Regression II
Vanderbilt University
Prof. Sean P. Corcoran **84 total points**

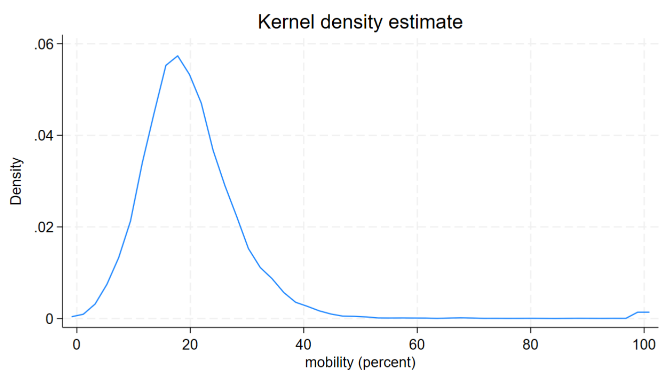
Problem Set 4 *Solutions*

Question 1. This problem will use the panel dataset of Texas elementary schools used in class (*texas_elementary_panel.dta*) to estimate the effects of student mobility on school average performance on standardized tests. (**38 points**)

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/Texas_elementary_panel_2004_2007.dta
```

(a) The variable *cpemallp* is defined as the percentage of students in a school who were enrolled less than 83% of the school year (i.e., were not present 6 or more weeks at that school). Rename this variable *mobility*, report the overall mean and standard deviation for this variable, and produce a kernel density plot for this variable (use the `kdensity` command). Describe what this distribution looks like. (**3 points**)

```
. rename cpemallp mobility

. sum mobility avgpassing

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    mobility |     16,072    20.15625    9.892888          0        100
  avgpassing |     16,225     75.4024    13.83064          5         99

. kdensity mobility
```



The mean *mobility* share is 20.2%, meaning in the average school about 1 in 5 students were enrolled less than 83% of the school year. The standard deviation is 9.9 percentage points (based on the school-by-year observations). The kernel density shows this is very right-skewed distribution, with some schools having unusually high mobility rates.

(b) Declare this dataset to be a panel using `xtset`. Use the same cross-sectional unit and time dimension variables used in class. Use `xtsum` to get a set of descriptive statistics for *mobility*. Does it appear that school mobility is primarily a between-school phenomenon, or something that varies more within schools over time? Explain how you know, and explain in words how the standard deviations (overall, within, and between) are calculated. (**4 points**)

```
. xtset campus year
       panel variable:  campus (unbalanced)
        time variable:  year, 2004 to 2007, but with gaps
                delta:  1 unit

. xtsum mobility

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
mobility overall |  20.15625   9.892888          0        100 |     N =   16072
         between |             10.56573          0        100 |     n =    4302
         within  |             2.926674  -7.643754   70.18124 | T-bar = 3.73594
```

At 10.6, the between-school standard deviation of *mobility* is considerably larger than the within-school standard deviation of 2.9. The latter is calculated using deviations from school-specific means, while the former is calculated using deviations of school-specific means from the grand mean. The "overall" standard deviation uses deviations of each data point from the grand mean. The finding that school mobility is primarily a between-school phenomenon is not surprising. Some schools likely suffer from persistently high mobility year after year. Annual deviations from this long-run average are likely to be smaller.

(c) Estimate a simple regression of the average TAKS exam passing rate (*avgpassing*) on mobility (refer to the lecture notes for the *avgpassing* variable). How are these variables related? Report your results and interpret your coefficient estimate in words. Is the coefficient statistically significant? Practically significant? (**4 points**)

Results are below. There is a strong negative relationship between school mobility and the average passing rate on state tests. The estimated coefficient is statistically (p<0.001) and practically significant. A one-standard deviation increase in school mobility rates (9.9) is associated with a 7.5 percentage point lower passing rate. When benchmarked against the standard deviation in passing rates in the data (13.8), this is a large effect.

```
. rename ca311tar avgpassing

. reg avgpassing mobility
```

```
      Source |       SS           df       MS      Number of obs   =    15,831
-------------+----------------------------------   F(1, 15829)     =   3341.83
       Model |  525812.917         1  525812.917   Prob > F        =    0.0000
    Residual |  2490582.58    15,829  157.343015   R-squared       =    0.1743
-------------+----------------------------------   Adj R-squared   =    0.1743
       Total |   3016395.5    15,830  190.549305   Root MSE        =    12.544

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mobility |  -.7570524   .0130959   -57.81   0.000    -.7827218    -.731383
       _cons |   90.29461    .276085   327.05   0.000     89.75345    90.83576
------------------------------------------------------------------------------
```

(d) Should the coefficient estimated in part (c) be interpreted as the causal effect of student mobility on school performance? Briefly explain why or why not, with reference to potential outcomes. (**3 points**)

For the regression in (c) to have a causal interpretation, we have to believe that the coefficient on *mobility* estimates the difference in potential outcomes under varying levels of student mobility for some well-defined population of schools. This seems unlikely if there are omitted variables correlated with both mobility and passing rates. Chances are, schools with high mobility rates are disadvantaged in other ways that would lead us to predict lower achievement in those schools.

(e) Add the following explanatory variables to your regression in (c): percent black, white, Hispanic, Asian or Pacific Islander (API), Limited English Proficient (LEP), and economically disadvantaged. Also include year effects and a dummy variable for charter schools (*charter*, which may need to be encoded as numeric). How does the inclusion of these covariates affect your estimated coefficient on *mobility*? Is it still statistically significant? Does the change make sense to you (explain)? Finally, provide a written interpretation of the estimated coefficients for the three year dummies (2005, 2006 and 2007). (**4 points**)

Results shown below. Perhaps not surprisingly, the coefficient on *mobility* is much smaller in absolute value (-0.152). This was anticipated given our answer in part (d). Omitted variables were likely positively correlated with *mobility* and negatively correlated with *avgpassing*, suggesting our "short" regression coefficient was upwardly biased. That is, it likely over-stated the negative relationship between *mobility* and *avgpassing*.

```
. encode charter, gen(charter2)

. reg avgpassing mobility cpetblap cpetwhip cpethisp cpetpacp cpetecop i.year i.charter2
```

```
      Source |       SS           df       MS      Number of obs   =    15,831
-------------+----------------------------------   F(10, 15820)    =   1480.98
       Model |  1458455.37        10  145845.537   Prob > F        =    0.0000
    Residual |  1557940.13    15,820  98.4791487   R-squared       =    0.4835
-------------+----------------------------------   Adj R-squared   =    0.4832
       Total |   3016395.5    15,830  190.549305   Root MSE        =    9.9237

