Homework #5

This homework is due on Monday, March 6th.

1. The data file WA BUILDING DATA has been placed on the class website. This is an unbalanced panel of all Washington State public K-12 schools between 2002 and 2011. It contains annual observations of building-level demographic data as well as some information required by the No Child Left Behind Act (NCLB). The NCLB required each building to have a certain fraction of their students pass a standardized math and reading tests. In this case, the percent of a building's students passing these exams are "*math_pass*" and "*reading_pass*."

   One argument against the NCLB is that the percent of high performing students is a function of school demographics—students from disadvantaged backgrounds are unlikely to do well on standardized exams. To make matters worse, the NCLB removed resources from schools which had too few students achieving passing scores on these standardized tests. You will explore these claims in this homework.

   a. Perform an OLS regression on the pooled data with your model being:

      (1) $math\_pass_i = \beta_0 + \beta_1 perwhitei + \beta_2 perfreelunch_i + \beta_3 avgexp_i + \beta_4 studperteacheri_i + \varepsilon_i$

      Comment on β_2. How do you interpret this coefficient?

```
. reg math_pass perwhite perfreelunch avgexp studperteacher

      Source |       SS           df       MS      Number of obs   =    17,763
-------------+----------------------------------   F(4, 17758)     =   2050.27
       Model |  2063536.61         4  515884.151   Prob > F        =    0.0000
    Residual |  4468223.29    17,758  251.617485   R-squared       =    0.3159
-------------+----------------------------------   Adj R-squared   =    0.3158
       Total |   6531759.9    17,762  367.737862   Root MSE        =    15.862

------------------------------------------------------------------------------
   math_pass | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     perwhite |   .0425097   .0065606     6.48   0.000     .0296503    .055369
perfreelunch |  -.4470066   .0064552   -69.25   0.000    -.4596595   -.4343538
      avgexp |  -.0184969   .0439272    -0.42   0.674    -.1045985    .0676047
studperteacher |  -.1801058   .0140013   -12.86   0.000    -.2075497   -.1526619
       _cons |   71.83993   .8395207    85.57   0.000     70.19439    73.48547
------------------------------------------------------------------------------
```

$\hat{\beta}_2$ = -0.447066, this means that on average, it is expected that the percentage of students passing the standardized math test decreases by 0.447066 when there is a 1% increase in the percentage of students qualifying for free or reduced lunches while holding all other variables constant.

   b. Perform an OLS regression on the pooled data with your model being:

      (2) $math\_pass_i = \beta_0 + \beta_1 perwhitei + \beta_2 perfreelunch_i + \beta_3 avgexp_i + \beta_4 studperteacheri_i + \alpha i + \varepsilon_i$

      Comment on β_2. How do you interpret this coefficient? How did your estimate of β_2 change relative to your pooled OLS estimates of (1)? Provide an explanation for this change.

```
. xtreg math_pass perwhite perfreelunch avgexp studperteacher, fe

Fixed-effects (within) regression               Number of obs     =     17,763
Group variable: bldg                            Number of groups  =      2,048

R-squared:                                      Obs per group:
     Within  = 0.0035                                         min =          1
     Between = 0.1886                                         avg =        8.7
     Overall = 0.1540                                         max =         10

                                                F(4,15711)        =      13.83
corr(u_i, Xb) = 0.3620                          Prob > F          =     0.0000

------------------------------------------------------------------------------
   math_pass | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     perwhite |   .0613373   .0101018     6.07   0.000     .0415365    .081138
perfreelunch |  -.0126137   .0120768    -1.04   0.296    -.0362855    .0110581
      avgexp |   .1939155   .0554206     3.50   0.000     .0852846    .3025463
studperteacher |  -.0098184   .0126395    -0.78   0.437    -.0345934    .0149565
       _cons |   46.40884   1.213953    38.23   0.000     44.02935    48.78833
-------------+----------------------------------------------------------------
     sigma_u |  17.280053
     sigma_e |  10.32602
         rho |  .73687165   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(2047, 15711) = 12.80                 Prob > F = 0.0000
```

$\hat{\beta}_2$ = -0.01261257, our estimate for β_2 increased by 0.46444743. Our new $\hat{\beta}_2$ is interpreted as follows, on average, the building's percentage of students who pass the standardized math test decreases by 0.0126137 per 1% increase of students who qualify for free or reduced lunches in that building. If α_i wasn't correlated to X_i, then our $\hat{\beta}_i$ would remain the same. Adding a fixed and school specific error term (that is an adjustment from the mean) influences the intercepts, which then requires our estimation line to adjust its slope to better fit the data. This could explain our change in $\hat{\beta}_2$.

c. When I estimate (1) with fixed effects, I find ρ = .736. What does this mean? Specifically, for schools attempting to raise the number of students passing the math exam, is a high ρ or a low ρ better?

- Our standard error of our alphas is sigma_u = 17.280053.
- Our standard error of our epsilons is sigma_e = 10.32602 is our standard error of our error terms.
- ρ = .73687165, is our percentage of variation in our error terms that is driven by our alphas.

This is interpreted as roughly 74% of our error terms are driven by our alphas while the other 26% is driven by our error terms. In other words, since our alphas are constant across schools while our variation of our error terms across time and within schools, the bigger variation is happening across schools. This means that big variation of math pass rates is not happening because of what's happening within the school but more so because of the students attending one school versus another school.

