Regression II
Vanderbilt University
Prof. Sean P. Corcoran **67 total points**

Problem Set 1 *Solutions*

---

1. For the following questions use the Stata dataset on Github 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. (**48 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 use https://github.com/spcorcor18/LPO-8852/raw/main/data/LUSD4_5.dta, clear

. 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, *using potential outcomes terminology.* (**5 points**)

We are using sample data to estimate a population regression function: the "best prediction" of Y given X in the population. This "best prediction" function may or may not be the CEF. By definition, the CEF tells us how the mean of Y varies with X in the population. This relationship may not be linear (or any other standard function), so the PRF estimated in part (a) need not be the CEF.

Even if it were a CEF, it is unlikely to represent a *causal* relationship. A causal CEF describes how mean potential outcomes vary with X for a given reference population. Suppose the following describes how mean potential outcomes vary with experience for a population of 5th graders:

$$E(Y_i|exp_i, A_i) = \gamma_0 + \gamma_1 exp_i + \gamma_2 A_i$$

where exp_i is the number of years of teacher experience and A_i represents some baseline characteristic of students like prior achievement, "ability," family resources, etc. Suppose we ignore A_i and use regression to estimate the difference in achievement for students with an experienced teacher ($exp_i = 10$) and students with a new teacher ($exp_i = 0$). The estimated regression function is:

$$Y_i = \beta_0 + \beta_1 exp_i + u_i$$

We could use this regression to estimate the difference in mean Y_i when $exp_i = 10$ and when $exp_i = 0$. This would estimate

$E(Y_i|exp_i = 10) - E(Y_i|exp_i = 0)$, but we know from the CEF that this is:

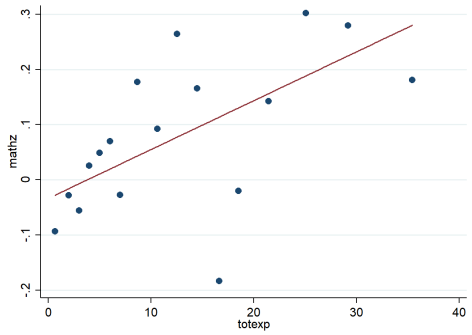
$$\gamma_0 + \gamma_1 10 + \gamma_2 E(A_i|exp = 10) - \gamma_0 - gamma_1 0 - \gamma_2 E(A_i|exp = 0)$$

or

$$10\gamma_1 + \gamma_2 \left[E(A_i|exp_i = 10) - E(A_i|exp_i = 0)\right]$$

or, the true causal effect of an additional 10 years of experience ($10\gamma_1$) plus selection bias. It's possible (and likely) that the mean A_i differs for students with more and less experience.

(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) 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, and *lep* (Limited English Proficient). Estimate the multivariate regression with *totexp, econdis, mathz_1*, and *lep*. 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. The outcome here is a z-score (math achievement in standard deviation units) so the slope coefficients represent the standard deviation change in math achievement from a one-unit change in the predictor variable. For example, a one-year increase in teacher experience is associated with a 0.0018 standard deviation increase in math achievement, holding other variables in the model constant. Economically disadvantaged students score 0.295 sd lower, on average, than non-economically disadvantaged students. LEP students score 0.163 sd lower than non-LEP students.

Not surprisingly, the estimated coefficient on *mathz_1* is large (0.651)—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 economically disadvantaged 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 a lot of unexplained variation in y.

```
. reg mathz totexp econdis mathz_1 lep

      Source |       SS           df       MS      Number of obs   =    11,755
-------------+----------------------------------   F(4, 11750)     =   2620.54
       Model |   5534.6132         4   1383.6533   Prob > F        =    0.0000
    Residual |  6204.02803    11,750  .528002386   R-squared       =    0.4715
-------------+----------------------------------   Adj R-squared   =    0.4713
       Total |  11738.6412    11,754  .998693316   Root MSE        =    .72664

------------------------------------------------------------------------------
       mathz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     totexp |   .0018447   .0006937     2.66   0.008     .0004849    .0032044
     econdis |  -.2948771   .0188054   -15.68   0.000    -.3317388   -.2580154
     mathz_1 |   .6509335   .0071455    91.10   0.000     .6369272    .6649398
         lep |  -.1631717   .0160643   -10.16   0.000    -.1946605    -.131683
       _cons |   .2530979    .019203    13.18   0.000     .2154568     .290739
------------------------------------------------------------------------------
```

