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

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

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:

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

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:

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

The unrestricted sum of squared residuals comes from the regression:

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

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.

MODULE THREE, PART TWO: PANEL DATA ANALYSIS IN ECONOMIC EDUCATION RESEARCH USING LIMDEP (NLOGIT)

Part Two of Module Three provides a cookbook-type demonstration of the steps required to use LIMDEP (NLOGIT) in panel data analysis. Users of this model need to have completed Module One, Parts One and Two, and Module Three, Part One. That is, from Module One users are assumed to know how to get data into LIMDEP, recode and create variables within LIMDEP, and run and interpret regression results. They are also expected to know how to test linear restrictions on sets of coefficients as done in Module One, Parts One and Two. Module Three, Parts Three and Four demonstrate in STATA and SAS what is done here in LIMDEP.

THE CASE

As described in Module Three, Part One, Becker, Greene and Siegfried (2009) examine the extent to which undergraduate degrees (BA and BS) in economics or Ph.D. 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. Here we retrace their analysis for the institutions that offer only the bachelor degree. We provide and demonstrate the LIMDEP (NLOGIT) code necessary to duplicate their results.

DATA FILE

The following panel data are provided in the **comma separated values** (CSV) text file “bachelors.csv”, which will automatically open in EXCEL by simply double clicking on it after it has been downloaded to your hard drive. Your EXCEL spreadsheet should look like this:

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

“Year” runs from 1996 through 2006.

“Degrees” is the number of BS or BA degrees awarded in each year by each college.

“DegreBar” 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.

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

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

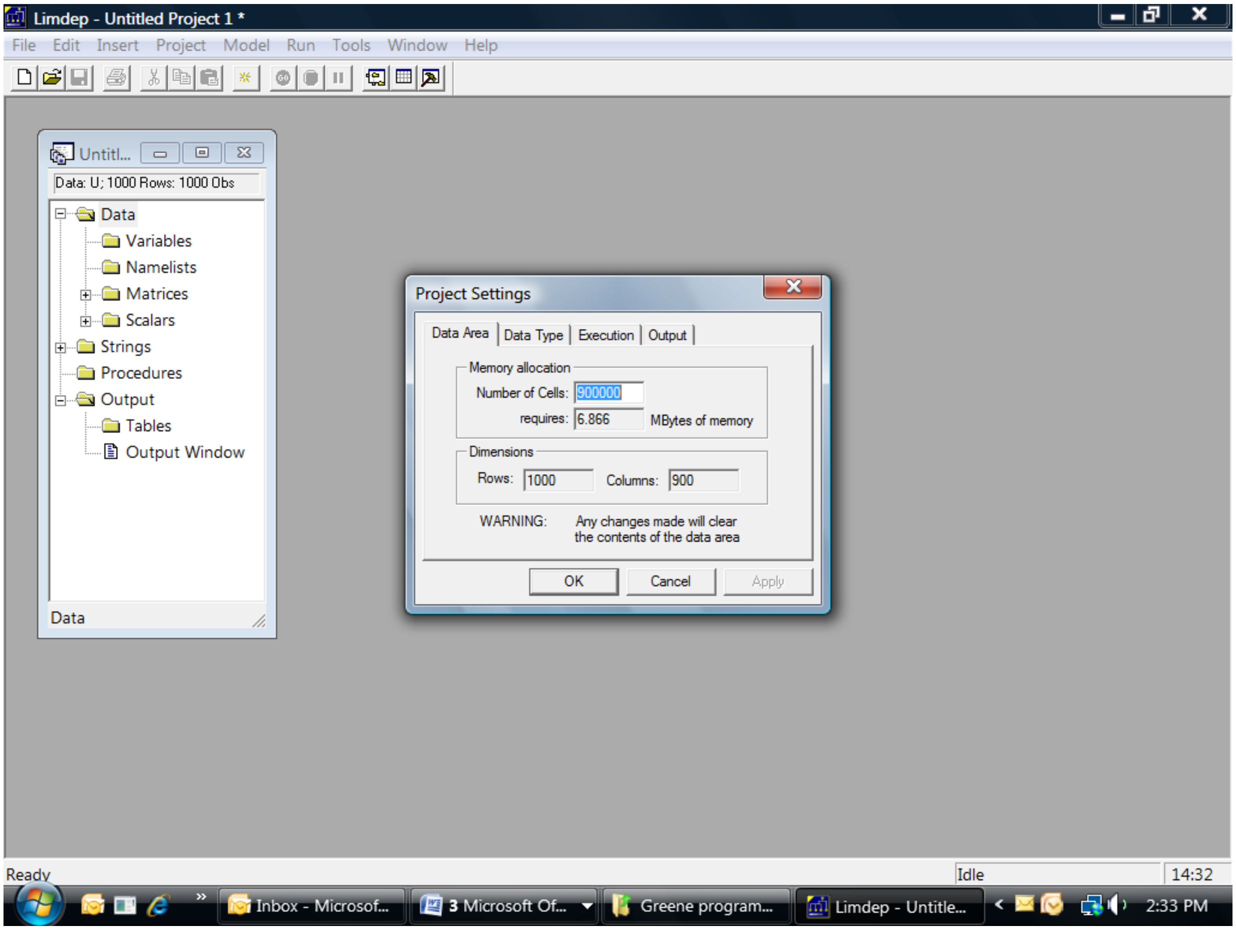
“T” is the time trend running from −7 to 8, corresponding to years from 1996 through 2006.

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

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

If you opened this CSV file in a word processor or text editing program, it would show that each of the 289 lines (including the headers) corresponds to a row in the EXCEL table, but variable values would be separated by commas and not appear neatly one on top of the other as in EXCEL.

As discussed in Module One, Part Two, older versions of LIMDEP (NLOGIT) have a data matrix default restriction of no more than 222 rows (records per variable), 900 columns (number of variables) and 200,000 cells. LIMDEP 9 and NLOGIT 4.0 automatically adjust the data constraints but in older versions the number of cells must be increased to accommodate work with our data set. After opening LIMDEP, the number of working cells can be increased by clicking the Project button on the top ribbon, going to Settings, and changing the number of cells. Going from the default 200,000 cells to 900,000 cells (1,000 Rows and 900 columns) is more than sufficient for this panel data set.



Limdep - Untitled Project 1 *

Data: U; 1000 Rows: 1000 Obs

Project Settings

Memory allocation

Number of Cells: 900000

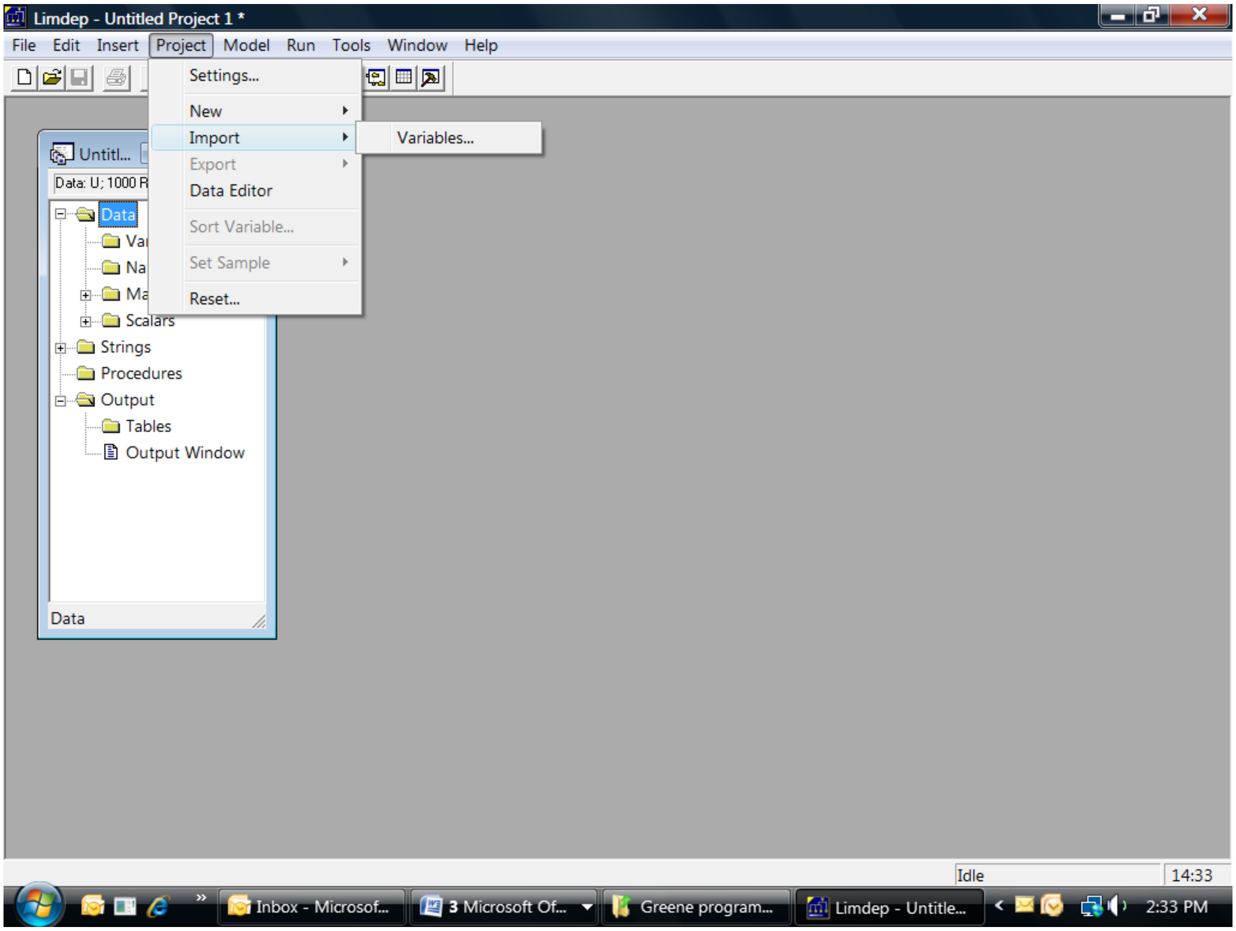
requires: 6.866 MBytes of memory

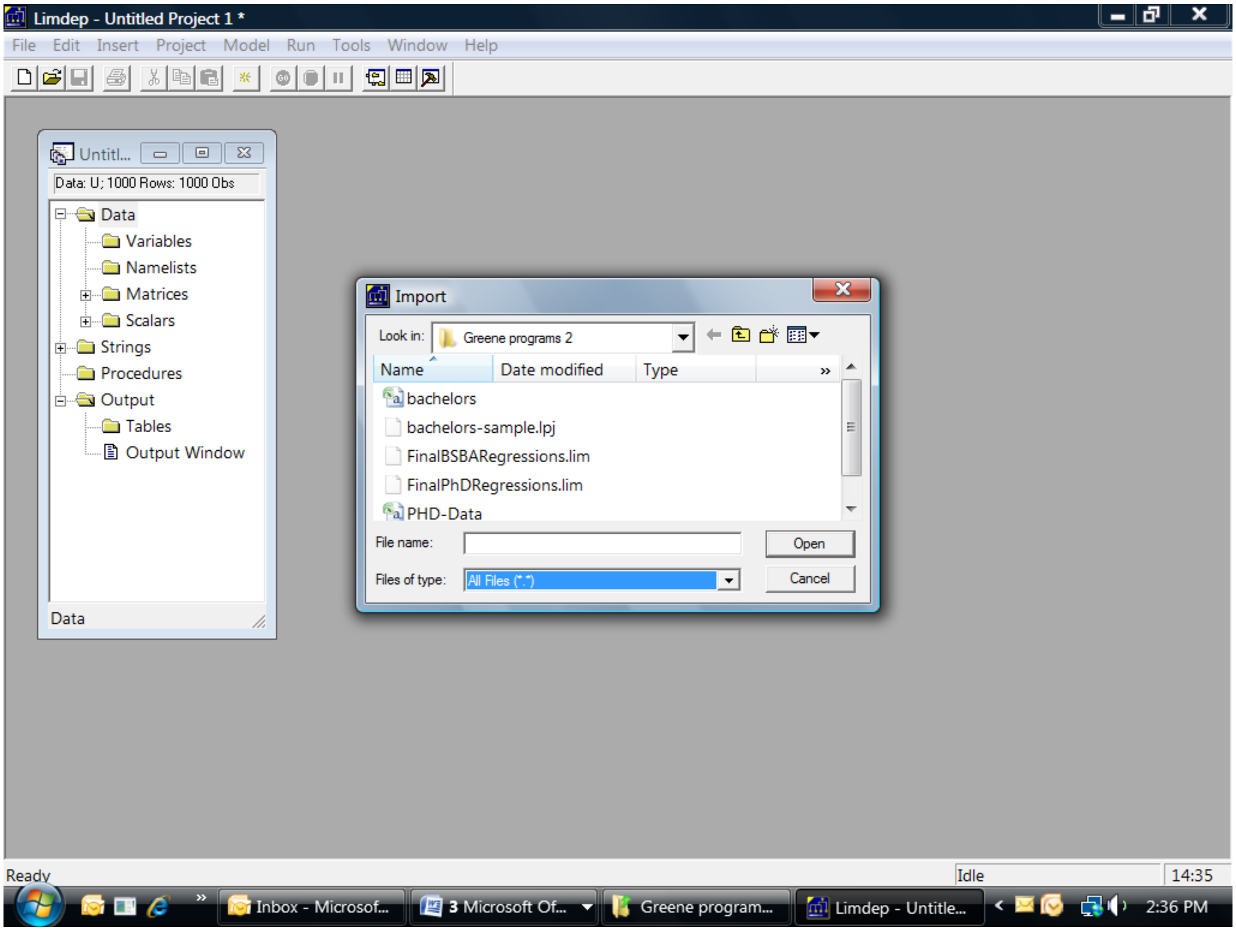
Dimensions

Rows: 1000 Columns: 900

WARNING: Any changes made will clear the contents of the data area

We could write a “READ” command to bring this text data file into LIMDEP but like EXCEL it can be imported into LIMDEP directly by clicking the Project button on the top ribbon, going to Import, and then clicking on Variables, from which the bachelors.cvs file can be located wherever it is stored (in our case in the “Greene programs 2” folder). Hitting the Open button will bring the data set into LIMDEP, which can be checked by clicking the “Activate Data Editor” button, which is second from the right on the tool bar or go to Data Editor in the Window’s menu, as described and demonstrated in Module One, Part Two.





