BINGHAM McCUTCHEN LLP
DONN P. PICKETT (SBN 72257)
GEOFFREY M. HOWARD (SBN 157468)
HOLLY A. HOUSE (SBN 136045)
ZACHARY J. ALINDER (SBN 209009)
BREE HANN (SBN 215695)
Three Embarcadero Center
San Francisco, CA 94111-4067
Telephone: (415) 393-2000
Facsimile: (415) 393-2286
donn.pickett@bingham.com
geoff.howard@bingham.com
holly.house@bingham.com
zachary.alinder@bingham.com
bree.hann@bingham.com

BOIES, SCHILLER & FLEXNER LLP
DAVID BOIES (Admitted *Pro Hac Vice*)
333 Main Street
Armonk, NY 10504
Telephone: (914) 749-8200
Facsimile: (914) 749-8300
dboies@bsfllp.com
STEVEN C. HOLTZMAN (SBN 144177)
FRED NORTON (SBN 224725)
1999 Harrison St., Suite 900
Oakland, CA 94612
Telephone: (510) 874-1000
Facsimile: (510) 874-1460
sholtzman@bsfllp.com
fnorton@bsfllp.com

DORIAN DALEY (SBN 129049)
JENNIFER GLOSS (SBN 154227)
500 Oracle Parkway, M/S 5op7
Redwood City, CA 94070
Telephone: (650) 506-4846
Facsimile: (650) 506-7114
dorian.daley@oracle.com
jennifer.gloss@oracle.com

Attorneys for Plaintiffs Oracle USA, Inc., *et al.*

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
OAKLAND DIVISION

ORACLE USA, INC., *et al.*,

Plaintiffs,

v.

SAP AG, *et al.*,

Defendants.

CASE NO. 07-CV-01658 PJH (EDL)

EXHIBIT 3 TO THE DECLARATION OF DANIEL S. LEVY, PH.D. IN SUPPORT OF MOTION NO. 1: TO EXCLUDE TESTIMONY OF DEFENDANTS' EXPERT STEPHEN CLARKE

FILED PURSUANT TO DKT. NO. 915

EXHIBIT 3

A GUIDE TO ECONOMETRICS

SIXTH EDITION

PETER KENNEDY
Simon Fraser University

Blackwell
Publishing

BLACKWELL PUBLISHING
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford OX4 2DQ, UK
550 Swanston Street, Carlton, Victoria 3053, Australia

First edition published 1979 by Martin Robertson and Company Ltd
Second and Third editions published by Blackwell Publishers Ltd
Fourth edition published 1998
Fifth edition published 2003
Sixth edition published 2008 by Blackwell Publishing Ltd

3 2009

Library of Congress Cataloging-in-Publication Data

Kennedy, Peter, 1943-
A guide to econometrics / Peter Kennedy. 6th ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4051-8258-4 (hardcover : alk. paper) ISBN 978-1-4051-8257-7 (pbk. : alk. paper)
1. Econometrics. I. Title.

HB139.K45 2008
330.015195—dc22
2007039113

A catalogue record for this title is available from the British Library.

Set in 10.5/12.5 pt Times
by Newgen Imaging Systems (P) Ltd, Chennai, India
Printed and bound in the United States of America
by Sheridan Books, Inc.

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

For further information on
Blackwell Publishing, visit our website:
www.blackwellpublishing.com

General Notes

2.2 Computational Cost

- Computational cost has been reduced significantly by the development of extensive computer software for econometricians. The more prominent of these are EVIEWS, GAUSS, LIMDEP, PC-GIVE, RATS, SAS, SHAZAM, SPSS, STATA, and TSP. For those wanting to code special estimation procedures themselves, this can be done using features of these software packages, or specialized software such as GAUSS, MATLAB, and OX. The *Journal of Applied Econometrics* and the *Journal of Economic Surveys* both publish software reviews regularly. All these packages are very comprehensive, encompassing most of the econometric techniques discussed in textbooks. For applications that they do not cover, in most cases, specialized programs exist. These packages should only be used by those well versed in econometric theory, however. Misleading or even erroneous results can easily be produced if these packages are used without a full understanding of the circumstances in which they are applicable, their inherent assumptions, and the nature of their output; sound research cannot be produced merely by feeding data to a computer and saying SHAZAM.
- The rapid drop in the cost of computer-intensive analysis has markedly changed econometrics. Now there is much more analysis using graphics, nonparametrics, simulation, bootstrapping, Monte Carlo, Bayesian statistics, and data exploration/mining, all discussed in later chapters.
- Problems with the accuracy of computer calculations are ignored in practice, but can be considerable, as discussed at length by McCullough and Vinod (1999). See also Aigner (1971, pp. 99–101) and Rhodes (1975).