------------------------------------------------------------------------------
  avgpassing |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mobility |  -.1524116   .0129286   -11.79   0.000    -.1777531     -.12707
    cpetblap |  -.0949501   .1328527    -0.71   0.475    -.3553565    .1654564
    cpetwhip |    .021062   .1338113     0.16   0.875    -.2412234    .2833475
    cpethisp |  -.0208526   .1328268    -0.16   0.875    -.2812084    .2395031
    cpetpacp |   .1701193   .1345458     1.26   0.206    -.0936058    .4338443
    cpetecop |  -.2318625   .0062852   -36.89   0.000    -.2441822   -.2195428
             |
        year |
       2005  |   5.127777   .2243313    22.86   0.000     4.688062    5.567492
       2006  |   6.194567   .2246185    27.58   0.000      5.75429    6.634845
       2007  |   8.183221   .2229574    36.70   0.000     7.746199    8.620243
             |
    charter2 |
          Y  |  -9.452535   .5655251   -16.71   0.000    -10.56103   -8.344042
       _cons |   89.32539    13.3053     6.71   0.000     63.24549    115.4053
------------------------------------------------------------------------------
```

(f) Should the coefficient estimated in (e) be interpreted as the causal effect of student mobility on school performance? Briefly explain why or why not. How might a regression model with school fixed effects improve upon the model in (e)? (**3 points**)

Again, for the regression in (e) to have a causal interpretation, we have to believe that the covariance between the population error term *u* and *mobility* is zero, conditional on the other explanatory variables. While we have now controlled for several school characteristics that made this assumption more plausible, there may be other unobserved school characteristics that are omitted from the regression that are systematically related to *mobility* and *avgpassing*.

(g) Estimate the regression in (e) with school fixed effects. (Show this using both `xtreg` and `areg`). Ensure the standard errors allow for clustering at the school level. How does this approach affect the estimated coefficient on *mobility*, if at all? Is it statistically significant? Does the change in the estimated coefficient make sense to you? Provide an intuitive explanation of the finding. Were any explanatory variables dropped from the model (or are there any that should have been dropped that weren't)? If so, why? (**5 points**)

Results for `xtreg` are shown below. Interestingly, the coefficient on *mobility*

is now very small and statistically insignificant. This change makes sense if we believe the school fixed effect is capturing unobserved school characteristics that are systematically associated with high mobility rates and low achievement. The fixed effects model relies entirely on *within-school* variation in mobility rates over time to estimate the slope coefficients. Note that charter status falls out of the model, since it is time-invariant.

```
.  xtreg avgpassing mobility cpetblap cpetwhip cpethisp cpetpacp ///
>                 cpetecop i.year i.charter2, fe cluster(campus)
note: 2.charter2 omitted because of collinearity.

Fixed-effects (within) regression               Number of obs     =     15,831
Group variable: campus                          Number of groups  =      4,230

R-squared:                                      Obs per group:
     Within  = 0.2684                                         min =          1
     Between = 0.3627                                         avg =        3.7
     Overall = 0.3351                                         max =          4

                                                F(9, 4229)        =     326.67
corr(u_i, Xb) = -0.1838                         Prob > F          =     0.0000

                            (Std. err. adjusted for 4,230 clusters in campus)
------------------------------------------------------------------------------
             |               Robust
  avgpassing | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    mobility |   .0002792   .0239146     0.01   0.991     -.046606    .0471645
    cpetblap |   .4700703    .195755     2.40   0.016     .0862877    .8538529
    cpetwhip |   .8211552   .1936187     4.24   0.000     .4415608     1.20075
    cpethisp |   .5235963   .1938826     2.70   0.007     .1434846    .9037079
    cpetpacp |   .4689925   .2050314     2.29   0.022     .0670233    .8709617
    cpetecop |   -.007252   .0291075    -0.25   0.803    -.0643179     .049814
             |
        year |
       2005  |   5.144047   .1222464    42.08   0.000     4.904379    5.383714
       2006  |   6.222979   .1558198    39.94   0.000     5.917491    6.528468
       2007  |   8.739497     .19092    45.78   0.000     8.365194    9.113801
             |
    charter2 |
         Y   |          0  (omitted)
       _cons |   8.935894   19.30347     0.46   0.643    -28.90904    46.78083
-------------+----------------------------------------------------------------
     sigma_u |  10.502657
     sigma_e |   5.575046
         rho |  .78016966   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

(h) What statistical assumptions must hold in order to interpret the coefficient estimate in (g) as causal? Are they likely to hold here? Explain your answer. (**4 points**)

The fixed effects assumptions as described in the Wooldridge text are required. These include FE1 (linear model), FE2 (cross-sectional units are a random sample), FE3 (variation in x over time, with no perfect collinearity), and FE4 (strict exogeneity). The last assumption is rather important: there can effectively be no relationship between the population error term u and the x in any time period. In this context, this assumption would be violated if, for example, unusually low achievement in one year affected the mobility rate in another year, perhaps through a changing composition of students in the school. Assumptions 5-6 in Wooldridge relate to the error variance, and thus the appropriate calculation of standard errors. It is preferable to adjust standard errors for clustering at the school level in this context.

(i) For parts (i)-(j), keep only four variables—*campus, year, avgpassing* and *mobility*—and drop any cases where *avgpassing* or *mobility* are missing. Create a scatterplot showing the relationship between *avgpassing* and *mobility* and calculate the sample mean for these two variables. (**4 points**)

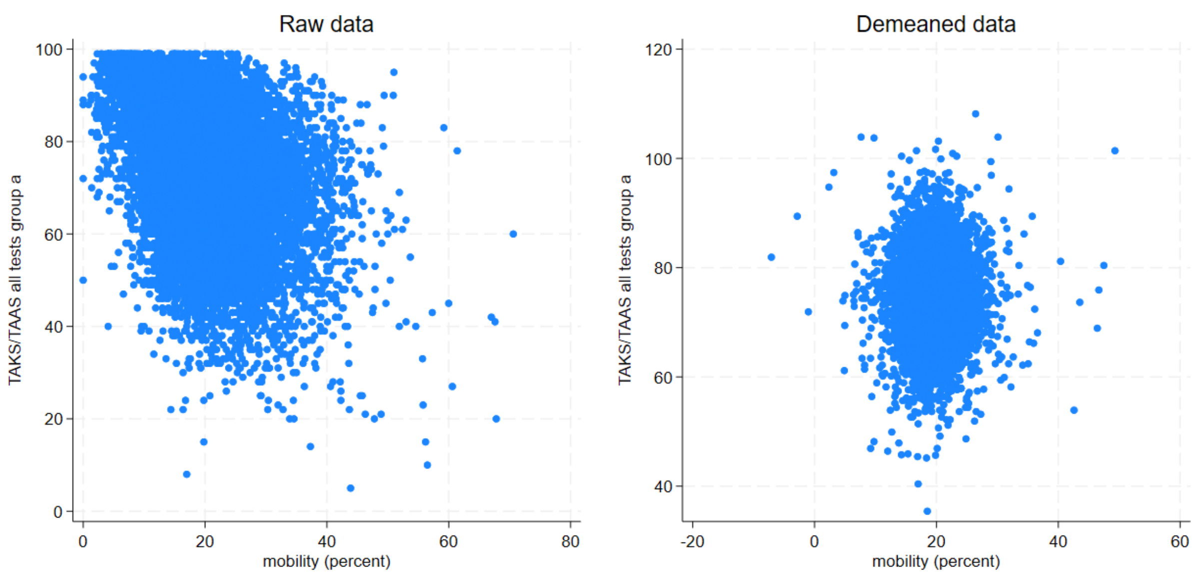
See scatterplots and output below.

