Statistics/Data Analysis

```
      name:  main
       log:  C:\Users\Conor\Documents\Conor\Grad School\TA Work\Econ 103 - Econometric
> s\Stata Work\Week 8\wk8_section_log.smcl
  log type:  smcl
 opened on:  14 Feb 2018, 14:51:34

1 .
2 . // Demonstration STATA code for week 8
3 . // Principles of Econometrics 4th Edition
4 . // Covered Problems: 7.9
5 .
6 . set more off

7 . clear all

8 . use star.dta, clear

9 .
10. ////////////////////////////////////////////\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
  > ////////////////////////////////// Question 7.9 \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
  > ////////////////////////////////////////////\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
  >
11. ********************************************************************************
12. *Setup: We consider data from an experiment where classrooms were divided into
13. * three groups: (1) small class (13-17 students), (2) regular-size classes (22-
14. * 25 students), and (3) regular-size classes with a teaching assistant. The data
15. * include test scores, along with some student, teacher, and school
16. * characteristics.
17. *
18. * Parts (A) - (G)
19. ********************************************************************************
20.
21. ********************************************************************************
22. *Part A: Calculate the average of TOTALSCORE for (i) students in regular-sized
23. * classrooms with full time teachers, but no aide (ii) stduents in regular-sized
24. * cassrooms with full time teachers, and an aide, and (iii) students in small
25. * classrooms. What do you observe about test scores in these three types of
26. * learning environments?
27. ********************************************************************************
28.
29. // Each of the three class types is associated with its own dummy variable.
30. // Despite the name, regular and aide are muutally exclusive (i.e. if
31. // regular = 1, then aide must be 0.) Together, each observation has a 1 for
32. // at least one of small, regular, and aide.
33.
34. sum totalscore if small == 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  totalscore |      1,738    931.9419    76.35863       747       1253

35. sum totalscore if regular == 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  totalscore |      2,005    918.0429    73.13799       635       1229

36. sum totalscore if aide == 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  totalscore |      2,043    918.3568    71.31358       719       1253
```

```
37.
38. // Looking at these simple, UNCONDITIONAL, averages, the small classrooms appear
39. // to have higher average scores than regular-size classes, while there is not
40. // much difference between regular classes with and without an aide.
41.
42. ********************************************************************************
43. *Part B: Estimate the regression model
44. *
45. *    TOTALSCORE_i = beta1 + beta2*SMALL_i + beta3*AIDE_i + e_i
46. *
47. * where AIDE in a indicator variable equaling one for classes taught by a
48. * teacher and an aide, and zero otherwise. What is the relation of the estimated
49. * coefficients from this regression to the sample means in part (a)? Test the
50. * significance of beta3 at the 5% level of significance.
51. ********************************************************************************
52.
53. reg totalscore small aide
```

Source	SS	df	MS		
Model	**229672.452**	**2**	**114836.226**	Number of obs	= **5,786**
Residual	**31232400.3**	**5,783**	**5400.72632**	F(2, 5783)	= **21.26**
				Prob > F	= **0.0000**
Total	**31462072.8**	**5,785**	**5438.56055**	R-squared	= **0.0073**
				Adj R-squared	= **0.0070**
				Root MSE	= **73.49**

totalscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
small	**13.89899**	**2.40854**	**5.77**	**0.000**	**9.177354**	**18.62063**
aide	**.3139354**	**2.310229**	**0.14**	**0.892**	**-4.214977**	**4.842848**
_cons	**918.0429**	**1.641228**	**559.36**	**0.000**	**914.8255**	**921.2603**

```
54.
55. /* Discussion:
  >
  > When we have only non-interacted dummy variables in the regression, the constant
  > in the regression refers to the average for the excluded group. In this case,
  > the excluded group is the non-small, non-aide classes - the "regular" group.
  > Compare the beta estimate for the constant to the mean totalscore for group 2
  > (regular-sized classes).
  >
  > The beta on the dummy variable tells us the gap between the average for the
  > dummy group (e.g. small or aide) and the excluded group. In addition, if we
  > wanted to know the average difference between the two dummy groups, small and
  > aide, we would take the difference between the beta estimates for small and aide
  > respectively. Compare the difference in means we calcuated in Part (A) to the
  > beta estimates here.
  >
  > Testing the null that beta3 = 0 is the same as asking if the average test score
  > differs between the "regular" group and the "aide" group. The STATA output
  > already gave us the information we need for this test. The p-value of 0.892
  > means we fail to reject the null that beta3 = 0.
  >
  > */
56.
57. ********************************************************************************
58. *Part C: To the regression in (b) add the additional explanatory variable
59. * TCHEXPER. Is this variable statistically significant? Does its addition to the
60. * model affect the estimates of beta2 and beta3?
```

```
61. ********************************************************************************
62.
63. reg totalscore small aide tchexper
```

