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).

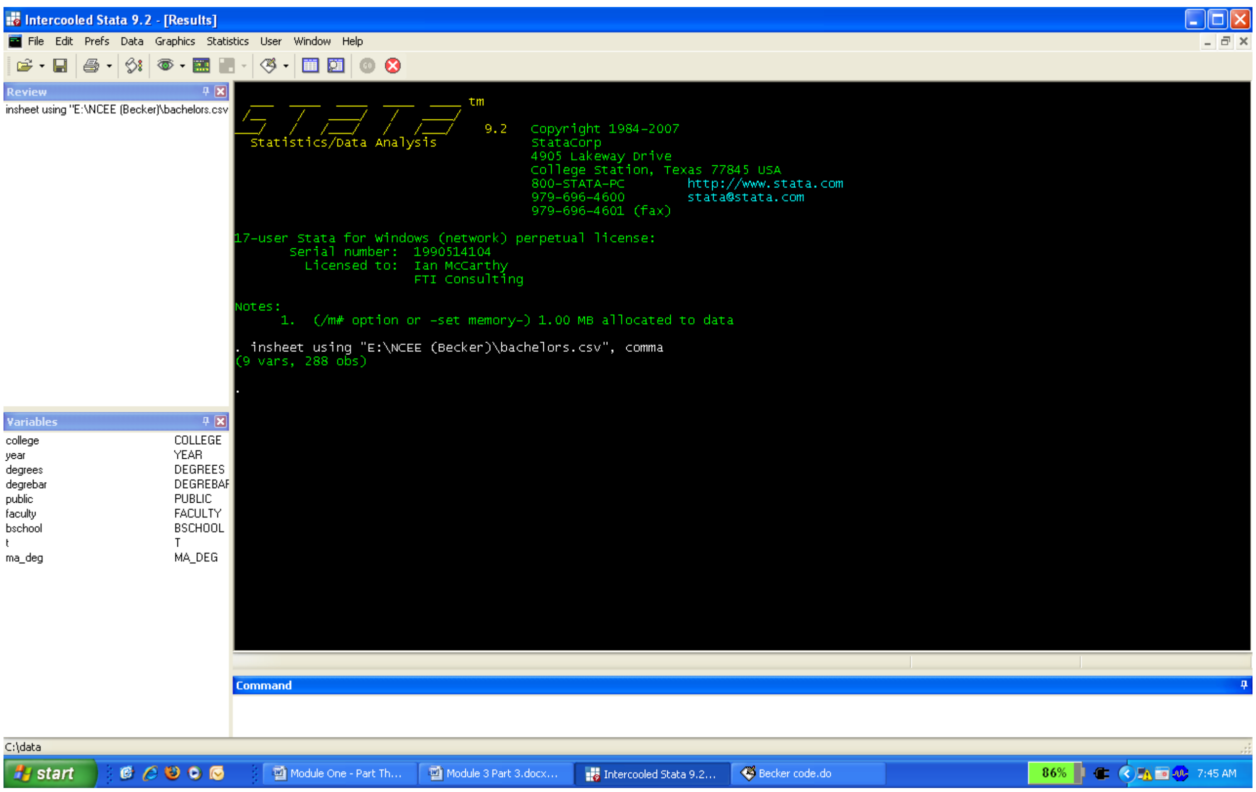
College	Year	Degrees	DegreBar	Public	Faculty	Bschol	T	MA_Deg
1	1991	50	47.375	2	11	1	-7	0
1	1992	32	47.375	2	8	1	-6	0
1	1993	31	47.375	2	10	1	-5	37.667
1	1994	35	47.375	2	9	1	-4	32.667
↓	↓	↓	↓	↓	↓	↓		↓
1	2003	57	47.375	2	7	1	5	56
1	2004	57	47.375	2	10	1	6	55.667
1	2005	57	47.375	2	10	1	7	57
1	2006	51	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:



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

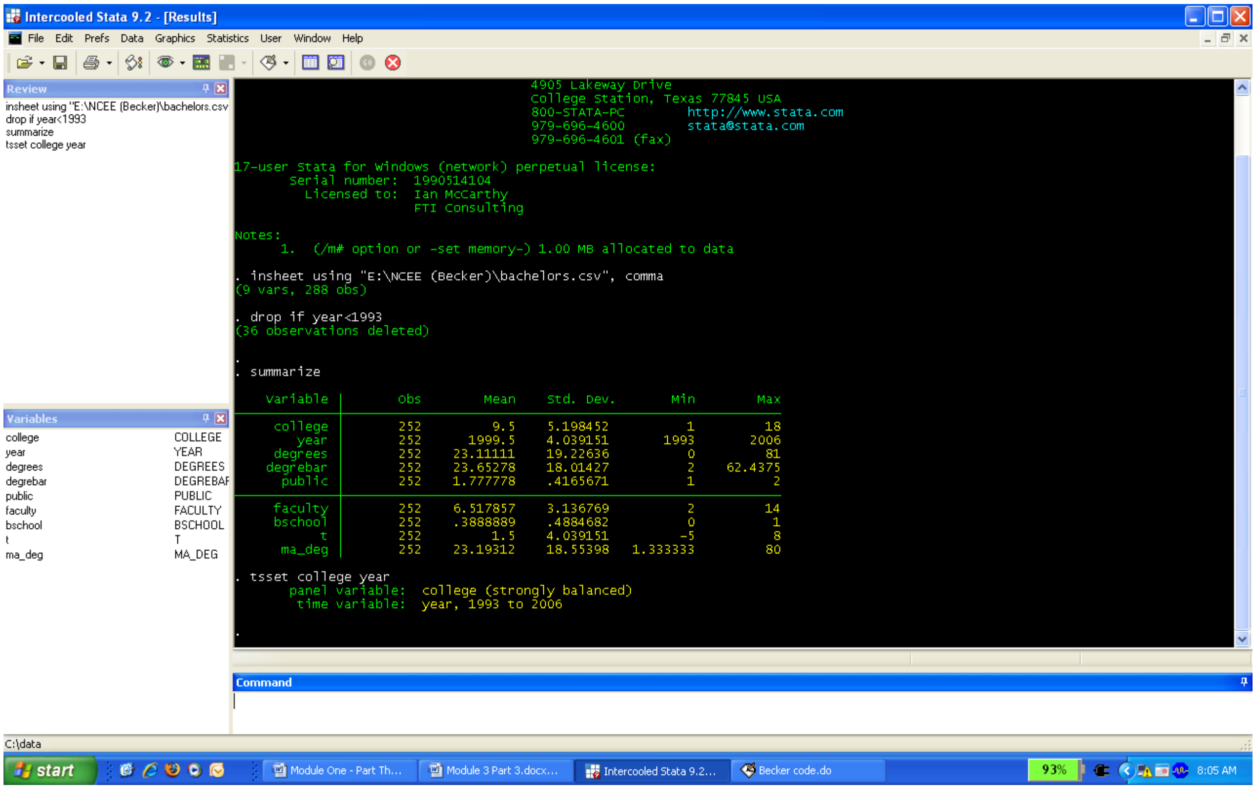
By default, STATA essentially considers all data as cross-sectional. Since we are working with panel data in this case, we need to indicate to STATA that there is a time-series component to our dataset. This is done with the "tsset" command. The general syntax for the "tsset" command with panel data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



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

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

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 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

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

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 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.