LSDV model. For simplicity assume that we want to regress total cost on output only. In Figure 16.1 we show this cost function estimated for two airline companies separately, as well as the cost function if we pool the data for the two

FIGURE 16.1
Bias from ignoring fixed effects.



companies; this is equivalent to neglecting the fixed effects.[4] You can see from Figure 16.1 how the pooled regression can bias the slope estimate.

How do we actually allow for the (fixed effect) intercept to vary among the airlines? We can easily do this by using the dummy variable technique, particularly the **differential intercept dummy technique,** which we learned in Chapter 9. Now we write Eq. (16.4.1) as:

$$\begin{aligned} C_{it} &= \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \alpha_5 D_{5i} + \alpha_6 D_{6i} \\ &\quad + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it} \end{aligned} \tag{16.4.2}$$

where $D_{2i} = 1$ for airline 2, 0 otherwise; $D_{3i} = 1$ for airline 3, 0 otherwise; and so on. Notice that since we have six airlines, we have introduced only five dummy variables to avoid falling into the **dummy-variable trap** (i.e., the situation of perfect collinearity). Here we are treating airline 1 as the base, or reference, category. Of course, you can choose any airline as the reference point. As a result, the intercept α_1 is the intercept value of airline 1 and the other α coefficients represent by how much the intercept values of the other airlines differ from the intercept value of the first airline. Thus, α_2 tells by how much the intercept value of the second airline differs from α_1. The sum $(\alpha_1 + \alpha_2)$ gives the actual value of the intercept for airline 2. The intercept values of the other airlines can be computed similarly. *Keep in mind that if you want to introduce a dummy for each airline, you will have to drop the (common) intercept; otherwise, you will fall into the dummy-variable trap.*

The results of the model (16.4.2) for our data are presented in Table 16.3.

The first thing to notice about these results is that all the differential intercept coefficients are individually highly statistically significant, suggesting that perhaps the six airlines are heterogeneous and, therefore, the pooled regression results given in Table 16.2 may be suspect. The values of the slope coefficients given in Tables 16.2 and 16.3 are also different, again casting some doubt on the results given in Table 16.2. It seems model (16.4.1) is better than model (16.3.1). In passing, note that OLS applied to a fixed effect model produces estimators that are called **fixed effect estimators.**

TABLE 16.3

```
Dependent Variable: TC
Method: Least Squares
Sample: 1-90
Included observations: 90

              Coefficient      Std. Error      t Statistic      Prob.

C (=α1)       -131236.0        350777.1        -0.374129        0.7093
   Q           3319023.        171354.1         19.36939        0.0000
   PF          0.773071        0.097319         7.943676        0.0000
   LF         -3797368.        613773.1        -6.186924        0.0000
  DUM2         601733.2        100895.7         5.963913        0.0000
  DUM3         1337180.        186171.0         7.182538        0.0000
  DUM4         1777592.        213162.9         8.339126        0.0000
  DUM5         1828252.        231229.7         7.906651        0.0000
  DUM6         1706474.        228300.9         7.474672        0.0000

R-squared            0.971642      Mean dependent var.    1122524.
Adjusted R-squared   0.968841      S.D. dependent var.    1192075.
S.E. of regression   210422.8      F-statistics           346.9188
Sum squared resid.   3.59E+12      Prob. (F-statistic)    0.000000
Log likelihood      -1226.082      Durbin-Watson stat.    0.693288
```

[4] Adapted from the unpublished notes of Alan Duncan.

We can provide a formal test of the two models. In relation to model (16.4.1), model (16.3.1) is a *restricted* model in that it imposes a common intercept for all the airlines. Therefore, we can use the **restricted *F* test** discussed in Chapter 8. Using formula (8.6.10), the reader can check that in the present case the F value is:

$$F = \frac{(0.971642 - 0.946093)/5}{(1 - 0.971642)/81} \approx 14.99$$

Note: The restricted and unrestricted R^2 values are obtained from Tables 16.1 and 16.2. Also note that the number of restrictions is 5 (why?).

The null hypothesis here is that all the differential intercepts are equal to zero. The computed F value for 5 numerator and 81 denominator df is highly statistically significant. Therefore, we reject the null hypothesis that all the (differential) intercepts are zero. If the F value were not statistically significant, we would have concluded that there is no difference in the intercepts of the six airlines. In this case, we would have pooled all 90 of the observations, as we did in the pooled regression given in Table 16.2.