2.3 Least Squares

- Experiments have shown that OLS estimates tend to correspond to the average of laymen's "freehand" attempts to fit a line to a scatter of data. See Mosteller *et al.* (1981).
- In Figure 2.1 the residuals were measured as the vertical distances from the observations to the estimated line. A natural alternative to this vertical measure is the orthogonal measure – the distance from the observation to the estimating line along a line perpendicular to the estimating line. This infrequently seen alternative is discussed in Malinvaud (1966, pp. 7–11); it is sometimes used when measurement errors plague the data, as discussed in section 10.2.

2.4 Highest R^2

- R^2 is called the coefficient of determination. It is the square of the correlation coefficient between y and its OLS estimate $\hat{y}$.
- The total variation of the dependent variable y about its mean, $\sum (y - \bar{y})^2$, is called SST (the total sum of squares); the "explained" variation, the sum of squared deviations of the estimated values of the dependent variable about their mean, $\sum (\hat{y} - \bar{\hat{y}})^2$ is called SSR (the regression sum of squares); and the "unexplained" variation, the sum of squared residuals, is called SSE (the error sum of squares). R^2 is then given by SSR/SST or by 1 – (SSE/SST).
- What is a high R^2? There is no generally accepted answer to this question. In dealing with time series data, very high R^2s are not unusual, because of common trends. Ames and Reiter (1961) found, for example, that on average the R^2 of a relationship between a randomly chosen variable and its own value lagged one period is about 0.7, and that an R^2 in excess of 0.5 could be obtained by selecting an economic time series and regressing it against two to six other randomly selected economic time series. For cross-sectional data, typical R^2s are not nearly so high. A more meaningful R^2 for time series data can be calculated by first removing the time trend by getting the residuals from regressing y on a time trend, and then regressing these residuals on the explanatory variables and a time trend. See Wooldridge (1991).
- The OLS estimator maximizes R^2. Since the R^2 measure is used as an index of how well an

estimator "fits" the sample data, the OLS estimator is often called the "best-fitting" estimator. A high R^2 is often called a "good fit."

- Because the R^2 and OLS criteria are formally identical, objections to the latter apply to the former. The most frequently voiced of these is that searching for a good fit is likely to generate parameter estimates tailored to the particular sample at hand rather than to the underlying "real world." Further, a high R^2 is not necessary for "good" estimates: R^2 could be low because of a high variance of the disturbance terms, and our estimate of β could be "good" on other criteria, such as those discussed in later sections of this chapter.
- The neat breakdown of the total variation into the "explained" and "unexplained" variations that allows meaningful interpretation of the R^2 statistic is valid only under three conditions. First, the estimator in question must be the OLS estimator. Second, the relationship being estimated must be linear. Thus the R^2 statistic only gives the percentage of the variation in the dependent variable explained *linearly* by variation in the independent variables. And third, the linear relationship being estimated must include a constant, or intercept, term. The formulas for R^2 can still be used to calculate an R^2 for estimators other than the OLS estimator, for nonlinear cases, and for cases in which the intercept term is omitted; it can no longer have the same meaning, however, and could possibly lie outside the 0–1 interval. The zero intercept case is discussed at length in Aigner (1971, pp. 85–90). An alternative R^2 measure, in which the variations in y and $\hat{y}$ are measured as deviations from zero rather than their means, is suggested.
- Running a regression without an intercept is the most common way of obtaining an R^2 outside the 0–1 range. To see how this could happen, draw a scatter of points in (x, y) space with an estimated OLS line such that there is a substantial intercept. Now draw in the OLS line that would be estimated if it were forced to go through the origin. In both cases SST is identical (because the same y observations are used). But in the second case the SSE and the SSR could be gigantic, because the $\hat{\varepsilon}$s and the $(\hat{y} - \bar{\hat{y}})$s could be huge. Thus if R^2 is calculated as 1 − SSE/SST, a negative number could result; if it is calculated as SSR/SST, a number greater than one could result.
- R^2 is sensitive to the range of variation of the dependent variable, so that comparisons of R^2s must be undertaken with care. The favorite example used to illustrate this is the case of the consumption function versus the savings function. If savings is defined as income less consumption, income will do exactly as well in explaining variations in consumption as in explaining variations in savings, in the sense that the sum of squared residuals, the unexplained variation, will be exactly the same for each case. But in *percentage* terms, the unexplained variation will be a higher percentage of the variation in savings than of the variation in consumption because the latter are larger numbers. Thus the R^2 in the savings function case will be lower than in the consumption function case.
- R^2 is also sensitive to the range of variation of the independent variable, basically because a wider range of the independent variables will cause a wider range of the dependent variable and so affect R^2 as described above. A consequence of this is that it makes no sense to compare R^2 across different samples – do not compare the R^2 for data from one country with the R^2 for data from another country, for example. Comparing estimates of the variance of the error term would make more sense.
- In general, econometricians are interested in obtaining "good" parameter estimates where "good" is not defined in terms of R^2. Consequently the measure R^2 is not of much importance in econometrics. Unfortunately, however, many practitioners act as though it is important, for reasons that are not entirely clear, as noted by Cramer (1987, p. 253):