(j) Use `xtdata` to transform your data using the fixed effects (within) transformation. Create another scatterplot showing the relationship between *avgpassing* and *mobility* and calculate the sample mean for these two variables. How do these compare with part (i), and what is the basic difference between these two? (**4 points**)

See scatterplots and output below. The means are the same for the raw and demeaned data. This is because Stata adds back the grand mean when demeaning the data: $(X_{it} - \bar{X}_i + \bar{X})$. It is easy to show that the average of these is the grand mean $\bar{X}$.

```
// Code for parts i-j:

.         keep campus year avgpassing mobility

.         drop if avgpassing==. | mobility==.
(1,239 observations deleted)

.         scatter avgpassing mobility, name(scatter1, replace) title(Raw data)

.         summ avgpassing mobility

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
```

```
avgpassing |     15,831    75.41141    13.80396          5         99
  mobility |     15,831     19.6594    7.612884          0       70.6

.         xtdata, fe clear

.         scatter avgpassing mobility, name(scatter2, replace) title(Demeaned data)

.         summ avgpassing mobility

  Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
avgpassing |     15,831    75.41141    5.577471   35.41141   108.1614
  mobility |     15,831     19.6594    2.765933  -7.090597    49.3094
```



Question 2. This problem will examine teacher effects on students' math and reading achievement using student-level data from a large urban school district. You will use methods that are closely related to those used in practice for estimating teacher "value-added." You can find the necessary data on Github under the name *LUSD4_5.dta*. All students in this database are in grades 4 and 5, and the test results are from 2005 and 2006. (**26 points**)

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/LUSD4_5.dta
```

Note, unlike Question 1, the regressions in this problem are not designed to estimate the causal effect of any particular input or intervention. Rather, we will be estimating fixed effects for individual teachers.

(a) First provide some descriptive information about the contents of this panel database. How many student observations are there in each grade and year? How many students appear in *both* grades 4 and 5 in this data? How many unique schools are in the data? How many unique teachers? The variable *school* is a unique school identifier, and *teacher* is the unique teacher identifier. Be clear in your Stata code how you answered these questions. (**3 points**)

See below. By using `xtset` with *id* and *grade* we can easily see how many students appear in both grades (N=9,728). There are other ways one can do this. There are 190 unique schools and 1,856 unique teachers.

```
. table grade year,row col

----------------------------------
grade      |     year (spring)
level      |   2005    2006   Total
----------+-----------------------
         4 | 12,116  11,556  23,672
         5 | 11,919  11,570  23,489
           |
     Total | 24,035  23,126  47,161
----------------------------------

. xtset id grade
       panel variable:  id (unbalanced)
        time variable:  grade, 4 to 5
                delta:  1 unit

. xtdescribe

      id:  9.000e+09, 9.000e+09, ..., 9.001e+09              n =      37433
   grade:  4, 5, ..., 5                                      T =          2
           Delta(grade) = 1 unit
           Span(grade)  = 2 periods
           (id*grade uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       1         1         2       2       2

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
    13944     37.25   37.25 |  1.
    13761     36.76   74.01 |  .1
     9728     25.99  100.00 |  11
 ---------------------------+---------
    37433    100.00         |  XX

. unique school
Number of unique values of school is  190
Number of records is  47161
```

```
. unique teacher
Number of unique values of teacher is  1856
Number of records is  47161
```

(b) Estimate four separate regressions: by grade (4 and 5) and by subject (math and reading). The dependent variable will be either the standardized math score (*mathz*) or standardized reading score (*readz*). Both are z-scores with a mean of zero and standard deviation of 1 (standardized for the grade, subject, and year). Use the following explanatory variables: age, female, LEP, special ed, immigrant, economically disadvantaged, black, Hispanic, Asian, and a year effect (i.e., a dummy variable for 2006). At this point, do not include any fixed effects. Provide a brief interpretation of your regression results. (**5 points**)

Results below. Across models, older students, special education, economically disadvantaged, LEP, black, and Hispanic students tend to score lower than their younger, non-special education, non-economically disadvantaged, non-LEP, white, and Asian counterparts. Girls tend to score lower in math than boys, but higher in reading. Scores tend to be higher in 2006 than in 2005. (This may seem unusual since these are standardized by year, but it may have to do with sample composition).

```
. foreach g in 4 5 {
  2. foreach s in math read {
  3.    reg ‘s’z age female lep speced immig econdis black hispanic asian i.year if grade==‘g’
  4.    }
  5.    }

      Source |       SS           df       MS      Number of obs   =    23,611
-------------+----------------------------------   F(10, 23600)    =    460.73
       Model |  3292.59677        10  329.259677   Prob > F        =    0.0000
    Residual |  16865.7702    23,600  .714651281   R-squared       =    0.1633
-------------+----------------------------------   Adj R-squared   =    0.1630
       Total |   20158.367    23,610  .853806311   Root MSE        =    .84537

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.2368717   .0104099   -22.75   0.000    -.2572758   -.2164675
      female |  -.0789751   .0110999    -7.11   0.000    -.1007316   -.0572187
         lep |  -.1231985   .0142853    -8.62   0.000    -.1511985   -.0951984
      speced |  -.7125033   .0255836   -27.85   0.000    -.7626488   -.6623578
       immig |   .0577393   .0327909     1.76   0.078    -.0065329    .1220115
     econdis |  -.3175577   .0181703   -17.48   0.000    -.3531727   -.2819427
       black |  -.6690733   .0238113   -28.10   0.000    -.7157449   -.6224016
    hispanic |   -.358638   .0237969   -15.07   0.000    -.4052814   -.3119945
       asian |   .2083883   .0362151     5.75   0.000     .1374043    .2793722
             |
```

```
        year |
        2006  |   .0770331   .0110199     6.99   0.000     .0554333    .0986328
             |
       _cons |    3.27814   .1069792    30.64   0.000     3.068454    3.487826
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =    22,963
-------------+----------------------------------   F(10, 22952)    =    372.57
       Model |  3163.10178        10  316.310178   Prob > F        =    0.0000
    Residual |  19486.2337    22,952  .848999379   R-squared       =    0.1397
-------------+----------------------------------   Adj R-squared   =    0.1393
       Total |  22649.3355    22,962  .986383395   Root MSE        =    .92141

------------------------------------------------------------------------------
       readz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1814648   .0116574   -15.57   0.000     -.204314   -.1586156
      female |   .1493099   .0122513    12.19   0.000     .1252966    .1733233
         lep |  -.1918733    .015795   -12.15   0.000    -.2228326   -.1609141
      speced |  -.4105074   .0342374   -11.99   0.000    -.4776151   -.3433998
       immig |   .3268251   .0380238     8.60   0.000     .2522958    .4013544
     econdis |  -.4380903   .0200412   -21.86   0.000    -.4773724   -.3988082
       black |  -.6574464    .026341   -24.96   0.000    -.7090765   -.6058163
    hispanic |  -.4320054   .0263343   -16.40   0.000    -.4836224   -.3803884
       asian |  -.0345248   .0398236    -0.87   0.386    -.1125817    .0435321
             |
        year |
        2006  |   .0089465   .0121804     0.73   0.463    -.0149278    .0328208
             |
       _cons |   2.709992   .1198535    22.61   0.000     2.475071    2.944912
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =    23,225
-------------+----------------------------------   F(10, 23214)    =    599.20
       Model |  3708.88513        10  370.888513   Prob > F        =    0.0000
    Residual |  14368.7779    23,214   .61897036   R-squared       =    0.2052
-------------+----------------------------------   Adj R-squared   =    0.2048
       Total |  18077.6631    23,224  .778404369   Root MSE        =    .78675

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.2383004   .0089102   -26.74   0.000    -.2557651   -.2208358
      female |   -.095322   .0104105    -9.16   0.000    -.1157274   -.0749167
         lep |  -.3295649    .013557   -24.31   0.000    -.3561376   -.3029922
      speced |  -.6832481   .0236204   -28.93   0.000    -.7295457   -.6369505
       immig |   .0375388   .0324508     1.16   0.247    -.0260669    .1011444
     econdis |  -.2781333   .0167387   -16.62   0.000    -.3109423   -.2453243
       black |  -.6007952   .0224095   -26.81   0.000    -.6447193   -.5568712
    hispanic |  -.2916257   .0222087   -13.13   0.000    -.3351562   -.2480952
       asian |   .2373296   .0338157     7.02   0.000     .1710486    .3036106
```

