Regression II
Vanderbilt University
Prof. Sean P. Corcoran **62 total points**

Problem Set 1 *Solutions*

---

1. For the following questions use the Stata dataset called *LUSD4_5.dta*. This dataset consists of 47,161 observations of 4th and 5th graders from a large urban school district ("LUSD") in 2005 and 2006. For now, keep only 5th grade observations from 2005. Assume these observations are random draws from the population. (**41 points**)

   (a) Estimate a simple regression relating student z-scores in math (*mathz*) to their teachers' years of experience (*totexp*). Interpret the slope and intercept in words. Is the coefficient for teacher experience statistically significant? Is the estimated coefficient *practically* significant? (Hint: consider a one standard deviation change in the explanatory variable). Explain your answers. (**7 points**)

The results are shown below. Keep in mind that *mathz* has mean zero and standard deviation 1. The intercept of -0.033 means we predict a math score 0.033 sd below the average for a student with a new teacher (*totexp* $= 0$). The slope of 0.0088 means we predict an increase in a student's math score of 0.0088 sd for every 1 year increase in their teacher's experience. The estimated slope coefficient is statistically significant (using the p-value or t-statistic). It is also practically significant. For example, 1 sd in the distribution of teacher experience is 9.8 years. A 1 sd increase in teacher experience is associated with a $9.8 \times 0.0088 = 0.087$ sd increase in math scores. In education research, a 0.10 sd effect is a large one, so this is a practically meaningful effect.

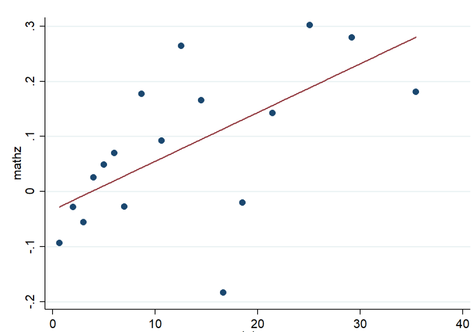
```
. use LUSD4_5.dta

. keep if grade==5 & year==2005
(35,242 observations deleted)

. reg mathz totexp

      Source |       SS           df       MS      Number of obs   =    11,759
-------------+----------------------------------   F(1, 11757)     =     89.91
       Model |  89.1137402         1  89.1137402   Prob > F        =    0.0000
    Residual |  11653.2051    11,757  .991171654   R-squared       =    0.0076
-------------+----------------------------------   Adj R-squared   =    0.0075
       Total |  11742.3189    11,758  .998666345   Root MSE        =    .99558

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totexp |   .0088442   .0009327     9.48   0.000     .0070159    .0106726
       _cons |  -.0334428   .0137211    -2.44   0.015    -.0603384   -.0065473
------------------------------------------------------------------------------
```

```
. scalar b=_b[totexp]

.
. summ totexp

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      totexp |     11,919    10.93338    9.846894          0         45

. display b*r(sd)
.08708838
```

(b) Applying the terminology used in class, is part (a) estimating a *population regression function*? Is it estimating a *conditional expectation function* (CEF)? Is it estimating a *causal* "ceteris paribus" relationship in the population? Defend your answers. (**5 points**)

> Yes, we are estimating a population regression function: the linear prediction of Y given X. This may or may not be a CEF. By definition the CEF tells us how the mean of Y varies with X in the population. This relationship may not be linear, so the PRF estimated in part (a) may not represent the CEF. Even if this were a CEF, it is unlikely to represent a causal relationship. That is, the slope coefficient on *totexp* is unlikely to tell us how the mean of Y changes with a (ceteris paribus) change in X. See part (d) below for more on this.

(c) Install the user-written .ado file called `binscatter`. Use this command to produce a binned scatter plot showing the relationship between math z-scores on the vertical axis and teacher experience on the horizontal axis. Bearing in mind this is sample data, do your findings suggest that the population CEF is linear? Provide an intuitive explanation for why the CEF might not be linear. (**5 points**)



The binned scatter plot is shown above, and suggests a nonlinear relationship between teacher experience and math scores. The slope appears to be initially steep at low levels of experience but then diminishes with higher levels of experience.

