
Problem Set 1 *Solutions*

1. Use the Stata syntax below to create a dataset of potential outcomes for 600 students. The data include four “types” of students, indicated by the X variable. The indicator variable $D = 1$ if the student participated in an educational intervention (and $D = 0$ otherwise). **(28 points—4 each)**

```
clear all
set seed 3791
set obs 100
    gen x = 1
    gen y0 = 25
    gen y1 = 35
    gen d = runiform()<=0.20
set obs 250
    replace x = 2 if d==.
    replace y0= 50 if d==.
    replace y1= 90 if d==.
    replace d = runiform()<=0.80 if d==.
set obs 450
    replace x = 3 if d==.
    replace y0= 40 if d==.
    replace y1= 60 if d==.
    replace d = runiform()<=0.50 if d==.
set obs 600
    replace x = 4 if d==.
    replace y0= 30 if d==.
    replace y1= 45 if d==.
    replace d = runiform()<=0.40 if d==.
```

- (a) Use this dataset to calculate the ATE, ATT, and ATU (show your syntax). How do they compare? Show that the ATE is a weighted average of ATT and ATU.
- (b) How would you describe (in words) the four student “types” in this dataset, in terms of their potential outcomes, treatment effects, and propensity to be treated? Speculate on the direction of selection bias and heterogeneous treatment effect bias (if any) if you were to use the simple difference in observed means ($\bar{Y}_1 - \bar{Y}_0$) to estimate the ATE.
- (c) What is the simple difference in mean *observed* outcomes between the treated and untreated cases? Given what you know about potential outcomes for these students, calculate the selection bias and heterogeneous treatment effect bias.

- (d) As an alternative to the naïve estimator in (c), calculate the difference in mean *observed* outcomes separately for each student type. Then, take the simple average of these four differences. How does it compare to your answer in (c)? To the (known) ATE? ATT? Why is this better (or is it) than the mean in (c)?
- (e) As another alternative, calculate the *weighted* average of the four differences found in part (d), using the number of students of each type as weights. How does it compare to your answer in parts (c)-(d)? To the (known) ATE? ATT? Why is this better (or is it) than the means in (c)-(d)?
- (f) Estimate an OLS regression of Y on D and include dummy variable indicators for student type (use Stata factor variables, and exclude the first type). What is your estimated coefficient on D ? How does it compare to the (known) ATE? ATT? ATU? To your earlier treatment effect estimates?
- (g) Suppose D were randomly assigned to the students in this dataset. Will this guarantee that the simple difference in means equals the ATE? Why or why not?

See attached do-file for code and responses to Question 1.

2. Suppose you conduct a randomized controlled trial in which 50% of your study population is assigned to the treatment condition and 50% is untreated. Unfortunately, 1/3 of your treated subjects fail to comply and do not actually receive the treatment. Explain (using potential outcomes terminology) why the ATE cannot be estimated in this case. **(5 points)**

Suppose you have 120 subjects, with 60 assigned to treatment and 60 assigned to control. If the randomization holds, these two groups should be equivalent in expectation—each group can provide the counterfactual outcome for the other, and you can estimate the ATE, ATT, and ATU. However, if 1/3 of the treated subjects (20) fail to comply, the mean outcomes of the compliers no longer represent the (randomly assigned) treatment group. The non-compliers are a selected sample, as are the compliers. Without a clean estimate of the mean outcomes for the treated group, the ATE cannot be estimated.

3. 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. NOTE: I also recommend keeping only observations that have nonmissing *mathz*, *totexp* and *econdis*. **(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. **(5 points)**

The results are shown below. Note I have retained only cases with nonmissing $mathz$, $totexp$, and $econdis$, the three variables used in parts (a)-(e). (Not doing so will create a small problem in that the long and short regressions will have different numbers of nonmissing observations.

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.

