PAIRWISE CORRELATION. *See* CORRELATION

PAIRWISE DELETION

Pairwise deletion is a term used in relation to computer software programs such as SPSS in connection with the handling of MISSING DATA. Pairwise deletion of missing data means that only cases relating to each pair of variables with missing data involved in an analysis are deleted. Consider the following scenario: We have a sample of 200 cases and want to produce a set of PEARSON CORRELATIONS for 10 variables. Let us take the first 3 variables. Variable 1 has 4 missing cases, Variable 2 has 8 missing cases, and Variable 3 has 2 missing cases. Let us also assume that the missing cases are different for each of the variables; that is, Variable 1's 4 missing cases are not among Variable 2's 8 missing cases. (In practice, this will not always be the case because if a person does not answer questions relating to both Variable 1 and Variable 2, he or she will be a case with missing data on both variables.) When we analyze the 10 variables, the correlation between Variables 1 and 2 will be based on 188 cases (because between them, Variable 1 and Variable 2 have 12 missing cases). The correlation between Variable 1 and Variable 3 will be based on 194 cases (because between them, they have 6 missing cases), and between Variable 2 and Variable 3, it will be based on 190 cases (because between them, they have 10 missing cases).

—Alan Bryman

See also DELETION, LISTWISE DELETION

PANEL

The term *panel* refers to a RESEARCH DESIGN in which the DATA are gathered on at least two occasions on the same units of analysis. Most commonly, that unit of analysis is an individual in a SURVEY. In a two-wave panel survey, respondents at time t are reinterviewed at time $t + 1$; in a three-wave panel, they would be reinterviewed yet again at time $t + 2$. Sometimes, the units of analysis of the panel are aggregates, such as nations. For example, a sample of European countries might be measured at time t and again at $t + 1$. A special value of a panel design is that it incorporates the temporality required for strong CAUSAL inference because the potential causal variables, the Xs, can actually be measured before Y occurs. In addition, panel studies allow social change to be gauged. For example, with repeated INTERVIEWING over a long time, panel surveys (unlike COHORT designs) have the potential to distinguish age effects, PERIOD EFFECTS, and cohort effects.

A chief disadvantage of the panel approach is the data ATTRITION from time point to time point, especially with panel surveys in which individuals drop out of the sample. A related problem is the issue of addition to the panel, in that demographic change may, over time, suggest that the original panel is no longer representative of the POPULATION initially SAMPLED. For example, as a result of an influx of immigrants, it may be that the panel should have respondents added to it to better reflect the changed population.

—Michael S. Lewis-Beck

PANEL DATA ANALYSIS

Panel data refer to data sets consisting of multiple observations on each sampling unit. This could be generated by pooling time-series observations across a variety of cross-sectional units, including countries, states, regions, firms, or randomly sampled individuals or households. This encompasses longitudinal data analysis in which the primary focus is on individual histories. Two well-known examples of U.S. panel data are the Panel Study of Income Dynamics (PSID), collected by the Institute for Social Research at the University of Michigan, and the National Longitudinal Surveys of Labor Market Experience (NLS) from the Center for Human Resource Research at Ohio State University. An inventory of national studies using panel data is given at http://www.ceps.lu/Cher/Cherpres.htm. These include the Belgian Household Panels, the German Socioeconomic Panel, the French Household Panel, the British Household Panel Survey, the Dutch Socioeconomic Panel, the Luxembourg Household Panel, and, more recently, the European Community household panel. The PSID began in 1968 with 4,802

families and includes an oversampling of poor households. Annual interviews were conducted and socioeconomic characteristics of each family and roughly 31,000 individuals who had been in these or derivative families were recorded. The list of variables collected is more than 5,000. The NLS followed five distinct segments of the labor force. The original samples include 5,020 men ages 45 to 59 years in 1966, 5,225 men ages 14 to 24 years in 1966, 5,083 women ages 30 to 44 years in 1967, 5,159 women ages 14 to 24 years in 1968, and 12,686 youths ages 14 to 21 years in 1979. There was an oversampling of Blacks, Hispanics, poor Whites, and military in the youths survey. The variables collected run into the thousands. Panel data sets have also been constructed from the U.S. Current Population Survey (CPS), which is a monthly national household survey conducted by the Census Bureau. The CPS generates the unemployment rate and other labor force statistics. Compared with the NLS and PSID data sets, the CPS contains fewer variables, spans a shorter period, and does not follow movers. However, it covers a much larger sample and is representative of all demographic groups.