In addition to a visual inspection of the data via the "Activate Data Editor," we use the "dstat" command to check the descriptive statistics. First, however, we need to remove the two years (1991 and 1992) for which no data are available for the degree moving average measure. This is done with the "Reject" command. In our "File:New Text/Command Document" (which was described in Module One, Part Two), we have

```
reject ; year < 1993 $
dstat;rhs=*$
```

which upon highlighting and pressing "Go" yields

```
--> reject ; year < 1993 $
--> dstat;rhs=*$
Descriptive Statistics
All results based on nonmissing observations.
==============================================================================
Variable     Mean        Std.Dev.     Minimum      Maximum        Cases Missing
==============================================================================
All observations in current sample
--------+---------------------------------------------------------------------
COLLEGE |   9.50000      5.19845      1.00000      18.0000          252       0
YEAR    |   1999.50      4.03915      1993.00      2006.00          252       0
DEGREES |   23.1111      19.2264      .000000      81.0000          252       0
DEGREBAR|   23.6528      18.0143      2.00000      62.4375          252       0
PUBLIC  |   1.77778      .416567      1.00000      2.00000          252       0
FACULTY |   6.51786      3.13677      2.00000      14.0000          252       0
BSCHOOL |   .388889      .488468      .000000      1.00000          252       0
T       |   1.50000      4.03915     -5.00000      8.00000          252       0
MA_DEG  |   23.1931      18.5540      1.33333      80.0000          252       0
```

CONSTANT COEFFICIENT REGRESSION

The constant coefficient panel data model for the faculty size data-generating process for bachelor degree-granting undergraduate departments is given by

$$Faculty\ size_{it} = \beta_1 + \beta_2 T_t + \beta_3 BA\&S_{it} + \beta_4 MEANBA\&S_i + \beta_5 PUBLIC_i$$
$$+ \beta_6 Bschl + \beta_7 MA\_Deg_{it} + \varepsilon_{it}$$

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 records. The LIMDEP OLS regression command that needs to be entered into the command document (again, following the procedure for opening the command document window shown in Module One, Part Two), including the standard error adjustment for clustering is

```
reject ; year < 1993 $
regress
;lhs=faculty;rhs=one,t,degrees,degrebar,public,bschool,MA_deg
;cluster=14$
```

Upon highlighting and hitting the “Go” button the Output file shows the following results

```
--> reject ; year < 1993 $
--> regress;lhs=faculty;rhs=one,t,degrees,degrebar,public,bschool,MA_deg
    ;cluster=14$

+----------------------------------------------------+
| Ordinary    least squares regression               |
| LHS=FACULTY  Mean                 =   6.517857     |
|              Standard deviation   =   3.136769     |
|              Number of observs.   =        252     |
| Model size   Parameters           =          7     |
|              Degrees of freedom   =        245     |
| Residuals    Sum of squares       =   868.4410     |
|              Standard error of e  =   1.882726     |
| Fit          R-squared            =   .6483574     |
|              Adjusted R-squared   =   .6397458     |
| Model test   F[  6,   245] (prob) =  75.29 (.0000) |
| Diagnostic   Log likelihood       =  -513.4686     |
|              Restricted(b=0)      =  -645.1562     |
|              Chi-sq [  6]  (prob) = 263.38 (.0000) |
| Info criter. LogAmemiya Prd. Crt. =   1.292840     |
|              Akaike Info. Criter. =   1.292826     |
|              Bayes Info. Criter.  =   1.390866     |
| Autocorrel   Durbin-Watson Stat.  =   .3295926     |
|              Rho = cor[e,e(-1)]   =   .8352037     |
| Model was estimated Jul 16, 2009 at 04:21:28PM     |
+----------------------------------------------------+
+---------------------------------------------------------------------+
| Covariance matrix for the model is adjusted for data clustering.    |
| Sample of    252 observations contained     18 clusters defined by  |
|     14 observations (fixed number) in each cluster.                 |
| Sample of    252 observations contained      1 strata defined by    |
|    252 observations (fixed number) in each stratum.                 |
+---------------------------------------------------------------------+
+--------+-------------+----------------+--------+--------+----------+
|Variable| Coefficient  | Standard Error |t-ratio |P[|T|>t]| Mean of X|
+--------+-------------+----------------+--------+--------+----------+
|Constant|    10.1397***       .91063       11.135   .0000            |
|T       |    -.02809          .02227       -1.261   .2083     1.50000|
|DEGREES |    -.01636          .01866        -.877   .3814     23.1111|
|DEGREBAR|     .10832***       .03378        3.206   .0015     23.6528|
|PUBLIC  |   -3.86239***       .56950       -6.782   .0000     1.77778|
|BSCHOOL |     .58112          .94253         .617   .5381      .38889|
|MA_DEG  |     .03780**        .01810        2.089   .0377     23.1931|
+--------+------------------------------------------------------------+
| Note: ***, **, * = Significance at 1%, 5%, 10% level.               |
+---------------------------------------------------------------------+
```

Contemporaneous degrees have little to do with current faculty size but both overall number of degrees awarded (the school means) and the moving average of degrees (MA_DEG) have significant effects. It takes an increase of 26 or 27 bachelor degrees in the moving average to expect just one more faculty position. Whether it is a public or a private college is highly significant. Moving from a public to a private college lowers predicted faculty size by nearly four members for otherwise comparable institutions. There is an insignificant erosion of tenured and tenure-track faculty size over time. Finally, while economics departments in colleges with a business school tend to have a larger permanent faculty, ceteris paribus, the effect is small and insignificant.

FIXED-EFFECTS REGRESSION

To estimate the fixed-effects model we can either insert seventeen (0,1) covariates to capture the unique effect of each of the 18 colleges (where each of the 17 dummy coefficients are measured relative to the constant term) or the insert of 18 dummy variables with no overall constant term in the OLS regression. The results for the other coefficients and for R^2 will be identical either way, while the the constant terms, since they measure the difference of each college from the 18th in the first case, or the difference of all 18 from zero in the second, will differ. This difference is inconsequential for the regression of interest. Which way the model is estimated is purely a matter of convenience and preference.

An important implication of the fixed effects specification is that no time invariant variables can be included in the equation because they would be perfectly correlated with the respective college dummies. Thus, the overall school mean number of degrees, the public or private dummy variables, and business school dummy variables must all be excluded from the fixed effects model.

A LIMDEP (NLOGIT) program to be run from the Test/Command Document, including the commands to create the dummy variables then run the regression is shown below. (An alternative, more compact way to create the dummies and run the regression is shown in the Appendix.)

```
reject ; year < 1993 $

create
;Col1=college=1
;Col2=college=2
;Col3=college=3
;Col4=college=4
;Col5=college=5
;Col6=college=6$
create
;Col7=college=7
;Col8=college=8
;Col9=college=9
;Col10=college=10
;Col11=college=11
;Col12=college=12$
create
;Col13=college=13
;Col14=college=14
;Col15=college=15
;Col16=college=16
;Col17=college=17
;Col18=college=18$

regress;lhs=faculty;rhs=one,t,degrees,MA_deg,
Col1,Col2,Col3,Col4,Col5,Col6,Col7,Col8,Col9,
Col10,Col11,Col12,Col13,Col14,Col15,Col16,Col17; cluster=14$
```

The resulting regression information appearing in the output window is

```
+---------------------------------------------------------------------+
| Covariance matrix for the model is adjusted for data clustering.    |
| Sample of    252 observations contained     18 clusters defined by  |
|     14 observations (fixed number) in each cluster.                 |
+---------------------------------------------------------------------+

----------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=FACULTY  Mean                 =        6.51786
             Standard deviation   =        3.13677
             Number of observs.   =            252
Model size   Parameters           =             21
             Degrees of freedom   =            231
Residuals    Sum of squares       =      146.63709
             Standard error of e  =         .79674
Fit          R-squared            =         .94062
             Adjusted R-squared   =         .93548
Model test   F[ 20,   231] (prob) =   183.0(.0000)
Diagnostic   Log likelihood       =     -289.34751
             Restricted(b=0)      =     -645.15625
             Chi-sq [ 20]  (prob) =   711.6(.0000)
Info criter. LogAmemiya Prd. Crt. =        -.37441
             Akaike Info. Criter. =        -.37480
             Bayes Info. Criter.  =        -.08068
Model was estimated on Sep 23, 2009 at 06:44:38 PM
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  t-ratio  P[|T|>t]   Mean of X
--------+-------------------------------------------------------------
Constant|    2.69636***       .15109       17.846   .0000
       T|    -.02853          .02245       -1.271   .2051      1.50000
 DEGREES|    -.01608          .01521       -1.058   .2913      23.1111
  MA_DEG|     .03985***       .01485        2.683   .0078      23.1931
    COL1|    5.77747***       .76816        7.521   .0000       .05556
    COL2|     .15299***       .01343       11.392   .0000       .05556
    COL3|    4.29759***       .55420        7.755   .0000       .05556
    COL4|    6.28973***       .65533        9.598   .0000       .05556
    COL5|    4.91094***       .56987        8.618   .0000       .05556
    COL6|    5.02016***       .02561      196.041   .0000       .05556
    COL7|    1.21384***       .01321       91.876   .0000       .05556
    COL8|     .77797***       .06785       11.466   .0000       .05556
    COL9|    3.16474***       .06270       50.478   .0000       .05556
   COL10|    2.86345***       .15540       18.427   .0000       .05556
   COL11|    5.15181***       .02403      214.385   .0000       .05556
   COL12|    -.06802***       .02153       -3.160   .0018       .05556
   COL13|    3.98895***      1.01415        3.933   .0001       .05556
   COL14|    -.63196***       .11986       -5.272   .0000       .05556
   COL15|    8.25859***       .47255       17.477   .0000       .05556
   COL16|    8.00970***       .55461       14.442   .0000       .05556
   COL17|     .43544          .59258         .735   .4632       .05556
--------+-------------------------------------------------------------
Note: ***, **, * = Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------
```

Once again, contemporaneous degrees is not a driving force in faculty size. An F test is not needed to assess if at least one of the 17 colleges differ from college 18. With the exception of college 17, each of the other colleges are significantly different. The moving average of degrees is again significant.

The preceding approach, of computing all the dummy variables and building them into the regression, is likely to become unduly cumbersome if the number of colleges (units) is very large. Most contemporary software, including LIMDEP will do this computation automatically without explicitly computing the dummy variables and including them in the equation. As an alternative to specifying all the dummies in the regression command, the same results can be obtained with the simpler "FixedEffects" command:

```
regress;lhs=faculty;rhs=one,t,degrees,MA_deg
;Panel;Str=College
;FixedEffects;Robust$

----------------------------------------------------------------------
Least Squares with Group Dummy Variables..........
Ordinary     least squares regression ............
LHS=FACULTY  Mean                 =       6.51786
             Standard deviation   =       3.13677
             Number of observs.   =           252
Model size   Parameters           =            21
             Degrees of freedom   =           231
Residuals    Sum of squares       =     146.63709
             Standard error of e  =        .79674
Fit          R-squared            =        .94062
             Adjusted R-squared   =        .93548
Model test   F[ 20,   231] (prob) =   183.0(.0000)
Diagnostic   Log likelihood       =    -289.34751
             Restricted(b=0)      =    -645.15625
             Chi-sq [ 20]  (prob) =   711.6(.0000)
Info criter. LogAmemiya Prd. Crt. =       -.37441
             Akaike Info. Criter. =       -.37480
             Bayes Info. Criter.  =       -.08068
Model was estimated on Sep 23, 2009 at 06:44:38 PM
Estd. Autocorrelation of e(i,t)   =       .293724
Robust cluster corrected covariance matrix used
Panel:Groups Empty      0,     Valid data       18
             Smallest  14,     Largest          14
             Average group size in panel     14.00
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  t-ratio  P[|T|>t]   Mean of X
--------+-------------------------------------------------------------
       T|    -.02853          .02245       -1.271   .2050      1.50000
 DEGREES|    -.01608          .01521       -1.058   .2912      23.1111
  MA_DEG|     .03985***       .01485        2.683   .0078      23.1931
--------+-------------------------------------------------------------
Note: ***, **, * = Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------
```

RANDOM-EFFECTS REGRESSION

Finally, consider the random-effects model in which we employ Mundlak's (1978) approach to estimating panel data. The Mundlak model posits that the fixed effects in the equation, β_{1i} , can be projected upon the group means of the time-varying variables, so that

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

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. Logically, adding the means to the equations picks up the correlation between the school effects and the other variables. We could not incorporate the mean number of degrees awarded in the fixed-effects model (because it was time invariant) but this variable plays a critical role in the Mundlak approach to panel data modeling and estimation.