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

.      keep if mathz~= . & totexp~= . & econdis~= .
(160 observations deleted)

.      reg mathz totexp
```

Source	SS	df	MS	Number of obs	=	11,759
Model	89.1137402	1	89.1137402	F(1, 11757)	=	89.91
Residual	11653.2051	11,757	.991171654	Prob > F	=	0.0000
				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 |      11,759      10.93214      9.843389      0      45
.
      display b*r(sd)
      .08705738

```

- (b) Do you think the regression in part (a) estimates differences in potential outcomes for a population of interest? Explain why or why not. (4 points)

This regression is unlikely to estimate differences in potential outcomes for any population of interest. 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 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

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

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) 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. (4 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.

- (d) Using these variables (*mathz*, *totexp*, and *econdis*, an indicator variable for economically disadvantaged students), demonstrate the omitted variables bias formula shown in class ($\beta_s = \beta_l + \pi_1\gamma$), where the parameters are as defined in the lecture notes. Do these results conform with your answer in part (c)? Provide an interpretation in words of the auxiliary regression coefficient π_1 . (5 points)

The results are below. The slope coefficient of -0.005 ($\hat{\pi}_1$) means that a one-year increase in teacher experience is associated with a 0.5 percentage point lower propensity for the student to be economically disadvantaged. (Less experienced teachers are more likely to teach economically disadvantaged students).

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

Source	SS	df	MS	Number of obs	=	11,759
Model	993.398767	2	496.699383	F(2, 11756)	=	543.24
Residual	10748.9201	11,756	.914334817	Prob > F	=	0.0000
Total	11742.3189	11,758	.998666345	R-squared	=	0.0846
				Adj R-squared	=	0.0844
				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	=	11,759
Model	30.6888446	1	30.6888446	F(1, 11757)	=	217.36
Residual	1659.97057	11,757	.141189977	Prob > F	=	0.0000
Total	1690.65941	11,758	.143788009	R-squared	=	0.0182
				Adj R-squared	=	0.0181
				Root MSE	=	.37575

```
-----+-----
```

econdis	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
totexp	-.0051901	.000352	-14.74	0.000	-.0058802 -.0045001
_cons	.8826599	.0051786	170.44	0.000	.8725089 .8928108

```
-----+-----
```

```
.      scalar pi1=_b[totexp]
```

```
.      // "short" regression
```

```
.      reg mathz totexp
```

Source	SS	df	MS	Number of obs	=	11,759
Model	89.1137402	1	89.1137402	F(1, 11757)	=	89.91
Residual	11653.2051	11,757	.991171654	Prob > F	=	0.0000
Total	11742.3189	11,758	.998666345	R-squared	=	0.0076
				Adj R-squared	=	0.0075
				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) /* this should be the same as the OLS slope */
```

```
.00884425
```

- (e) Another formula that is useful in regression is called “regression anatomy,” below. It looks similar to—but is not the same as—the OVB formula. 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. Show that this fomrula holds in your data. (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. **(5 points)**

The results are below.

```
.          // get residual from regressing totexp on econdis
.          reg totexp econdis
```

Source	SS	df	MS	Number of obs	=	11,759
Model	20679.8413	1	20679.8413	F(1, 11757)	=	217.36
Residual	1118580	11,757	95.1416181	Prob > F	=	0.0000
Total	1139259.85	11,758	96.8923155	R-squared	=	0.0182
				Adj R-squared	=	0.0181
				Root MSE	=	9.7541

totexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
econdis	-3.497402	.2372232	-14.74	0.000	-3.962399 -3.032405
_cons	13.82071	.2155889	64.11	0.000	13.39812 14.2433


```
.          predict uhat, resid
.
.          // regression anatomy formula: b1 = COV(Y,RESID)/VAR(RESID)
.
.          corr mathz uhat, cov
(obs=11,759)
```

	mathz	uhat
mathz	.998666	
uhat	.476954	95.1335


```
.          scalar cov=r(cov_12)
.          display cov
.4769539

.          summ uhat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
uhat	11,759	-2.52e-08	9.753642	-13.82071	34.67669


```
.          scalar vuhat=r(Var)
.          display uhat
3.6766887

.          display cov/vuhat
.00501352
.
.          reg mathz totexp econdis
```

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

- (f) 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. (5 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