Some of the benefits and limitations of using panel data are given in Hsiao (1986). Obvious benefits include a much larger data set because panel data are multiple observations on the same individual. This means that there will be more variability and less COLLINEARITY among the variables than is typical of cross-section or time-series data. For example, in a demand equation for a given good (say, gasoline) price and income may be highly correlated for annual time-series observations for a given country or state. By stacking or pooling these observations across different countries or states, the variation in the data is increased and collinearity is reduced. With additional, more informative data, one can get more reliable estimates and test more sophisticated behavioral models with less restrictive assumptions. Another advantage of panel data is their ability to control for individual heterogeneity. Not controlling for these unobserved individual specific effects leads to BIAS in the resulting estimates. For example, in an earnings equation, the wage of an individual is regressed on various individual attributes, such as education, experience, gender, race, and so on. But the error term may still include unobserved individual characteristics, such as ability, which is correlated with some of the regressors, such as education. Cross-sectional studies attempt to control for this unobserved ability by collecting hard-to-get data on twins. However, using individual panel data, one can, for example, difference the data over time and wipe out the unobserved individual invariant ability. Panel data sets are also better able to identify and estimate effects that are not detectable in pure cross-section or pure time-series data. In particular, panel data sets are better able to study complex issues of dynamic behavior. For example, with cross-section data, one can estimate the rate of unemployment at a particular point in time. Repeated cross-sections can show how this proportion changes over time. Only panel data sets can estimate what proportion of those who are unemployed in one period remains unemployed in another period.

Limitations of panel data sets include the following: problems in the design, data collection, and data management of panel surveys (see Kasprzyk, Duncan, Kalton, & Singh, 1989). These include the problems of coverage (incomplete account of the population of interest), nonresponse (due to lack of cooperation of the respondent or because of interviewer error), recall (respondent not remembering correctly), frequency of interviewing, interview spacing, reference period, the use of bounding to prevent the shifting of events from outside the recall period into the recall period, and time-in-sample bias. Another limitation of panel data sets is the distortion due to measurement errors. Measurement errors may arise because of faulty response due to unclear questions, memory errors, deliberate distortion of responses (e.g., prestige bias), inappropriate informants, misrecording of responses, and interviewer effects. Although these problems can occur in cross-section studies, they are aggravated in panel data studies. Panel data sets may also exhibit bias due to sample selection problems. For the initial wave of the panel, respondents may refuse to participate, or the interviewer may not find anybody at home. This may cause some bias in the inference drawn from this sample. Although this nonresponse can also occur in cross-section data sets, it is more serious with panels because subsequent waves of the panel are still subject to nonresponse. Respondents may die, move, or find that the cost of responding is high. The rate of attrition differs across panels and usually increases from one wave to the next, but the rate of increase declines over time. Typical panels involve annual data covering a short span of time for each individual. This means that asymptotic arguments rely crucially on the number of individuals in the panel tending to infinity. Increasing the time span of the panel is not without cost either. In

fact, this increases the chances of attrition with every new wave, as well as the degree of computational difficulty in estimating qualitative limited dependent variable panel data models (see Baltagi, 2001).

Although RANDOM-COEFFICIENT MODELS can be used in the estimation and specification of panel data models (Hsiao, 1986), most panel data applications have been limited to a simple regression with error components disturbances, such as the following:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mu_i + \nu_{it}, \quad i = 1, \ldots, N; t = 1, \ldots, T$$

