MODULE THREE, PART ONE:
PANEL DATA ANALYSIS IN ECONOMIC EDUCATION RESEARCH

William E. Becker, William H. Greene and John J. Siegfried*

As discussed in Modules One and Two, most of the empirical economic education research is based on a "value-added," "change-score" or a "difference-in-differences" model specifications in which the expected improvement in student performance from a pre-program measure (pretest) to its post-program measurement (posttest) is estimated and studied for a cross section of subjects. Other than the fact that testing occurs at two different points in time, there is no time dimension, as seen in the data sets employed in Modules One and Two. Panel data analysis provides an alternative structure in which measurements on the cross section of subjects are taken at regular intervals over multiple periods of time.[i] Collecting data on the cross section of subjects over time enables a study of change. It opens the door for economic education researchers to address unobservable attributes that lead to biased estimators in cross-section analysis.[ii] As demonstrated in this module, it also opens the door for economic education researchers to look at things other than test scores that vary with time.

This module provides an introduction to panel data analysis with specific applications to economic education. The data structure for a panel along with constant coefficient, fixed effects and random effects representations of the data generating processes are presented. Consideration is given to different methods of estimation and testing. Finally, as in Modules One and Two, contemporary estimation and testing procedures are demonstrated in Parts Two, Three and Four using LIMDEP (NLOGIT), STATA and SAS.

THE PANEL DATA SET

As an example of a panel data set, consider our study (Becker, Greene and Siegfried, Forthcoming) that examines the extent to which undergraduate degrees (BA and BS) in economics or PhD degrees (PhD) in economics drive faculty size at those U.S. institutions that offer only a bachelor degree and those that offer both bachelor degrees and PhDs.

----

*William Becker is Professor of Economics, Indiana University, Adjunct Professor of Commerce, University of South Australia, Research Fellow, Institute for the Study of Labor (IZA) and Fellow, Center for Economic Studies and Institute for Economic Research (CESifo). William Greene is Professor of Economics, Stern School of Business, New York University, Distinguished Adjunct Professor at American University and External Affiliate of the Health Econometrics and Data Group at York University. John Siegfried is Professor of Economics, Vanderbilt University, Senior Research Fellow, University of Adelaide, South Australia, and Secretary-Treasurer of the American Economic Association. Their e-mail addresses are <beckerw@indiana.edu>, <wgreene@stern.nyu.edu> and <john.siegfried@vanderbilt.edu>.

We obtained data on the number of full-time tenured or tenure-track faculty and the number of undergraduate economics degrees per institution per year from the American Economic Association's Universal Academic Questionnaire (UAQ). The numbers of PhD degrees in economics awarded by department were obtained from the Survey of Earned Doctorates, which is sponsored by several U.S. federal government agencies. These sources provided data on faculty size and degree yearly data for each institution for 16 years from 1990-91 through 2005-06. For each year, we had data from 18 bachelor degree-granting institutions and 24 institutions granting both the PhD and bachelor degrees. Pooling the cross-section observations on each of the 18 bachelor only institutions, at a point in time, over the 16 years, implies a panel of 288 observations on each initial variable. Pooling the cross-section observations on each of the 24 PhD institutions, at a point in time, over the 16 years, implies a panel of 384 observations on each initial variable. Subsequent creation of a three-year moving average variable for degrees granted at each type of institution reduced the length of each panel in the data set to 14 years of usable data.

Panel data are typically laid out in sequential blocks of cross-sectional data. For example, the bachelor degree institution data observations for each of the 18 colleges appear in blocks of 16 rows for years 1991 through 2006:

"College" identifies the bachelor degree-granting institution by a number 1 through 18.

"Year" runs from 1996 through 2006.

"*BA&S*" is the number of BS or BA degrees awarded in each year by each college.

"*MEANBA&S*" is the average number of degrees awarded by each college for the 16-year period.

"Public" equals 1 if the institution is a public college and 2 if it is a private college.

"Bschol" equals 1 if the college has a business program and 0 if not.

"Faculty" is the number of tenured or tenure-track economics department faculty members.

"T" is a time trend running from −7 to 8, corresponding to years from 1996 through 2006.

"MA_Deg" is a three-year moving average of degrees (unknown for the first two years).