The random effects model for BA and BS degree-granting undergraduate departments is

$$FACULTY\ size_{it} = \beta_1 + \beta_2 T_t + \beta_3 BA\&S_{it} + \beta_4 MEANBA\&S_i + \beta_5 MOVAVBA\&BS$$
$$+ \beta_6 PUBLIC_i + \beta_7 Bschl + \varepsilon_{it} + u_i$$

where error term ε is *iid* over time, $E(\varepsilon_{it}^2|\mathbf{x}_{it}) = \sigma^2$ for $I = 18$ and $T_i = 14$ and $E[u_i^2] = \theta^2$ for $I = 18$. The LIMDEP program to be run from the Text/Command Document (with 1991 and 1992 data suppressed) is

```
regress
;lhs=faculty
;rhs=one,t,degrees,degrebar,public,bschool,MA_deg
;pds=14
;panel
;random
;robust$
```

The resulting regression information appearing in the output window is

```
--> regress
    ;lhs=faculty;rhs=one,t,degrees,degrebar,public,bschool,MA_deg
    ;pds=14;panel;random;robust$
```

```
-----------------------------------------------------------------------
OLS Without Group Dummy Variables.................
Ordinary     least squares regression ............
LHS=FACULTY  Mean                 =       6.51786
             Standard deviation   =       3.13677
             Number of observs.   =           252
Model size   Parameters           =             7
             Degrees of freedom   =           245
Residuals    Sum of squares       =     868.44104
             Standard error of e  =       1.88273
Fit          R-squared            =        .64836
             Adjusted R-squared   =        .63975
Model test   F[  6,   245] (prob) =    75.3(.0000)
Diagnostic   Log likelihood       =    -513.46861
             Restricted(b=0)      =    -645.15625
             Chi-sq [  6]  (prob) =   263.4(.0000)
Info criter. LogAmemiya Prd. Crt. =       1.29284
             Akaike Info. Criter. =       1.29283
             Bayes Info. Criter.  =       1.39087
Model was estimated on Sep 23, 2009 at 07:17:22 PM
Panel Data Analysis of FACULTY           [ONE way]
               Unconditional ANOVA (No regressors)
Source         Variation  Deg. Free.   Mean Square
Between       2312.22321         17.     136.01313
Residual       157.44643        234.        .67285
Total         2469.66964        251.       9.83932
--------+--------------------------------------------------------------
Variable| Coefficient    Standard Error  t-ratio  P[|T|>t]   Mean of X
--------+--------------------------------------------------------------
       T|    -.02809         .03030        -.927   .3549      1.50000
 DEGREES|    -.01636         .02334        -.701   .4839      23.1111
DEGREBAR|     .10832***      .02047        5.293   .0000      23.6528
  PUBLIC|   -3.86239***      .29652      -13.026   .0000      1.77778
 BSCHOOL|     .58112**       .25115        2.314   .0215       .38889
  MA_DEG|     .03780         .02907        1.300   .1947      23.1931
Constant|    10.1397***      .52427       19.341   .0000
--------+--------------------------------------------------------------
Note: ***, **, * = Significance at 1%, 5%, 10% level.
-----------------------------------------------------------------------
```

```
+----------------------------------------------------+
| Panel:Groups   Empty       0,   Valid data      18 |
|                Smallest  14,    Largest         14 |
|                Average group size            14.00 |
| There are  3 vars. with no within group variation. |
| DEGREBAR PUBLIC   BSCHOOL                          |
+----------------------------------------------------+

----------------------------------------------------------------------
Random Effects Model: v(i,t)    = e(i,t) + u(i)
Estimates:  Var[e]              =       .643145
            Var[u]              =      2.901512
            Corr[v(i,t),v(i,s)] =       .818559
Lagrange Multiplier Test vs. Model (3) =1096.30
( 1 degrees of freedom, prob. value =  .000000)
(High values of LM favor FEM/REM over CR model)
Baltagi-Li form of LM Statistic =       1096.30
            Sum of Squares           868.488173
            R-squared                   .648338
Robust cluster corrected covariance matrix used
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
       T|    -.02853          .02146       -1.329   .1838      1.50000
 DEGREES|    -.01609          .01793        -.897   .3696      23.1111
DEGREBAR|     .10610***       .03228        3.287   .0010      23.6528
  PUBLIC|   -3.86365***       .54685       -7.065   .0000      1.77778
 BSCHOOL|     .58176          .90497         .643   .5203       .38889
  MA_DEG|     .03981**        .01728        2.305   .0212      23.1931
Constant|    10.1419***       .87456       11.597   .0000
--------+-------------------------------------------------------------
Note: ***, **, * = Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------
```

The marginal effect of an additional economics major is again insignificant but slightly negative within the sample. Both the short-term moving average number and long-term average number of bachelor degrees are significant. A long-term increase of about 10 students earning degrees in economics is required to predict that one more tenured or tenure-track faculty member is in a department. Ceteris paribus, economics departments at private institutions are smaller than comparable departments at public schools by a large and significant number of four members. Whether there is a business school present is insignificant. There is no meaningful trend in faculty size.

CONCLUDING REMARKS

The goal of this hands-on component of this third of four modules is to enable economic education researchers to make use of panel data for the estimation of constant coefficient, fixed-effects and random-effects panel data models in LIMDEP (NLOGIT). It was not intended to explain all of the statistical and econometric nuances associated with panel data analysis. For this an intermediate level econometrics textbook (such as Jeffrey Wooldridge, *Introductory Econometrics*) or advanced econometrics textbook (such as William Greene, *Econometric Analysis*) should be consulted.

APPENDIX:

ALTERNATIVES FOR CREATING COLLEGE DUMMY VARIABLES AND RUNNING REGRESSIONS IN LIMDEP (NLOGIT)

There are two alternative ways to create the college dummy variables for use in the regression.

First is the “`create ; expand(college)$`” command, where COLLEGE is expanded as _COLLEG_, with the following resulting output:

```
COLLEGE  was expanded as _COLLEG_.
Largest value =  18.  18 New variables were created.
Category
  1  New variable = COLLEG01    Frequency=      14
  2  New variable = COLLEG02    Frequency=      14
  3  New variable = COLLEG03    Frequency=      14
  4  New variable = COLLEG04    Frequency=      14
  5  New variable = COLLEG05    Frequency=      14
  6  New variable = COLLEG06    Frequency=      14
  7  New variable = COLLEG07    Frequency=      14
  8  New variable = COLLEG08    Frequency=      14
  9  New variable = COLLEG09    Frequency=      14
 10  New variable = COLLEG10    Frequency=      14
 11  New variable = COLLEG11    Frequency=      14
 12  New variable = COLLEG12    Frequency=      14
 13  New variable = COLLEG13    Frequency=      14
 14  New variable = COLLEG14    Frequency=      14
 15  New variable = COLLEG15    Frequency=      14
 16  New variable = COLLEG16    Frequency=      14
 17  New variable = COLLEG17    Frequency=      14
 18  New variable = COLLEG18    Frequency=      14
Note, this is a complete set of dummy variables.  If
you use this set in a regression, drop the constant.
```

The second method for creating dummies and running the regression is an even more condensed, and as yet an undocumented feature in the LIMDEP (NLOGIT) manual:

```
regress;lhs=faculty;rhs=one,t,degrees,MA_deg,expand(college)
       ;cluster=college$
```

which produces the same results as the fixed effects regression command we used at the beginning of this duscussion.

REFERENCES

Becker, William, William Greene and John Siegfried (2009). “Does Teaching Load Affect Faculty Size? ” Working Paper (July).

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

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

Wooldridge, Jeffrey (2009). *Introductory Econometrics*. 4th Edition, Mason OH: South-Western.

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

Part Three of Module Three provides a cookbook-type demonstration of the steps required to use STATA in panel data analysis. Users of this model need to have completed Module One, Parts One and Three, and Module Three, Part One. That is, from Module One users are assumed to know how to get data into STATA, recode and create variables within STATA, and run and interpret regression results. They are also expected to know how to test linear restrictions on sets of coefficients as done in Module One, Parts One and Three. Module Three, Parts Two and Four demonstrate in LIMDEP and SAS what is done here in STATA.

THE CASE

As described in Module Three, Part One, Becker, Greene and Siegfried (2009) examine the extent to which undergraduate degrees (BA and BS) in economics or Ph.D. 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. Here we retrace their analysis for the institutions that offer only the bachelor degree. We provide and demonstrate the STATA code necessary to duplicate their results.

DATA FILE

The following panel data are provided in the **comma separated values** (CSV) text file "bachelors.csv", which will automatically open in EXCEL by simply double clicking on it after it has been downloaded to your hard drive. Your EXCEL spreadsheet should look like this:

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

"Year" runs from 1996 through 2006.

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

"DegreBar" 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.

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

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

"T" is the 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).

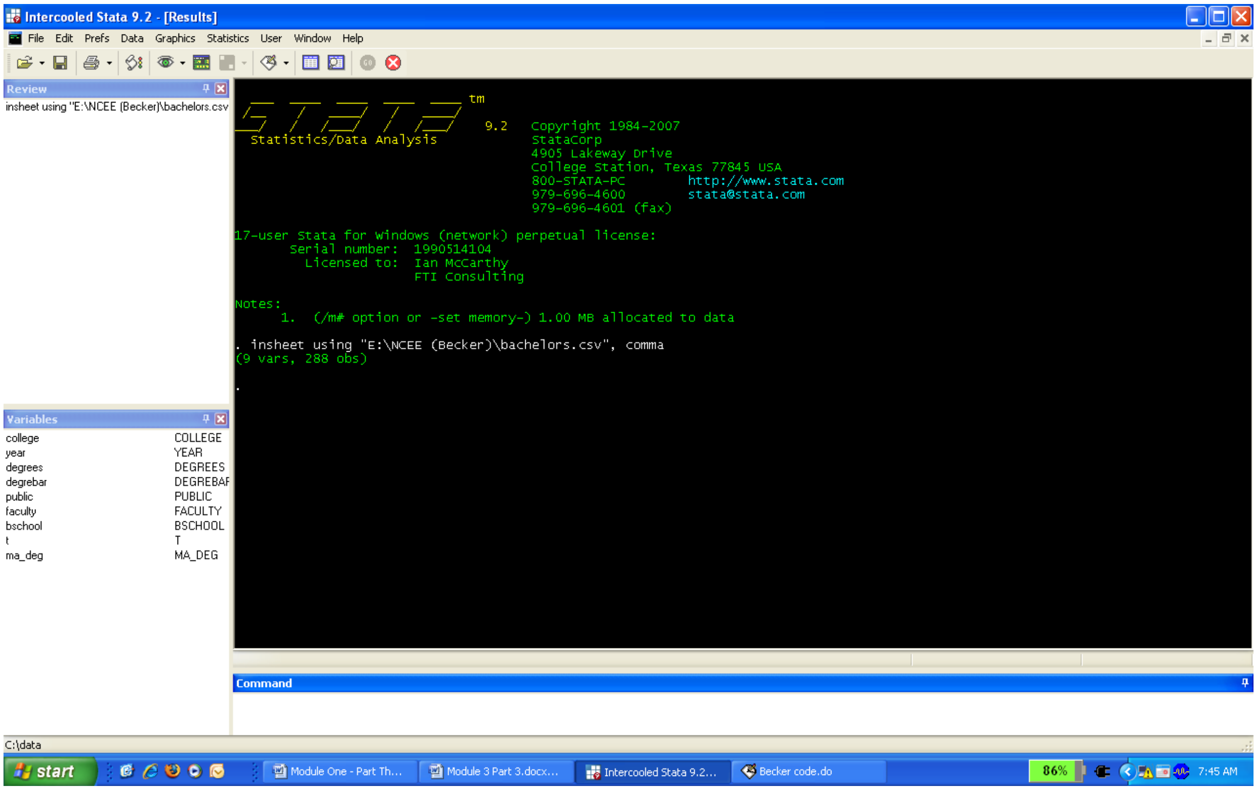
College	Year	Degrees	DegreBar	Public	Faculty	Bschol	T	MA_Deg
1	1991	50	47.375	2	11	1	-7	0
1	1992	32	47.375	2	8	1	-6	0
1	1993	31	47.375	2	10	1	-5	37.667
1	1994	35	47.375	2	9	1	-4	32.667
↓	↓	↓	↓	↓	↓	↓		↓
1	2003	57	47.375	2	7	1	5	56
1	2004	57	47.375	2	10	1	6	55.667
1	2005	57	47.375	2	10	1	7	57
1	2006	51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

If you opened this CSV file in a word processor or text editing program, it would show that each of the 289 lines (including the headers) corresponds to a row in the EXCEL table, but variable values would be separated by commas and not appear neatly one on top of the other as in EXCEL.

As discussed in Module One, Part Three, you can read the CSV file into STATA by typing the following command into the command window and pressing enter:

```
insheet using “E:\NCEE (Becker)\bachelors.csv”, comma
```

In this case, the “bachelors.csv” file is saved in the file “E:\NCEE (Becker)” but this will vary by user. For these data, the default memory allocated by STATA should be sufficient. After entering the above command in the command window and pressing enter, you should see the following screen:



insheet using "E:\NCEE (Becker)\bachelors.csv

```
                                          tm
  ___  ____  ____  ____  ____
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/       9.2   Copyright 1984-2007
  Statistics/Data Analysis                StataCorp
                                          4905 Lakeway Drive
                                          College Station, Texas 77845 USA
                                          800-STATA-PC        http://www.stata.com
                                          979-696-4600        stata@stata.com
                                          979-696-4601 (fax)

17-user Stata for Windows (network) perpetual license:
       Serial number:  1990514104
         Licensed to:  Ian McCarthy
                       FTI Consulting

Notes:
      1.  (/m# option or -set memory-) 1.00 MB allocated to data

. insheet using "E:\NCEE (Becker)\bachelors.csv", comma
(9 vars, 288 obs)

.
```

college	COLLEGE
year	YEAR
degrees	DEGREES
degrebar	DEGREBAF
public	PUBLIC
faculty	FACULTY
bschool	BSCHOOL
t	T
ma_deg	MA_DEG