> These measures of goodness of fit have a fatal attraction. Although it is generally conceded among insiders that they do not mean a thing, high values are still a source of pride and satisfaction to their authors, however hard they may try to conceal these feelings.

(c) *Simultaneous equation estimation* – situations in which the dependent variables are determined by the simultaneous interaction of several relationships.

5. The *fifth assumption* of the CLR model is that the number of observations is greater than the number of independent variables and that there are no exact linear relationships between the independent variables. Although this is viewed as an assumption for the general case, for a specific case it can easily be checked, so that it need not be assumed. The problem of *multicollinearity* (two or more independent variables being approximately linearly related in the sample data) is associated with this assumption. This is discussed in chapter 12.

All this is summarized in Table 3.1, which presents these five assumptions of the CLR model, shows the appearance they take when dressed in mathematical notation, and lists the econometric problems most closely associated with violations of these assumptions. Later chapters in this book comment on the meaning and significance of these assumptions, note implications of their violation for the OLS estimator, discuss ways of determining whether or not they are violated, and suggest new estimators appropriate to situations in which one of these assumptions must be replaced by an alternative assumption. Before we move on to this, however, more must be said about the character of the OLS estimator in the context of the CLR model, because of the central role it plays in the econometrician's "catalog."

Table 3.1 The assumptions of the CLR model.

Assumption	*Mathematical expression*		*Violations*	*Chapter in which discussed*
	Bivariate	*Multivariate*		
1. Dependent variable a linear function of a specific set of independent variables, plus a disturbance	$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, $t = 1, \ldots, N$	$Y = X\beta + \varepsilon$	Wrong regressors Nonlinearity Changing parameters	6
2. Expected value of disturbance term is zero	$E\varepsilon_t = 0$, for all t	$E\varepsilon = 0$	Biased intercept	7
3. Disturbances have uniform variance and are uncorrelated	$E\varepsilon_t\varepsilon_r = 0$, $t \neq r$ $= \sigma^2$, $t = r$	$E\varepsilon\varepsilon' = \sigma^2 I$	Heteroskedasticity Autocorrelated errors	8
4. Observations on independent variables can be considered fixed in repeated samples	x_t fixed in repeated samples	X fixed in repeated samples	Errors in variables Autoregression Simultaneous equations	10 11
5. No exact linear relationships between independent variables and more observations than independent variables	$\sum_{t=1}^{N}(x_t - \bar{x})^2 \neq 0$	Rank of $X = K \leq N$	Perfect multicollinearity	12

The mathematical terminology is explained in the technical notes to this section. The notation is as follows: Y is a vector of observations on the dependent variable; X is a matrix of observations on the independent variables; ε is a vector of disturbances; σ^2 is the variance of the disturbances; I is the identity matrix; K is the number of independent variables; N is the number of observations.

the degrees of freedom of the sample variance statistic is $(N - 1)$.