Model (16.4.1) is known as a **one-way fixed effects** model because we have allowed the intercepts to differ between airlines. But we can also allow for **time effect** if we believe that the cost function changes over time because of factors such as technological changes, changes in government regulation and/or tax policies, and other such effects. Such a time effect can be easily accounted for if we introduce time dummies, one for each year from 1970 to 1984. Since we have data for 15 years, we can introduce 14 time dummies (why?) and extend model (16.4.1) by adding these variables. If we do that, the model that emerges is called a **two-way fixed effects model** because we have allowed for both individual and time effects.

In the present example, if we add the time dummies, we will have in all 23 coefficients to estimate--the common intercept, five airlines dummies, 14 time dummies, and three slope coefficients. As you can see, we will consume several degrees of freedom. Furthermore, if we decide to allow the slope coefficients to differ among the companies, we can interact the five firm (airline) dummies with each of the three explanatory variables and introduce **differential slope dummy coefficients.** Then we will have to estimate 15 additional coefficients (five dummies interacted with three explanatory variables). As if this is not enough, if we interact the 14 time dummies with the three explanatory variables, we will have in all 42 additional coefficients to estimate. As you can see, we will not have any degrees of freedom left.

A Caution in the Use of the Fixed Effect LSDV Model

As the preceding discussion suggests, the LSDV model has several problems that need to be borne in mind:

First, if you introduce too many dummy variables, you will run up against the degrees of freedom problem. That is, you will lack enough observations to do a meaningful statistical analysis. *Second*, with many dummy variables in the model, both individual and interactive or multiplicative, there is always the possibility of multicollinearity, which might make precise estimation of one or more parameters difficult.

Third, in some situations the LSDV may not be able to identify the impact of time-invariant variables. Suppose we want to estimate a wage function for a group of workers using panel data. Besides wage, a wage function may include age, experience, and education as explanatory variables. Suppose we also decide to add sex, color, and ethnicity as additional variables in the model. Since these variables will not change over time for an individual subject, the LSDV approach may not be able to identify the impact of such time-invariant variables on wages. To put it differently, the subject-specific intercepts absorb all heterogeneity that may exist in the dependent and explanatory variables. Incidentally, the time-invariant variables are sometimes called **nuisance variables** or **lurking variables.**

Fourth, we have to think carefully about the error term u_{it}. The results we have presented in Eqs. (16.3.1) and (16.4.1) are based on the assumption that the error term follows the classical assumptions, namely, $u_{it} \sim N(0, \sigma^2)$. Since the index i refers to cross-section observations and t to time series observations, the classical assumption for u_{it} may have to be modified. There are several possibilities, including:

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.[5]
2. For each entity, we can assume that there is no autocorrelation over time. Thus, in our illustrative example, we can assume that the error term of the cost function for airline #1 is non-autocorrelated, or we can assume that it is autocorrelated, say, of the AR(1) type.
3. For a given time, it is possible that the error term for airline #1 is correlated with the error term for, say, airline #2.[6] Or we can assume that there is no such correlation.

There are also other combinations and permutations of the error term. As you will quickly realize, allowing 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. The references in footnote 1 discuss some of these topics.) Some of these problems may be alleviated, however, if we consider the alternatives discussed in the next two sections.

16.5 The Fixed-Effect Within-Group (WG) Estimator

One way to estimate a pooled regression is to eliminate the fixed effect, β_{1i}, by expressing the values of the dependent and explanatory variables for each airline as deviations from their respective mean values. Thus, for airline #1 we will obtain the sample mean values of TC, Q, PF, and LF, ($\overline{TC}$, $\overline{Q}$, $\overline{PF}$, and $\overline{LF}$, respectively) and subtract them from the individual values of these variables. The resulting values are called "de-meaned" or *mean-corrected* values. We do this for each airline and then pool all the (90) mean-corrected values and run an OLS regression.

Letting tc_{it}, q_{it}, pf_{it}, and lf_{it} represent the mean-corrected values, we now run the regression:

$$tc_{it} = \beta_2 q_{it} + \beta_3 pf_{it} + \beta_4 lf_{it} + u_{it} \tag{16.5.1}$$

where $i = 1, 2, \ldots, 6$, and $t = 1, 2, \ldots, 15$. Note that Eq. (16.5.1) *does not* have an intercept term (why?).

Returning to our example, we obtain the results in Table 16.4. *Note:* The prefix DM means that the values are mean-corrected or expressed as deviations from their sample means.

Note the difference between the pooled regression given in Table 16.2 and the pooled regression in Table 16.4. The former simply ignores the heterogeneity among the six airlines, whereas the latter takes it into account, not by the dummy variable method, but by eliminating it by differencing sample observations around their sample means. The difference between the two is obvious, as shown in Figure 16.2.