STATA indicates that the data consist of 9 variables and 288 observations. In addition to a visual inspection of the data via the “browse” command, you can use the "summarize" command to check the descriptive statistics. First, however, we need to remove the two years (1991 and 1992) for which no data are available for the degree moving average measure. This is done with the "drop if" command. In the command window, type:

```
drop if year < 1993
summarize
```

which upon pressing enter yields the following summary statistics:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     college |       252         9.5    5.198452          1         18
        year |       252      1999.5    4.039151       1993       2006
     degrees |       252    23.11111    19.22636          0         81
    degrebar |       252    23.65278    18.01427          2    62.4375
      public |       252    1.777778    .4165671          1          2
-------------+--------------------------------------------------------
     faculty |       252    6.517857    3.136769          2         14
     bschool |       252    .3888889    .4884682          0          1
           t |       252         1.5    4.039151         -5          8
      ma_deg |       252    23.19312    18.55398   1.333333         80
```

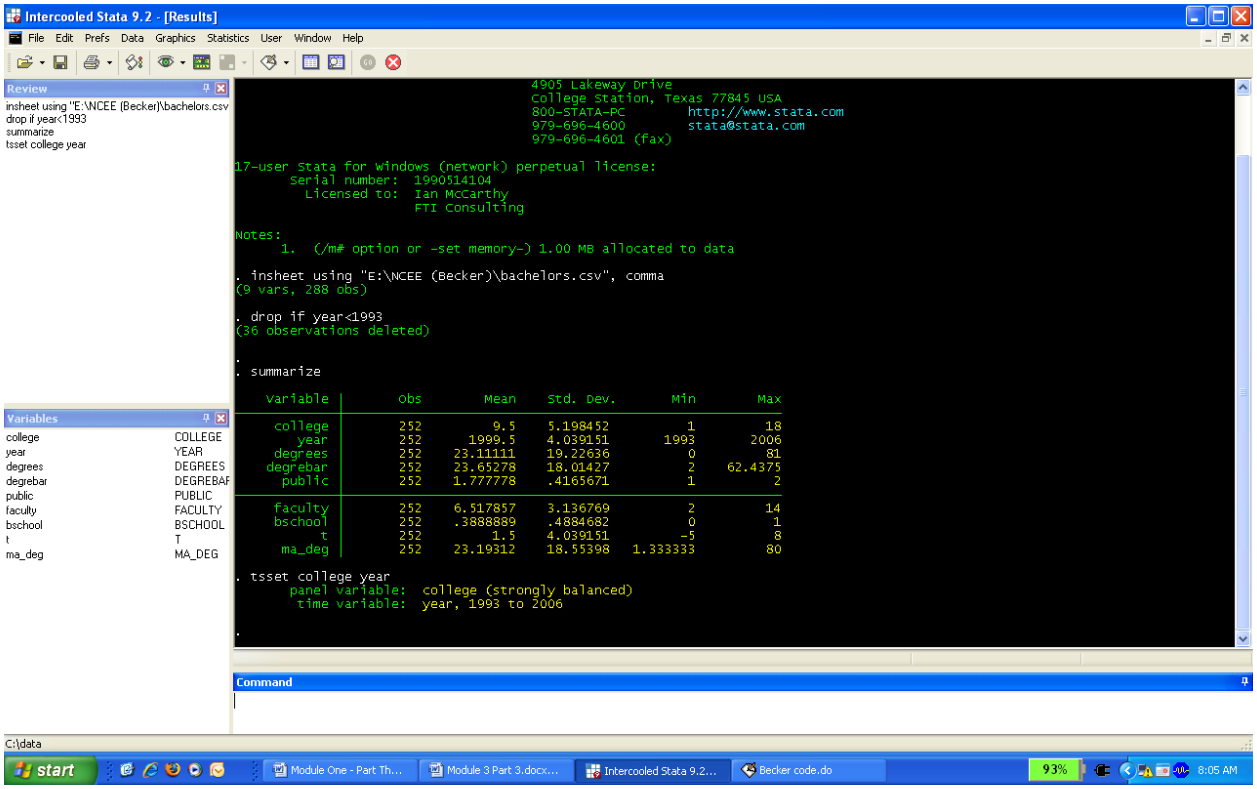
By default, STATA essentially considers all data as cross-sectional. Since we are working with panel data in this case, we need to indicate to STATA that there is a time-series component to our dataset. This is done with the “tsset” command. The general syntax for the “tsset” command with panel data is:

tsset “panel variable” “time variable”

In this case, our panel variable is college and our time variable is year, so the relevant command is:

tsset college year

After typing the above command into STATA’s command window and pressing enter, you should see the following screen:



```
insheet using "E:\NCEE (Becker)\bachelors.csv
drop if year<1993
summarize
tsset college year
```

```
college          COLLEGE
year             YEAR
degrees          DEGREES
degrebar         DEGREBAF
public           PUBLIC
faculty          FACULTY
bschool          BSCHOOL
t                T
ma_deg           MA_DEG
```

```
                                        4905 Lakeway Drive
                                        College Station, Texas 77845 USA
                                        800-STATA-PC        http://www.stata.com
                                        979-696-4600        stata@stata.com
                                        979-696-4601 (fax)

17-user Stata for Windows (network) perpetual license:
       Serial number:  1990514104
         Licensed to:  Ian McCarthy
                       FTI Consulting

Notes:
      1.  (/m# option or -set memory-) 1.00 MB allocated to data

. insheet using "E:\NCEE (Becker)\bachelors.csv", comma
(9 vars, 288 obs)

. drop if year<1993
(36 observations deleted)

.
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     college |       252         9.5    5.198452          1         18
        year |       252      1999.5    4.039151       1993       2006
     degrees |       252    23.11111    19.22636          0         81
    degrebar |       252    23.65278    18.01427          2    62.4375
      public |       252    1.777778    .4165671          1          2
-------------+--------------------------------------------------------
     faculty |       252    6.517857    3.136769          2         14
     bschool |       252    .3888889    .4884682          0          1
           t |       252         1.5    4.039151         -5          8
      ma_deg |       252    23.19312    18.55398   1.333333         80

. tsset college year
       panel variable:  college (strongly balanced)
        time variable:  year, 1993 to 2006

.
```

This indicates that STATA recognizes a strongly balanced panel (i.e., the same number of years for each college) with observations for each panel from 1993 through 2006. Note that we could also use the variable "t" as our time variable.

In general, we ***must*** "tsset" the data before we can utilize any of STATA's time-series or panel data commands (for example, the "xtreg" command presented below). Our time variable should also be appropriately spaced. For example, if we have yearly data, but our time variable was recorded in a daily format (e.g., 1/1/1999, 1/1/2000, 1/1/2002, etc.), we would want to reformat this variable as a yearly variable rather than daily. Correctly formatting the time variable is important to ensure the various time-series commands in STATA work properly. For more detail on formats and other options for the "tsset" command type "help tsset" into STATA's command window.

CONSTANT COEFFICIENT REGRESSION

The constant coefficient panel data model for the faculty size data-generating process for bachelor degree-granting undergraduate departments is given by

$$\begin{aligned}&Faculty\ size_{it} = \beta_1 + \beta_2 T_t + \beta_3 BA\&S_{it} + \beta_4 MEANBA\&S_i + \beta_5 PUBLIC_i\\&+ \beta_6 Bschl + \beta_7 MA\_Deg_{it} + \varepsilon_{it}\end{aligned}$$

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 records. The STATA OLS regression command that needs to be entered into the command window, including the standard error adjustment for clustering is

```
regress faculty t degrees degrebar public bschool ma_deg, cluster(college)
```

After typing the above command into the command window and pressing enter, the output window shows the following results:

```
. regress faculty t degrees degrebar public bschool ma_deg, cluster(college)

Linear regression                                      Number of obs =     252
                                                       F(  6,    17) =   27.70
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.6484
Number of clusters (college) = 18                      Root MSE      =  1.8827

------------------------------------------------------------------------------
             |               Robust
     faculty |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.0280875   .0222654    -1.26   0.224    -.0750634    .0188885
     degrees |  -.0163611   .0186579    -0.88   0.393    -.0557259    .0230037
    degrebar |   .1083201   .0337821     3.21   0.005     .0370461    .1795942
      public |  -3.862393   .5694961    -6.78   0.000    -5.063925   -2.660862
     bschool |   .5811154   .9425269     0.62   0.546    -1.407443    2.569673
      ma_deg |   .0378038   .0180966     2.09   0.052    -.0003767    .0759842
       _cons |   10.13974   .9106264    11.13   0.000     8.218486    12.06099
------------------------------------------------------------------------------
```

Contemporaneous degrees have little to do with current faculty size but both overall number of degrees awarded (the school means) and the moving average of degrees (MA_DEG) have significant effects. It takes an increase of 26 or 27 bachelor degrees in the moving average to expect just one more faculty position. Whether it is a public or a private college is highly significant. Moving from a public to a private college lowers predicted faculty size by nearly four members for otherwise comparable institutions. There is an insignificant erosion of tenured and tenure-track faculty size over time. Finally, while economics departments in colleges with a business school tend to have a larger permanent faculty, ceteris paribus, the effect is small and insignificant.

FIXED-EFFECTS REGRESSION

The fixed-effects model requires either the insertion of 17 (0,1) covariates to capture the unique effect of each of the 18 colleges (where each of the 17 dummy coefficients are measured relative to the constant term) or the insertion of 18 dummy variables with no constant term in the OLS regression. In addition, no time invariant variables can be included because they would be perfectly correlated with the respective college dummies. Thus, the overall mean number of degrees, the public or private dummy, and business school dummy cannot be included as regressors.

The STATA code, including the commands to create the dummy variables, is (two additional ways to estimate fixed-effects models in STATA are presented in the Appendix):

```
gen Col1=(college==1)
gen Col2=(college==2)
gen Col3=(college==3)
gen Col4=(college==4)
gen Col5=(college==5)
gen Col6=(college==6)
gen Col7=(college==7)
gen Col8=(college==8)
gen Col9=(college==9)
gen Col10=(college==10)
gen Col11=(college==11)
gen Col12=(college==12)
gen Col13=(college==13)
gen Col14=(college==14)
gen Col15=(college==15)
gen Col16=(college==16)
gen Col17=(college==17)
gen Col18=(college==18)

regress faculty t degrees ma_deg Col1-Col17, cluster(college)
```

The resulting regression information appearing in the output window is:

```
Linear regression                                      Number of obs =     252
                                                       F(  2,    17) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.9406
Number of clusters (college) = 18                      Root MSE      =  .79674

------------------------------------------------------------------------------
             |               Robust
     faculty |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.0285342    .022453    -1.27   0.221    -.0759059    .0188374
     degrees |  -.0160847   .0152071    -1.06   0.305    -.0481689    .0159995
      ma_deg |    .039847   .0148528     2.68   0.016     .0085103    .0711837
        Col1 |   5.777467   .7681565     7.52   0.000     4.156799    7.398136
        Col2 |   .1529889   .0134293    11.39   0.000     .1246555    .1813222
        Col3 |   4.297591   .5541956     7.75   0.000     3.128341    5.466842
        Col4 |   6.289728   .6553347     9.60   0.000     4.907093    7.672363
        Col5 |   4.910941   .5698701     8.62   0.000     3.708621    6.113262
        Col6 |   5.020157   .0256077   196.04   0.000     4.966129    5.074185
        Col7 |   1.213842   .0132117    91.88   0.000     1.185967    1.241716
        Col8 |   .7779701   .0678475    11.47   0.000     .6348244    .9211157
        Col9 |   3.164737   .0626958    50.48   0.000      3.03246    3.297013
       Col10 |   2.863453   .1553986    18.43   0.000      2.53559    3.191315
       Col11 |   5.151815   .0240307   214.39   0.000     5.101115    5.202515
       Col12 |  -.0680152   .0215257    -3.16   0.006    -.1134304      -.0226
       Col13 |   3.988947   1.014148     3.93   0.001     1.849282    6.128611
       Col14 |   -.631956   .1198635    -5.27   0.000    -.8848458   -.3790662
       Col15 |   8.258587   .4725524    17.48   0.000     7.261588    9.255585
       Col16 |   8.009696   .5546092    14.44   0.000     6.839573    9.179819
       Col17 |   .4354377   .5925837     0.73   0.472    -.8148046     1.68568
       _cons |   2.696364   .1510869    17.85   0.000     2.377598    3.015129
------------------------------------------------------------------------------
```

Once again, contemporaneous degrees is not a driving force in faculty size. There is no need to do an F test to assess if at least one of the 17 colleges differ from college 18. With the exception of college 17, each of the other colleges are significantly different. The moving average of degrees is again significant.

RANDOM-EFFECTS REGRESSION

Finally, consider the random-effects model in which we employ Mundlak's (1978) approach to estimating panel data. The Mundlak model posits that the fixed effects in the equation, β_{1i} , can be projected upon the group means of the time-varying variables, so that

$$\beta_{1i} \; = \; \beta_1 + \delta' \overline{x}_i + w_i$$

where $\overline{x}_i$ 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. Logically, adding the means to the equations picks up the correlation between the school effects and the other variables. We could not incorporate the mean number of degrees awarded in the fixed-effects model (because it was time invariant) but this variable plays a critical role in the Mundlak approach to panel data modeling and estimation.

The random effects model for BA and BS degree-granting undergraduate departments is

$$FACULTY\ size_{it} = \beta_1 + \beta_2 T_t + \beta_3 BA\&S_{it} + \beta_4 MEANBA\&S_i + \beta_5 MOVAVBA\&BS \\ + \beta_6 PUBLIC_i + \beta_7 Bschl + \varepsilon_{it} + u_i$$

where error term ε is *iid* over time, $E(\varepsilon_{it}^2|\mathbf{x}_{it}) = \sigma^2$ for $I = 18$ and $T_i = 14$ and $E[u_i^2] = \theta^2$ for $I = 18$. The STATA command to estimate this model is

```
xtreg faculty t degrees degrebar public bschool ma_deg, re cluster(college)
```

The resulting regression information appearing in the output window is[1]

```
. xtreg faculty t degrees degrebar public bschool ma_deg, re cluster(college)

Random-effects GLS regression                   Number of obs      =       252
Group variable (i): college                     Number of groups   =        18

R-sq:  within  = 0.0687                         Obs per group: min =        14
       between = 0.6878                                        avg =      14.0
       overall = 0.6483                                        max =        14

Random effects u_i ~ Gaussian                   Wald chi2(7)       =   1273.20
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

                               (Std. Err. adjusted for 18 clusters in college)
------------------------------------------------------------------------------
             |               Robust
     faculty |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.0285293   .0218015    -1.31   0.191    -.0712594    .0142007
     degrees |  -.0160879   .0147378    -1.09   0.275    -.0449734    .0127976
    degrebar |   .1060891   .0312801     3.39   0.001     .0447811     .167397
      public |  -3.863652   .5662052    -6.82   0.000    -4.973394    -2.75391
     bschool |   .5817666   .9406433     0.62   0.536     -1.26186    2.425394
      ma_deg |   .0398252     .01444     2.76   0.006     .0115233    .0681271
       _cons |   10.14196   .9033207    11.23   0.000     8.371485    11.91244
-------------+----------------------------------------------------------------
     sigma_u |  2.0564748
     sigma_e |  .79673873
         rho |  .86948846   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The marginal effect of an additional economics major is again insignificant but slightly negative within the sample. Both the short-term moving average number and long-term average number

[1] Note that the Wald statistic of 1273.20 is based on a test of all coefficients in the model (including the constant). This is inconsistent with the default Wald statistic reported in other regression results, including random-effects models without robust or clustered standard errors, where the default statistic is based on a test of all slope coefficients in the model. In the model estimated here, the Wald statistic based on a test of all slope coefficients equal to 0 is 198.55. I understand that the current version of STATA (STATA 11) now consistently presents the Wald statistic based on a test of all slope coefficients.

of bachelor degrees are significant. A long-term increase of about 10 students earning degrees in economics is required to predict that one more tenured or tenure-track faculty member is in a department. Ceteris paribus, economics departments at private institutions are smaller than comparable departments at public schools by a large and significant number of four members. Whether there is a business school present is insignificant. There is no meaningful trend in faculty size.