- A special case of the F statistic is automatically reported by most regression packages – the F statistic for the "overall significance of the regression." This F statistic tests the hypothesis that all the slope coefficients are zero. The constrained regression in this case would have only an intercept.
- To clarify further how one runs a constrained regression, suppose, for example, that $y = \alpha + \beta x + \delta w - \varepsilon$ and we wish to impose the constraint that $\beta + \delta = 1$. Substitute $\beta = 1 - \delta$ and rearrange to get $y - x = \alpha + \delta(w - x) + \varepsilon$. The restricted SSE is obtained from regressing the constructed variable $(y - x)$ on a constant and the constructed variable $(w - x)$. Note that because the dependent variable has changed R^2 will not be meaningful to compare the R^2 of this regression with that of the original regression.
- In the preceding example it should be clear that it is easy to construct an F test of the hypothesis that $\beta + \delta = 1$. The resulting F statistic will be the square of the t statistic that could be used to test this same hypothesis (described in the preceding section, involving a messy computation of the required standard error). This reflects the general result that the square of a t statistic is an F statistic (with numerator degrees of freedom equal to one and denominator degrees of freedom equal to the t test degrees of freedom). With the exception of testing a single coefficient equal to a specific value, it is usually easier to perform an F test than a t test. Note that the square root of an F statistic is not equal to a t statistic unless the degrees of freedom of the numerator is one.
- By dividing the numerator and denominator of the F statistic by SST (total sum of squares), the total variation in the dependent variable Y can be written in terms of R^2 and $\overline{R}^2$. This method is not recommended, however, because often the restricted SSE is obtained by running a regression with a different dependent variable than that used by the regression run to obtain the unrestricted SSE (as in the example above), implying different SSTs and incomparable R^2s.
- An F statistic with p and n degrees of freedom is the ratio of two independent chi-square statistics, each divided by its degrees of freedom, p for the numerator and n for the denominator. For the standard F statistic that we have been discussing, the chi-square on the denominator is SSE, the sum of squared OLS residuals, with degrees of freedom $T - K$, divided by σ^2. Asymptotically, $SSE/(T - K)$ equals σ^2, so the denominator becomes unity, leaving F equal to the numerator chi-square divided by its degrees of freedom p. Thus, asymptotically pF is distributed as a chi-square with degrees of freedom p. This explains why test statistics derived on asymptotic arguments are invariably expressed as chi-square statistics rather than as F statistics. In small samples it cannot be said that this approach, calculating the chi-square statistic and using critical values from the chi-square distribution, is definitely preferred to calculating the F statistic and using critical values from the F distribution; the choice of chi-square statistic here is an econometric ritual.
- One application of the F test is in testing for causality. It is usually assumed that movements in the dependent variable are caused by movements in the independent variable(s), but the existence of a relationship between these variables proves neither the existence of causality nor its direction. Using the dictionary meaning of causality, it is impossible to test for causality. Granger developed a special definition of causality which econometricians use in place of the dictionary definition; strictly speaking, econometricians should say "Granger-cause" in place of "cause," but usually they do not. A variable x is said to Granger-cause y if prediction of the current value of y is enhanced by using past values of x. This definition is implemented for empirical testing by regressing y on past, current, and future values of x; if causality runs one way, from x to y, the set of coefficients of the future values of x should test insignificantly different from the zero vector (via an F test), and the set of coefficients of the past values of x should test significantly different from zero. Before running this regression both data sets are transformed (using the same transformation), so as to eliminate any autocorrelation

7. *Tests for exogeneity* These tests, often referred to as Hausman tests, test for contemporaneous correlation between regressors and the error. They are discussed in chapter 9.
8. *Data transformation tests* These tests, which do not have any specific alternative hypothesis, are considered variants of the Hausman test. Examples are the grouping test and the differencing test, discussed later in this chapter.
9. *Non-nested tests* When testing rival models that are not nested, as might arise when testing for encompassing, non-nested tests must be employed. Examples are the non-nested F test and the J test, discussed later in this chapter.
10. *Conditional moment tests* These tests are based on a very general testing methodology which in special cases gives rise to most of the tests listed above. Beyond serving as a unifying framework for existing tests, the value of this testing methodology is that it suggests how specification tests can be undertaken in circumstances in which alternative tests are difficult to construct. More discussion is provided later in this chapter.

Categorizing tests in this way is awkward, for several reasons.

1. Such a list will inevitably be incomplete. For example, it could be expanded to incorporate tests for specification encountered in more advanced work. Should there be categories for unit root and cointegration tests (see chapter 19), identification tests (see chapter 11), and selection bias tests (see chapter 17), for example? What about Bayesian "tests"?
2. It is common for practitioners to use a selection criterion, such as the Akaike information criterion, or adjusted R^2, to aid in model specification, particularly for determining things like the number of lags to include. Should this methodology be classified as a test?
3. These categories are not mutually exclusive. There are non-nested variants of tests for nonspherical errors and of functional form tests, some tests for functional form are just variants of tests for structural break, and the RESET is a special case of an OV test, for example.
4. Tests take different forms. Some are Lagrange multiplier (LM) tests, some are Likelihood ratio (LR) tests, and some are Wald (W) tests. Some use F tables, some use t tables, some use χ^2 tables, and some require their own special tables. Some are exact tests, whereas some rest on an asymptotic justification.
5. Some tests are "specification" tests, involving a specific alternative, whereas others are "misspecification" tests, with no specific alternative.

This last distinction is particularly relevant for this chapter. A prominent feature of the list of general principles given earlier is the use of misspecification tests, the more common of which are often referred to as diagnostics. These tests are designed to detect an inadequate specification (as opposed to "specification" tests, which examine the validity of a specific alternative). There have been calls for researchers to submit their models to misspecification tests as a matter of course, and it is becoming common for econometric software packages automatically to print out selected diagnostics.