It can be shown that the WG estimator produces *consistent estimates* of the slope coefficients, whereas the ordinary pooled regression may not. It should be added, however,

[5] STATA provides heteroscedasticity-corrected standard errors in the panel data regression models.

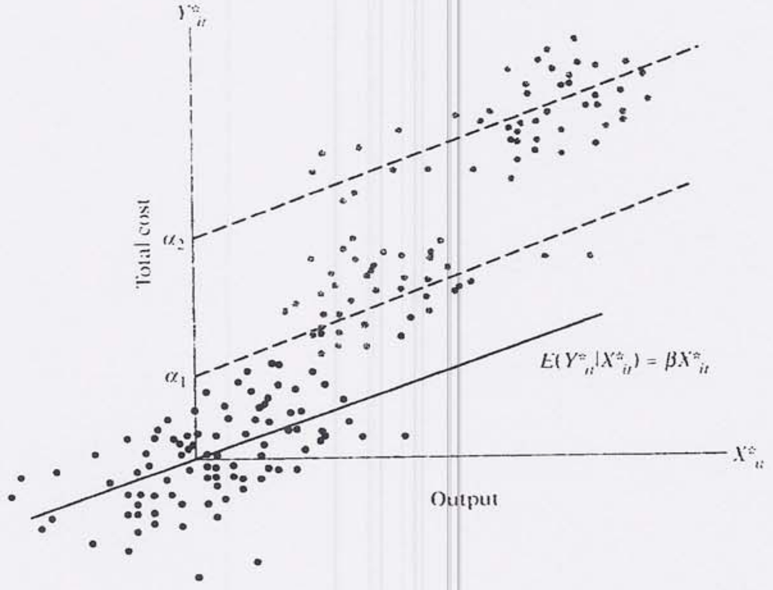
[6] This leads to the so-called **seemingly unrelated regression (SURE) model,** originally proposed by Arnold Zellner. See A. Zellner, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, vol. 57, 1962, pp. 348–368.

TABLE 16.4

```
Dependent Variable: DMTC
Method: Least Squares
Sample: 1-90
Included observations: 90

              Coefficient       Std. Error       t Statistic     Prob.

DMQ           3319023.          165339.8         20.07396        0.0000
DMPF          0.773071          0.093903         8.232630        0.0000
DMLF          -3797368.         592230.5         -6.411976       0.0000

R-squared            0.929366      Mean dependent var.   2.59E-11
Adjusted R-squared   0.927743      S.D. dependent var.   755325.8
S.E. of regression   203037.2      Durbin-Watson stat.   0.693287
Sum squared resid.   3.59E+12
```

FIGURE 16.2
The within-groups estimator.

Source: Alan Duncan, "Cross-Section and Panel Data Econometrics," unpublished lecture notes (adapted).



that WG estimators, although consistent, are inefficient (i.e., have larger variances) compared to the ordinary pooled regression results.[7] Observe that the slope coefficients of the *Q*, *PF*, and *LF* are identical in Tables 16.3 and 16.4. *This is because mathematically the two models are identical.* Incidentally, the regression coefficients estimated by the WG method are called *WG estimators*.

One disadvantage of the WG estimator can be explained with the following wage regression model:

$$W_{it} = \beta_{1i} + \beta_2\text{Experience}_{it} + \beta_3\text{Age}_{it} + \beta_4\text{Gender}_{it} + \beta_5\text{Education}_{it} + \beta_6\text{Race}_{it} \tag{16.5.2}$$

In this wage function, variables such as gender, education, and race are time-invariant. If we use the WG estimators, these time-invariant variables will be wiped out (because of

[7] The reason for this is that when we express variables as deviations from their mean values, the variation in these mean-corrected values will be much smaller than the variation in the original values of the variables. In that case, the variation in the disturbance term u_{it} may be relatively large, thus leading to higher standard errors of the estimated coefficients.

differencing). As a result, we will not know how wage reacts to these time-invariant variables.[8] But this is the price we have to pay to avoid the correlation between the error term (α_i included in v_{it}) and the explanatory variables.

Another disadvantage of the WG estimator is that, ". . . it may distort the parameter values and can certainly remove any long run effects."[9] *In general, when we difference a variable, we remove the long-run component from that variable.* What is left is the short-run value of that variable. We will discuss this further when we discuss time series econometrics later in the book.

In using LSDV we obtained direct estimates of the intercepts for each airline. How can we obtain the estimates of the intercepts using the WG method? For the airlines example, they are obtained as follows:

$$\hat{\alpha}_i = \overline{C}_i - \hat{\beta}_2\overline{Q_i} - \hat{\beta}_3\overline{PF_i} - \hat{\beta}_4\overline{LF} \qquad \textbf{(16.5.3)}$$

where bars over the variables denote the sample mean values of the variables for the ith airline.