College	Year	*BA&S*	*MEANBA&S*	Public	Bschol	Faculty	T	MA_Deg
1	1991	50	47.375	2	1	11	-7	Missing
1	1992	32	47.375	2	1	8	-6	Missing
1	1993	31	47.375	2	1	10	-5	37.667
1	1994	35	47.375	2	1	9	-4	32.667
↓	↓	↓	↓	↓	↓	↓		↓
1	2003	57	47.375	2	1	7	5	56
1	2004	57	47.375	2	1	10	6	55.667
1	2005	57	47.375	2	1	10	7	57
1	2006	51	47.375	2	1	10	8	55
2	1991	16	8.125	2	1	3	-7	Missing
2	1992	14	8.125	2	1	3	-6	Missing
2	1993	10	8.125	2	1	3	-5	13.333
↓	↓	↓	↓	↓	↓	↓		↓
2	2004	10	8.125	2	1	3	6	12.667
2	2005	7	8.125	2	1	3	7	11.333
2	2006	6	8.125	2	1	3	8	7.667
3	1991	40	35.5	2	1	8	-7	Missing
3	1992	31	37.125	2	1	8	-6	Missing
↓	↓	↓	↓	↓	↓	↓		↓
17	2004	64	39.3125	2	0	5	6	54.667
17	2005	37	39.3125	2	0	4	7	51.333
17	2006	53	39.3125	2	0	4	8	51.333
18	1991	14	8.4375	2	0	4	-7	Missing
18	1992	10	8.4375	2	0	4	-6	Missing
18	1993	10	8.4375	2	0	4	-5	11.333
18	1994	7	8.4375	2	0	3.5	-4	9
↓	↓	↓	↓	↓	↓	↓		↓
18	2005	4	8.4375	2	0	2.5	7	7.333
18	2006	7	8.4375	2	0	3	8	6

In a few years for some colleges, faculty size was missing. We interpolated missing data on the number of faculty members in the economics department from the reported information in the years prior and after a missing observation; thus, giving rise to the prospect for a half person in those cases. If a panel data set such as this one has missing values that cannot be meaningfully interpolated, it is an "**unbalanced panel**," in which the number of usable observations differs across the units. If there are no missing values and there are the same number of periods of data for every group (college) in the sample, then the resulting pooled cross-section and time-series data set is said to be a "**balanced panel**." Typically, the cross-section dimension is designated the *i* dimension and the time-series dimension is the *t* dimension. Thus, panel data studies are sometimes referred to as "***it*** " **studies**.

THE PANEL DATA-GENERATING PROCESS

There are three ways in which we consider the effect of degrees on faculty size. Here we will consider only the bachelor degree-granting institutions.

First, the decision makers might set the permanent faculty based on the most current available information, as reflected in the number of contemporaneous degrees ($BA\&S_{it}$). That is, the decision makers might form a type of rational expectation by setting the faculty size based on the anticipated number of majors to receive degrees in the future, where that expectation for that future number is forecasted by this year's value. Second, we included the overall mean number of degrees awarded at each institution ($MEANBA\&S_i$) to reflect a type of historical steady state. That is, the central administration or managers of the institution may have a target number of permanent faculty relative to the long-term expected number of annual graduates from the department that is desired to maintain the department's appropriate role within the institution.[iii] Third, the central authority might be willing to marginally increase or decrease the permanent faculty size based on the near term trend in majors, as reflected in a three-year moving average of degrees awarded (MA_Deg_{it}).

We then assume the faculty size data-generating process for bachelor degree-granting undergraduate departments to be

$$FACULTY\ size_{it} = \beta_1 + \beta_2 T_t + \beta_3 BA\&S_{it} + \beta_4 MEANBA\&S_i + \beta_5 PUBLIC_i + \beta_6 Bschl + \beta_7 \text{MA\_Deg}_{it} + \varepsilon_{it} \qquad (1)$$

where the error term ε_{it} is independent and identically distributed (*iid)* across institutions and over time and $E(\varepsilon_{it}^2|\mathbf{x}_{it}) = \sigma^2$, for $I = 18$ colleges and $T = 14$ years (−5 through 8) for 252 complete observations. Notice that there is no time subscript on the mean number of degrees, public/private and B school regressors because they do not vary with time.