Chapter 7
Violating Assumption Two: Nonzero Expected Disturbance

The second assumption of the classical linear regression (CLR) model states that the population from which the disturbance or error term is drawn has mean zero. Violation of this assumption may or may not be of concern, depending on specific circumstances.

Constant Nonzero Mean

The disturbance may have a nonzero mean because of systematically positive or systematically negative errors of measurement in calculating the dependent variable. This problem is most easily analyzed if the estimating equation is rearranged by removing the nonzero mean from the error term and adding it to the intercept term. This creates an estimating equation obeying all the CLR model assumptions; in particular, the mean of the new error term is zero. The only problem is that ordinary least squares (OLS) estimation gives an unbiased estimate of the *new* intercept, which is the sum of the original intercept and the mean of the original error term; it is therefore a *biased* estimate of the original intercept (the bias being exactly equal to the mean of the original error term). Thus the only implication of this violation of the second assumption of the CLR model is that the OLS estimate of the intercept is biased; the slope coefficient estimates are unaffected. This biased estimate is often welcomed by the econometrician since, for prediction purposes, he or she would want to incorporate the mean of the error term into the prediction.

Zero Intercept

Sometimes economic theory suggests that the intercept in a regression is zero. An example is transforming for heteroskedasticity (discussed in chapter 8) resulting in a regression on transformed variables without an intercept. Practitioners usually include an intercept, however. Why? It is possible that a relevant explanatory variable was omitted,

creating bias. This bias can be alleviated (but not eliminated) by including an intercept term; no bias is created by including an unnecessary intercept.

Limited Dependent Variable

When the nonzero expected value of the error term is not constant, problems can arise. Consider, for example, the case of a limited dependent variable, discussed in chapter 17. Suppose an observation is included in the sample only if the dependent variable y is less than K. For example, data may have been gathered only on people whose income fell below some poverty level K. This means that the data will not contain errors large enough to cause the dependent variable to be greater than K. Thus, in this example the right-hand tail of the distribution of the error terms is chopped off (the error comes from a "truncated" distribution), implying that the expected value of the error term is negative, rather than zero. But this negative expected value of the error term is not the same for all observations. People with characteristics such that their expected y values are greater than K cannot have positive errors – they are only included in the sample if their error terms are sufficiently negative, so for these observations the expected value of the error is a relatively high negative number. On the other hand, people whose characteristics are such that their expected y values are well below K will be included in the sample if their error terms are negative or positive numbers, excepting only very high positive errors, so for these observations the expected value of the error term is a low negative number.

This suggests that the expected value of the error term varies from observation to observation, and in a way that is affected by the values of the explanatory variables (characteristics of the individuals). The impact of this on the OLS estimator can be deduced by viewing the expected value of the error term as an omitted explanatory variable, discussed in chapter 6. Since this "omitted variable" is correlated with the other explanatory variables, the OLS estimator for all coefficients, not just the intercept, is biased.

Frontier Production Function

In economic theory, a frontier production function determines the maximum output that can be produced with given inputs. Firms could be less than fully efficient and thus produce inside the production frontier, but they cannot produce more than the output given by this frontier. This suggests that the error should be negative, or at best zero, causing its expected value to be negative.

Econometricians model this by specifying two error terms. The first of these error terms is a traditional error (with both positive and negative values) reflecting errors in measuring output or factors over which the firm has no control such as weather. When added to the frontier production function formula, it creates a stochastic frontier production function, saying in effect that not all observations have exactly the same frontier production function. The second error is a non-positive error reflecting the degree to which a firm is inside its stochastic frontier. The two errors together form a

category deviates from some base (the "omitted" category). Whenever there exist more than two categories, the presentation of these results can be awkward, especially when laymen are involved; a more relevant, easily understood base might make the presentation of these results more effective. For example, suppose household energy consumption is determined by income and the region in which the household lives. Rather than, say, using the South as a base and comparing household energy consumption in the North East, North Central, and West to consumption in the South, it may be more effective, as a means of presenting these results to laymen, to calculate dummy variable coefficients in such a way as to compare consumption in each region with the national average. A simple adjustment permits this. See Suits (1984) and Kennedy (1986).

- Goodman and Dubin (1990) note that alternative specifications containing different dummy variable specifications may not be nested, implying that a non-nested testing procedure should be employed to analyze their relative merits.