CONCLUDING REMARKS

The goal of this hands-on component of this third of four modules is to enable economic education researchers to make use of panel data for the estimation of constant coefficient, fixed-effects and random-effects panel data models in STATA. It was not intended to explain all of the statistical and econometric nuances associated with panel data analysis. For this an intermediate level econometrics textbook (such as Jeffrey Wooldridge, *Introductory Econometrics*) or advanced econometrics textbook (such as William Greene, *Econometric Analysis*) should be consulted.

APPENDIX: Alternative commands to estimate fixed-ffects models in STATA

Method 1 – Alternative Method of Creating Dummy variables

We estimated the above fixed-effects model after explicitly creating 18 different dummy variables. STATA also has a built in command ("xi") to create a sequence of dummy variables from a single categorical variable. To be consistent with the above model, we can first indicate to STATA which category it should omit when creating the college dummy variables by typing the following command into the command window and pressing enter:

```
char college[omit] 18
```

We can now automatically create the relevant college dummy variables and estimate the fixed-effects model all through one command:

```
xi: regress faculty t degrees ma_deg i.college, cluster(college)
```

The resulting regression information appearing in the output window is

```
. xi: regress faculty t degrees ma_deg i.college, cluster(college)

i.college         _Icollege_1-18      (naturally coded; _Icollege_18 omitted)

Linear regression                                      Number of obs =     252
                                                       F(  2,    17) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.9406
Number of clusters (college) = 18                      Root MSE      =  .79674

------------------------------------------------------------------------------
             |               Robust
     faculty |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.0285342    .022453    -1.27   0.221    -.0759059    .0188374
     degrees |  -.0160847   .0152071    -1.06   0.305    -.0481689    .0159995
      ma_deg |    .039847   .0148528     2.68   0.016     .0085103    .0711837
 _Icollege_1 |   5.777467   .7681565     7.52   0.000     4.156799    7.398136
 _Icollege_2 |   .1529889   .0134293    11.39   0.000     .1246555    .1813222
 _Icollege_3 |   4.297591   .5541956     7.75   0.000     3.128341    5.466842
 _Icollege_4 |   6.289728   .6553347     9.60   0.000     4.907093    7.672363
 _Icollege_5 |   4.910941   .5698701     8.62   0.000     3.708621    6.113262
 _Icollege_6 |   5.020157   .0256077   196.04   0.000     4.966129    5.074185
 _Icollege_7 |   1.213842   .0132117    91.88   0.000     1.185967    1.241716
 _Icollege_8 |   .7779701   .0678475    11.47   0.000     .6348244    .9211157
 _Icollege_9 |   3.164737   .0626958    50.48   0.000      3.03246    3.297013
_Icollege_10 |   2.863453   .1553986    18.43   0.000      2.53559    3.191315
_Icollege_11 |   5.151815   .0240307   214.39   0.000     5.101115    5.202515
_Icollege_12 |  -.0680152   .0215257    -3.16   0.006    -.1134304      -.0226
_Icollege_13 |   3.988947   1.014148     3.93   0.001     1.849282    6.128611
_Icollege_14 |   -.631956   .1198635    -5.27   0.000    -.8848458   -.3790662
_Icollege_15 |   8.258587   .4725524    17.48   0.000     7.261588    9.255585
_Icollege_16 |   8.009696   .5546092    14.44   0.000     6.839573    9.179819
_Icollege_17 |   .4354377   .5925837     0.73   0.472    -.8148046     1.68568
       _cons |   2.696364   .1510869    17.85   0.000     2.377598    3.015129
------------------------------------------------------------------------------
```

Method 2 – xtreg fe

STATA's "xtreg" command allows for various panel data models to be estimated. A random-effects model was presented above, but "xtreg" also estimates a fixed-effects model, a between-effects model, and various other models. The basic syntax for the "xtreg" command is:

```
xtreg "dependent variable" "independent variables", "model to be estimated" "other
options"
```

To estimate a random-effects model, the "model to be estimated" is "re." Similarly, to estimate a fixed-effects model, the "model to be estimated" is "fe." When using "xtreg" to estimate a fixed-effects model, STATA does not estimate the panel-specific dummy variables. This is a by-product of the type of estimator used by STATA. However, the coefficient estimates for the remaining independent variables are identical to those estimated by OLS with panel specific dummy variables. For example, using the "xtreg" command to estimate the fixed-effects model presented above, STATA provides the following output:

```
. xtreg faculty t degrees ma_deg, fe cluster(college) dfadj

Fixed-effects (within) regression               Number of obs      =       252
Group variable (i): college                     Number of groups   =        18

R-sq:  within  = 0.0687                         Obs per group: min =        14
       between = 0.4175                                        avg =      14.0
       overall = 0.3469                                        max =        14

                                                F(3,17)            =      2.66
corr(u_i, Xb)  = 0.4966                         Prob > F           =    0.0815

                               (Std. Err. adjusted for 18 clusters in college)
------------------------------------------------------------------------------
             |               Robust
     faculty |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.0285342    .022453    -1.27   0.221    -.0759059    .0188374
     degrees |  -.0160847   .0152071    -1.06   0.305    -.0481689    .0159995
      ma_deg |    .039847   .0148528     2.68   0.016     .0085103    .0711837
       _cons |   6.008218   .4400811    13.65   0.000     5.079728    6.936708
-------------+----------------------------------------------------------------
     sigma_u |  2.8596636
     sigma_e |  .79673873
         rho |  .92796654   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The additional "dfadj" option adjusts the cluster-robust standard error estimates to account for the transformation used by STATA in estimating the fixed-effects model (called the within transform). Although estimating the fixed-effects model with xtreg no longer provides estimates of the dummy variable coefficients, we see that the coefficient estimates and standard errors for the remaining variables are identical to those of an OLS regression with panel-specific dummies and cluster-robust standard errors.

REFERENCES

Becker, William, William Greene and John Siegfried (2009). “Does Teaching Load Affect Faculty Size? ” Working Paper (July).

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

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

Wooldridge, Jeffrey (2009). *Introductory Econometrics*. 4th Edition, Mason OH: South-Western.

MODULE THREE, PART FOUR: PANEL DATA ANALYSIS IN ECONOMIC EDUCATION RESEARCH USING SAS

Part Four of Module Three provides a cookbook-type demonstration of the steps required to use SAS in panel data analysis. Users of this model need to have completed Module One, Parts One and Four, and Module Three, Part One. That is, from Module One users are assumed to know how to get data into SAS, recode and create variables within SAS, and run and interpret regression results. They are also expected to know how to test linear restrictions on sets of coefficients as done in Module One, Parts One and Two. Module Three, Parts Two and Three demonstrate in LIMDEP and STATA what is done here in SAS.

THE CASE

As described in Module Three, Part One, Becker, Greene and Siegfried (2009) examine the extent to which undergraduate degrees (BA and BS) in economics or Ph.D. 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. Here we retrace their analysis for the institutions that offer only the bachelor degree. We provide and demonstrate the SAS code necessary to duplicate their results.

DATA FILE

The following panel data are provided in the **comma separated values** (CSV) text file "bachelors.csv", which will automatically open in EXCEL by simply double clicking on it after it has been downloaded to your hard drive. Your EXCEL spreadsheet should look like this:

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

"Year" runs from 1996 through 2006.

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

"DegreBar" 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.

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

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

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

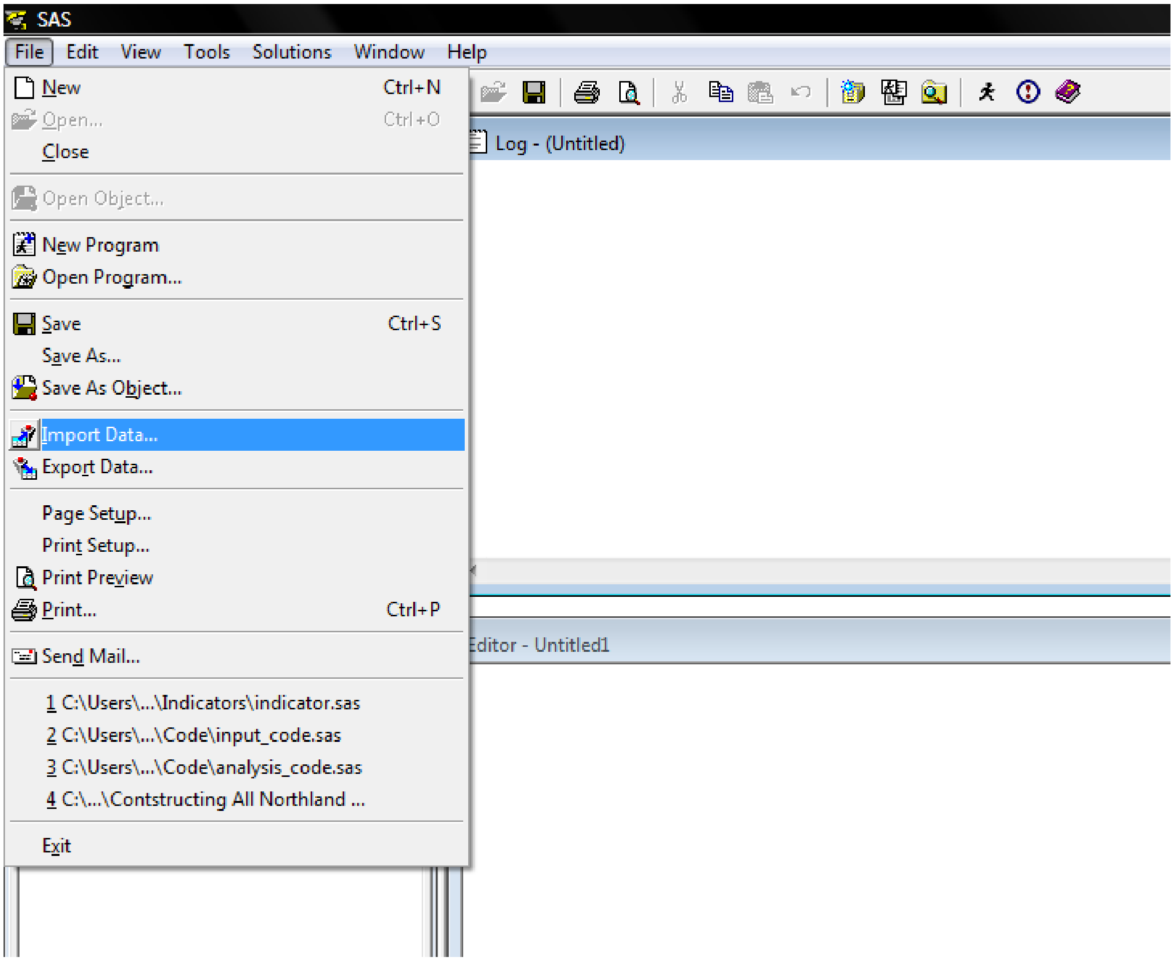
"MA_Deg" is a three-year moving average of degrees (unknown for the first two years).

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

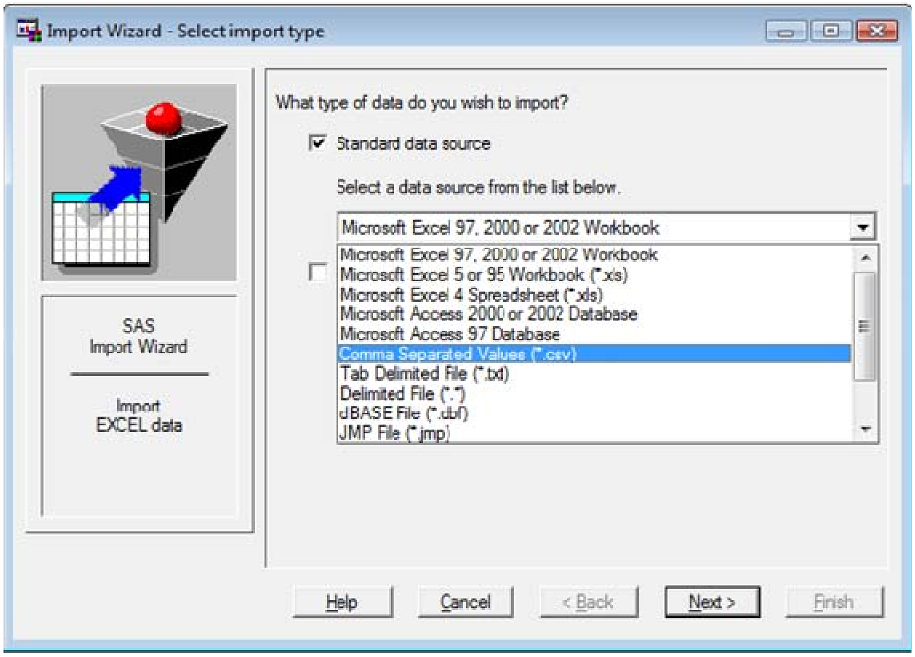
If you opened this CSV file in a word processor or text editing program, it would show that each of the 289 lines (including the headers) corresponds to a row in the EXCEL table, but variable values would be separated by commas and not appear neatly one on top of the other as in EXCEL.

As discussed in Module One, Part Two, SAS has a data matrix default restriction. This data set is sufficiently small, so there is no need to adjust the size of the matrix. We could write a "READ" command to bring this text data file into SAS similar to Module 1, Part 4, but like EXCEL, it can be imported into SAS directly by using the import wizard.

To import the data into SAS, click on 'File' at the top left corner of your screen in SAS, and then click 'Import Data'.



This will initialize the Import Wizard pop-up screen. Since the data is comma separated values, scroll down under the 'Select data source below.' tab and click on 'Comma Separated Values (*.csv)' as shown below.



Import Wizard - Select import type

What type of data do you wish to import?

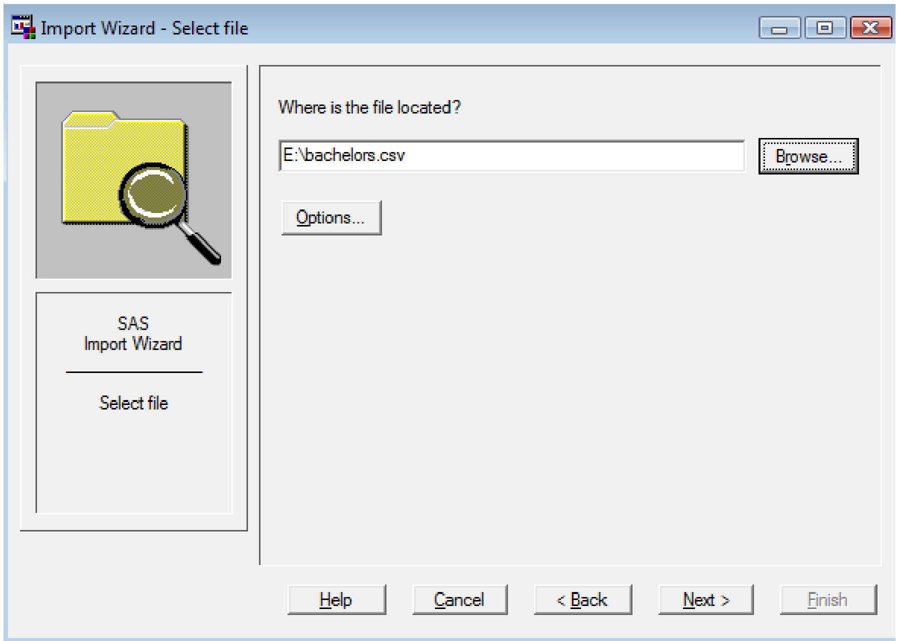
☑ Standard data source

Select a data source from the list below.

Microsoft Excel 97, 2000 or 2002 Workbook

SAS
Import Wizard
Import
EXCEL data
□
Microsoft Excel 97, 2000 or 2002 Workbook
Microsoft Excel 5 or 95 Workbook (*.xls)
Microsoft Excel 4 Spreadsheet (*.xls)
Microsoft Access 2000 or 2002 Database
Microsoft Access 97 Database
Comma Separated Values (*.csv)
Tab Delimited File (*.txt)
Delimited File (*.*)
dBASE File (*.dbf)
JMP File (*.jmp)
Help
Cancel
< Back
Next >
Finish

Click 'Next', and then provide the location from which the file bachelor.cvs can be located wherever it is stored (in our case in "e:\bachelor.csv").

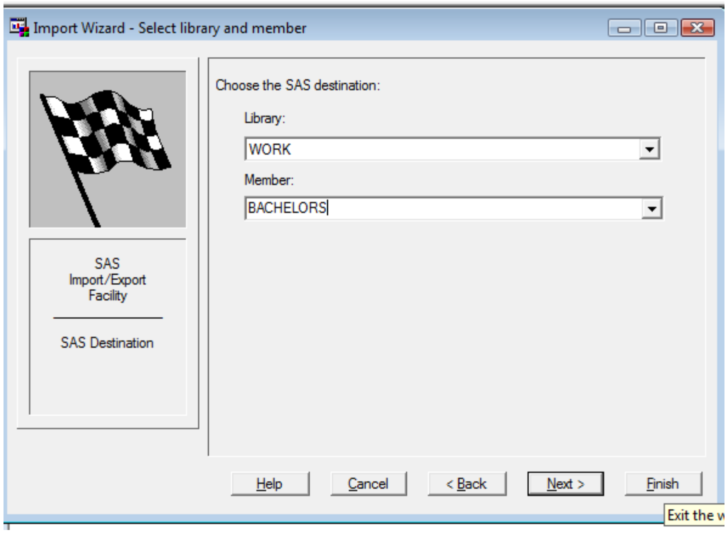


Import Wizard - Select file

Where is the file located?

E:\bachelors.csv

To finish importing the data, click ‘Next’, and then name the dataset, known as a member in SAS, to be stored in the temporary library called ‘WORK’. Recall that a library is simply a folder to store datasets and output. I named the file ‘BACHELORS’ as seen below. Hitting the Finish button will bring the data set into SAS.

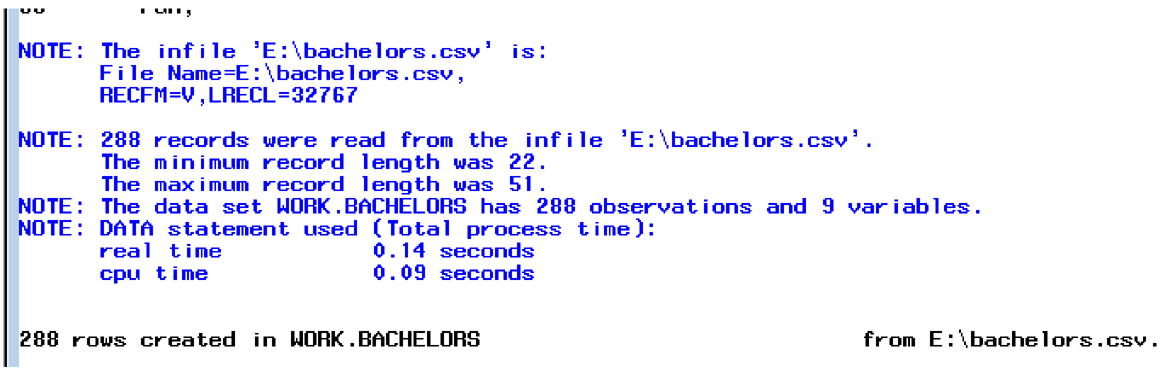


Choose the SAS destination:

Library: WORK

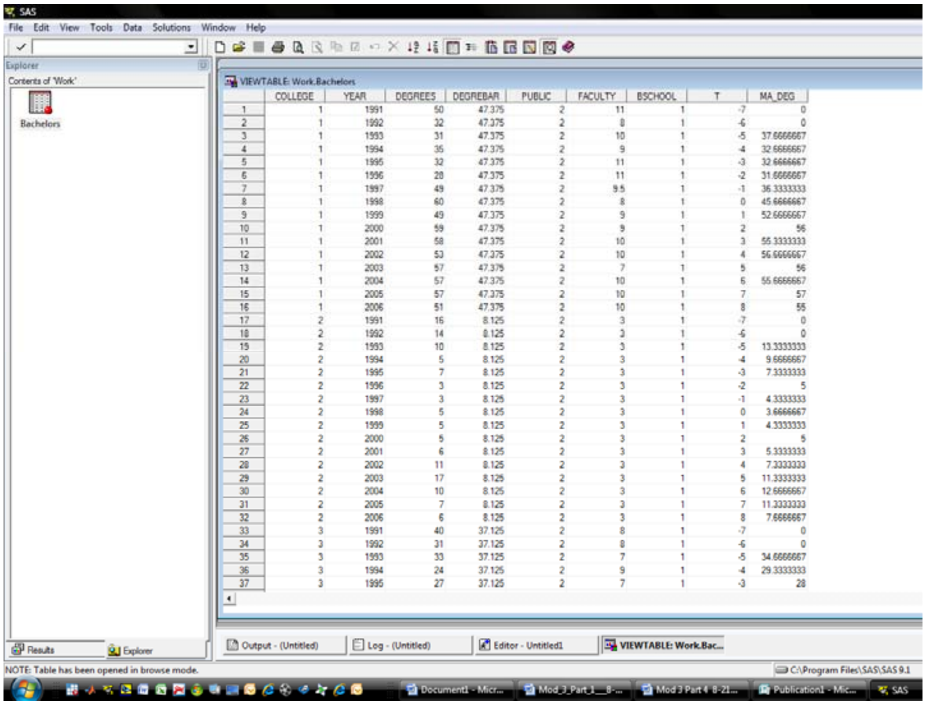
Member: BACHELORS

To verify that the wizard imported the data correct, review the Log file and physically inspect the dataset. When SAS is opened, the default panels are the ‘Log’ window at the top right, the ‘Editor’ window in the bottom right and the ‘Explorer/Results’ window on the left. Scrolling through the Log reveals that the dataset was successfully imported. The details of the data step procedure are provided along with a few summary statistics of how many observations and variables were imported.