In a more general and nondescript algebraic form for any *it* study, in which all explanatory variables are free to vary with time and the error term is of the simple *iid* form with $E(\varepsilon_{it}^2|x_{it}) = \sigma^2$, the model would be written

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \ldots + \beta_k X_{kit} + \varepsilon_{it}\,, \text{ for } i = 1, 2, \ldots I \text{ and, } t = 1, 2, \ldots T. \quad (2)$$

This is a **constant coefficient model** because the intercept and slopes are assumed not to vary within a cross section (not to vary across institutions) or over time. If this assumption is true, the parameters can be estimated without bias by ordinary least squares applied directly to the panel data set. Unfortunately, this assumption is seldom true, requiring us to consider the fixed-effect and random-effects models.

FIXED-EFFECTS MODEL

The *fixed effects model* allows the intercept to vary across institutions (or among whatever cross-section categories that are under consideration in other studies), while keeping the slope coefficients the same for all institutions (or categories). The model could be written as

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_{it}\,. \quad (3)$$

Where β_{1i} suggests that there is a separate intercept for each unit. No restriction is placed on how the intercepts vary, except, of course, that they do so independently of ε_{it}. The model can be made explicit for our application by inserting a 0-1 covariate or dummy variable for each of the institutions except the one for which comparisons are to be made. In our case, there are 18 colleges; thus, 17 dummy variables are inserted and each of their coefficients is interpreted as the expected change in faculty size for a movement from the excluded college to the college of interest. Alternatively, we could have a separate dummy variable for each college and drop the overall intercept. Both approaches give the same results for the other coefficients and for the R^2 in the regression. (A moment's thought will reveal, however, that in this setting, either way it is formulated, it is not possible to have variables, such as type of school, that do not vary through time. In the fixed effects model, such a variable would just be a multiple of the school specific dummy variable.)

To clarify, the fixed-effects model for our study of bachelor degree institutions is written (where Collge*i* = 1 if college *i* and 0 if not, for *i* = 1, 2, ... 18) as

$$\begin{aligned} FACULTY\ size_{it} = {} & \beta_1 + \beta_2 YEAR_t + \beta_3 BA\&S_{it} + \beta_4 MEANBA\&S_i + \beta_5 PUBLIC_i + \\ & \beta_6 Bschl + \beta_7 \text{MA\_Deg}_{it} + \beta_8 College1 + \beta_9 College2 + \\ & \beta_{10} College3 + \ldots + \beta_{23} College16 + \beta_{24} College17 + \varepsilon_{it}. \end{aligned} \quad (4)$$

Here a dummy for college 18 is omitted and its effects are reflected in the constant term β_1 when College1 = College2 =…= College16 = College17 = 0. For example, β_9 is the expected change in faculty size for a movement from college 18 to college 2. Which college is omitted is arbitrary, but one must be omitted to avoid perfect collinearity in the data set. In general, if *i* goes from 1 to *I* categories, then only *I* − 1 dummies are used to form the fixed-effects model:

$$\begin{aligned} Y_{it} = {} & \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \ldots + \beta_k X_{kit} \\ & + \beta_{k+1} D_1 + \beta_{k+2} D_2 + \ldots + \beta_{k+1} D_{I-1} + \beta_{k+2} D_2 + \ldots \beta_{k+(I-)1} D_{I-1} + \varepsilon_{it}\,, \\ & \text{for } i = 1, 2, \ldots I \text{ and, } t = 1, 2, \ldots T. \end{aligned} \quad (5)$$

After creating the relevant dummy variables (Ds), the parameters of this fixed-effects model can be estimated without bias using by ordinary least squares.[iv]

If one has sufficient observations, the categorical dummy variables can be interacted with the other time-varying explanatory variables to enable the slopes to vary along with the intercept over time. For our study with 18 college categories this would be laborious to write out in equation form. In many cases there simply are not sufficient degrees of freedom to accommodate all the required interactions.[v]

To demonstrate a parsimonious model setup with both intercept and slope variability consider a hypothetical situation involving three categories (represented by dummies D_1, D_2 and D_3) and two time-varying explanatory variables (represented by X_{2it} and X_{3it}):