Source	SS	df	MS		
Model	**638736.792**	**3**	**212912.264**	Number of obs	= **5,766**
Residual	**30777099.3**	**5,762**	**5341.39175**	F(3, 5762)	= **39.86**
				Prob > F	= **0.0000**
Total	**31415836.1**	**5,765**	**5449.40782**	R-squared	= **0.0203**
				Adj R-squared	= **0.0198**
				Root MSE	= **73.085**

totalscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
small	**14.00613**	**2.395304**	**5.85**	**0.000**	**9.310438**	**18.70183**
aide	**-.6005832**	**2.306456**	**-0.26**	**0.795**	**-5.122104**	**3.920937**
tchexper	**1.46903**	**.167235**	**8.78**	**0.000**	**1.141187**	**1.796874**
_cons	**904.7212**	**2.227989**	**406.07**	**0.000**	**900.3535**	**909.0889**

```
64.
65. /* Discussion:
  >
  > Now that we have non-dummy variables in the regression, the interpretation of
  > the betas for the dummies (small and aide) changes slightly. Now, we should
  > think of these as telling us the change in the average test score CONDITIONAL on
  > teacher experience. The beta estimates changed a little bit, falling 0.9 points
  > for aide and rising 0.2 points for small. The standard errors and t-statistics
  > also changed, but only slightly. This all suggests that teacher experience is
  > only weakly correlated with the class types.
  >
  > What do we mean by the average difference CONDITIONAL on tecaher experience?
  > Imagine if we lined up all the classes and only took the difference between
  > small, regular, and aide classes where the teacher had the same experience.
  > Then, after taking these experience-by-experience differences, we average over
  > the gaps. This is in essense what OLS is doing: we pick a slope for tchexper
  > that works well across all three types, and given this slope we adjust the
  > average test score in each group to best fit the data. Mechanically, suppose we
  > fix the level of tchexper at some number X and want to find the expected
  > difference in student scores between a small class and a regular class. Then
  > we have:
  >
  > score_hat(small = 1, tchexper = X) - score_hat(reg = 1, tchexper = X) =
  > beta1 + beta2 + beta4*X - (beta1 + beta4*X) = beta2
  >
  > While the beta estimates for small and aide changed only slightly, the beta
  > for tchexper is significant (t = 8.78) and the R2 of the regression has increased
  > from 0.007 to 0.02, a modest improvement. The improved fit of the regression
  > also helps to lower the standard errors of our estimates.
  >
  > While the betas on the dummy variables have changed only slightly, the estimate
  > for the constant term is very different. The constant is no longer measuring
  > the average for the "regular" group. Rather, the constant ensures that the
  > regression line will go through the average of totalscore and tchexper among
  > regular-group classes. That is, if score_reg and tchexper_reg are the values we
  > get for the mean when we put in the command "sum totalscore if reg == 1" and
  > "sum tchexper if reg == 1", then the constant in the regression ensure that
  >
  > score_hat(reg=1, tchexper = tchexper_reg) =
  > score_reg = beta1 + (tchexper_reg)*beta4
  >
  > Similarly, the regression line when small = 1 or when aide = 1 will go through
  > (score_small, tchexper_small) and (score_aide, tchexper_aide).
  >
  > Also - a minor note: the sample size changed because there are 20 observations
  > that lack data on tchexper. It turns out that all of these observations are also
  > for classrooms with teacher aides. However, if we re-run the regression in
  > part (b) using only the data points that have tchexper data, there is little
  > change in our estimates.
  >
  > */
```

```
66.
67. ******************************************************************************
68. *Part D: To the regression in (c) add the additional explanatory variables BOY,
69. * FREELUNCH, and WHITE_ASIAN. Are any of these variables statistically
70. * significant? Does their addition to the model affect the estimates of beta2 and
71. * beta3?
72. ******************************************************************************
73.
74. reg totalscore small aide tchexper boy freelunch white_asian
```