(d) Your co-author is concerned that the regression in part (a) does not have a causal interpretation. Specifically, she thinks that experienced teachers are less likely to work with low-income students, who perform worse on tests in general. What does this say about the likely direction of omitted variables bias? Explain. (**3 points**)

The omitted variables bias formula is $\beta_s = \beta_l + \pi_1\gamma$ where π_1 is the slope coefficient from a regression of the omitted variable on the included, and γ is slope coefficient on the omitted variable in the "long" regression. Suppose student poverty is the omitted variable. If experienced teachers are less likely to work with poor students then $\pi_1 < 0$. It is also likely that, other things being equal, poor students have lower math achievement ($\gamma < 0$). The OVB term is the product of two negative numbers and thus positive. By omitting student poverty status we are likely overstating the effect of teacher experience.

(e) Using these variables (*mathz, totexp,* and *econdis*, an indicator variable for economically disadvantaged students), demonstrate the omitted variables bias formula shown in class ($\beta_s = \beta_\ell + \pi_1\gamma$), where the parameters are as defined in the lecture notes. Do these results conform with your answer in part (d)? Provide an interpretation in words of the auxiliary regression coefficient π_1. (**7 points**)

The results are below. The calculated β_s using the OVB formula is slightly different from the OLS estimate since the sample sizes differ a bit between the short and long regressions. To be precise, we should have limited the analysis to the observations for which there were no missing values on any variables.

```
. // "long" regression
. reg mathz totexp econdis

      Source |       SS           df       MS      Number of obs   =    11,759
-------------+----------------------------------   F(2, 11756)     =    543.24
       Model |  993.398767         2  496.699383   Prob > F        =    0.0000
    Residual |  10748.9201    11,756  .914334817   R-squared       =    0.0846
-------------+----------------------------------   Adj R-squared   =    0.0844
       Total |  11742.3189    11,758  .998666345   Root MSE        =    .95621

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
      totexp |   .0050135   .0009041     5.55   0.000     .0032413    .0067857
     econdis |  -.7380784   .0234694   -31.45   0.000    -.7840823   -.6920744
       _cons |   .6180293   .0245521    25.17   0.000     .5699032    .6661555
------------------------------------------------------------------------------

. scalar gamma=_b[econdis]
. scalar b = _b[totexp]

. // "auxiliary regression"
. reg econdis totexp

      Source |       SS           df       MS      Number of obs   =    11,919
-------------+----------------------------------   F(1, 11917)     =    218.18
       Model |  30.6823129         1  30.6823129   Prob > F        =    0.0000
    Residual |   1675.8971    11,917  .140630788   R-squared       =    0.0180
-------------+----------------------------------   Adj R-squared   =    0.0179
       Total |  1706.57941    11,918  .143193439   Root MSE        =    .37501

------------------------------------------------------------------------------
     econdis |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totexp |  -.0051528   .0003489   -14.77   0.000    -.0058366    -.004469
       _cons |   .8831686   .0051329   172.06   0.000     .8731074    .8932299
------------------------------------------------------------------------------

. scalar pi1=_b[totexp]

. // "short" regression
. reg mathz totexp

      Source |       SS           df       MS      Number of obs   =    11,759
-------------+----------------------------------   F(1, 11757)     =     89.91
       Model |  89.1137402         1  89.1137402   Prob > F        =    0.0000
    Residual |  11653.2051    11,757  .991171654   R-squared       =    0.0076
-------------+----------------------------------   Adj R-squared   =    0.0075
       Total |  11742.3189    11,758  .998666345   Root MSE        =    .99558

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totexp |   .0088442   .0009327     9.48   0.000     .0070159    .0106726
       _cons |  -.0334428   .0137211    -2.44   0.015    -.0603384   -.0065473
------------------------------------------------------------------------------

. display b + (pi1*gamma)
.00881669
```