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \beta_4 D_1 + \beta_5 D_2 + \beta_6(X_{2it}\, D_1) + \beta_7(X_{3it}\, D_1) + \beta_8(X_{2it}\, D_2) + \beta_9(X_{3it}\, D_2) + \varepsilon_{it}. \qquad (6)$$

In this model, β_1 is the intercept for category three, where $D_3 = 0$. The intercept for category one is $\beta_1 + \beta_4$ and for category 2 it is $\beta_1 + \beta_5$. The change in the expected value of Y given a change in X_2 is $\beta_2 + \beta_6 D_1 + \beta_8 D_2$; thus for category 1 it is $\beta_2 + \beta_6$ and for category 2 it is $\beta_2 + \beta_8$. The change in the expected value of Y for a movement from category two to category three is

$$(\beta_5 - \beta_4) + (\beta_8 - \beta_6)X_{2it} + (\beta_9 - \beta_7)X_{3it}\ . \qquad (7)$$

Individual coefficients are tested in fixed-effects models as in any other model with the z ratio (with asymptotic properties) or t ratio (finite sample properties). There could be category-specific heteroscedasticity or autocorrelation over time. As described and demonstrated in Module One, where students were grouped or clustered into classes, a type of White heteroscedasticity consistent covariance estimator can be used in fixed-effects models with ordinary least squares to obtain standard errors robust to unequal variances across the groups. Correlation of residuals from one period to the next within a panel can also be a problem. If this serial correlation is of the first-order autoregressive type, a Prais-Winston transformation transformation might be considered to first partial difference the data to remove the serial correlation problem. In general, because there are typically few time-series observations, it is difficult to both correctly identify the nature of the time-series error term process and appropriately address with a least-squares estimator.[vi] Contemporary treatments typically rely on robust, "cluster" corrections that accommodate more general forms of correlation across time.

Hypotheses tests about sets of coefficients related to the categories in fixed-effects models are conducted as tests of linear restrictions for a subset of coefficients as described and demonstrated in Module One. For instance, as a starting point one might want to test if there is any difference in intercepts or slopes. For our hypothetical parsimonious model the null and alternative hypotheses are:

$\mathrm{H_O}$: $\beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ *vs.* (8)
$\mathrm{H_A}$: at least one of these six βs is not zero.

The unrestricted sum of squared residuals comes from the regression:

$$\hat{y}_{it} = b_1 + b_2x_{2it} + b_3x_{3it} + b_4D_1 + b_5D_2 + b_6(x_{2it}\,D_1) + b_7(x_{3it}\,D_1) + b_8(x_{2it}\,D_2) + b_9(x_{3it}\,D_2). \qquad (9)$$

The restricted sum of squared residuals comes from the regression:

$$\hat{y}_{it} = b_1 + b_2x_{2it} + b_3x_{3it}\,. \qquad (10)$$

The relevant *F* statistic is

$$F = \frac{[\text{Restricted ResSS}(\beta subset = 0) - \text{Unrestricted ResSS}]\ /\ (9-3)}{\text{Unrestricted ResSS}\,/(\Sigma T_i - 9)}\ . \qquad (11)$$

RANDOM-EFFECTS MODELS

The *random effects model*, like the fixed effects model, allows the intercept term to vary across units. The difference is an additional assumption, not made in the fixed effects case, that this variation is independent of the values of the other variables in the model. Recall, in the fixed effects case, we placed no restriction on the relationship between the intercepts and the other independent variables. In essence, a random-effects data generating process is a regression with an intercept that is subject to purely random perturbations; it is a category-specific random variable (β_{1i}). The realization of the random variable intercept β_{1i} is assumed to be formed by the overall mean plus the i^{th} category-specific random term v_i. In the case of our hypothetical, parsimonious two explanatory variable model, the relevant random-effects equations are

$$Y_{it} = \beta_{1i} + \beta_2X_{2it} + \beta_3X_{3it} + \varepsilon_{it}\,. \qquad (12)$$
$$\beta_{1i} = \alpha + v_i \text{ with } \mathrm{Cov}[\mathrm{v_i},(\mathrm{X_{1i2}},\mathrm{X_{2it}})] = 0.$$