4. 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.

```

-----
> -----
      name: <unnamed>
      log: C:\Users\corcorssp\Dropbox\_TEACHING\Regression II\Problem sets\Problem se
> t 1 - Potential outcomes\PS1.txt
      log type: text
      opened on: 5 Sep 2023, 17:02:33

.
.
. // *****
. //
. // Problem set 1
. // Last updated: September 5, 2023
. //
. // *****
.
. // *****
. // Question 1
. // *****
.
. // Set up data
. clear all

. set seed 3791

. set obs 100
number of observations (_N) was 0, now 100

. gen x = 1

. gen y0 = 25

. gen y1 = 35

. gen d = runiform()<=0.20

. set obs 250
number of observations (_N) was 100, now 250

. replace x = 2 if d==.
(150 real changes made)

. replace y0= 50 if d==.
(150 real changes made)

. replace y1= 90 if d==.
(150 real changes made)

. replace d = runiform()<=0.80 if d==.
(150 real changes made)

. set obs 450
number of observations (_N) was 250, now 450

. replace x = 3 if d==.
(200 real changes made)

. replace y0= 40 if d==.
(200 real changes made)

```

```

.      replace y1= 60 if d==.
(200 real changes made)

.      replace d = runiform()<=0.50 if d==.
(200 real changes made)

.      set obs 600
number of observations (_N) was 450, now 600

.      replace x = 4 if d==.
(150 real changes made)

.      replace y0= 30 if d==.
(150 real changes made)

.      replace y1= 45 if d==.
(150 real changes made)

.      replace d = runiform()<=0.40 if d==.
(150 real changes made)

```

```

.      table x

-----
      x |      Freq.
-----+-----
      1 |          100
      2 |          150
      3 |          200
      4 |          150
-----

```

```

.      tabstat d, by(x)

```

```

Summary for variables: d
by categories of: x

```

```

      x |      mean
-----+-----
      1 |          .27
      2 |      .7933333
      3 |          .505
      4 |      .4133333
-----+-----
Total |          .515
-----

```

```

.
.
. // *****
. // 1a - treatment effects
. // *****
.
.      // Individual treatment effects
.      gen te = y1 - y0
.
.      // ATE
.      summ te

```

```

Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
      te |      600      22.08333      10.89836      10      40

```

```

.      scalar ate=r(mean)
.
.      // ATT
.      summ te if d==1
.
.      Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
.      te |      309     25.82524     11.58881      10      40
.
.      scalar att=r(mean)
.
.      // ATU
.      summ te if d==0
.
.      Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
.      te |      291     18.10997      8.481314      10      40
.
.      scalar atu=r(mean)
.
.      // ATE is a weighted average of ATT and ATU
.      qui summ d
.
.      scalar p=r(mean)
.
.      display (p*att)+((1-p)*atu)
22.083333
.
.      // As seen above, ATT > ATE > ATU. The last line above demonstrates that
.      // ATE = p*ATT + (1-p)*ATU
.
.      // *****
.      // 1b - differences in student types
.      // *****
.
.      // Treatment effects and potential outcomes by type
.      tabstat te, by(x)

```

Summary for variables: te
by categories of: x

x	mean
1	10
2	40
3	20
4	15
Total	22.08333

```

.      tabstat y0, by(x)

```

Summary for variables: y0
by categories of: x

x	mean
1	25
2	50
3	40
4	30
Total	37.5

```
.      // Probability of treatment varies by type
.      tabstat d, by(x)
```

```
Summary for variables: d
by categories of: x
```

x	mean
1	.27
2	.7933333
3	.505
4	.4133333
Total	.515

```
.      // Treated are most likely to be group 2, followed by 3
.      tabulate x if d==1
```