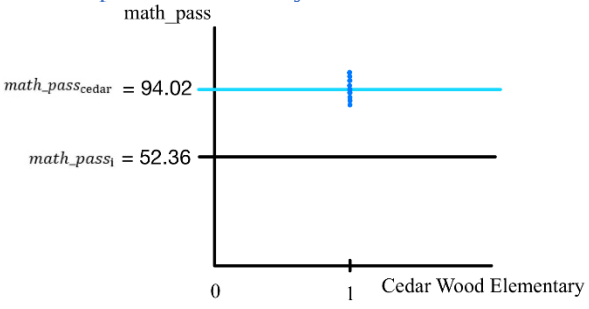
This implies that a low rho is better for schools because the school can influence the factors within their school. Since the rho is high, its percentage of students who pass the math test is more influenced by the school itself, which could be related to exogenous variables such as school wealth or location.

d. Does including fixed effects explain a statistically significant amount of the variation in *math_pass*? Given your answer, what does this mean for schools attempting to increase the number of students passing the math exam?

From the Fixed Effect regression, we look at our F Test for all u_i or $\alpha = 0$. Here, we have F(2047, 15711) = 12.80 with Prob > F = 0.00. We reject the null hypothesis that all u_i or $\alpha = 0$. This concludes that we are dealing with panel data and should include the alphas and including the fixed effects explains a statistically significant amount of variation in math_pass.

This means that there are a number of unobserved exogenous variables affecting the percentage of students passing the math test for that school. This suggests that schools attempting to increase the number of students passing the math exam should focus on external factors attributing to their ability to produce higher passing percentages rather than internal factors. Possible external factors could include school wealth, location, zoning perimeters, etc.

e. When I examine the fixed effects from (1), I find that Cedar Wood Elementary School in the Everett School District has an alpha = 40.7 meaning that 40.7% more of this school's children pass the math exam than would be predicted by *perwhite*, *perfreelunch*, *avgexp*, and *studperteacher*. For a while, the State of Washington gave out awards to schools that had high alphas. Comment on this practice. Would you want to reward Cedar Wood Elementary?



Awarding the schools with high alphas seems like a misinterpretation of what alpha represents. Above is a visualization of the average line of all schools over time and math_pass as well as the average line of Cedar Wood Elementary's average math_pass over time. In our fixed effect regression, Cedar Wood's alpha is the difference between the two lines and is the school's adjustment from the mean.

Cedar Wood's average percentage of students who pass the math test is around 40% higher than the average. Awarding the school with a high alpha without accounting for the average rho value, makes this award meaningless. Without the interpretation of the on average high rho value, it may seem as if the school is highly successful at producing students capable of passing the math test. However, with accounting for the on average high rho value, giving this award based on a high alpha seems biased and is like giving the gold medal to the winner of a race without accounting for the head start they had. I would not want to reward Cedar Wood Elementary solely based on its alpha value.

f. Is the fixed effects approach appropriate in this case? Should random effects be used? Test this.

```
. xtreg math_pass perwhite perfreelunch avgexp studperteacher, re

Random-effects GLS regression                   Number of obs     =     17,763
Group variable: bldg                            Number of groups  =      2,048

R-squared:                                      Obs per group:
     Within  = 0.0018                                         min =          1
     Between = 0.3494                                         avg =        8.7
     Overall = 0.2933                                         max =         10

                                                Wald chi2(4)      =     726.18
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000

------------------------------------------------------------------------------
   math_pass | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     perwhite |   .0993381   .0087349    11.37   0.000     .0822181    .1164581
perfreelunch |  -.1861077   .0097141   -19.16   0.000     -.205147   -.1670684
      avgexp |   .0932802   .0507406     1.84   0.066    -.0061695     .19273
studperteacher |  -.0346009   .0121711    -2.84   0.004    -.0584558   -.0107459
       _cons |   51.77964   1.119447    46.25   0.000     49.58556    53.97371
-------------+----------------------------------------------------------------
     sigma_u |  13.483943
     sigma_e |   10.32602
         rho |   .63033759   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. hausman fe re

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       fe           re         Difference          Std. err.
-------------+----------------------------------------------------------------
     perwhite |   .0613373    .0993381       -.0380008        .0050744
perfreelunch |  -.0126137   -.1861077         .173494        .0071753
      avgexp |   .1939155    .0932802        .1006352        .0222899
studpertea~r |  -.0098184   -.0346009        .0247824         .003409
------------------------------------------------------------------------------
                           b = Consistent under H0 and Ha; obtained from xtreg.
            B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

Test of H0: Difference in coefficients not systematic

    chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            = 672.29
Prob > chi2 = 0.0000
```

$$H_0: cov(\alpha, x) = 0$$
$$H_1: cov(\alpha, x) \neq 0$$

Using the Hausman test, our null hypothesis is the difference in coefficients is not systematic or random Effect model is more appropriate and our alternative hypothesis is the fixed model is more appropriate. Our result is rejecting the null, meaning $cov(\alpha, x) \neq 0$. In other words, the alpha's and x's are correlated.