```
NOTE: The infile 'E:\bachelors.csv' is:
      File Name=E:\bachelors.csv,
      RECFM=V,LRECL=32767

NOTE: 288 records were read from the infile 'E:\bachelors.csv'.
      The minimum record length was 22.
      The maximum record length was 51.
NOTE: The data set WORK.BACHELORS has 288 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.14 seconds
      cpu time            0.09 seconds


288 rows created in WORK.BACHELORS                            from E:\bachelors.csv.
```

To view the dataset, click on the "Libraries" folder, which is in the top left of the 'Explorer' panel, and then click on the 'Work' library. This reveals all of the members in the 'Work' library. In this case, the only member is the dataset 'Bachelors'. To view the dataset, click on the dataset icon 'Bachelors'.



VIEWTABLE: Work.Bachelors

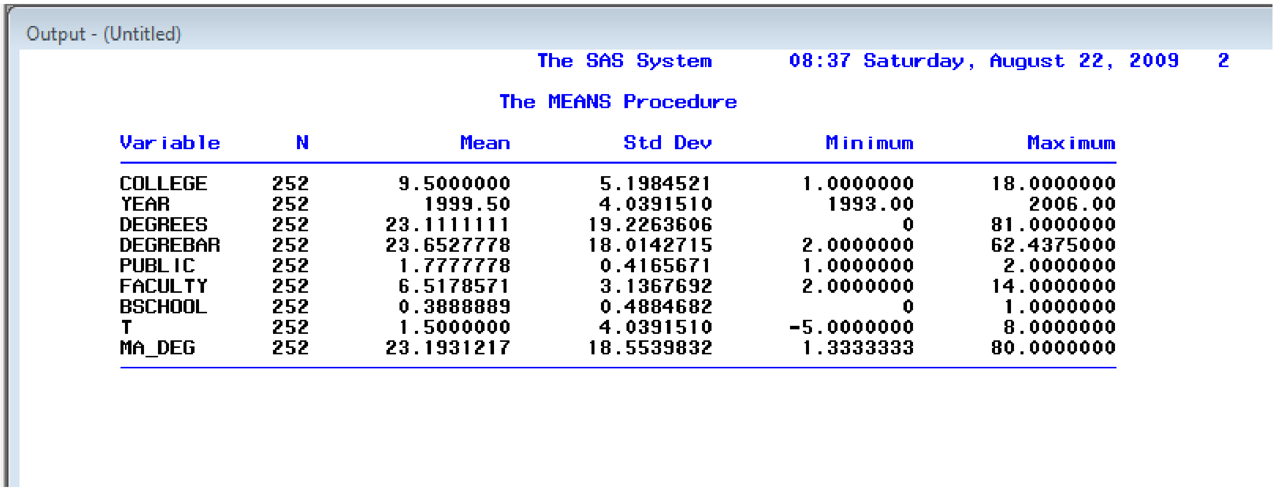
	COLLEGE	YEAR	DEGREES	DEGREBAR	PUBLIC	FACULTY	BSCHOOL	T	MA_DEG
1	1	1991	50	47.375	2	11	1	-7	0
2	1	1992	32	47.375	2	8	1	-6	0
3	1	1993	31	47.375	2	10	1	-5	37.6666667
4	1	1994	35	47.375	2	9	1	-4	32.6666667
5	1	1995	32	47.375	2	11	1	-3	32.6666667
6	1	1996	28	47.375	2	11	1	-2	31.6666667
7	1	1997	49	47.375	2	9.5	1	-1	36.3333333
8	1	1998	60	47.375	2	8	1	0	45.6666667
9	1	1999	49	47.375	2	9	1	1	52.6666667
10	1	2000	59	47.375	2	9	1	2	56
11	1	2001	58	47.375	2	10	1	3	55.3333333
12	1	2002	53	47.375	2	10	1	4	56.6666667
13	1	2003	57	47.375	2	7	1	5	56
14	1	2004	57	47.375	2	10	1	6	55.6666667
15	1	2005	57	47.375	2	10	1	7	57
16	1	2006	51	47.375	2	10	1	8	55
17	2	1991	16	8.125	2	3	1	-7	0
18	2	1992	14	8.125	2	3	1	-6	0
19	2	1993	10	8.125	2	3	1	-5	13.3333333
20	2	1994	5	8.125	2	3	1	-4	9.6666667
21	2	1995	7	8.125	2	3	1	-3	7.3333333
22	2	1996	3	8.125	2	3	1	-2	5
23	2	1997	3	8.125	2	3	1	-1	4.3333333
24	2	1998	5	8.125	2	3	1	0	3.6666667
25	2	1999	5	8.125	2	3	1	1	4.3333333
26	2	2000	5	8.125	2	3	1	2	5
27	2	2001	6	8.125	2	3	1	3	5.3333333
28	2	2002	11	8.125	2	3	1	4	7.3333333
29	2	2003	17	8.125	2	3	1	5	11.3333333
30	2	2004	10	8.125	2	3	1	6	12.6666667
31	2	2005	7	8.125	2	3	1	7	11.3333333
32	2	2006	6	8.125	2	3	1	8	7.6666667
33	3	1991	40	37.125	2	8	1	-7	0
34	3	1992	31	37.125	2	8	1	-6	0
35	3	1993	33	37.125	2	7	1	-5	34.6666667
36	3	1994	24	37.125	2	9	1	-4	29.3333333
37	3	1995	27	37.125	2	7	1	-3	28

NOTE: Table has been opened in browse mode.

In addition to a visual inspection of the data, we use the "means" command to check the descriptive statistics. Since we don't list any variables in the command, by default, SAS runs the 'means' command on all variables in the dataset. First, however, we need to remove the two years (1991 and 1992) for which no data are available for the degree moving average measure. Since we may need the full dataset later, it is good practice to delete the observations off of a copy of the dataset (called bachelors2). This is done in a data step using an 'if then' command.

