Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Lecture 1 Exercise Solutions

1.1 **Potential outcomes and treatment effects.** See Lecture 1 exercises do file for code.

1.2 **Estimating treatment effects with randomization.** See Lecture 1 exercises do file for code.

1.3 **Simulated data with selection into treatment on *X*.** See Lecture 1 exercises do file for code.

1.4 **RCT of private school vouchers.** In a well-known study, Howell and Peterson (2006) evaluated the effects of a private school voucher in NYC from the School Choice Scholarships Foundation (SCSF). This program provided scholarships of up to $1,400 for 1,300 children from low-income families to attend a private elementary school. There were more applicants to the program than vouchers, so a random lottery was used to award the scholarships. Ultimately, 1,300 families received the voucher and 960 didn't.

(a) Let $D_i = 1$ if the student was offered a voucher and $D_i = 0$ if not. Suppose we wanted to estimate the simple population regression function below, where Y_i represents student achievement after three years of the program:

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

Under what conditions would this regression describe "differences in average potential outcomes for a well-defined population" (our criteria for causal interpretation)? Do those conditions hold here? How would you describe the relevant population? What is our *estimand* of interest (ATE, ATT, ATU, something else)?

In the population, $\beta_1 = \mathbf{E}[\mathbf{Y}(1)|\mathbf{D_i} = 1] - \mathbf{E}[\mathbf{Y}(0)|\mathbf{D_i} = 0]$. If the $\mathbf{D_i} = 1$ and $\mathbf{D_i} = 0$ groups have the same distribution of potential outcomes (e.g., $(\mathbf{Y_{i1}}, \mathbf{Y_{0i}}) \perp\!\!\!\perp \mathbf{D_i}$) then this simplifies to $\beta_1 = \mathbf{E}[\mathbf{Y}(1) - \mathbf{Y}(0)]$. If the randomization was successful, this condition should hold here—that is, mean potential outcomes for the treated and untreated groups should be the same. The relevant population is low-income families in NYC who applied for a private school voucher. (Note the population of applicants may differ from the general population of low-income families in NYC). The estimand of interest here would probably be considered an ATT since the focus is on applicants to a voucher program (families

that presumably would use the voucher if given the opportunity). We should also be careful about how we define "treatment." Students were *offered* a voucher at random but did not necessarily use the voucher to attend a private school. So the treatment here is the voucher "offer," not attending a private school. One could think of this exercise as estimating the "intent-to-treat" effect of attending a private school with a voucher.

(b) Read the following dataset from Github which contains a subsample of 521 African-American students who participated in the lottery:

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/nyvoucher.dta, clear
```

(c) Use `ttest` and the simple regression model above to estimate the effects of the voucher (*voucher*) on student achievement after three years of the program (*post_ach*). Is the estimated effect statistically significant? Practically significant? (The outcome variable is a composite measure of reading and math achievement, expressed as a national percentile score).

See output below. Students offered the private school voucher scored 4.9 percentile points higher, on average, than students not offered the voucher. (Note the point estimate, standard error, *t*-statistic and *p*-value are the same in this case whether one uses a t-test or simple regression.) The difference is statistically significant ($p < 0.004$). To assess practical significance, it is useful to compare the magnitude of the difference (4.9 points) to the overall standard deviation in the outcome (19.2). This is an effect size of 0.255, a rather large effect size in education.

```
. ttest post_ach, by(voucher) rev

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       1 |     291    26.02921       1.158      19.754    23.75006    28.30836
       0 |     230    21.13043    1.198249    18.17234    18.76943    23.49144
---------+--------------------------------------------------------------------
Combined |     521     23.8666     .841557     19.2089    22.21333    25.51987
---------+--------------------------------------------------------------------
    diff |            4.898775    1.682719                1.592998    8.204552
------------------------------------------------------------------------------
    diff = mean(1) - mean(0)                                      t =   2.9112
H0: diff = 0                                     Degrees of freedom =      519

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9981         Pr(|T| > |t|) = 0.0038          Pr(T > t) = 0.0019
```

```
. reg post_ach voucher

      Source |       SS           df       MS      Number of obs   =       521
-------------+----------------------------------   F(1, 519)       =      8.48
       Model |  3082.89021         1  3082.89021   Prob > F        =    0.0038
    Residual |  188787.589       519  363.752579   R-squared       =    0.0161
-------------+----------------------------------   Adj R-squared   =    0.0142
       Total |  191870.479       520   368.98169   Root MSE        =    19.072