Inserting the second equation into the first produces the “random effects” model,

$$Y_{it} = \alpha + \beta_2X_{2it} + \beta_3X_{3it} + \varepsilon_{it} + v_i\,. \qquad (13)$$

Deviations from the main intercept, α, as measured in the category specific part of the error term, v_i, must be uncorrelated with the time-varying regressors (that is, v_i, is uncorrelated with X_{2it} and X_{3it}) and have zero mean. Because v_i does not vary with time, it is reasonable to assume its variance is fixed given the explanatory variables.[vii] Thus,

$$\mathrm{E}(v_i|\,X_{2it}, X_{3it}) = 0 \text{ and } \mathrm{E}(v_i^2|\,X_{2it}, X_{3it}) = \theta^2\,. \qquad (14)$$

An important difference between the fixed and random effects models is that time-invariant regressors, such as type of school, can be accommodated in the random effects but not in the fixed effects model. The surprising result ultimately follows from the assumption that the variation of the constant terms is independent of the other variables, which allows us to put the term v_i in the disturbance of the equation, rather than build it into the regression with a set of dummy variables.)

In a random-effects model, disturbances for a given college (in our case) or whatever entity is under study will be correlated across periods whereas in the fixed-effects model this correlation is assumed to be absent. However, in both settings, correlation between the panel error term effects and the explanatory variables is also likely. Where it occurs, this correlation will reflect the effect of substantive influences on the dependent variable that have been omitted from the equation – the classic "missing variables" problem. The now standard Mundlak (1978) approach is a method of accommodating this correlation between the effects and means of the regressors. The approach is motivated by the suggestion that the correlation can be explained by the overall levels (group means) of the time variables. By this device, the effect, β_{1i}, is projected upon the group means of the time-varying variables, so that

$$\beta_{1i} = \beta_1 + \delta' \bar{x}_i + w_i \tag{15}$$

where $\bar{x}_i$ is the set of group (school) means of the time-varying variables and w_i is a (now) random effect that is uncorrelated with the variables and disturbances in the model, $w_i \sim N(0, \sigma_w^2)$.

In fact, the random effects model as described here departs from an assumption that the school effect, v_i, actually is uncorrelated with the other variables. If true, the projection would be unnecessary However, in most cases, the initial assumption of the random-effects model, that the effects and the regressors are uncorrelated, is considered quite strong. In the fixed effects case, the assumption is not made. However, it remains a useful extension of the fixed effects model to think about the "effect," β_{1i}, in terms of a projection such as suggested above – perhaps by the logic of a "hierarchical model," for the regression. That is, although the fixed effects model allows for an unrestricted effect, freely correlated with the time varying variables, the Mundlak projection adds a layer of explanation to this effect. The Mundlak approach is a useful approach in either setting. Note that after incorporating the Mundlak "correction" in the fixed effects specification, the resulting equation becomes a random effects equation.

Adding the unit means to the equations picks up the correlation between the school effects and the other variables as well as reflecting an expected long-term steady state. Our random effects models for BA and BS degree-granting undergraduate economics departments is

$$\begin{aligned} FACULTY\,size_{it} &= \beta_1 + \beta_2 YEAR_t + \beta_3 BA\&S_{it} + \\ &\quad \beta_4 MEANBA\&S_i + \beta_5 PUBLIC_i + \beta_6 Bschl + \beta_7 \mathrm{MA\_Deg}_{it} + \varepsilon_{it} + w_i \end{aligned} \tag{16}$$

where error term ε is *iid* over time and $E(\varepsilon_{it}^2|\mathbf{x}_{it}) = \sigma_i^2$ for $I = 18$ colleges and $T = 14$ years and $E[u_i^2] = \theta^2$.

FIXED EFFECTS VERSUS RANDOM EFFECTS

Fixed-effects models can be estimated efficiently by ordinary least squares whereas random-effects models are usually estimated using some type of generalized least-squares procedure. GLS should yield more asymptotically efficient estimators *if the assumptions for the random-effects model are correct*. Current practice, however, favors the fixed-effects approach for estimating standard errors because of the likelihood that the stronger assumptions behind the GLS estimator are likely not satisfied, implying poor finite sample properties (Angrist and Pischke, 2009, p. 223). This creates a bit of a dilemma, because the fixed effects approach is, at least potentially, very inefficient if the random effects assumptions are actually met. The fixed effects approach could lead to estimation of K+1+n rather than K+2 parameters (including σ^2).