where y_{it} may denote log(wage) for the ith individual at time t, and $\mathbf{x}_{it}$ is a vector of observations on k explanatory variables such as education, experience, race, sex, marital status, union membership, hours worked, and so on. In addition, $\boldsymbol{\beta}$ is a k vector of unknown coefficients, μ_i is an unobserved individual specific effect, and ν_{it} is a zero mean random disturbance with variance σ_ν^2. The error components disturbances follow a one-way analysis of variance (ANOVA). If μ_i denote fixed parameters to be estimated, this model is known as the FIXED-EFFECTS (FE) MODEL. The $\mathbf{x}_{it}$s are assumed independent of the ν_{it}s for all i and t. Inference in this case is conditional on the particular N individuals observed. Estimation in this case amounts to including $(N-1)$ individual dummies to estimate these individual invariant effects. This leads to an enormous loss in degrees of freedom and attenuates the problem of MULTICOLLINEARITY among the regressors. Furthermore, this may not be computationally feasible for large N panels. In this case, one can eliminate the μ_is and estimate $\boldsymbol{\beta}$ by running least squares of $\tilde{y}_{it} = y_{it} - \bar{y}_{i.}$ on the $\tilde{\mathbf{x}}_{it}$s similarly defined, where the dot indicates summation over that index and the bar denotes averaging. This transformation is known as the within transformation, and the corresponding estimator of $\boldsymbol{\beta}$ is called the within estimator or the FE estimator. Note that the FE estimator cannot estimate the effect of any time-invariant variable such as gender, race, religion, or union participation. These variables are wiped out by the within transformation. This is a major disadvantage if the effect of these variables on earnings is of interest. Ignoring the individual unobserved effects (i.e., running ordinary least squares [OLS] without individual dummies) leads to biased and inconsistent estimates of the regression coefficients.

If μ_i denotes independent random variables with zero mean and constant variance σ_μ^2, this model is known as the random-effects model. The preceding moments are conditional on the $\mathbf{x}_{it}$s. In addition, μ_i and ν_{it} are assumed to be conditionally independent. The random-effects (RE) model can be estimated by generalized least squares (GLS), which can be obtained using a least squares regression of $y_{it}^* = y_{it} - \theta\bar{y}_{i.}$ on $\mathbf{x}_{it}^*$ similarly defined, where θ is a simple function of the variance components σ_μ^2 and σ_ν^2 (Baltagi, 2001). The corresponding GLS estimator of $\boldsymbol{\beta}$ is known as the RE estimator. Note that for this RE model, one can estimate the effects of individual-invariant variables. The best quadratic unbiased (BQU) estimators of the variance components are ANOVA-type estimators based on the true disturbances, and these are minimum variance unbiased (MVU) under normality of the disturbances. One can obtain feasible estimates of the variance components by replacing the true disturbances by OLS or fixed-effects residuals. For the random-effects model, OLS is still unbiased and consistent but not efficient.

Fixed versus random effects has generated a lively debate in the biometrics and econometrics literature. In some applications, the random- and fixed-effects models yield different estimation results, especially if T is small and N is large. A specification test based on the difference between these estimates is given by Hausman (1978). The null hypothesis is that the individual effects are not correlated with the $\mathbf{x}_{it}$s. The basic idea behind this test is that the fixed-effects estimator $\tilde{\boldsymbol{\beta}}_{\text{FE}}$ is consistent, whether or not the effects are correlated with the $\mathbf{x}_{it}$s. This is true because the fixed-effects transformation described by $\tilde{y}_{it}$ wipes out the μ_i effects from the model. In fact, this is the modern econometric interpretation of the FE model—namely, that the μ_is are random but hopelessly correlated with all the $\mathbf{x}_{it}$s. However, if the null hypothesis is true, the fixed-effects estimator is not efficient under the random-effects specification because it relies only on the within variation in the data. On the other hand, the random-effects estimator $\hat{\boldsymbol{\beta}}_{\text{RE}}$ is efficient under the null hypothesis but is biased and inconsistent when the effects are correlated with the $\mathbf{x}_{it}$s. The difference between these estimators $\hat{q} = \tilde{\boldsymbol{\beta}}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}$ tends to zero in probability limits under the null hypothesis and is nonzero under the alternative. The variance of this difference is equal to the difference in variances, $\text{var}(\hat{q}) = \text{var}(\tilde{\boldsymbol{\beta}}_{\text{FE}}) - \text{var}(\hat{\boldsymbol{\beta}}_{\text{RE}})$ because $\text{cov}(\hat{q}, \hat{\boldsymbol{\beta}}_{\text{RE}}) = 0$ under the null hypothesis. Hausman's test statistic is based on $m = \hat{q}'[\text{var}(\hat{q})]^{-1}\hat{q}$ and is asymptotically distributed as a chi-square with k degrees of freedom under the null hypothesis.