```
data bachelors2;
      set bachelors;
if year = 1991 then delete;
if year = 1992 then delete;
run;

PROC MEANS DATA=bachelors2;
RUN;
```

Typing the following commands into the 'Editor' window and then clicking the run bottom (recall this is the running man at the top) yields the following screen.



```
                                   The SAS System       08:37 Saturday, August 22, 2009   2

                                 The MEANS Procedure

     Variable        N            Mean         Std Dev         Minimum         Maximum
     ------------------------------------------------------------------------------
     COLLEGE       252       9.5000000       5.1984521       1.0000000      18.0000000
     YEAR          252         1999.50       4.0391510         1993.00         2006.00
     DEGREES       252      23.1111111      19.2263606               0      81.0000000
     DEGREBAR      252      23.6527778      18.0142715       2.0000000      62.4375000
     PUBLIC        252       1.7777778       0.4165671       1.0000000       2.0000000
     FACULTY       252       6.5178571       3.1367692       2.0000000      14.0000000
     BSCHOOL       252       0.3888889       0.4884682               0       1.0000000
     T             252       1.5000000       4.0391510      -5.0000000       8.0000000
     MA_DEG        252      23.1931217      18.5539832       1.3333333      80.0000000
     ------------------------------------------------------------------------------
```

CONSTANT COEFFICIENT REGRESSION

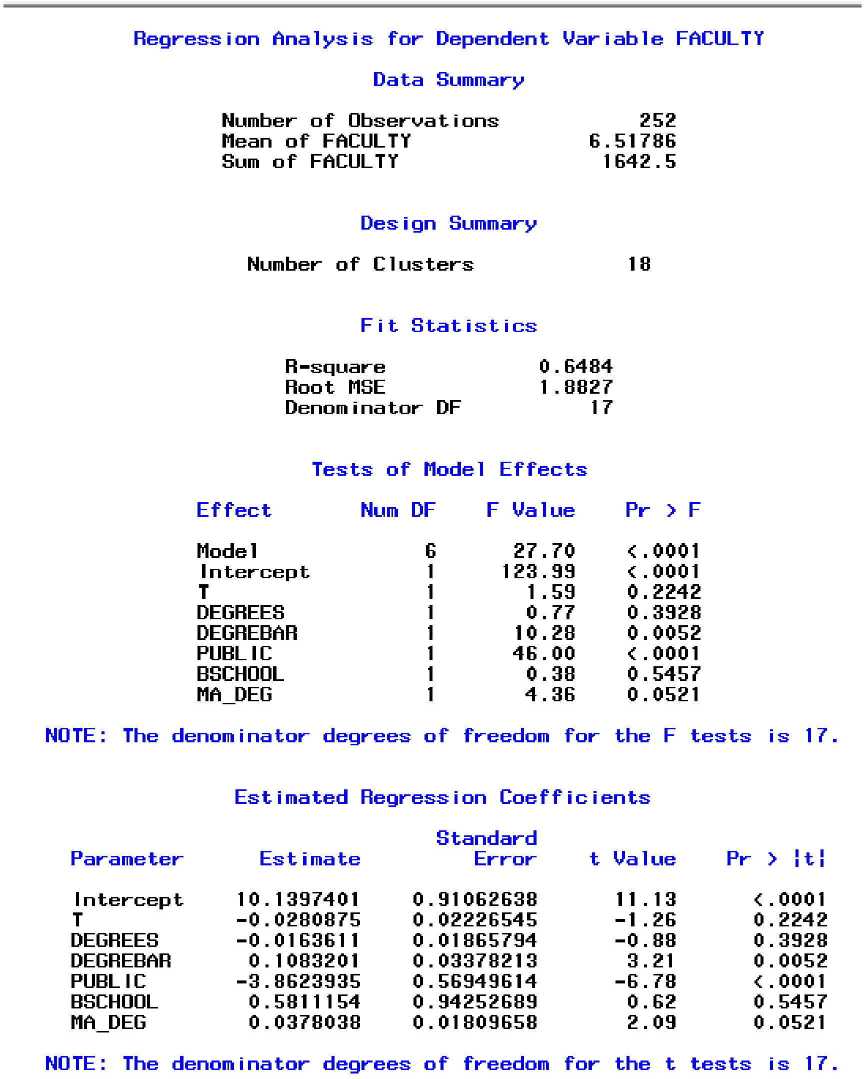
The constant coefficient panel data model for the faculty size data-generating process for bachelor degree-granting undergraduate departments is given by

$$\begin{aligned} &Faculty\ size_{it} = \beta_1 + \beta_2 T_t + \beta_3 BA\&S_{it} + \beta_4 MEANBA\&S_i + \beta_5 PUBLIC_i \\ &+ \beta_6 Bschl + \beta_7 MA\_Deg_{it} + \varepsilon_{it} \end{aligned}$$

where the error term ε_{it} is independent and identically distributed (*iid*) across institutions and over time and E($\varepsilon_{it}^2|\mathbf{x}_{it}$) = σ^2 , for I = 18 colleges and T = 14 years (−5 through 8) for 252 complete records. To take into account clustering, include the cluster option with the cluster being on the colleges. The SAS OLS regression command that needs to be entered into the editor, including the standard error adjustment for clustering is

```
proc surveyreg data=bachelors2;
      cluster college;
      model faculty = t degrees degrebar public bschool ma_deg;
run;
```

Upon highlighting and hitting the “run” button, the Output panel shows the following results



```
                 Regression Analysis for Dependent Variable FACULTY

                                  Data Summary

                      Number of Observations          252
                      Mean of FACULTY             6.51786
                      Sum of FACULTY               1642.5

                                 Design Summary

                        Number of Clusters             18

                                 Fit Statistics

                           R-square            0.6484
                           Root MSE            1.8827
                           Denominator DF          17

                             Tests of Model Effects

                  Effect        Num DF    F Value    Pr > F

                  Model              6      27.70    <.0001
                  Intercept          1     123.99    <.0001
                  T                  1       1.59    0.2242
                  DEGREES            1       0.77    0.3928
                  DEGREBAR           1      10.28    0.0052
                  PUBLIC             1      46.00    <.0001
                  BSCHOOL            1       0.38    0.5457
                  MA_DEG             1       4.36    0.0521

     NOTE: The denominator degrees of freedom for the F tests is 17.

                        Estimated Regression Coefficients

                                         Standard
      Parameter      Estimate             Error     t Value     Pr > |t|

      Intercept    10.1397401     0.91062638       11.13       <.0001
      T            -0.0280875     0.02226545       -1.26       0.2242
      DEGREES      -0.0163611     0.01865794       -0.88       0.3928
      DEGREBAR      0.1083201     0.03378213        3.21       0.0052
      PUBLIC       -3.8623935     0.56949614       -6.78       <.0001
      BSCHOOL       0.5811154     0.94252689        0.62       0.5457
      MA_DEG        0.0378038     0.01809658        2.09       0.0521

     NOTE: The denominator degrees of freedom for the t tests is 17.
```

Contemporaneous degrees have little to do with current faculty size but both overall number of degrees awarded (the school means) and the moving average of degrees (MA_DEG) have significant effects. It takes an increase of 26 or 27 bachelor degrees in the moving average to expect just one more faculty position. Whether it is a public or a private college is highly significant. Moving from a public to a private college lowers predicted faculty size by nearly four members for otherwise comparable institutions. There is an insignificant erosion of tenured and tenure-track faculty size over time. Finally, while economics departments in colleges with a business school tend to have a larger permanent faculty, ceteris paribus, the effect is small and insignificant.

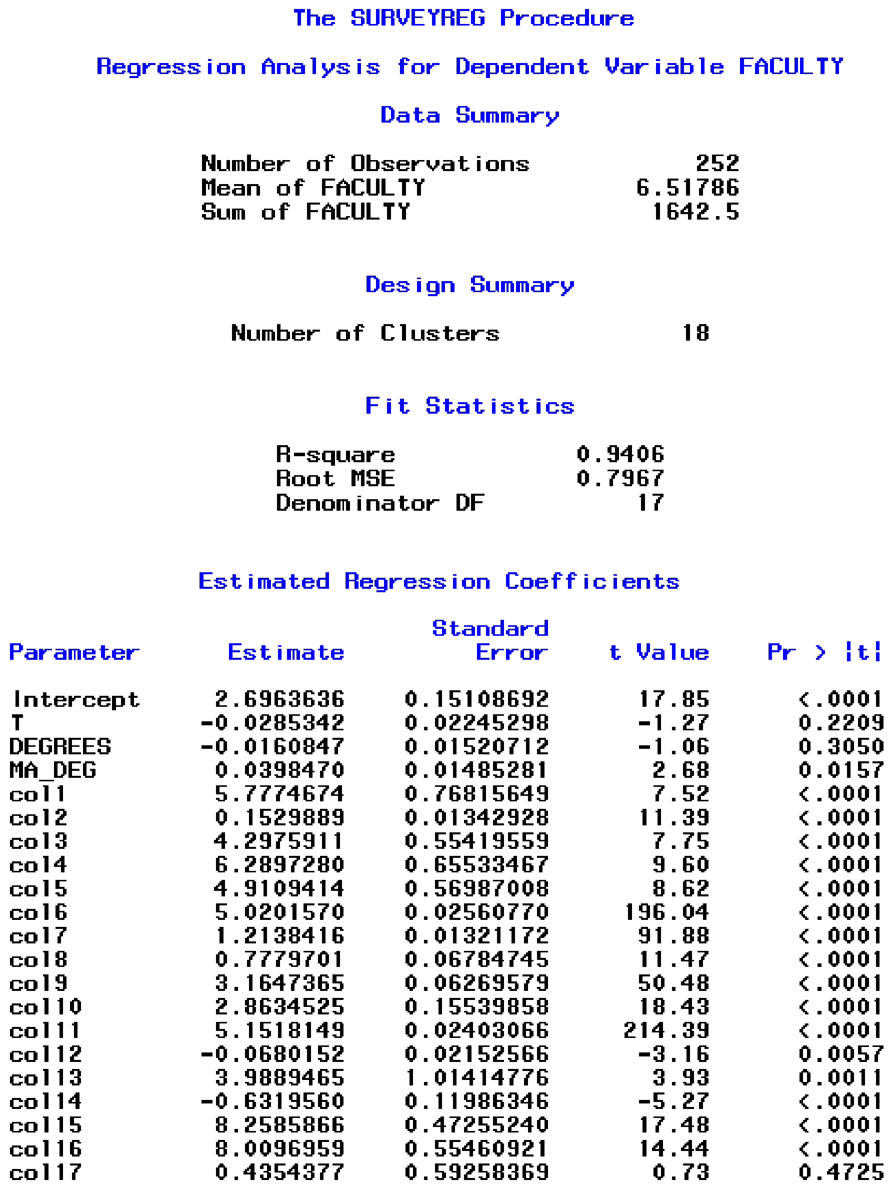
FIXED-EFFECTS REGRESSION

The fixed-effects model requires either the insertion of 17 (0,1) covariates to capture the unique effect of each of the 18 colleges (where each of the 17 dummy coefficients are measured relative to the constant term) or the insertion of 18 dummy variables with no constant term in the OLS regression. In addition, no time invariant variables can be included because they would be perfectly correlated with the respective college dummies. Thus, the overall mean number of degrees, the public or private dummy, and business school dummy cannot be included as regressors.

The SAS code to be run from the editor window, including the commands to create the dummy variables is:

```
data bachelors2;
      set bachelors2;

 col1 = 0; col2 = 0;  col3 = 0; col4 = 0;  col5=0;    col6=0;
 col7 = 0; col8 = 0;  col9 = 0; col10 = 0; col11 = 0; col12 = 0;
col13 = 0; col14 =0; col15 = 0; col16 = 0; col17 = 0; col18 = 0;

if college = 1 then col1=1;         if college = 2 then col2=1;
if college = 3 then col3=1;         if college = 4 then col4=1;
if college = 5 then col5=1;         if college = 6 then col6=1;
if college = 7 then col7=1;         if college = 8 then col8=1;
if college = 9 then col9=1;         if college = 10 then col10=1;
if college = 11 then col11=1;       if college = 12 then col12=1;
if college = 13 then col13=1;       if college = 14 then col14=1;
if college = 15 then col15=1;       if college = 16 then col16=1;
if college = 17 then col17=1;       if college = 18 then col18=1;

run;

proc surveyreg data=bachelors2;
      cluster college;
      model faculty = t degrees ma_deg col1 col2 col3 col4 col5
                              col6 col7 col8 col9 col10 col11 col12
                              col13 col14 col15 col16 col17;
quit;
```

The resulting regression information appearing in the output window is



```
                          The SURVEYREG Procedure

              Regression Analysis for Dependent Variable FACULTY

                               Data Summary

                 Number of Observations            252
                 Mean of FACULTY               6.51786
                 Sum of FACULTY                 1642.5

                              Design Summary

                   Number of Clusters            18

                              Fit Statistics

                      R-square              0.9406
                      Root MSE              0.7967
                      Denominator DF            17