Whether we treat the effects as fixed (with a constant intercept β_1 and dummy category variables) or random (with a stochastic intercept β_{1i}) makes little difference when there are a large number of time periods (Hsiao, 2007, p. 41). But, the typical case is one for which the time series is short, with many cross-section units,

The Hausman (1978) test has become the standard approach for assessing the appropriateness of the fixed-effects versus random-effects model. Ultimately, the question is whether there is strong correlation between the unobserved case-specific random effects and the explanatory variables. If this correlation is significant, the random-effects model is inappropriate and the fixed-effects model is supported. On the other hand, insignificant correlation between the specific random-effects errors and the regressors implies that the more efficient random-effects coefficient estimators trump the consistent fixed-effects estimators. The correlation cannot be assessed directly. But, indirectly, it has a testable implication for the estimators. If the effects are correlated with the time varying variable, then, in essence, the dummy variables will have been left out of the random effects model/estimator. The classic left out variable result then implies that the random effects estimator will be biased because of this problem, but the fixed effects estimator will not be biased because it includes the dummy variables. If the random effects model is appropriate, the fixed effects approach will still be unbiased, though it will fail to use the information that the extra dummy variables in the model are not needed. Thus, an indirect test for the presence of this correlation is based on the empirical difference between the fixed and random effects estimators.

Let $\boldsymbol{\beta}_{FE}$ and $\boldsymbol{\beta}_{RE}$ be the vectors of coefficients from a fixed-effects and random-effects specification. The null hypothesis for purpose of the Hausman test is that under the random effects assumption, estimators of both of these vectors are consistent, but the estimator for $\boldsymbol{\beta}_{RE}$ is more efficient (with a smaller asymptotic variance) than that of $\boldsymbol{\beta}_{FE}$. Hausman's alternative hypothesis is that the random-effects estimator is inconsistent (with coefficient distributions not settling on the correct parameter values as sample size goes to infinity) under the hypothesis of the fixed-effects model, but is consistent under the hypothesis of the random-effects model. The fixed-effects estimator is consistent in both cases. The Hausman test statistic is based on the difference between the estimated covariance matrix for least-squares dummy variable coefficient estimates ($\boldsymbol{b}_{FE}$) and that for the random-effects model:

$$H = (\boldsymbol{b}_{FE} - \boldsymbol{b}_{RE})' [\text{Var}(\boldsymbol{b}_{FE}) - \text{Var}(\boldsymbol{b}_{RE})]^{-1} (\boldsymbol{b}_{FE} - \boldsymbol{b}_{RE})$$

where H is distributed Chi square, with K (number in $\boldsymbol{b}$) degrees of freedom.

If the Chi-square statistic p value < 0.05, reject the Hausman null hypotheis and do not use random effects. If the Chi-square statistic p value > 0.05, do not reject the Hausman null hypothesis and use random effects. An intuitively appealing, and fully equivalent (and usually more convenient) way to carry out the Hausman test is to test the null hypothesis in the context of the random-effects model that the coefficients on the group means in the Mundlak-augmented regression are jointly zero.

CONCLUDING COMMENTS

As stated in Module One, "theory is easy, data are hard – hard to find and hard to get into a computer program for statistical analysis." This axiom is particularly true for those wishing to do panel data analysis on topics related to the teaching of economics where data collected for only the cross sections is the norm. As stated in Endnote One, a recent exception is Stanca (2006), in which a large panel data set for students in Introductory Microeconomics is used to explore the effects of attendance on performance. As with Modules One and Two, Parts Two, Three and Four of this module provide the computer code to conduct a panel data analysis with LIMDEP (NLOGIT), STATA and SAS, using the Becker, Greene and Siegfried (2009) data set.

REFERENCES

Angrist, Joshua D. and Jorn-Steffen Pischke (2009). *Mostly Harmless Econometrics.* Princeton New Jersey: Princeton University Press.

Becker, William, William Greene and John Siegfried (Forthcoming). "Does Teaching Load Affect Faculty Size?" *American Economists*.