15.4 Interacting with Quantitative Variables

- Dummy variables play an important role in structuring Chow tests for testing if there has been a change in a parameter value from one data set to another. Suppose Y is a linear function of X and Z and the question at hand is whether the coefficients are the same in period 1 as in period 2. A dummy variable D is formed such that D takes the value 0 for observations in period 1 and the value 1 for observations in period 2. "Product" dummy variables DX and DZ are also formed (i.e., DX takes the value X in period 2 and is 0 otherwise). Then the equation

$$Y = \beta_0 + \alpha_0 D + \beta_1 X + \alpha_1 (DX) + \beta_2 Z + \alpha_2 (DZ) + \varepsilon \qquad (15.9)$$

is formed.

Running regression (1) as is allows the intercept and slope coefficients to differ from period 1 to period 2. This produces SSE unrestricted. Running regression (1) forcing α_0, α_1, and α_2 to be 0 forces the intercept and slope coefficients to be identical in both periods. An F test, structured in the usual way, can be used to test whether or not the vector with elements α_0, α_1, and α_2 is equal to the zero vector. The resulting F statistic is

$$\frac{\left[\text{SSE(constrained)} - \text{SSE(unconstrained)}\right] K}{\text{SSE(unconstrained)}\ (N_1 + N_2 - 2K)}$$

where K is the number of parameters, N_1 is the number of observations in the first period and N_2 is the number of observations in the second period. If there were more than two periods and we wished to test for equality across all periods, this methodology can be generalized by adding extra dummies in the obvious way.

Whenever the entire set of parameters is being tested for equality between two data sets the SSE unconstrained can be obtained by summing the SSEs from the two separate regressions and the SSE constrained can be obtained from a single regression using all the data; the Chow test often appears in textbooks in this guise. In general, including dummy variables to allow the intercept and all slopes to differ between two data sets produces the same coefficient estimates as those obtained by running separate regressions, but estimated variances differ because the former method constrains the estimated variance to be the same in both equations.

- The advantage of the dummy variable variant of the Chow test is that it can easily be modified to test subsets of the coefficients. Suppose, for example, that it is known that, in equation (15.9) above, β_2 changed from period 1 to period 2 and that it is desired to test whether or not the other parameters (β_0 and β_1) changed. Running regression (1) as is gives the unrestricted SSE for the required F statistic, and running (1) without D and DX gives the restricted SSE. The required degrees of freedom are 2 for the numerator and $N - 6$ for the denominator, where N is the total number of observations.

Notice that a slightly different form of this test must be used if, instead of knowing

Chapter 18
Panel Data

18.1 Introduction

Modern econometrics is divided into two branches: microeconometrics and time series analysis. The latter is covered in chapter 19. The former has many elements, of which we have discussed several examples, such as qualitative dependent variables, duration models, count data, and limited dependent variables, all of which primarily involve different types of cross-sectional data. In light of this it would seem natural to call microeconometrics cross-sectional data analysis. We do not, however, because a major category of microeconometrics involves longitudinal or panel data in which a cross-section (of people, firms, countries, etc.) is observed over time. Thanks to the computer revolution, such data sets, in which we have observations on the same units in several different time periods, are more common and have become more amenable to analysis.

Two prominent examples of panel data are the PSID (Panel Study of Income Dynamics) data and the NLS (National Longitudinal Surveys of Labor Market Experience) data, both of which were obtained by interviewing several thousand people over and over again through time. These data sets were designed to enable examination of the causes and nature of poverty in the United States, by collecting information on such things as employment, earnings, mobility, housing, and consumption behavior. Indeed, thousands of variables were recorded. These data are typical of panel data in that they are short and wide, consisting of a very large number of cross-sectional units observed over a small number of time periods. Such data are expensive to obtain, involving tracking large numbers of people over extended time periods. Is this extra expense warranted?

Panel data have several attractive features that justify this extra cost, four of which are noted below.

1. Panel data can be used to deal with heterogeneity in the micro units. In any cross-section there is a myriad of unmeasured explanatory variables that affect

the behavior of the people (firms, countries, etc.) being analyzed. (Heterogeneity means that these micro units are all different from one another in fundamental unmeasured ways.) Omitting these variables causes bias in estimation. The same holds true for omitted time series variables that influence the behavior of the micro units uniformly, but differently in each time period. Panel data enable correction of this problem. Indeed, some would claim that the ability to deal with this omitted variable problem is the main attribute of panel data.