```
             |
        year |
       2006  |   .2088527   .0103401    20.20   0.000     .1885853    .2291201
             |
       _cons |   3.474444   .1005276    34.56   0.000     3.277403    3.671485
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =    22,699
-------------+----------------------------------   F(10, 22688)    =    822.96
       Model |  5973.09754        10  597.309754   Prob > F        =    0.0000
    Residual |  16467.1657    22,688  .725809489   R-squared       =    0.2662
-------------+----------------------------------   Adj R-squared   =    0.2659
       Total |  22440.2632    22,698  .988644957   Root MSE        =    .85194

------------------------------------------------------------------------------
       readz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.221383   .0098717   -22.43   0.000    -.2407321   -.2020338
      female |   .0124986   .0113864     1.10   0.272    -.0098195    .0348166
         lep |  -.7163557   .0148404   -48.27   0.000     -.745444   -.6872674
      speced |  -.4139044   .0316657   -13.07   0.000    -.4759713   -.3518375
       immig |   .1305244   .0404587     3.23   0.001     .0512226    .2098262
     econdis |  -.4631401   .0182894   -25.32   0.000    -.4989887   -.4272916
       black |  -.6213506   .0244451   -25.42   0.000    -.6692646   -.5734365
    hispanic |  -.4472469   .0242336   -18.46   0.000    -.4947465   -.3997474
       asian |  -.0723524   .0369619    -1.96   0.050    -.1448002    .0000954
             |
        year |
       2006  |   .0445679   .0113193     3.94   0.000     .0223813    .0667546
             |
       _cons |   3.548013   .1114301    31.84   0.000     3.329602    3.766423
------------------------------------------------------------------------------
```

(c) Now estimate the same regressions as in part (b), but add as an additional control the lagged math score (in the math regressions) and the lagged reading score (in the reading regressions). These variables are already in the dataset as *mathz_1* and *readz_1*. How do the results change, and how should our interpretation of these results change, given the inclusion of lagged (prior grade) achievement? (**5 points**)

Results shown below. Not surprisingly, the coefficient on the lagged score is positive and highly significant. (A student's score in the prior grade is a strong predictor of their score in the current grade). The interpretation of the other slope coefficients now differs since achievement in the prior grade is being controlled for. For example, the coefficient on *econdis* is now the predicted difference between the average scores of economically disadvantaged students and non-economically disadvantaged students, holding constant the other predictor variables in the model and prior achievement. For example, 4th grade students who are economically disadvantaged do

worse in math than their prior year's math score would predict. Some analysts think of this in terms of "gains," although we are not strictly modeling year-to-year gains.

```
. foreach g in 4 5 {
  2. foreach s in math read {
  3.    reg `s'z `s'z_1 age female lep speced immig econdis black hispanic asian i.year if grade==
  4.    }
  5.    }

      Source |       SS           df       MS      Number of obs   =    23,453
-------------+----------------------------------   F(11, 23441)    =   1752.16
       Model |  8622.61122        11  783.873747   Prob > F        =    0.0000
    Residual |  10486.9181    23,441  .447375029   R-squared       =    0.4512
-------------+----------------------------------   Adj R-squared   =    0.4510
       Total |  19109.5293    23,452  .814835804   Root MSE        =    .66886

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathz_1 |    .542333   .0047903   113.21   0.000     .5329437    .5517223
         age |  -.1578989     .00832   -18.98   0.000    -.1742066   -.1415912
      female |  -.0366597   .0088193    -4.16   0.000     -.053946   -.0193734
         lep |  -.0777164   .0113459    -6.85   0.000    -.0999552   -.0554776
      speced |  -.2977279   .0211226   -14.10   0.000    -.3391295   -.2563262
       immig |   .0804139   .0260006     3.09   0.002     .0294509    .1313769
     econdis |  -.1438634   .0144808    -9.93   0.000    -.1722466   -.1154801
       black |  -.3174066   .0191312   -16.59   0.000     -.354905   -.2799083
    hispanic |  -.1517818   .0189552    -8.01   0.000    -.1889352   -.1146283
       asian |   .1002384   .0287294     3.49   0.000     .0439268    .1565499
             |
        year |
       2006  |    .092809   .0087483    10.61   0.000     .0756618    .1099561
             |
       _cons |   1.992013   .0859064    23.19   0.000     1.823631    2.160395
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =    22,792
-------------+----------------------------------   F(11, 22780)    =   1508.25
       Model |  9446.96228        11  858.814753   Prob > F        =    0.0000
    Residual |  12971.1558    22,780  .569409826   R-squared       =    0.4214
-------------+----------------------------------   Adj R-squared   =    0.4211
       Total |  22418.1181    22,791  .983639073   Root MSE        =    .75459

------------------------------------------------------------------------------
       readz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     readz_1 |   .5962596   .0056649   105.26   0.000      .585156    .6073632
         age |  -.1141489   .0096121   -11.88   0.000    -.1329893   -.0953085
      female |   .0781508   .0100935     7.74   0.000     .0583668    .0979347
         lep |  -.1622692   .0129857   -12.50   0.000    -.1877221   -.1368162
```

```
      speced |  -.1671733   .0285109    -5.86   0.000    -.2230566     -.11129
       immig |   .2316611   .0330387     7.01   0.000      .166903    .2964193
     econdis |  -.2313162   .0165799   -13.95   0.000    -.2638139   -.1988185
       black |   -.403942   .0217743   -18.55   0.000    -.4466211   -.3612628
    hispanic |  -.2514092   .0217181   -11.58   0.000    -.2939782   -.2088403
       asian |   .0187075   .0328078     0.57   0.569    -.0455981    .0830131
             |
        year |
       2006  |  -.0005656   .0100142    -0.06   0.955     -.020194    .0190629
             |
       _cons |   1.609512   .0991515    16.23   0.000     1.415169    1.803856
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =    23,152
-------------+----------------------------------   F(11, 23140)    =   1853.74
       Model |  8238.08421        11  748.916746   Prob > F        =    0.0000
    Residual |  9348.65566    23,140  .404004134   R-squared       =    0.4684
-------------+----------------------------------   Adj R-squared   =    0.4682
       Total |  17586.7399    23,151  .759653573   Root MSE        =    .63561

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathz_1 |    .511132   .0047122   108.47   0.000     .5018958    .5203682
         age |  -.1221437   .0072934   -16.75   0.000    -.1364392   -.1078482
      female |  -.0350648   .0084422    -4.15   0.000    -.0516119   -.0185176
         lep |  -.1784876   .0110541   -16.15   0.000    -.2001544   -.1568209
      speced |   -.321753   .0197192   -16.32   0.000    -.3604039   -.2831022
       immig |   .0903732    .026243     3.44   0.001     .0389351    .1418112
     econdis |  -.1033876   .0136281    -7.59   0.000    -.1300997   -.0766755
       black |  -.2482319   .0184122   -13.48   0.000    -.2843211   -.2121428
    hispanic |  -.1219343   .0180299    -6.76   0.000    -.1572741   -.0865946
       asian |   .1048426   .0273801     3.83   0.000     .0511757    .1585095
             |
        year |
       2006  |   .2150489   .0083652    25.71   0.000     .1986525    .2314452
             |
       _cons |   1.680642   .0830348    20.24   0.000     1.517888    1.843396
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =    22,595
-------------+----------------------------------   F(11, 22583)    =   2078.42
       Model |  11231.9672        11  1021.08792   Prob > F        =    0.0000
    Residual |  11094.5981    22,583  .491280966   R-squared       =    0.5031
-------------+----------------------------------   Adj R-squared   =    0.5028
       Total |  22326.5652    22,594   .98816346   Root MSE        =    .70091