(f) Now use the same data to demonstrate the "regression anatomy" formula below. In this expression, β_1 is the coefficient on teacher experience from the "long" regression on teacher experience and *econdis*. $\tilde{X}_{1i}$ is the estimated residual after regressing teacher experience on *econdis*. $C()$ is covariance and $V()$ is variance. (Hint: you can easily get the covariance using `corr`).

$$\beta_1 = \frac{C(Y_i, \tilde{X}_{1i})}{V(\tilde{X}_{1i})}$$

This formula has a simple interpretation: the multivariate regression coefficient on X_1 (here, teacher experience) can be written as the *simple* regression coefficient from a regression of Y on $\tilde{X}_{1i}$, teacher experience that has been "purged" of all correlation with the other explanatory variables in the model. (**7 points**)

The results are below. Again there are slight differences between the "regression anatomy" calculation and the OLS slope because of differences in sample size. It is preferable to repeat the below for the set of observations with no missing values.

```
. reg totexp econdis

      Source |       SS           df       MS      Number of obs   =    11,919
-------------+----------------------------------   F(1, 11917)     =    218.18
       Model |  20776.0761         1  20776.0761   Prob > F        =    0.0000
    Residual |  1134809.03    11,917  95.2260662   R-squared       =    0.0180
-------------+----------------------------------   Adj R-squared   =    0.0179
       Total |  1155585.11    11,918   96.961328   Root MSE        =    9.7584

------------------------------------------------------------------------------
      totexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     econdis |  -3.489141   .2362189   -14.77   0.000    -3.952169   -3.026113
       _cons |   13.81831   .2147945    64.33   0.000     13.39728    14.23935
------------------------------------------------------------------------------

. predict uhat, resid

. corr mathz uhat, cov
(obs=11,759)

             |    mathz     uhat
-------------+------------------
       mathz |  .998666
        uhat |  .477851  95.1335

. scalar cov=r(cov_12)

. summ uhat

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        uhat |     11,919    4.32e-08    9.757975  -13.81831   34.67083

. scalar vuhat=r(Var)

. display cov/vuhat
.00501849

. reg mathz totexp econdis

      Source |       SS           df       MS      Number of obs   =    11,759
-------------+----------------------------------   F(2, 11756)     =    543.24
       Model |  993.398767         2  496.699383   Prob > F        =    0.0000
    Residual |  10748.9201    11,756  .914334817   R-squared       =    0.0846
-------------+----------------------------------   Adj R-squared   =    0.0844
       Total |  11742.3189    11,758  .998666345   Root MSE        =    .95621

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totexp |   .0050135   .0009041     5.55   0.000     .0032413    .0067857
```

```
     econdis |  -.7380784   .0234694   -31.45   0.000    -.7840823   -.6920744
       _cons |   .6180293   .0245521    25.17   0.000     .5699032    .6661555
------------------------------------------------------------------------------
```

(g) Finally, your co-author remains unsatisfied with this regression specification and recommends you also control for *mathz_1*, the student's math score in the prior grade. Estimate the multivariate regression with *totexp, econdis,* and *mathz_1*. Provide an interpretation, in words, of the three regression coefficients. How did the two regression coefficients on *totexp* and *econdis* change from the case in which these were the only two explanatory variables? What happened to their standard errors? Provide some intuition behind both changes. (**7 points**)