2. Panel data create more variability, through combining variation across micro units with variation over time, alleviating multicollinearity problems. With this more informative data, more efficient estimation is possible.
3. Panel data can be used to examine issues that cannot be studied using time series or cross-sectional data alone. As an example, consider the problem of separating economies of scale from technological change in the analysis of production functions. Cross-sectional data can be used to examine economies of scale, by comparing the costs of small and large firms, but because all the data come from one time period there is no way to estimate the effect of technological change. Things are worse with time series data on a single firm: we cannot separate the two effects because we cannot tell if a change in that firm's costs over time is due to technological change or due to a change in the size of the firm. As a second example, consider the distinction between temporary and long-term unemployment. Cross-sectional data tell us who is unemployed in a single year, and time series data tell us how the unemployment level changed from year to year. But neither can tell us if the same people are unemployed from year to year, implying a low turnover rate, or if different people are unemployed from year to year, implying a high turnover rate. Analysis using panel data can address the turnover question because these data track a common sample of people over several years.
4. Panel data allow better analysis of dynamic adjustment. Cross-sectional data can tell us nothing about dynamics. Time series data need to be very lengthy to provide good estimates of dynamic behavior, and then typically relate to aggregate dynamic behavior. Knowledge of individual dynamic reactions can be crucial to understanding economic phenomena. Panel data avoid the need for a lengthy time series by exploiting information on the dynamic reactions of each of several individuals.

18.2 Allowing for Different Intercepts

Suppose an individual's consumption y is determined linearly by his or her income x and we have observations on a thousand individuals ($N = 1000$) in each of four time periods ($T = 4$). A plot of all the data produces a scatter shown in simplified form (only a few observations are shown, not all 4000 observations!) in Figure 18.1. (Ignore the ellipses for the moment.) If we were to run ordinary least squares (OLS), we would produce a slope estimate shown by the line AA drawn through these data. But now suppose we identify these data by the cross-sectional unit (person, firm, or country, for example) to which they belong, in this case a person. This is shown in Figure 18.1 by drawing an ellipse for each person, surrounding all four time series observations

Figure 18.1

on that perso
with roughly
way of view
same slope,
that this cros
unmeasured
intercept for
influence of
with the inch
suggested, a
intercept for
The first
Doing this a
all these dun
model gives
effects mod
because (in
that a compu
This transfo
values withi
them the ave
data produc
The fixed

1. By impli
dom (by
some w
common

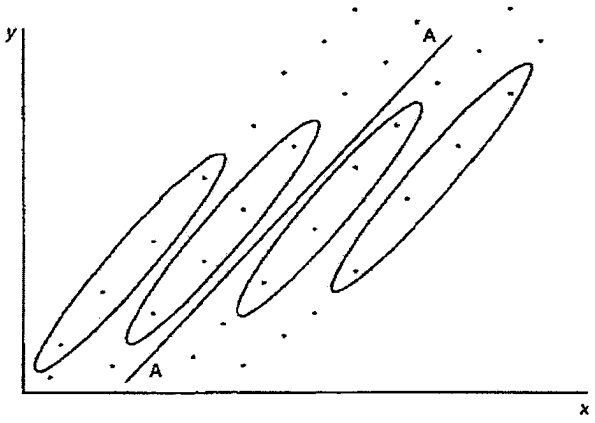

Figure 18.1 Panel data showing four observations on each of four individuals.

on that person. (There would be a thousand such ellipses in the actual data scatterplot, with roughly half above and half below *AA*; only four are drawn in Figure 18.1.) This way of viewing the data reveals that although each person in this example has the same slope, these people all have different intercepts. Most researchers would agree that this cross-sectional heterogeneity is the normal state of affairs – there are so many unmeasured variables that determine y that their influence gives rise to a different intercept for each individual. This phenomenon suggests that OLS is biased unless the influence of these omitted variables (embodied in different intercepts) is uncorrelated with the included explanatory variables. Two ways of improving estimation have been suggested, associated with two different ways of modeling the presence of a different intercept for each cross-sectional unit.

The first way is to put in a dummy for each individual (and omit the intercept). Doing this allows each individual to have a different intercept, and so OLS including all these dummies should guard against the bias discussed above. This "fixed effect" model gives rise to what is called the *fixed effects estimator* – OLS applied to the fixed effects model. At first glance this seems as though it would be difficult to estimate because (in our example above) we would require a thousand dummies. It turns out that a computational trick avoids this problem via an easy transformation of the data. This transformation consists of subtracting from each observation the average of the values within its ellipse – the observations for each individual have subtracted from them the averages of all the observations for that individual. OLS on these transformed data produces the desired slope estimate.

The fixed effects model has two major drawbacks:

1. By implicitly including a thousand dummy variables we lose 999 degrees of freedom (by dropping the intercept we save one degree of freedom). If we could find some way of avoiding this loss, we could produce a more efficient estimate of the common slope.