------------------------------------------------------------------------------
       readz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     readz_1 |   .5459355   .0052757   103.48   0.000     .5355948    .5562762
```

```
         age |   -.123624   .0082074   -15.06   0.000     -.139711    -.107537
      female |  -.0428927   .0094032    -4.56   0.000    -.0613237   -.0244617
         lep |  -.5386067   .0123582   -43.58   0.000    -.5628296   -.5143838
      speced |  -.1946905   .0267903    -7.27   0.000    -.2472012   -.1421797
       immig |   -.061676   .0345765    -1.78   0.074    -.1294483    .0060963
     econdis |  -.2332101   .0152689   -15.27   0.000    -.2631381    -.203282
       black |  -.3151816   .0204038   -15.45   0.000    -.3551746   -.2751887
    hispanic |  -.2482304   .0201228   -12.34   0.000    -.2876726   -.2087882
       asian |  -.0199688   .0305672    -0.65   0.514    -.0798826     .039945
             |
        year |
       2006  |   .0383332   .0093364     4.11   0.000     .0200332    .0566333
             |
       _cons |   1.972282    .093289    21.14   0.000     1.789429    2.155135
------------------------------------------------------------------------------
```

(d) Next, estimate the regressions in part (c) (with the lagged score), but this time use `xtreg` and include a fixed effect for the classroom teacher. (Instead of using `xtset`, you can include the options `fe` and `i(teacher)` in the `xtreg` command. This is equivalent to `xtset` without officially setting the panel variables). How should our interpretation of the coefficients change, if at all, given the inclusion of teacher fixed effects? (**5 points**)

Results below. The interpretations of the slope coefficients do not have a fundamentally different interpretation, but it is important to keep in mind that they are estimated using *within-teacher* variation in the covariates and achievement. So, for example, the achievement of girls is effectively compared with the achievement of boys in the same class.

```
. foreach g in 4 5 {
  2. foreach s in read math {
  3.    xtreg ‘s’z ‘s’z_1 age female lep speced immig econdis black hispanic asian ///
>    i.year if grade==‘g’, fe i(teacher)
  4.    }
  5.    }
warning: existing panel variable is not teacher

Fixed-effects (within) regression               Number of obs     =     22,792
Group variable: teacher                         Number of groups  =      1,065

R-sq:                                           Obs per group:
     within  = 0.3308                                         min =          1
     between = 0.5839                                         avg =       21.4
     overall = 0.4132                                         max =         46

                                                F(11,21716)       =     975.82
corr(u_i, Xb)  = 0.1508                         Prob > F          =     0.0000

------------------------------------------------------------------------------
```

```
       readz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     readz_1 |   .5593767    .005819    96.13   0.000      .547971    .5707824
         age |  -.0927569   .0092881    -9.99   0.000    -.1109623   -.0745514
      female |   .0797181   .0096228     8.28   0.000     .0608567    .0985794
         lep |  -.2897263   .0254108   -11.40   0.000    -.3395333   -.2399193
      speced |  -.1660317   .0276888    -6.00   0.000    -.2203038   -.1117597
       immig |   .2069517   .0318912     6.49   0.000     .1444425    .2694608
     econdis |  -.1093277   .0178174    -6.14   0.000    -.1442511   -.0744043
       black |  -.2515821   .0247615   -10.16   0.000    -.3001164   -.2030479
    hispanic |  -.1448184   .0235422    -6.15   0.000    -.1909628   -.0986741
       asian |    .010661   .0328858     0.32   0.746    -.0537977    .0751196
             |
        year |
       2006  |  -.0085693   .0117874    -0.73   0.467    -.0316735    .0145349
             |
       _cons |    1.24184   .0978645    12.69   0.000     1.050019    1.433662
-------------+----------------------------------------------------------------
     sigma_u |  .35924206
     sigma_e |  .70436283
         rho |  .20642771   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(1064, 21716) = 4.16                 Prob > F = 0.0000

Fixed-effects (within) regression               Number of obs     =     23,453
Group variable: teacher                         Number of groups  =      1,069

R-sq:                                           Obs per group:
     within  = 0.3763                                         min =          1
     between = 0.5623                                         avg =       21.9
     overall = 0.4478                                         max =         47

                                                F(11,22373)       =    1227.36
corr(u_i, Xb)  = 0.1465                         Prob > F          =     0.0000

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathz_1 |   .5197333    .004928   105.46   0.000      .510074    .5293926
         age |  -.1334734   .0078374   -17.03   0.000    -.1488352   -.1181116
      female |  -.0396879   .0082003    -4.84   0.000    -.0557612   -.0236146
         lep |  -.1174284   .0207741    -5.65   0.000    -.1581472   -.0767097
      speced |  -.2785083   .0200028   -13.92   0.000    -.3177153   -.2393014
       immig |   .0548364   .0245967     2.23   0.026     .0066251    .1030477
     econdis |  -.0509026   .0151735    -3.35   0.001    -.0806437   -.0211614
       black |  -.2164042   .0210678   -10.27   0.000    -.2576985   -.1751098
    hispanic |  -.1043176   .0199644    -5.23   0.000    -.1434492   -.0651861
       asian |    .063301   .0280419     2.26   0.024     .0083369    .1182651
             |
        year |
       2006  |   .0813267   .0100288     8.11   0.000     .0616696    .1009837
```

```
             |
       _cons |   1.636226   .0826756    19.79   0.000     1.474176    1.798276
-------------+----------------------------------------------------------------
     sigma_u |  .36165127
     sigma_e |  .60862724
         rho |  .26094723   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(1068, 22373) = 5.56                 Prob > F = 0.0000

Fixed-effects (within) regression               Number of obs     =     22,595
Group variable: teacher                         Number of groups  =        894

R-sq:                                           Obs per group:
     within  = 0.3515                                         min =          1
     between = 0.7229                                         avg =       25.3
     overall = 0.4917                                         max =         60

                                                F(11,21690)       =    1068.63
corr(u_i, Xb)  = 0.3050                         Prob > F          =     0.0000

------------------------------------------------------------------------------
       readz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     readz_1 |   .5321018   .0055098    96.57   0.000     .5213022    .5429015
         age |  -.0998047   .0080284   -12.43   0.000     -.115541   -.0840684
      female |  -.0481933   .0090702    -5.31   0.000    -.0659715   -.0304151
         lep |   -.327592   .0191981   -17.06   0.000    -.3652217   -.2899622
      speced |  -.1791101   .0261651    -6.85   0.000    -.2303956   -.1278247
       immig |  -.0770656   .0337926    -2.28   0.023    -.1433015   -.0108296
     econdis |  -.1130444   .0162896    -6.94   0.000    -.1449732   -.0811155
       black |  -.1930209   .0227613    -8.48   0.000    -.2376347   -.1484072
    hispanic |  -.1361477   .0216871    -6.28   0.000    -.1786559   -.0936394
       asian |  -.0317498   .0305681    -1.04   0.299    -.0916655    .0281659
             |
        year |
       2006  |   .0172887   .0109689     1.58   0.115     -.004211    .0387885
             |
       _cons |   1.464472   .0928889    15.77   0.000     1.282403    1.646541
-------------+----------------------------------------------------------------
     sigma_u |  .33181699
     sigma_e |  .65984112
         rho |  .20184045   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(893, 21690) = 4.25                  Prob > F = 0.0000