The results are below. Not surprisingly, the estimated coefficient on *mathz_1* is large—math achievement in the prior year is a strong predictor of math achievement in the current year. The estimated coefficients on *totexp* and *econdis* are now smaller. This might have been predicted if we think students with less-experienced teachers and poor students came into the classroom with lower levels of math achievement. The standard errors on these coefficients are smaller. This is also to be expected since inclusion of *mathz_1* reduced unexplained variation in y.

```
. reg mathz mathz_1 totexp econdis

      Source |       SS           df       MS      Number of obs   =    11,755
-------------+----------------------------------   F(3, 11751)     =   3429.85
       Model |  5480.13775         3  1826.71258   Prob > F        =    0.0000
    Residual |  6258.50348    11,751  .532593267   R-squared       =    0.4668
-------------+----------------------------------   Adj R-squared   =    0.4667
       Total |  11738.6412    11,754  .998693316   Root MSE        =    .72979

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathz_1 |   .6566555   .0071541    91.79   0.000     .6426322    .6706788
      totexp |   .0027675   .0006907     4.01   0.000     .0014136    .0041214
     econdis |  -.3361978   .0184398   -18.23   0.000    -.3723428   -.3000528
       _cons |   .2345797   .0191992    12.22   0.000     .1969461    .2722133
------------------------------------------------------------------------------
```

2. A researcher estimates a bivariate regression of the form $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ but confides to a colleague that she believes $C(\epsilon_i, x_i) \neq 0$ and therefore $\hat{\beta}_1$ is biased. The colleague then asks whether one can test whether $C(\epsilon_i, x_i) \neq 0$. The colleague suggests that the researcher construct $\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ and then run a regression of $\hat{\epsilon}_i$ on x_i, that is a regression of the form $\hat{\epsilon}_i = \gamma_0 + \gamma_1 x_i + \nu_i$ and then test the null $H_0 : \gamma_1 = 0$ to see if ϵ_i and x_i are correlated. Is this a good idea, or not? Explain. (**5 points**)

This is not a good idea. The OLS model chooses an intercept and slope such that x_i is, by construction, uncorrelated with $\hat{\epsilon}_i$. Therefore, in the estimate for γ_1 the numerator will by construction be equal to zero. Therefore, this approach will tell us nothing about whether x_i and ϵ_i are correlated in the population. It helps to reflect a bit on what the researcher was suggesting when she revealed her concern about $C(\epsilon_i, x_i) \neq 0$. She is interested in interpreting $\hat{\beta}_1$ as a "ceteris paribus" relationship between x_i and y_i, which leads one to inquire about $C(\epsilon_i, x_i)$. When thinking about "bias," it is helpful to ask "biased for what?"

3. Demonstrate that you understand how bootstrapping works by doing the steps below. Use the *HSLS-09 extract* dataset available on GitHub This is a sample of 500 students from the High School Longitudinal Study of 2009. (**16 points**)

   (a) Estimate the following simple regression which relates the student's standardized math score to a measure of their family's socioeconomic status: $x1txmtscor = \beta_0 + \beta_1 x1ses + u$. Interpret the estimated coefficient on $x1ses$ (call this $\hat{\beta}_{1,OLS}$). Is this a large effect size? Explain your rationale for assessing the effect size. (**3 points**)

Regression results are shown below. The OLS slope coefficient of 4.98 means that a 1 unit change in SES is associated with a 4.98 point increase in tested math achievement (on average). As the descriptive statistics show, a 1-unit change in SES is quite large, while a 1-point change on the math test is not especially large. To put the slope coefficient in context, we can calculate how much of a change in math achievement is predicted—in standard deviation units—from a one standard deviation change in SES (0.771). This calculation is shown below. A 1 SD change in SES is associated with a 0.396 SD change in math achievement, which is a large effect.

```
. reg x1txmtscor x1ses

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(1, 498)       =     92.59
       Model |  7374.92905         1  7374.92905   Prob > F        =    0.0000
    Residual |  39667.2196       498  79.6530513   R-squared       =    0.1568
-------------+----------------------------------   Adj R-squared   =    0.1551
       Total |  47042.1486       499  94.2728429   Root MSE        =    8.9249

------------------------------------------------------------------------------
  x1txmtscor |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       x1ses |    4.98339   .5179015     9.62   0.000     3.965849    6.000931
       _cons |   51.12905   .4019788   127.19   0.000     50.33927    51.91884
```