x	Freq.	Percent	Cum.
1	27	8.74	8.74
2	119	38.51	47.25
3	101	32.69	79.94
4	62	20.06	100.00
Total	309	100.00	

```
.
.      // As shown above, there are heterogeneous treatment effects, with group
.      // 2 having the largest te=40, and group 1 having the smallest te=10. The
.      // groups also differ by potential outcomes, with y0 varying from 25
.      // (group 1) to 50 (group 2). The probability of treatment also varies,
.      // from 27% of group 1 treated to 80% of group 2. By design (for sake of
.      // this example), the group with the largest expected effect from treatment
.      // is also the one most likely to be treated. There is likely to be positive
.      // selection bias in the naive simple difference estimator in that the mean
.      // y0 is larger for the groups most likely to be treated (2 and 3). There
.      // is also likely to be positive heterogeneous treatment effect bias, since
.      // the treated tend to have higher tes on average than the untreated.
.
.      // Another way to see the latter points:
.      tabstat y0, by(d)
```

```
Summary for variables: y0
by categories of: d
```

d	mean
0	34.27835
1	40.53398
Total	37.5

```
.      tabstat te, by(d)
```

```
Summary for variables: te
by categories of: d
```

d	mean
0	18.10997
1	25.82524
Total	22.08333

```

.
.
. // *****
. // lc - simple diff in means
. // *****
.
. // Observed Y ("switching equation")
. gen y=(d*y1) + (1-d)*y0
.
. // Simple difference in means--will have selection bias
. ttest y, by(d)

Two-sample t test with equal variances
-----+-----
      Group |      Obs      Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
          0 |      291   34.27835   .4719713    8.051227    33.34943    35.20727
          1 |      309   66.35922   1.148007   20.18012    64.10029    68.61815
-----+-----
combined |      600    50.8     .9112938   22.32205    49.01028    52.58972
-----+-----
      diff |          -32.08087   1.268596          -34.57232   -29.58943
-----+-----
      diff = mean(0) - mean(1)                                t = -25.2885
Ho: diff = 0                                                  degrees of freedom = 598

      Ha: diff < 0              Ha: diff != 0              Ha: diff > 0
Pr(T < t) = 0.0000          Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

.      scalar sdo = r(mu_2) - r(mu_1)
.      display sdo
32.080873

.
. // Selection bias: difference in y0 for D=1 and D=0 groups
. qui ttest y0, by(d)
.
.      scalar selbias = r(mu_2) - r(mu_1)
.      display selbias
6.2556301

.
. // Heterogeneous treatment effect bias: difference in te for D=1 and D=0
. // NOTE: ATT, ATU, and ATE were calculated in part 1a
. qui summ d
.
.      scalar ptreat=r(mean)
.
.      scalar htebias = (1-ptreat)*(att-atu)
.      display htebias
3.7419094

.
. // Note that SDO = ATE + selbias + htebias
. display sdo - selbias - htebias
22.083333

```

```

.      display ate
22.083333

.
.
. // *****
. // 1d - calculate diff in means separately by group, then average
. // *****
.
.      qui ttest y if x==1, by(d)
.
.      scalar te1 = r(mu_2) - r(mu_1)
.
.
.      qui ttest y if x==2, by(d)
.
.      scalar te2 = r(mu_2) - r(mu_1)
.
.
.      qui ttest y if x==3, by(d)
.
.      scalar te3 = r(mu_2) - r(mu_1)
.
.
.      qui ttest y if x==4, by(d)
.
.      scalar te4 = r(mu_2) - r(mu_1)
.
.
.      // simple average of these four estimates
.      di (te1 + te2 + te3 + te4) / 4
21.25

.
.      // The average te across these four groups is 21.25, which differs a bit
.      // from the simple difference in outcomes (32.08), known ATT (25.8), and
.      // known ATE (22.1). It is arguably better than the simple difference since
.      // it compares treated and untreated within group. All units with the same
.      // group have the same y0, so this is removing the selection bias.
.
.
. // *****
. // 1e - calculate a weighted average of the above group differences
. // *****
.
.      // weighted average of these four estimates (using # in each type as weights
> )
.      di ((te1*100)+(te2*150)+(te3*200)+(te4*150))/600
22.083333

.      di ate
22.083333

.
.      // This is 22.1, equal to the ATE. This makes sense as we are calculating
.      // the ATE separately for each group (without concern for selection bias)
.      // and then weighting each group according to the number of units in each
.      // group. This improves on the mean in part d since the groups vary in size.
.      // The straight average of the four groups weights these groups equally.
.