Fixed-effects (within) regression               Number of obs     =     23,152
Group variable: teacher                         Number of groups  =        898

R-sq:                                           Obs per group:
     within  = 0.3882                                         min =          1
     between = 0.6437                                         avg =       25.8
```

```
     overall = 0.4660                                     max =         59

                                                F(11,22243)      =    1283.27
corr(u_i, Xb)  = 0.1237                         Prob > F         =     0.0000

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathz_1 |   .5014138   .0048905   102.53   0.000     .4918281    .5109994
         age |  -.1038846   .0069652   -14.91   0.000     -.117537   -.0902323
      female |  -.0390795   .0079636    -4.91   0.000    -.0546887   -.0234704
         lep |  -.1188986   .0163934    -7.25   0.000    -.1510308   -.0867663
      speced |  -.2934953     .01881   -15.60   0.000    -.3303642   -.2566264
       immig |   .0893871   .0252091     3.55   0.000     .0399754    .1387988
     econdis |  -.0630879   .0142334    -4.43   0.000    -.0909864   -.0351894
       black |  -.2064401   .0200474   -10.30   0.000    -.2457344   -.1671457
    hispanic |   -.088959   .0190129    -4.68   0.000    -.1262257   -.0516923
       asian |   .0940341   .0267766     3.51   0.000     .0415501    .1465181
             |
        year |
       2006  |   .1579761     .00962    16.42   0.000     .1391201    .1768321
             |
       _cons |   1.424638   .0807365    17.65   0.000     1.266389    1.582887
-------------+----------------------------------------------------------------
     sigma_u |  .32092983
     sigma_e |  .58513872
         rho |  .23125241   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(897, 22243) = 5.64                 Prob > F = 0.0000
```

(e) Teacher fixed effects—systematic variation in achievement after controlling for prior student achievement and other student characteristics—are often referred to as the teacher's "value added." How much of the variance in achievement is due to the teacher effect? (This is reported as the "rho" in the `xtreg` output). (**3 points**)

The values of *rho* in the above regressions are 0.206, 0.261, 0.202, and 0.231. After controlling for prior achievement and other student characteristics, roughly 20-25% of the variation in achievement is attributable to variation across teachers. This provides some indication of the "importance" of teachers to student outcomes.

(f) Save the estimated teacher fixed effects using `predict`, as shown in class. Keep one observation per teacher (you can use `duplicates drop` to do this) and create a histogram of the estimated teacher fixed effects. What is the standard deviation of these teacher fixed effects? What is the difference between a teacher at the 75th percentile of the teacher effect distribution and a teacher at the 25th percentile? (**5 points**)

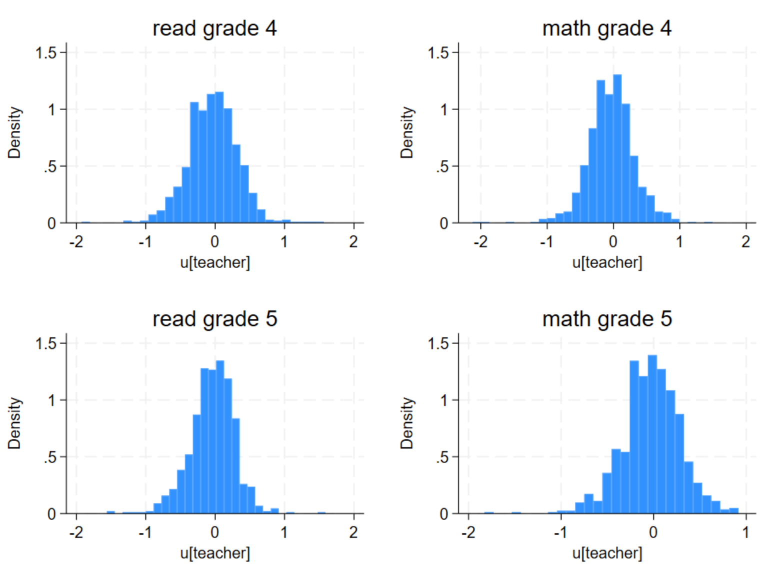
Stata syntax, output, and histograms are shown below. The standard devi-

ation in teacher effects ranges from 0.32 - 0.34, depending on the grade and subject. The difference between the 25th and 75th percentiles ranges from 0.38 to 0.45, depending on the grade and subject. What do these numbers mean? Recall that the fixed effects are estimates of unique intercepts for each teacher. In the case of 4th grade reading, a standard deviation of 0.35 means the students of a teacher one standard deviation above average perform 0.35 better than average than the students of the average teacher.

```
.         // Add teacher fixed effect
.         foreach g in 4 5 {
  2.         foreach s in read math {
  3.            qui xtreg ‘s’z ‘s’z_1 age female lep speced immig econdis black
> ///
>               hispanic asian i.year if grade==‘g’, fe i(teacher)
  4.
.            // Get estimated teacher fixed effects and keep one obs per teacher
.            predict tcheff‘s’‘g’, u
  5.            preserve
  6.                   duplicates drop teacher, force
  7.                   summ tcheff‘s’‘g’, detail
  8.                   tabstat tcheff‘s’‘g’, stat(p25 p75 iqr)
  9.                   histogram tcheff‘s’‘g’, name(q2f‘s’‘g’, replace) title("‘
> s’ grade ‘g’")
 10.            restore
 11.            }
 12.            }
(24,369 missing values generated)

Duplicates in terms of teacher

(45,305 observations deleted)

                         u[teacher]
-------------------------------------------------------------
      Percentiles      Smallest
 1%     -.863707      -1.925172
 5%    -.5911111       -1.28669
10%    -.4549403      -1.102798       Obs                 974
25%    -.2620801      -1.024002       Sum of Wgt.         974

50%    -.0418055                      Mean          -.0292627
                        Largest       Std. Dev.      .3520821
75%     .1851209        1.11441
90%     .3914686       1.243015       Variance       .1239618
95%     .5294719       1.418795       Skewness       .0488429
99%     .9337133       1.566302       Kurtosis       4.580015

    variable |       p25       p75       iqr
-------------+------------------------------
 tcheffread4 | -.2620801  .1851209   .447201
```

```
--------------------------------------------
(bin=29, start=-1.9251719, width=.12039566)
(23,708 missing values generated)

Duplicates in terms of teacher

(45,305 observations deleted)

                         u[teacher]
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.9152396      -1.986226
 5%    -.5875053      -1.594639
10%    -.4370334      -1.163883       Obs               1,004
25%    -.2358474      -1.062243       Sum of Wgt.       1,004

50%    -.0311564                      Mean           -.0344432
                        Largest       Std. Dev.       .3442303
75%      .169738        .890529
90%     .3790837       .9583023       Variance        .1184945
95%     .5267823       1.196018       Skewness       -.1834778
99%     .8022034       1.493365       Kurtosis        4.888053

    variable |       p25       p75       iqr
-------------+------------------------------
 tcheffmath4 | -.2358474   .169738  .4055854
--------------------------------------------
(bin=30, start=-1.9862257, width=.11598635)
(24,566 missing values generated)

Duplicates in terms of teacher

(45,305 observations deleted)

                         u[teacher]
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.8895572      -1.868376
 5%    -.5953411      -1.559686
10%    -.4524955      -1.512358       Obs                 806
25%     -.219957      -1.300198       Sum of Wgt.         806

50%    -.0207986                      Mean           -.0439087
                        Largest       Std. Dev.       .3286723
75%     .1577055       .8488992
90%     .3190411       .8810328       Variance        .1080255
95%     .4412906       1.081512       Skewness       -.5181577
99%     .6588145       1.588887       Kurtosis        5.845002

    variable |       p25       p75       iqr
-------------+------------------------------
```