Source	SS	df	MS	Number of obs	=	5,766
				F(6, 5759)	=	109.32
Model	3212337.11	6	535389.519	Prob > F	=	0.0000
Residual	28203499	5,759	4897.29102	R-squared	=	0.1023
				Adj R-squared	=	0.1013
Total	31415836.1	5,765	5449.40782	Root MSE	=	69.981

totalscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
small	13.89596	2.293583	6.06	0.000	9.399671	18.39224
aide	.6982668	2.209336	0.32	0.752	-3.632862	5.029396
tchexper	1.113963	.1612504	6.91	0.000	.7978521	1.430075
boy	-14.04522	1.845673	-7.61	0.000	-17.66344	-10.42701
freelunch	-34.11702	2.063872	-16.53	0.000	-38.16299	-30.07106
white_asian	11.83727	2.210759	5.35	0.000	7.503356	16.17119
_cons	923.2498	3.12096	295.82	0.000	917.1316	929.3681

```
75.
76. /* Discussion:
  >
  > As before, including the new variables has only a slight change on the estimates
  > of the coefficients for small and aide. AIDE jumps to a positive level again
  > but the new estimate is well within the confidence bounds of the estimates in
  > parts (b) and (c). The standard errors for small and aide continue to decline,
  > in part reflecting an increase in R2 from 0.02 to 0.10.
  >
  > The estimates for boy, freelunch, and white_asian are all significant, with the
  > t-statistics for a null of zero all having large absolute values.
  >
  > */
77.
78. ******************************************************************************
79. *Part E: To the regression in (d) add the additional explanatory variables
80. * TCHWHITE, TCHMASTERS, SCHURBAN, and SCHRURAL. Are any of these variables
81. * statistically significant? Does their addition to the model affect the
82. * estimates of beta2 and beta3?
83. ******************************************************************************
84.
85. reg totalscore small aide tchexper boy freelunch white_asian tchwhite tchmaster schu
  > rban schrural
```

Source	SS	df	MS	Number of obs	=	5,766
				F(10, 5755)	=	68.14
Model	3325998.13	10	332599.813	Prob > F	=	0.0000
Residual	28089837.9	5,755	4880.94491	R-squared	=	0.1059
				Adj R-squared	=	0.1043
Total	31415836.1	5,765	5449.40782	Root MSE	=	69.864

totalscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
small	13.98027	2.302332	6.07	0.000	9.466832	18.49371
aide	1.002339	2.216588	0.45	0.651	-3.343008	5.347686
tchexper	1.156161	.1656677	6.98	0.000	.8313898	1.480932
boy	-14.00826	1.842613	-7.60	0.000	-17.62047	-10.39604
freelunch	-32.53246	2.126002	-15.30	0.000	-36.70023	-28.3647
white_asian	16.23264	2.780438	5.84	0.000	10.78193	21.68335
tchwhite	-7.668267	2.842013	-2.70	0.007	-13.23968	-2.096852
tchmasters	-3.559812	2.019344	-1.76	0.078	-7.518486	.3988614
schurban	-5.74988	2.85797	-2.01	0.044	-11.35258	-.1471824