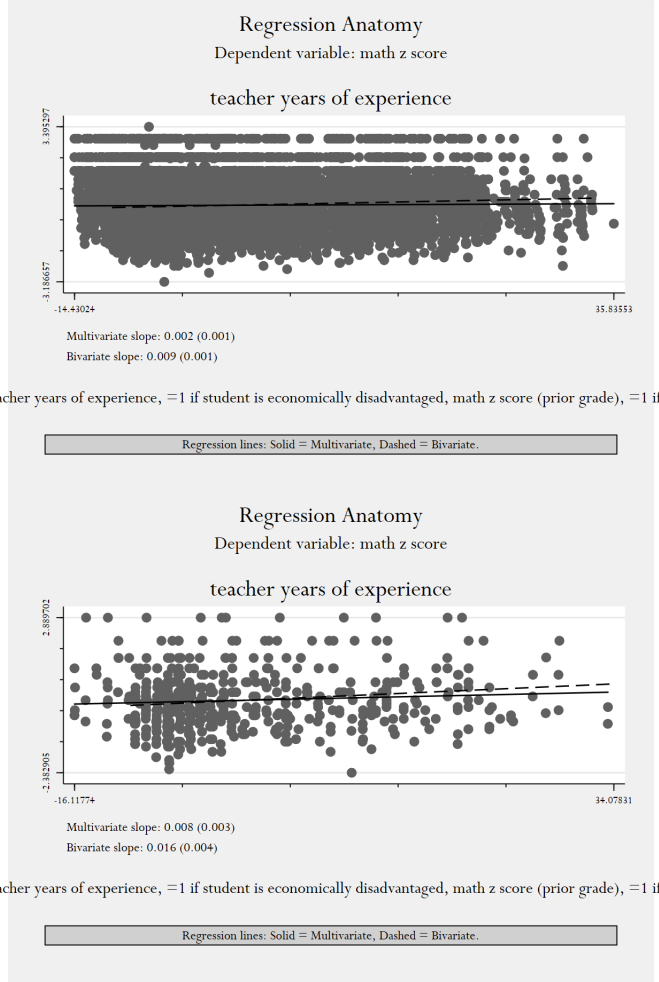
(h) Finally, the user-written command `reganat` (short for "regression anatomy") allows you to visually compare bivariate relationships with those that account for the correlation between the X of interest and other covariates. Install this command using the syntax below, and then use it to show the relationship between *mathz* and *totexp* after "purging" *totexp* of its correlation with the other covariates *econdis, mathz_1,* and *lep*. Provide a written interpretation of what this graph shows you.

The results are below. Note the plot shows both the bivariate slope and the multivariate slope. The latter shows the relationship between the outcome (math achievement) and the residualized teacher experience (what remains after purging teacher experience of its correlation with other predictors).

```
ssc install reganat
reganat mathz totexp econdis mathz_1 lep, dis(totexp) biline
graph save reganat1, replace

sample 500, count
reganat mathz totexp econdis mathz_1 lep, dis(totexp) biline
graph save reganat2, replace

graph combine reganat1.gph reganat2.gph, ysize(6) xsize(4) col(1)
graph export reganat_combine.png, as(png) replace
```



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 this regression model suffers from omitted variables bias. 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 $\hat{\epsilon}_i$. Your estimate for γ_1 will thus be zero. Clearly, 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 omitted variables bias. She is probably interested in the causal relationship between x_i and y_i, which leads one to ask whether this model will be informative about differences in mean potential outcomes for a fixed population.

3. Use the syntax below to create a "toy" dataset of potential outcomes for 10 individuals. (Note: this example dataset was taken from the *Mixtape* Potential Outcomes chapter). **(14 points)**

   (a) What is the ATE? The ATT? The ATU? Show in your syntax how you calculated these values, and show that the ATE is a weighted average of the ATT and ATU. **(5 points)**

   See code below. $ATE = 0.6$, $ATT = 4.4$ and $ATU = -3.2$. The ATE is an equally-weighted average of the ATT and ATU in this case, since there are 5 persons in each group. Note in this example there was intended to be an extreme form of selection bias. Those in the treatment group really benefit from the treatment, while those not treated would actually be worse off had they gotten the treatment (-3.2).

   (b) What is the simple difference in mean observed outcomes between the treated and untreated cases? If this estimator were used for the ATT, what would the selection bias be? If this estimator were used for the ATE, what would the selection bias be? **(5 points)**

See code below. The simple difference in means is -0.4: it actually appears as if the treatment "causes" lower outcomes. However, we know from part (a) that the ATT is actually 4.4, implying a selection bias of -4.8. The ATE is actually 0.6, implying a section bias of -1.

(c) Suppose D were randomly assigned. Would this guarantee that the simple differences in means gives you the ATE? Why or why not? **(4 points)**

The key word here is "guarantee." If treatment were randomly assigned, then *as the sample size grows to infinity* there will be no different in the mean potential outcomes for the $D = 0$ and $D = 1$ group. In this case we have only 10 observations. Even if there were random assignment, there would likely be chance differences in the two groups that would lead to an answer that differs from ATE. Over *repeated* samples we would get the ATE on average, but in the real world we only have one sample! The *Mixtape* chapter has a nice simulation of random sampling from these 10 observations.

```
clear
set obs 10

gen  y1 = 7 in 1
replace y1 = 5 in 2
replace y1 = 5 in 3
replace y1 = 7 in 4
replace y1 = 4 in 5
replace y1 = 10 in 6
replace y1 = 1 in 7
replace y1 = 5 in 8
replace y1 = 3 in 9
replace y1 = 9 in 10

gen  y0 = 1 in 1
replace y0 = 6 in 2
replace y0 = 1 in 3
replace y0 = 8 in 4
replace y0 = 2 in 5
replace y0 = 1 in 6
replace y0 = 10 in 7
replace y0 = 6 in 8
replace y0 = 7 in 9
replace y0 = 8 in 10
```

```
gen d = 1 if inlist(_n, 1, 3, 5, 6, 10)
replace d = 0 if d==.

gen y = y1*d + y0*(1-d)

// part a
// the ATE is the mean TE for all units
gen te=y1-y0
summ te

// the ATT is the mean TE for treated units
// the ATU is the mean TE for untreated units
summ te if d==1
scalar att = r(mean)
summ te if d==0
scalar atu = r(mean)

// the ATE is a weighted average of the ATT and ATU
display (5/10)*att + (5/10)*atu

// part b
// the simple difference in means
summ y if d==1
scalar ybar1 = r(mean)
summ y if d==0
scalar ybar0 = r(mean)
display ybar1 - ybar0

// selection bias vis a vis ATT is
display (ybar1 - ybar0) - att
// selection bias vis a vis ATE is
display (ybar1 - ybar0) - ((5/10)*att + (5/10)*atu)
```