```
 tcheffread5 |  -.219957  .1577055  .3776625
---------------------------------------------
(bin=28, start=-1.8683757, width=.12347366)
(24,009 missing values generated)

Duplicates in terms of teacher

(45,305 observations deleted)

                          u[teacher]
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.8771342      -1.827223
 5%    -.5433015      -1.528425
10%    -.4240153      -1.386065       Obs                 823
25%    -.2257731      -1.339902       Sum of Wgt.         823

50%    -.0290482                      Mean          -.0414026
                        Largest       Std. Dev.      .3205586
75%     .1641627       .8242272
90%     .3249792       .8543559       Variance       .1027578
95%     .4475881       .8997599       Skewness      -.5513376
99%     .6830953       .9163185       Kurtosis       5.204493

    variable |       p25       p75       iqr
-------------+-------------------------------
 tcheffmath5 | -.2257731  .1641627  .3899358
---------------------------------------------
(bin=28, start=-1.8272233, width=.09798364)
```



Question 3. This problem will use the same student-level data from a large urban school district to estimate the impact of having a same-race teacher on achievement. (That is, how a student performs when they share the same race/ethnicity as their teacher, relative to when they don't.) For a study that tackles this very question see Dee (2004). (**20 points**)

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/LUSD4_5.dta
```

(a) Create a variable called *same_race* that equals zero unless the student and teacher share the same race/ethnicity, in which case *same_race* should be coded as one. Use the white, black, Hispanic, and Asian categories, but not the "other" race category. In what percent of cases (i.e., student-year observations) are students assigned to a teacher of the same race/ethnicity? How does this rate of same race exposure vary by student race/ethnicity? (**4 points**)

Results below. In about 52% of cases (student x year observations) the student had a teacher with the same race or ethnicity. This percentage was higher for black and white students (at 73-74%) and lower for Hispanic (42%) and Asian (6%) students.

```
. gen same_race = 0

. replace same_race = 1 if tch_black==1 & black==1
(8,708 real changes made)

. replace same_race = 1 if tch_white==1 & white==1
(3,295 real changes made)

. replace same_race = 1 if tch_hisp==1 & hisp==1
(12,341 real changes made)

. replace same_race = 1 if tch_asian==1 & asian==1
(87 real changes made)

. tabulate same_race

  same_race |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     22,730       48.20       48.20
          1 |     24,431       51.80      100.00
------------+-----------------------------------
      Total |     47,161      100.00

. foreach j in black white hisp asian {
  2.    tabulate same_race if `j'==1
  3.    }

  same_race |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,221       27.00       27.00
```

```
          1 |      8,708       73.00      100.00
------------+-----------------------------------
      Total |     11,929      100.00

  same_race |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,182       26.40       26.40
          1 |      3,295       73.60      100.00
------------+-----------------------------------
      Total |      4,477      100.00

  same_race |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |     16,915       57.82       57.82
          1 |     12,341       42.18      100.00
------------+-----------------------------------
      Total |     29,256      100.00

  same_race |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,391       94.11       94.11
          1 |         87        5.89      100.00
------------+-----------------------------------
      Total |      1,478      100.00
```

(b) Estimate two regressions where the dependent variables are the math and reading z-scores, respectively, and *same_race* is the explanatory variable. Explain why the estimated coefficient on *same_race* should not be interpreted as causal. (**4 points**)

Results below, with separate models by subject and grade. In all cases, students with a same race/ethnicity teacher tend to perform worse, on average, than students who do not. For these regressions to have a causal interpretation, we have to believe that the covariance between the population error term *u* and *same_race* is zero. This seems unlikely if there are omitted variables correlated with both test scores and a match with a same race/ethnicity teachers. As the correlation matrix shows, black and LEP students are more likely to have a same race teacher. But these students also tend to have lower achievement, on average.

```
. foreach g in 4 5 {
  2. foreach s in math read {
  3.    reg `s'z same_race if grade==`g'
  4.    }
  5.    }

      Source |       SS           df       MS      Number of obs   =    23,611
-------------+----------------------------------   F(1, 23609)     =     40.26
       Model |  34.3198109         1  34.3198109   Prob > F        =    0.0000
    Residual |  20124.0472    23,609    .8523888   R-squared       =    0.0017
```

```
-------------+----------------------------------   Adj R-squared   =    0.0017
       Total |   20158.367    23,610  .853806311   Root MSE        =    .92325

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   same_race |  -.0767708   .0120988    -6.35   0.000    -.1004852   -.0530563
       _cons |    .170034   .0090384    18.81   0.000     .1523181    .1877499
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =    22,963
-------------+----------------------------------   F(1, 22961)     =     23.42
       Model |  23.0795803         1  23.0795803   Prob > F        =    0.0000
    Residual |  22626.2559    22,961   .98542119   R-squared       =    0.0010
-------------+----------------------------------   Adj R-squared   =    0.0010
       Total |  22649.3355    22,962  .986383395   Root MSE        =    .99268

------------------------------------------------------------------------------
       readz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   same_race |  -.0638795   .0131995    -4.84   0.000    -.0897515   -.0380075
       _cons |   .1045238   .0098844    10.57   0.000     .0851497    .1238979
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =    23,225
-------------+----------------------------------   F(1, 23223)     =    121.08
       Model |  93.7662193         1  93.7662193   Prob > F        =    0.0000
    Residual |  17983.8969    23,223  .774400244   R-squared       =    0.0052
-------------+----------------------------------   Adj R-squared   =    0.0051
       Total |  18077.6631    23,224  .778404369   Root MSE        =       .88

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   same_race |  -.1272158   .0115611   -11.00   0.000    -.1498764   -.1045552
       _cons |   .2204888   .0079835    27.62   0.000     .2048406     .236137
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =    22,699
-------------+----------------------------------   F(1, 22697)     =    159.84
       Model |  156.923336         1  156.923336   Prob > F        =    0.0000
    Residual |  22283.3399    22,697  .981774679   R-squared       =    0.0070
-------------+----------------------------------   Adj R-squared   =    0.0069
       Total |  22440.2632    22,698  .988644957   Root MSE        =    .99085

------------------------------------------------------------------------------
       readz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   same_race |  -.1664715   .0131675   -12.64   0.000    -.1922807   -.1406624
       _cons |   .1438336   .0090919    15.82   0.000     .1260129    .1616542
------------------------------------------------------------------------------
```

```
. corr same_race black white hisp asian lep speced econdis
(obs=47,161)

             | same_r~e    black    white hispanic    asian      lep   speced  econdis
-------------+------------------------------------------------------------------------
   same_race |   1.0000
       black |   0.2468   1.0000
       white |   0.1413  -0.1884   1.0000
    hispanic |  -0.2461  -0.7438  -0.4140   1.0000
       asian |  -0.1653  -0.1047  -0.0583  -0.2299   1.0000
         lep |   0.2429  -0.3901  -0.2158   0.5182  -0.1054   1.0000
      speced |  -0.0164   0.0115   0.0511  -0.0335  -0.0205  -0.0273   1.0000
     econdis |   0.0224   0.0237  -0.5321   0.3632  -0.1759   0.2816  -0.0146   1.0000
```

(c) Briefly explain how a regression model with *student fixed effects* might improve upon the regressions in part (b). What problem might this solve? (**2 points**)

There are likely to be observable and unobservable factors correlated with achievement and assignment to a same-race teacher. Some of this may have do with geography and the local teacher labor market—that is, whether or not teachers share the same demographics as their students. Student fixed effects estimate the "same race" effect using *within-student* variation over time. Students would effectively be compared against themselves, in states in which they are and are not exposed to a same-race teacher. Importantly, students that experience no variation in this explanatory variable do not contribute to the coefficient estimates. This is relevant if we are concerned about generalizing to the full population of students.