                     Estimated Regression Coefficients

                                   Standard
    Parameter       Estimate          Error    t Value    Pr > |t|

    Intercept      2.6963636     0.15108692      17.85      <.0001
    T             -0.0285342     0.02245298      -1.27      0.2209
    DEGREES       -0.0160847     0.01520712      -1.06      0.3050
    MA_DEG         0.0398470     0.01485281       2.68      0.0157
    col1           5.7774674     0.76815649       7.52      <.0001
    col2           0.1529889     0.01342928      11.39      <.0001
    col3           4.2975911     0.55419559       7.75      <.0001
    col4           6.2897280     0.65533467       9.60      <.0001
    col5           4.9109414     0.56987008       8.62      <.0001
    col6           5.0201570     0.02560770     196.04      <.0001
    col7           1.2138416     0.01321172      91.88      <.0001
    col8           0.7779701     0.06784745      11.47      <.0001
    col9           3.1647365     0.06269579      50.48      <.0001
    col10          2.8634525     0.15539858      18.43      <.0001
    col11          5.1518149     0.02403066     214.39      <.0001
    col12         -0.0680152     0.02152566      -3.16      0.0057
    col13          3.9889465     1.01414776       3.93      0.0011
    col14         -0.6319560     0.11986346      -5.27      <.0001
    col15          8.2585866     0.47255240      17.48      <.0001
    col16          8.0096959     0.55460921      14.44      <.0001
    col17          0.4354377     0.59258369       0.73      0.4725
```

Once again, contemporaneous degrees is not a driving force in faculty size. An F test is not needed to assess if at least one of the 17 colleges differ from college 18. With the exception of college 17, each of the other colleges are significantly different. The moving average of degrees is again significant.

RANDOM-EFFECTS REGRESSION

Finally, consider the random-effects model in which we employ Mundlak's (1978) approach to estimating panel data. The Mundlak model posits that the fixed effects in the equation, β_{1i} , can be projected upon the group means of the time-varying variables, so that

$$\beta_{1i} = \beta_1 + \delta' \overline{x}_i + w_i$$

where $\overline{x}_i$ 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. Logically, adding the means to the equations picks up the correlation between the school effects and the other variables. We could not incorporate the mean number of degrees awarded in the fixed-effects model (because it was time invariant) but this variable plays a critical role in the Mundlak approach to panel data modeling and estimation.

The random effects model for BA and BS degree-granting undergraduate departments is

$$FACULTY\ size_{it} = \beta_1 + \beta_2 YEAR_t + \beta_3 BA\&S_{it} + \beta_4 MEANBA\&S_i + \beta_5 MOVAVBA\&BS$$
$$+ \beta_6 PUBLIC_i + \beta_7 Bschl + \varepsilon_{it} + u_i$$

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

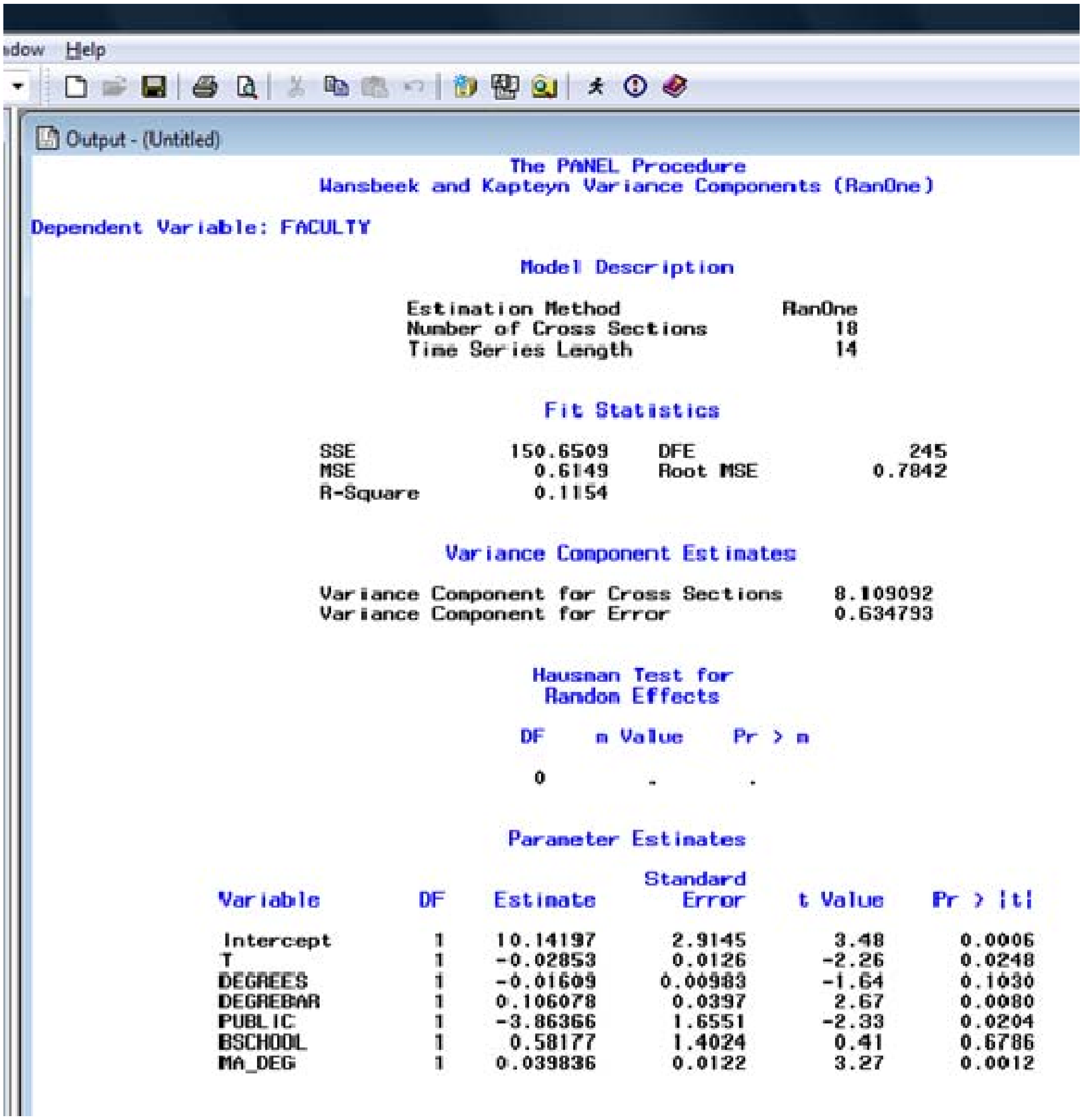
In SAS 9.1, there are no straightforward procedures to estimate this model. In the appendix, I do provide a lengthy procedure that estimates the random effects model by OLS regression on a transformed model. This is quite complex and is not recommended for beginners. See Cameron and Trivedi (2005) for further details. SAS 9.2 has a new command called the PANEL procedure to estimate panel data. For our model, we need to attach the / RANONE option to specify that a one-way random-effects model be estimated. We also need to correct for the clustering of the data. Unlike simple commands in LIMPDEP and STATA, SAS does not have an option for one-way random effects with clustered errors.

This new SAS 9.2 procedure has more options for specific error term structures in panel data. Although SAS does not allow the CLUSTER option, there is a VCOMP option that specifies the type of variance component estimate to use. For balanced data, the default is VCOMP=FB. However, the FB method does not always obtain nonnegative estimates for the cross section (or group) variance. In the case of a negative estimate, a warning is printed and the estimate is set to zero. Because we have to address clustering, WK option is specified, which is close to groupwise heteroscedastic regression.

The SAS code to be run from the Editor panel (with 1991 and 1992 data suppressed) is

```
PROC SORT DATA=bachelors2;
BY college year;

PROC panel DATA=bachelors2;
ID college year;
MODEL faculty = t degrees degrebar public bschool MA_deg /RANONE VCOMP=WK;
RUN;
```

The resulting regression information appearing in the output window is



```
                          The PANEL Procedure
           Wansbeek and Kapteyn Variance Components (RanOne)

Dependent Variable: FACULTY

                           Model Description

                 Estimation Method                  RanOne
                 Number of Cross Sections               18
                 Time Series Length                     14

                            Fit Statistics

           SSE               150.6509    DFE                  245
           MSE                 0.6149    Root MSE          0.7842
           R-Square            0.1154

                     Variance Component Estimates

           Variance Component for Cross Sections      8.109092
           Variance Component for Error               0.634793

                           Hausman Test for
                            Random Effects

                          DF     m Value    Pr > m

                           0           .         .

                         Parameter Estimates

                                                Standard
      Variable          DF      Estimate           Error     t Value     Pr > |t|

       Intercept         1      10.14197          2.9145        3.48       0.0006
      T                  1      -0.02853          0.0126       -2.26       0.0248
      DEGREES            1      -0.01609         0.00983       -1.64       0.1030
      DEGREBAR           1      0.106078          0.0397        2.67       0.0080
      PUBLIC             1      -3.86366          1.6551       -2.33       0.0204
      BSCHOOL            1       0.58177          1.4024        0.41       0.6786
      MA_DEG             1      0.039836          0.0122        3.27       0.0012
```

The marginal effect of an additional economics major is again insignificant but slightly negative within the sample. Both the short-term moving average number and long-term average number of bachelor degrees are significant. A long-term increase of about 10 students earning degrees in economics is required to predict that one more tenured or tenure-track faculty member is in a department. Ceteris paribus, economics departments at private institutions are smaller than comparable departments at public schools by a large and significant number of four members. Whether there is a business school present is insignificant. There is no meaningful trend in faculty size.

It should be clear that this regression is NOT identical to similar one-way random effect models controlling for clustering in LIMDEP or STATA. The standard errors are adjusted for a general groupwise heteroscedastic error structure. The difference does not alter the significance and the standard errors are, for the most part, very comparable.

CONCLUDING REMARKS

The goal of this hands-on component of this third of four modules is to enable economic education researchers to make use of panel data for the estimation of constant coefficient, fixed-effects and random-effects panel data models in SAS. It was not intended to explain all of the statistical and econometric nuances associated with panel data analysis. For this an intermediate level econometrics textbook (such as Jeffrey Wooldridge, *Introductory Econometrics*) or advanced econometrics textbook (such as William Greene, *Econometric Analysis*) should be consulted.

APPENDIX: Alternative Means to Estimate Random-Effects Model with Clustered Data.

The following code provides a necessary code to estimate the random-effect models with clustering. The estimation procedure is two-step feasible GLS. In the first step, the variance matrix is estimated. In the second step, this variance matrix is used to transform the equation.

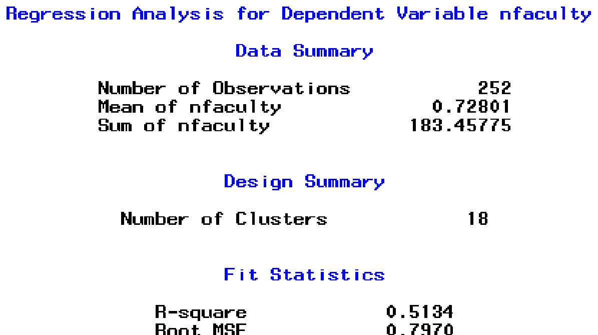
Because the variance matrix is *estimated* and not the true variance, this causes the standard errors to be slight different than the standard errors provided by LIMPDEP or STATA when estimating a random effects model with clustering.

The code to be run in the editor window is:

```
/* get SSE and SSU */

proc sort data= bachelors2;
by college year; quit;

proc tscsreg data=bachelors2 outest=covvc;
id college year;
model faculty = t degrees degrebar public bschool MA_deg / ranone;
quit;


/* find number of years */

data numobs (keep = year);
set bachelors2;
run;

proc sort nodupkey;
by year;
quit;

proc means data = numobs
      max;
output out = num;
      quit;

/* create lamda */
proc iml;
use covvc;
read all var {_VARERR_ _VARCS_} into x;
use num;
read var {_freq_} into y;
print y;
sesq = x[1,1];
susq = x[1,2];
lamda = 1 - sqrt( sesq / (y[1,1]*susq + sesq) );
print x y lamda;
cname = {"lamda"};
```

```
create out from lamda [ colname=cname];
append from lamda;
quit;

/* find averages of each variable grouped by college #*/
proc MEANS NOPRINT
data=bachelors2;
class college;
output out=stats
mean= avg_year avg_degrees avg_degrebar avg_public avg_faculty avg_bschool
avg_t avg_ma_deg;
run;

data bachelors3 (drop = _type_ _freq_);
      merge bachelors2 stats;
      by college;
      if _type_ = 0 then delete;
      one = 1;
      run;

DATA bachelors4;
    if _N_ = 1 then set out;
    SET bachelors3;
      l = one*lamda;
run;

/* transform data */
data clean (keep = college con nfaculty nt ndegrees ndegrebar npublic
nbschool nMA_deg year);
      set bachelors4;
nfaculty =  faculty - lamda*avg_faculty;
nt =  t - lamda*avg_t;
ndegrees = degrees - lamda*avg_degrees ;
ndegrebar = degrebar - lamda*avg_degrebar;
npublic = public - lamda*avg_public;
nbschool = bschool - lamda*avg_bschool;
nMA_deg = ma_deg - lamda*avg_ma_deg;
con = 1 - lamda*1;
run;

/* run regression on transformed equation assuming clustering */
/* Since intercept is included in transformed equation, use noint option*/

proc surveyreg data=clean;
      cluster college;
      model nfaculty = con nt ndegrees ndegrebar npublic nbschool nMA_deg /
      noint;
quit;
```

The output for this regression is:



```
                Regression Analysis for Dependent Variable nfaculty

                                  Data Summary

                      Number of Observations              252
                      Mean of nfaculty                0.72801
                      Sum of nfaculty               183.45775


                                 Design Summary

                        Number of Clusters              18


                                 Fit Statistics

                           R-square              0.5134
                           Root MSE              0.7970
```

The standard errors associated with this regression are much closer to the standard errors from LIMPDEP and STATA. However, this is a complex sequence of codes which should not be attempted by beginners.

REFERENCES

Becker, William, William Greene and John Siegfried (2009). “Does Teaching Load Affect Faculty Size? ” Working Paper (July).

Cameron, Colin and Pravin Trivedi (2005). *Microeconometrics*. 1st Edition, New York, Cambridge University Press.

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

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

Wooldridge, Jeffrey (2009). *Introductory Econometrics*. 4th Edition, Mason OH: South-Western.