2. The transformation involved in this estimation process wipes out all explanatory variables that do not vary within an individual. This means that any explanatory variable that is time-invariant, such as gender, race, or religion, disappears, and so we are unable to estimate a slope coefficient for that variable. (This happens because within the ellipse in Figure 18.1, the values of these variables are all the same so that when we subtract their average they all become zero.)

The second way of allowing for different intercepts, the "random effects" model, is designed to overcome these two drawbacks of the fixed effects model. This model is similar to the fixed effects model in that it postulates a different intercept for each individual, but it interprets these differing intercepts in a novel way. This procedure views the different intercepts as having been drawn from a bowl of possible intercepts, so they may be interpreted as random (usually assumed to be normally distributed) and treated as though they were a part of the error term. As a result, we have a specification in which there is an overall intercept, a set of explanatory variables with coefficients of interest, and a composite error term. This composite error has two parts. For a particular individual, one part is the "random intercept" term, measuring the extent to which this individual's intercept differs from the overall intercept. The other part is just the traditional random error with which we are familiar, indicating a random deviation for that individual in that time period. For a particular individual the first part is the same in all time periods; the second part is different in each time period.

The trick to estimation using the random effects model is to recognize that the variance–covariance matrix of this composite error is nonspherical (i.e., not all off-diagonal elements are zero). In the example above, for all four observations on a specific individual, the random intercept component of the composite error is the same, so these composite errors will be correlated in a special way. Observations on different individuals are assumed to have zero correlation between their composite errors. This creates a variance–covariance matrix with a special pattern. The *random effects estimator* estimates this variance–covariance matrix and performs estimated generalized least squares (EGLS). The EGLS calculation is done by finding a transformation of the data that creates a spherical variance–covariance matrix and then performing OLS on the transformed data. In this respect it is similar to the fixed effects estimator except that it uses a different transformation.

18.3 Fixed Versus Random Effects

By saving on degrees of freedom, the random effects model produces a more efficient estimator of the slope coefficients than the fixed effects model. Furthermore, the transformation used for the random effects estimation procedure does not wipe out the explanatory variables that are time-invariant, allowing estimation of coefficients on variables such as gender, race, and religion. These results suggest that the random effects model is superior to the fixed effects model. So should we always use the

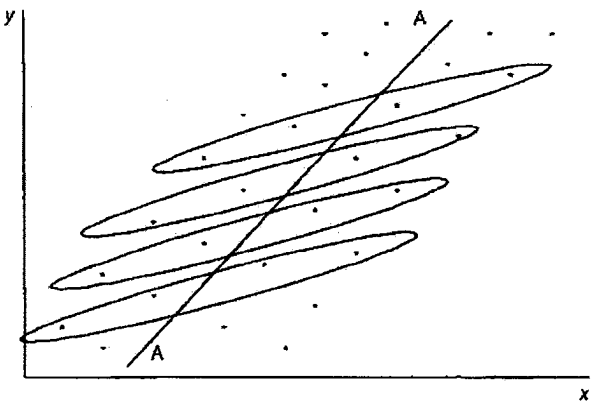

Figure 18.2 Panel data showing four observations on each of four individuals, with positive correlation between x and the intercept.

random effects model? Unfortunately, the random effects model has a major qualification that makes it applicable only in special circumstances.

This qualification is illustrated in Figure 18.2, where the data look exactly the same as in Figure 18.1, but the ellipses are drawn differently, to reflect a different allocation of observations to individuals. All persons have the same slope and different intercepts, just as before, but there is a big difference now – the common slope is not the same as the slope of the *AA* line, as it was in Figure 18.1. The main reason for this is that *the intercept for an individual is larger the larger is that individual's x value*. (Lines drawn through the observations in ellipses associated with higher x values cut the y axis at larger values.) This causes the OLS estimate using all the data to produce the *AA* line, clearly an overestimate of the common slope. This happens because as we move toward a higher x value, the y value increases for two reasons. First, it increases because the x value increases, and second, because there is likely to be a higher intercept. OLS estimation is biased upward because when x changes, OLS gives it credit for both of these y changes.

This bias does not characterize the fixed effects estimator because as described earlier the different intercepts are explicitly recognized by putting in dummies for them. But it is a problem for the random effects estimator because rather than being explicitly recognized, the intercepts are incorporated into the (composite) error term. As a consequence, the composite error term will tend to be bigger whenever the x value is bigger, creating correlation between x and the composite error term. Correlation between the error and an explanatory variable creates bias. As an example, suppose that wages are being regressed on schooling for a large set of individuals, and that a missing variable, ability, is thought to affect the intercept. Since schooling and ability are likely to be correlated, modeling this as a random effect will create correlation between the composite error and the regressor schooling, causing the random effects estimator to be biased. The bottom line here is that the random effects estimator should only be used