------------------------------------------------------------------------------
    post_ach | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     voucher |   4.898775   1.682719     2.91   0.004     1.592998    8.204552
       _cons |   21.13043    1.25759    16.80   0.000     18.65984    23.60103
------------------------------------------------------------------------------

. summ post_ach

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
    post_ach |        521     23.8666     19.2089          0         89

.  display b/r(sd)
.25502636
```

(d) Randomization in theory should prevent omitted variables bias. However, in finite samples, there may be *incidental* (chance) correlation between treatment assignment and other predictors of the outcome. The first step in the analysis of any RCT is to "check for balance" between the treated and untreated group on a host of baseline predictors. (This can also be revealing about whether the randomization "worked.") The only other variable in this dataset is a measure of baseline achievement, *pre_ach*. How does this measure differ between the treated and untreated group? (You can compare both means and other features of the distribution).

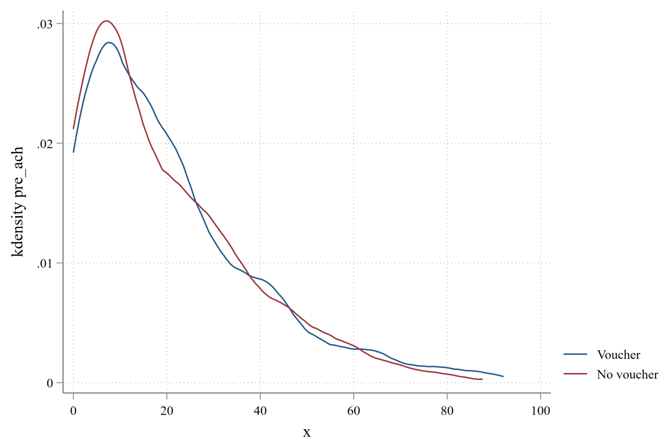
See output below. At *baseline,* students offered the voucher scored 1.17 percentile points higher than students not offered the voucher. The difference is not statistically significant, however (p = 0.4698). In larger samples, one would not expect to see a mean difference between these groups. However, in finite samples, there may be chance differences between them. It is good practice to compare more than just the means of the two groups. The code below includes an overlapping kernel densities for the voucher and no voucher groups.

```
. ttest pre_ach, by(voucher) rev
```

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       1 |     291    20.67869    1.097621    18.72401    18.51838    22.83901
       0 |     230    19.51304    1.164686    17.66333    17.21817    21.80791
---------+--------------------------------------------------------------------
Combined |     521    20.16411     .799776    18.25523    18.59292     21.7353
---------+--------------------------------------------------------------------
    diff |            1.165651    1.611368               -1.999956    4.331257
------------------------------------------------------------------------------
    diff = mean(1) - mean(0)                                      t =   0.7234
H0: diff = 0                                     Degrees of freedom =      519

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7651         Pr(|T| > |t|) = 0.4698          Pr(T > t) = 0.2349

. twoway (kdensity pre_ach if voucher==1) (kdensity pre_ach if voucher==0)
```



(e) Add the *pre_ach* measure to the regression function below (as X). What purpose does this serve? How does this additional covariate change your point estimate for β_1 (if at all)? How does it change the standard error for β_1 (if at all)?

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

See results below. If the randomization was successful, inclusion of *pre_ach* as a covariate should have little effect on the point estimate of *voucher*. However, as noted above, in finite samples there can be incidental correlation between treatment status and the baseline covariates. Controlling for these can help "purge" any chance correlation.

Inclusion of baseline covariates should also increase the *precision* of your estimator of the treatment effect. Note that the standard error fell from 1.68 (without covariates) to 1.27 (with the *pre_ach* control).

```
. reg post_ach voucher pre_ach

      Source |       SS           df       MS      Number of obs   =       521
-------------+----------------------------------   F(2, 518)       =    205.40
       Model |  84863.1705         2  42431.5852   Prob > F        =    0.0000
    Residual |  107007.308       518  206.577815   R-squared       =    0.4423
-------------+----------------------------------   Adj R-squared   =    0.4401
       Total |  191870.479       520   368.98169   Root MSE        =    14.373

------------------------------------------------------------------------------
    post_ach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     voucher |   4.097609    1.26873     3.23   0.001      1.60512    6.590097
     pre_ach |   .6873125   .0345439    19.90   0.000      .619449    .7551759
       _cons |   7.718877   1.162978     6.64   0.000     5.434143    10.00361
------------------------------------------------------------------------------
```