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

. summ x1txmtscor x1ses

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  x1txmtscor |        500     51.5886     9.70942    26.6308    82.1876
       x1ses |        500    .0922156    .7714429    -1.7526     2.5668

. di (0.7714429*4.98339)/9.70942
.39594547
```

(b) Now, bootstrap the sampling distribution of $\hat{\beta}_1$, using 250 replications. Each replication should be a bootstrap sample of size N=500. Do this manually by writing a loop, <u>not</u> with the `bootstrap` command. (Hint: the command `bsample` will be helpful here). Save your estimates from each replication, report your bootstrapped standard error for $\hat{\beta}_1$, and a 90% percentile interval for $\hat{\beta}_1$. Give a written interpretation of these two things, and provide a histogram of your coefficient estimates. (**7 points**)

The syntax below performs the following loop 250 times: (1) sample 500 observations, with replacement, from the original dataset; (2) estimate the OLS regression from part (a); and (3) save the iteration number, slope coefficient estimate (`_b[x1ses]`), and standard error estimate (`_se[x1ses]`). The postfile and postclose commands collect these results into a file called *reg_table*.

The statistics for *beta* describe the sampling distribution over 250 bootstrap samples. The bootstrapped standard error is 0.539 (the standard deviation of *beta*). 90% of the values of *beta* lie between 4.1039 and 5.9504 (the 5th and 95 percentile of *beta*).

```
set seed 1234
tempname reg_results
tempfile reg_table

postfile `reg_results' iter beta se using `reg_table', replace

local i 1
quietly while `i'<251 {
   preserve
   bsample 500
   reg x1txmtscor x1ses
   post `reg_results' (`i') (_b[x1ses]) (_se[x1ses])
   restore
   local i=`i'+1
   }
postclose `reg_results'

use `reg_table', clear

. summ beta, detail

                            beta
```

```
------------------------------------------------------------
      Percentiles      Smallest
 1%     3.658587       3.560383
 5%       4.1039       3.601388
10%      4.41069       3.658587       Obs                 250
25%     4.686244       3.763867       Sum of Wgt.         250

50%      5.01397                      Mean           5.030252
                        Largest       Std. Dev.      .5392142
75%     5.352006       6.274694
90%     5.712456       6.311025       Variance       .2907519
95%     5.950438       6.427313       Skewness       .1038691
99%     6.311025       6.852541       Kurtosis       3.359372
```

(c) Next, use the `bootstrap` command with `regress` to obtain the bootstrapped standard error. This will be much easier than doing it manually! (**3 points**)

The bootstrap prefix below tells Stata to bootstrap *beta* and the standard error estimate *se* using 250 replications. The bootstrapped standard error of *beta* is 0.542, not far from what was found in part (b). It will differ somewhat since these represent 250 new random samples with replacement.

```
. bootstrap _b _se, reps(250) saving(results, replace): reg x1txmtscor x1ses
(running regress on estimation sample)
(note: file results.dta not found)

Bootstrap replications (250)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250

Linear regression                               Number of obs     =        500
                                                Replications      =        250

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
b            |
       x1ses |    4.98339   .5415519     9.20   0.000     3.921968    6.044812
       _cons |   51.12905   .4054016   126.12   0.000     50.33448    51.92362
-------------+----------------------------------------------------------------
se           |
       x1ses |   .5179015   .0213991    24.20   0.000     .4759601    .5598429
       _cons |   .4019788   .0108878    36.92   0.000     .3806392    .4233185
------------------------------------------------------------------------------
```

(d) State whether the following statement is true or false. If false, explain why. (**3 points**)

Bootstrapping is a useful procedure, but relies on an assumption of normality for the underlying sampling distribution of $\hat{\beta}$.

This is false! In fact one of the chief advantages of bootstrapping is that does not rely on any underlying assumption about the sampling distribution of $\hat{\beta}$. Bootstrapping lets the data reveal to you what the sampling distribution looks like.