```

```
.
. // *****
. // 1f - OLS regression controlling for group/type
. // *****
```

```
.
. // regression of y on d controlling for type
. reg y d i.x
```

Source	SS	df	MS	Number of obs	=	600
Model	285845.879	4	71461.4697	F(4, 595)	=	3369.19
Residual	12620.1213	595	21.210288	Prob > F	=	0.0000
				R-squared	=	0.9577
				Adj R-squared	=	0.9574
Total	298466	599	498.27379	Root MSE	=	4.6055

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d	20.864	.4028864	51.79	0.000	20.07275 21.65525
x					
2	43.11451	.6308407	68.34	0.000	41.87556 44.35345
3	17.49696	.5719426	30.59	0.000	16.37369 18.62023
4	5.509494	.5973605	9.22	0.000	4.336302 6.682685
_cons	22.06672	.4732186	46.63	0.000	21.13734 22.9961

```
.
. // The coefficient on d is 20.9, again different from all of the above
. // estimates and known ATT and ATE. A regression like this provides a
. // variance-weighted average treatment effect, in which groups with more
. // treatment variance get more weight.
```

```
.
. // *****
. // 1g - What if d were randomly assigned? Would this guarantee that
. // the SDO would equal the ATE?
. // *****
```

```
.
. // Randomization will eliminate selection bias and heterogeneous treatment
. // effect bias in EXPECTATION, but it is possible that the SDO will differ
. // from the ATE simply due to chance (sampling error).
```

```
.
. gen drand = runiform()<=0.50
.
. gen y2 = (drand*y1) + ((1-drand)*y0)
.
. ttest y2, by(drand)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	315	36.92063	.5147998	9.136791	35.90774 37.93353
1	285	60.80702	1.153064	19.46597	58.53738 63.07665
combined	600	48.26667	.7809597	19.12953	46.73291 49.80042
diff		-23.88638	1.222981		-26.28824 -21.48452

```
diff = mean(0) - mean(1)
Ho: diff = 0
t = -19.5313
degrees of freedom = 598
```

```
Ha: diff < 0
Pr(T < t) = 0.0000
Ha: diff != 0
Pr(|T| > |t|) = 0.0000
Ha: diff > 0
Pr(T > t) = 1.0000
```



```

.
.      // In the above random assignment of d, the simple difference in means is
.      // 23.9. This is not too far from the ATE, but it is not exactly right.
.
.
.
.      // *****
.      // Question 3
.      // *****
.
.      // Set up data
.      use https://github.com/spcorcor18/LPO-8852/raw/main/data/LUSD4_5.dta, clear
.
.      // NOTE: keep grade 5 and year 2005 as instructed
.      keep if grade==5 & year==2005
(35,242 observations deleted)
.
.      // NOTE: I am keeping only cases with nonmissing mathz, totexp, and econdis,
.      // the three variables used below. (Not doing so will create a small problem
.      // below where the long and short regressions have different numbers of
.      // observations.
.      keep if mathz~=. & totexp~=. & econdi~=.
(160 observations deleted)
.
.      // part a
.      reg mathz totexp

```

Source	SS	df	MS	Number of obs	=	11,759
Model	89.1137402	1	89.1137402	F(1, 11757)	=	89.91
Residual	11653.2051	11,757	.991171654	Prob > F	=	0.0000
Total	11742.3189	11,758	.998666345	R-squared	=	0.0076
				Adj R-squared	=	0.0075
				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	11,759	10.93214	9.843389	0	45