Greene, William (2008). *Econometric Analysis*. 6th Edition, New Jersey: Prentice Hall.

Hausman, J. A. (1978). "Specification Tests in Econometrics," *Econometrica*. Vol. 46, No. 6. (November): 1251-1271.

Hsiao, Cheng (2007). *Analysis of Panel Data*. 2nd Edition (8th Printing), Cambridge: Cambridge University Press.

Johnson, William R. and Sarah Turner (2009). "Faculty Without Students: Resource Allocation in Higher Education, " *Journal of Economic Perspectives*. Vol. 23. No. 2 (Spring): 169-190.

Link, Charles R. and James G. Mulligan (1996). "The Value of Repeat Data for individual Students," in William E. Becker and William J. Baumol (eds), *Assessing Educational Practices: The Contribution of Economics,* Cambridge MA.: MIT Press.

Marburger, Daniel R. (2006). "Does Mandatory Attendance Improve Stduent Performance," *Journal of Economic Education*. Vol. 37. No. 2 (Spring)**:** 148-266155.

Mundlak, Yair (1978). "On the Pooling of Time Series and Cross Section Data, " *Econometrica.* Vol. 46. No. 1 (January): 69-85.

Stanca, Luca (2006). "The Effects of Attendance on Academic Performance: Panel Data Evidence for Introductory Microeconomics" *Journal of Economic Education*. Vol. 37. No. 3 (Summer)**:** 251-266.

ENDNOTES

[i] As seen in Stanca (2006), where a large panel data set for students in Introductory Microeconomics is used to explore the effects of attendance on performance, panel data analysis typically involved a dimension of time. However, panels can be set up in blocks that involve a dimension other than time. For example, Marburger (2006) uses a panel data structure (with no time dimension) to overcome endogeneity problems in assessing the effects of an enforced attendance policy on absenteeism and exam performance. Each of the *Q* multiple-choice exam questions was associated with specific course material that could be associated with the attendance pattern of *N* students, giving rise to *NQ* panel data records. A dummy variable for each student captured the fixed effects for the unmeasured attribute of students and thus eliminating any student specific sample selection problem.

[ii] Section V of Link and Mulligan (1996) outlines the advantage of panel analysis for a range of educational issues. They show how panel data analysis can be used to isolate the effect of individual teachers, schools or school districts on students' test scores.

[iii] One of us, as a member on an external review team for a well known economics department, was told by a high-ranking administrator that the department had received all the additional lines it was going to get because it now had too many majors for the good of the institution. Historically, the institution was known for turning out engineers and the economics department was attracting too many students away from engineering. This personal experience is consistent with Johnson and Turner's (2009, p. 170) assessment that a substantial part of the explanation for differences in student-faculty ratios across academic departments resides in politics or tradition rather than economic decision-making in many institutions of higher education.

[iv] As long as the model is static with all the explanatory variables exogenous and no lagged dependent variables used as explanatory variables, ordinary least-squares estimators are unbiased and consistent although not as efficient as those obtained by maximum likelihood routines. Unfortunately, this is not true if a lagged dependent variable is introduced as a regressor (as one might want to do if the posttest is to be explained by a pretest). The implied correlation between the lagged dependent variable and the individual specific effects and associated error terms bias the OLS estimators (Hsiao, 2007, pp. 73-74).

[v] Fixed-effects models can have too many categories, requiring too many dummies, for parameter estimation. Even if estimation is possible, there may be too few degrees of freedom and little power for statistical tests. In addition, problems of multicollinearity arise when many dummy variables are introduced.

[vi] Hsiao (2007, pp. 295-310) discusses panel data with a large number of time periods. When *T* is large serial correlation problems become a big issue, which is well beyond the scope of this introductory module.

[vii] Random-effects models in which the intercept error term v_i does not depend on time are referred to as one-way random-effects models. Two-way random-effects models have error terms of the form

$$\varepsilon_{it} = v_i + \varepsilon_t + u_{it}$$

where v_i is the cross-section-specific error, affecting only observations in the i^{th} panel; ε_t is the time-specific component, which is unique to all observations for the t^{th} period; and u_{it} is the random perturbation specific to the individual observation in the i^{th} panel at time t. These two-way random-effects models are also known as error component models and variance component models.