schrural	-7.006102	2.558522	-2.74	0.006	-12.02177	-1.990436
_cons	931.7553	3.940061	236.48	0.000	924.0312	939.4793

```
86.
87. /* Discussion:
  >
  > The new variables are modestly significant, with t-statistics in the range of
  > (absolute value) 1.74 to 2.74. This lowest p-value is for tchmaster, with a
  > p-value of 0.078, meaning we fail to the beta on tchmaster = 0 at 5%, but do
  > reject at 10%. The schurban coefficient is next, with a p-value of 0.044 (reject
  > at 5%, fail to reject at 1%) and the other two variables have p-values below
  > 0.01.
  >
  > The beta2 and beta3 estimates again move slightly. This time, however, the
  > standard errors increased slightly. This partly reflects the reduced degrees of
  > freedom together with the marginal increase in R2 (from 0.102 to 0.106).
  >
  >
  > */
88.
89. ********************************************************************************
90. *Part F: Discuss the importance of parts (c), (d), and (e) to our estimation of
91. * the "treatment" effects in part (b).
92. ********************************************************************************
93.
94. /* Discussion:
  >
  > The evidence presented above reflects that there is limited omitted variable
  > bias in our estimates. We can see this in the fact that including additional
  > RHS terms does little to move our point estimates for small and aide. The lack
  > of omitted variable bias despite the explanatory power of these RHS terms means
  > that there is little correlation between these RHS terms and the treatment - i.e.
  > assignment of a student/teacher to a classroom type. This is expected given the
  > experimental nature of the research design.
  >
  > Overall, in all our regression, we find that small classes have higher test
  > scores on the order of 13-14 points and regular-size classes with aides have no
  > significant improvement relative to a regular-size class without an aide.
  >
  > */
95.
96. ********************************************************************************
97. *Part G: Add to the models in (b) through (e) indicator variables for each
98. * school: SCHOOL_j = 1 if student is in school j, and = 0 otherwise. Test the
99. * joint significance of these school "fixed effects". Does the inclusion of
100 * these fixed effect indicator variables substantially alter the estimates of
101 * beta2 and beta3?
102 ********************************************************************************
103
104 /* Discussion:
  >
  > As we saw before, including additional RHS terms did little to change our beta
  > estimates. It turns out that including all the school fixed effects modestly
  > raises the point estimates for small and aide, but these adjustments are still
  > well within the confidence intervals for the regressions without fixed effects.
  > This once again aligns with our expectation that omitted variable bias will not
  > be a major concern given the experimental design.
  >
  > How does the fixed effect change our interpretation of the betas on small and
  > aide? As we mentioned previously, when we have additional RHS variables, we can
  > think of our betas as telling us the effect of class type CONDITIONAL on the
  > other RHS terms. By adding fixed effects, we've included ANYTHING THAT SHIFTS THE
  > SCHOOL-WIDE AVERAGE as the thing we're conditioning on. In other words, when we
  > have a school fixed-effect, all the other betas can only help to explain within-
  > school variation, since by definition the fixed effect estimates will make sure
  > we hit the school-wide average in each school.
  >
  > A couple semi-technical notes about fixed effects. First, recall the
  > dummy-variable trap, which tells us that we cannot include a dummy for every
  > schid value because every observations has a schid. STATA automatically drops
  > the smallest value for schid (112038) and includes a dummy for each of the
```

```
  > remaining 78 schools. Given this, the interpretation of the dummy on a given
  > fixed effect would be whether students at a school habe different average test
  > scores (conditional on other RHS variables) than the "base school" (i.e. the
  > school that didn't get a dummy variable, or 112038).
  >
  > Second, the full set of school id dummies will be collinear with any other
  > school-wide characteristic, such as the urban or rural status of the school. The
  > indicator for urban, for example, is just the simple sum of the indicators for
  > all urban schools. This shows the sense in which including all the fixed effects
  > picks up any school-wide characteristic that might affect scores. For any school
  > characteristic (that is constant over time) that we can imagine, the fixed
  > effect will pick this up (with the net effect of all these characteristics
  > determining the beta). In this experiment, we can have school-wide fixed effects
  > because the treatment is done on a classroom-by-classroom basis so we still have
  > within-school variation.
  >
  > */
105
106 quietly log off main

107 // OLS estimates for aide in regressions b-e, with and without school FEs
108 matrix list store_aide, format(%9.4f)

  store_aide[6,8]
                b         c         d         e       bFE       cFE       dFE       eFE
     beta   0.3139   -0.6006    0.6983    1.0023    1.6993    1.2152    1.7107    1.7819
   stdErr   2.3102    2.3065    2.2093    2.2166    2.0850    2.0929    2.0142    2.0249
        t   0.1359   -0.2604    0.3161    0.4522    0.8150    0.5806    0.8493    0.8800
        p   0.8919    0.7946    0.7520    0.6511    0.4151    0.5615    0.3957    0.3789
     y_r2   0.0073    0.0203    0.1023    0.1059    0.2307    0.2338    0.2908    0.2911
  FE_fstat        .         .         .         .   21.2351   20.3034   19.3694   19.1504

109
110 // OLS estimates for small in regressions b-e, with and without school FEs
111 matrix list store_small, format(%9.4f)

  store_small[6,8]
                b         c         d         e       bFE       cFE       dFE       eFE
     beta  13.8990   14.0061   13.8960   13.9803   16.0223   15.9331   15.8031   15.7455
   stdErr   2.4085    2.3953    2.2936    2.3023    2.1692    2.1673    2.0856    2.0962
        t   5.7707    5.8473    6.0586    6.0722    7.3863    7.3516    7.5771    7.5114
        p   0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
     y_r2   0.0073    0.0203    0.1023    0.1059    0.2307    0.2338    0.2908    0.2911
  FE_fstat        .         .         .         .   21.2351   20.3034   19.3694   19.1504

112
113
114 //Convert log file (smcl) to pdf
```