```

.      display b*r(sd)
.08705738

```

```

.      // part d
.      // "long" regression
.      reg mathz totexp econdis

```

Source	SS	df	MS	Number of obs	=	11,759
Model	993.398767	2	496.699383	F(2, 11756)	=	543.24
Residual	10748.9201	11,756	.914334817	Prob > F	=	0.0000
Total	11742.3189	11,758	.998666345	R-squared	=	0.0846
				Adj R-squared	=	0.0844
				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	=	11,759
Model	30.6888446	1	30.6888446	F(1, 11757)	=	217.36
Residual	1659.97057	11,757	.141189977	Prob > F	=	0.0000
				R-squared	=	0.0182
				Adj R-squared	=	0.0181
Total	1690.65941	11,758	.143788009	Root MSE	=	.37575

econdis	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
totexp	-.0051901	.000352	-14.74	0.000	-.0058802	-.0045001
_cons	.8826599	.0051786	170.44	0.000	.8725089	.8928108

```
. scalar pil=_b[totexp]
```

```
. // "short" regression
. reg mathz totexp
```

Source	SS	df	MS	Number of obs	=	11,759
Model	89.1137402	1	89.1137402	F(1, 11757)	=	89.91
Residual	11653.2051	11,757	.991171654	Prob > F	=	0.0000
				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 + (pil*gamma) /* this should be the same as the OLS slope */
.00884425
```

```
. // part e
. // get residual from regressing totexp on econdis
. reg totexp econdis
```

Source	SS	df	MS	Number of obs	=	11,759
Model	20679.8413	1	20679.8413	F(1, 11757)	=	217.36
Residual	1118580	11,757	95.1416181	Prob > F	=	0.0000
				R-squared	=	0.0182
				Adj R-squared	=	0.0181
Total	1139259.85	11,758	96.8923155	Root MSE	=	9.7541

totexp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
econdis	-3.497402	.2372232	-14.74	0.000	-3.962399	-3.032405
_cons	13.82071	.2155889	64.11	0.000	13.39812	14.2433

```
. predict uhat, resid
```

```
. // regression anatomy formula: b1 = COV(Y,RESID)/VAR(RESID)
. corr mathz uhat, cov
(obs=11,759)
```

```
-----+-----
      |      mathz      uhat
mathz |      .998666
uhat  |      .476954   95.1335
```

```
. scalar cov=r(cov_12)
```

```
. display cov
.4769539
```

```
. summ uhat
```

```
-----+-----
Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
uhat     |    11,759  -2.52e-08    9.753642  -13.82071   34.67669
```

```
. scalar vuhat=r(Var)
```

```
. display uhat
3.6766887
```

```
. display cov/vuhat
.00501352
```

```
. reg mathz totexp econdis
```

```
-----+-----
Source |      SS      df      MS      Number of obs      =      11,759
-----+-----
Model |  993.398767      2    496.699383      F(2, 11756)      =      543.24
Residual | 10748.9201    11,756    .914334817      Prob > F      =      0.0000
Total | 11742.3189    11,758    .998666345      R-squared      =      0.0846
                                           Adj R-squared   =      0.0844
                                           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   -1.7840823   -.6920744
_cons |   .6180293    .0245521     25.17    0.000    .5699032    .6661555
```

```
. // part f
. reg mathz totexp econdis mathz_1 lep
```

```
-----+-----
Source |      SS      df      MS      Number of obs      =      11,755
-----+-----
Model |  5534.6132      4    1383.6533      F(4, 11750)      =      2620.54
Residual | 6204.02803    11,750    .528002386      Prob > F      =      0.0000
Total | 11738.6412    11,754    .998693316      R-squared      =      0.4715
                                           Adj R-squared   =      0.4713
                                           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   -1.3317388   -.2580154
mathz_1 |   .6509335    .0071455     91.10    0.000    .6369272    .6649398
lep     |  -.1631717    .0160643    -10.16    0.000   -1.1946605   -.131683
_cons |   .2530979    .019203     13.18    0.000    .2154568    .290739
```

```
.  
.   
.   
. // Close log and convert to PDF  
. log close  
    name: <unnamed>  
    log: C:\Users\corcorssp\Dropbox\_TEACHING\Regression II\Problem sets\Problem se  
> t 1 - Potential outcomes\PS1.txt  
    log type: text  
    closed on: 5 Sep 2023, 17:02:35  
-----  
> -----
```