For maximum likelihood as well as generalized method of moments estimation of panel models, the reader is referred to Baltagi (2001). Space limitations

do not allow discussion of panel data models that include treatment of missing observations, dynamics, measurement error, qualitative limited dependent variables, endogeneity, and nonstationarity of the regressors. Instead, we focus on some frequently encountered special panel data sets—namely, pseudo-panels and rotating panels. Pseudo-panels refer to the construction of a panel from repeated cross sections, especially in countries where panels do not exist but where independent surveys are available over time. The United Kingdom Family Expenditure Survey, for example, surveys about 7,000 households annually. These are independent surveys because it may be impossible to track the same household across surveys, as required in a genuine panel. Instead, one can track cohorts and estimate economic relationships based on cohort means. Pseudo-panels do not suffer the attrition problem that plagues genuine panels and may be available over longer time periods. One important question is the optimal size of the cohort. A large number of cohorts will reduce the size of a specific cohort and the samples drawn from it. Alternatively, selecting few cohorts increases the accuracy of the sample cohort means, but it also reduces the effective sample size of the panel.

Rotating panels attempt to keep the same number of households in the survey by replacing the fraction of households that drop from the sample in each period with an equal number of freshly surveyed households. This is a necessity in surveys in which a high rate of attrition is expected from one period to the next. Rotating panels allow the researcher to test for the existence of time-in-sample bias effects. These correspond to a significant change in response between the initial interview and a subsequent interview when one would expect the same response.

With the growing use of cross-country data over time to study purchasing power parity, growth convergence, and international research and development spillovers, the focus of panel data econometrics has shifted toward studying the asymptotics of macro panels with large N (number of countries) and large T (length of the time series) rather than the usual asymptotics of micro panels with large N and small T. Researchers argue that the time-series components of variables such as per capita gross domestic product growth have strong nonstationarity. Some of the distinctive results that are obtained with nonstationary panels are that many test statistics and estimators of interest have Normal limiting distributions. This is in contrast to the nonstationary time-series literature in which the limiting distributions are complicated functionals of Weiner processes. Several unit root tests applied in the time-series literature have been extended to panel data (see Baltagi, 2001). However, the use of such panel data methods is not without their critics, who argue that panel data unit root tests do not rescue purchasing power parity (PPP). In fact, the results on PPP with panels are mixed depending on the group of countries studied, the period of study, and the type of unit root test used. More damaging is the argument that for PPP, panel data tests are the wrong answer to the low power of unit root tests in single time series. After all, the null hypothesis of a single unit root is different from the null hypothesis of a panel unit root for the PPP hypothesis. Similarly, panel unit root tests did not help settle the question of growth convergence among countries. However, it was useful in spurring much-needed research into dynamic panel data models.

Over the past 20 years, the panel data methodological literature has exhibited phenomenal growth. One cannot do justice to the many theoretical and empirical contributions to date. Space limitations prevented discussion of many worthy contributions. Some topics are still in their infancy but growing fast, such as nonstationary panels and semiparametric and nonparametric methods using panel data. It is hoped that this introduction will whet the reader's appetite and encourage more readings on the subject.

—Badi H. Baltagi

REFERENCES

Baltagi, B. H. (2001). *Econometric analysis of panel data.* Chichester, UK: Wiley.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46,* 1251–1271.

Hsiao, C. (1986). *Analysis of panel data.* Cambridge, UK: Cambridge University Press.

Kasprzyk, D., Duncan, G. J., Kalton, G., & Singh, M. P. (1989). *Panel surveys.* New York: John Wiley.

PARADIGM

In everyday usage, *paradigm* refers either to a model or an example to be followed or to an established system or way of doing things. The concept was introduced into the philosophy of science by Thomas