That is, we obtain the intercept value of the ith airline by subtracting from the mean value of the dependent variable the mean values of the explanatory variables for that airline times the estimated slope coefficients from the WG estimators. Note that the estimated slope coefficients remain the same for all of the airlines, as shown in Table 16.4. It may be noted that the intercept estimated in Eq. (16.5.3) is similar to the intercept we estimate in the standard linear regression model, which can be see from Eq. (7.4.21). We leave it for the reader to find the intercepts of the six airlines in the manner shown and verify that they are the same as the intercept values derived in Table 16.3, save for the rounding errors.

It may be noted that the estimated intercept of each airline represents the *subject-specific* characteristics of each airline, but we will not be able to identify these characteristics individually. Thus, the α_1 intercept for airline #1 represents the management philosophy of that airline, the composition of its board of directors, the personality of the CEO, the gender of the CEO, etc. All these heterogeneity characteristics are subsumed in the intercept value. As we will see later, such characteristics can be included in the *random effects model*.

In passing, we note that an alternative to the WG estimator is the **first-difference method.** In the WG method, we express each variable as a deviation from that variable's mean value. In the first-difference method, for each subject we take successive differences of the variables. Thus, for airline #1 we subtract the first observation of TC from the second observation of TC, the second observation of TC from the third observation of TC, and so on. We do this for each of the remaining variables and repeat this process for the remaining five airlines. After this process we have only 14 observations for each airline, since the first observation has no previous value. As a result, we now have 84 observations instead of the original 90 observations. We then regress the first-differenced values of the TC variable on the first-differenced values of the explanatory variables as follows:

$$\begin{aligned} \Delta TC_{it} &= \beta_2 \Delta Q_{it} + \beta_3 \Delta PF_{it} + \beta_4 \Delta LF_{it} + (u_{it} - u_{i,t-1}) \\ i &= 1, 2, \ldots, 6 \\ t &= 1, 2, \ldots, 84 \end{aligned} \qquad \textbf{(16.5.4)}$$

where $\Delta = (TC_{it} - TC_{i,\,t-1})$. As noted in Chapter 11, Δ is called the first difference operator.[10]

[8] This is also true of the LSDV model.

[9] Dimitrios Asteriou and Stephen G. Hall, *Applied Econometrics: A Modern Approach,* Palgrave Macmillan, New York, 2007, p. 347.

[10] Notice that Eq. (16.5.3) has no intercept term (why?), but we can include it if there is a trend variable in the original model.

In passing, note that the original disturbance term is now replaced by the difference between the current and previous values of the disturbance term. If the original disturbance term is not autocorrelated, the transformed disturbance *is*, and therefore it poses the kinds of estimation problems that we discussed in Chapter 11. However, if the explanatory variables are **strictly exogenous,** the first difference estimator is unbiased, given the values of the explanatory variables. Also note that the first-difference method has the same disadvantages as the WG method in that the explanatory variables that remain fixed over time for an individual are wiped out in the first-difference transformation.

It may be pointed out that the first difference and fixed effects estimators are the same when we have only two time periods, but if there are more than two periods, these estimators differ. The reasons for this are rather involved and the interested reader may consult the references.[11] It is left as an exercise for the reader to apply the first difference method to our airlines example and compare the results with the other fixed effects estimators.

16.6 The Random Effects Model (REM)

Commenting on fixed effect, or LSDV, modeling, Kmenta writes:[12]

> 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 *coverup of our ignorance*.

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? This is precisely the approach suggested by the proponents of the so-called **error components model (ECM)** or **random effects model (REM),** which we will now illustrate with our airline cost function.

The basic idea is to start with Eq. (16.4.1):

$$TC_{it} = \beta_{1i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it} \qquad (16.6.1)$$

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). The intercept value for an individual company can be expressed as

$$\beta_{1i} = \beta_1 + \varepsilon_i \qquad (16.6.2)$$

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

What we are essentially saying is that the six 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 ($= \beta_1$). The individual differences in the intercept values of each company are reflected in the error term ε_i.

Substituting Eq. (16.6.2) into Eq. (16.6.1), we obtain:

$$\begin{aligned} TC_{it} &= \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + \varepsilon_i + u_{it} \\ &= \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + w_{it} \end{aligned} \qquad (16.6.3)$$

where

$$w_{it} = \varepsilon_i + u_{it} \qquad (16.6.4)$$

[11] See in particular Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, Mass., 2002, pp. 279–283.

[12] Jan Kmenta, *Elements of Econometrics*, 2d ed., Macmillan, New York, 1986, p. 633.

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