(d) Use `xtset` to designate student as the panel variable, and year as the time dimension. Estimate the same regressions as in Question #3 part (d) (with student covariates and lagged score), and use `xtreg, fe` to include student fixed effects. Also include *same_race* among your explanatory variables. Do **not** run the model separately by grade; you need multiple observations per student for this model to make sense. Describe what you find for the "same race" coefficient. Is it statistically significant? Practically significant? Can one make a strong claim for causal inference in this case? Explain why or why not. (**6 points**)

Results below. Interestingly, in all cases the coefficient on *same_race* is positive and statistically significant. When students share the same race/ethnicity as their teacher, they score 0.09 sd higher in reading and 0.04 sd higher in math. Both are statistically and (I would argue) practically significant. It is easier to make a casual claim in this case. One would be concerned about omitted variables bias if there were a time-varying omitted variable that is correlated with changes in both *same_race* and test scores. (This would represent a violation of the strict exogeneity assumption). If, for example,

parents responded to a worse- or better-than-expected test result by purposefully moving their student into a classroom with a same-race teacher, this would be a violation of strict exogeneity. It's not clear whether this is likely to occur in practice, however.

```
. xtset id year
       panel variable:  id (unbalanced)
        time variable:  year, 2005 to 2006
                delta:  1 unit

.
. foreach s in read math {
  2.    xtreg `s'z `s'z_1 age female lep speced immig econdis black hispanic asian i.year same_rac
  3.    }

Fixed-effects (within) regression               Number of obs     =     45,387
Group variable: id                              Number of groups  =     35,987

R-sq:                                           Obs per group:
     within  = 0.1598                                         min =          1
     between = 0.1175                                         avg =        1.3
     overall = 0.1082                                         max =          2

                                                F(12,9388)        =     148.80
corr(u_i, Xb)  = -0.6685                        Prob > F          =     0.0000

------------------------------------------------------------------------------
       readz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     readz_1 |  -.3575557   .0087662   -40.79   0.000    -.3747392   -.3403721
         age |  -.3907786   .1608187    -2.43   0.015    -.7060181   -.0755392
      female |  -.0691433   .2348174    -0.29   0.768    -.5294362    .3911496
         lep |   .0530388   .0259626     2.04   0.041     .0021465    .1039311
      speced |  -.0738881    .072518    -1.02   0.308    -.2160391    .0682628
       immig |   .3530726    .058952     5.99   0.000     .2375138    .4686314
     econdis |  -.0803718   .0346397    -2.32   0.020    -.1482732   -.0124704
       black |   .2817545   .6501012     0.43   0.665    -.9925848    1.556094
    hispanic |   .1130556   .5752511     0.20   0.844    -1.014561    1.240672
       asian |   .8782935   .5752299     1.53   0.127    -.2492817    2.005869
             |
        year |
       2006  |   .3515197   .1610366     2.18   0.029      .035853    .6671863
             |
   same_race |   .0850675   .0148206     5.74   0.000      .056016     .114119
       _cons |   4.011723   1.734289     2.31   0.021     .6121403    7.411306
-------------+----------------------------------------------------------------
     sigma_u |  1.2400891
     sigma_e |  .52496864
         rho |  .84802577   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
F test that all u_i=0: F(35986, 9388) = 2.22                Prob > F = 0.0000

Fixed-effects (within) regression               Number of obs     =     46,605
Group variable: id                              Number of groups  =     37,022

R-sq:                                           Obs per group:
     within  = 0.2411                                         min =          1
     between = 0.4617                                         avg =        1.3
     overall = 0.3945                                         max =          2

                                                F(12,9571)        =     253.41
corr(u_i, Xb)  = -0.8502                        Prob > F          =     0.0000

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathz_1 |  -.4139616   .0081244   -50.95   0.000    -.4298872   -.3980361
         age |   .0713948   .1413035     0.51   0.613      -.20559    .3483795
      female |   .1176982   .2064662     0.57   0.569    -.2870193    .5224157
         lep |   .1201043   .0227077     5.29   0.000     .0755924    .1646162
      speced |  -.1588845   .0523051    -3.04   0.002    -.2614137   -.0563553
       immig |   .0420339   .0507581     0.83   0.408    -.0574628    .1415305
     econdis |  -.0212018   .0302013    -0.70   0.483    -.0804028    .0379992
       black |   .5786446   .5839093     0.99   0.322    -.5659413    1.723231
    hispanic |   .3287884   .5057374     0.65   0.516    -.6625641    1.320141
       asian |   -.188493   .5058222    -0.37   0.709    -1.180012    .8030258
             |
        year |
       2006  |   .0766177   .1414914     0.54   0.588    -.2007355    .3539709
             |
   same_race |   .0414423   .0128886     3.22   0.001     .0161779    .0667067
       _cons |  -1.042015   1.524875    -0.68   0.494    -4.031092    1.947062
-------------+----------------------------------------------------------------
     sigma_u |  1.2695828
     sigma_e |  .46158338
         rho |  .88324881   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(37021, 9571) = 2.29                Prob > F = 0.0000
```

(e) Are there any explanatory variables that are dropped in the models in (d)? Are there any explanatory variables that should be dropped that weren't? What does the latter indicate to you? (**2 points**)

There are no explanatory variables dropped in the above models. One would expect time-invariant variables such as gender and student race/ethnicity to fall out of the regression, but they appear not to have done so in this case. This suggests there is unexpected variation in these variables, perhaps due to miscoding or other errors.

(f) Finally, use the command `xttrans` to describe the frequency of changes in exposure to a same-race teacher over time. Interpret the results of this command. (**2 points**)

Results below. The panel used in the above regressions is unbalanced—some students are observed in two years, but many are only observed in one. Identification of the *same_race* coefficient only comes from students observed in more than one year, who experience a change in *same_race*. The `xttrans` output only pertains to the students observed in more than one year.

Note the row percentages of the `xttrans` output sum to 100, and cell frequencies sum to 9,728, the total number of students observed in both periods. Of the 4,336 students who do *not* have a same race teacher in year 1, 78% again do not have a same race teacher in year 2. 22% do. Of the 5,392 students who *do* have a same race teacher in year 1, 67% continue to do so in year 2. 33% do not. Taken together, only 973+1,795 of the students experienced a switch in the *same_race* variable, or about 28% of all students. If you were concerned that these students represent an unusual population, you could look descriptively at these students and contrast them with students that did not experience such a change. For example, are they more likely to live in urban areas? Did they change schools or districts?

```
. egen count=count(id),by(id)
. table year if count==2

----------------------
year      |
(spring)  |      Freq.
----------+-----------
     2005 |      9,728
     2006 |      9,728
----------------------

. tabulate same_race if year==2005 & count==2

  same_race |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      4,336       44.57       44.57
          1 |      5,392       55.43      100.00
------------+-----------------------------------
      Total |      9,728      100.00

. xttrans same_race, freq

           |       same_race
 same_race |         0          1 |     Total
-----------+----------------------+----------
         0 |     3,363        973 |     4,336
```

	77.56	22.44	100.00
1	1,795	3,597	5,392
	33.29	66.71	100.00
Total	5,158	4,570	9,728
	53.02	46.98	100.00