Regression II
Vanderbilt University
Prof. Sean P. Corcoran

September 9 2025
61 total points

Problem Set 2 *Solutions*

This problem set will use the National Education Longitudinal Study (NELS-88) data and matching methods to estimate the academic benefits, if any, to attending a Catholic high school. The variable definitions in this dataset should be self-explanatory, but if you have any questions, just ask.

You can read the data into Stata directly using this syntax:

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

1. Provide some basic descriptive information about students in this dataset. How many observations are there? What proportion attended a Catholic high school? What proportion graduated high school on time? What proportion entered post-secondary education after high school? What are the overall means and standard deviations for 12th grade math and reading scores, respectively? (**5 points**)

2. Create a few additional variables for the analysis (**5 points**):

   - The family income variable *faminc8* is an ordinal categorical variable with 12 categories. Create a "continuous" version of the family income variable *faminc8* by assigning a dollar amount equal to the midpoint of each interval. For example, \$4,000 for \$3,000-\$4,999.

   - Create a "collapsed" version of the family income variable *faminc8* in which 1=≤\$19,999, 2=\$20,000 to 34,999, and 3=\$35,000 to 74,900, corresponding to *Lo, Med,* and *Hi* income. This will allow you to replicate Tables 12.1 and 12.2 in Murnane & Willett.

   - Create a categorical version of the 8th grade math achievement variable (*math8*) with four categories corresponding to *Lo, MLo, MHi,* and *Hi* achievement. The cut points for these four categories should be 38, 44, and 51. Hint: I like to use the `egen` *`varname`*`=cut(`*`varname2`*`)` command for creating ordered categorical variables and quantiles. This will allow you to replicate Table 12.2 in Murnane & Willett.

   - Create dummy variables for each parent's *highest* level of education (<HS grad, HS grad, some college, college+). Also create dummy variables that indicate the maximum of the two parents' highest education.

3. Use this dataset to replicate the statistics found in Table 12.1 in Murnane & Willett (in the lecture notes and reproduced below). Specifically, report (**8 points**):

    - Mean (continuous) income by income strata, separately for public and Catholic school students. Also conduct t-tests for significant differences within each strata. Does income appear balanced within each strata? Note: M&W used the ordinal income variable here; you should use the continuous one you created.

    - Mean 12th grade math scores by income strata, separately for public and Catholic school students. Also conduct t-tests for significant differences within each strata.

    - The ATE and ATT estimates by calculating differences within each strata and weighting appropriately. Compare this to the simple difference in means.

4. Now replicate the statistics reported in Table 12.2 in Murnane & Willett (in the lecture notes and reproduced below), where the strata are income (3 categories) and 8th grade math achievement (4 categories). Specifically, report (**8 points**):

    - Mean 12th grade math scores by income and baseline achievement strata. Also conduct t-tests for significant differences within each strata.

    - The ATE and ATT estimates by calculating differences within each strata and weighting appropriately.

5. Use `teffects` to exact match on the 3-category family income variable used in #3 and calculate the ATE and ATT. How do these compare to your estimates in #3? What are the minimum and maximum number of exact matches? (**5 points**)

6. After exact matching in #5 use `tebalance summarize` to check for balance on your continuous family income measure (in dollars), and 8th grade math and reading scores. Note you *can* conduct balance checks on variables that were not part of your original exact matching algorithm. Explain how to read the results here. How do the Catholic and public schools students in the matched sample compare on their distributions of these variables? Note: do this after requesting the ATT, not ATE, as the results will differ. (**5 points**)

7. Do the same as #5 and #6 but exact match on both the 3-category family income variable and 4-category baseline math achievement variables used in #4. How do these compare to your answer in #4? What are the minimum and maximum number of exact matches? How do the Catholic and public school students in the matched sample compare now? (**5 points**)

8. Estimate the ATT of attending a Catholic school on two later outcomes: high school graduation and enrollment in post-secondary education. Use nearest neighbor matching (with Mahalanobis distance) on the following covariates: 8th grade math achievement, 8th grade reading achievement, family income (continuous), and the highest educational attainment of either parent. For now, just use the ordinal version of parent's educational attainment. Interpret the point estimates. What is the minimum and maximum number of nearest neighbors used? (**5 points**)

9. After nearest neighbor matching in #8 use `tebalance summarize` and `tebalance box` to check for balance on your matching variables. Use `tebalance density` to compare distributions of the two test score variables. How do the distributions compare? (**5 points**)

10. Repeat #8 but force an exact match on parent's educational attainment. Try `tebalance summarize` again. How did the exact match affect the balance, if at all? (**5 points**)

11. Repeat #8 but force an exact match on parent's educational attainment *and* increase the number of nearest neighbors to 5. Include the Abadie & Imbens bias correction for the continuous covariates. Try `tebalance summarize` again. How did the exact match affect the balance, if at all? What happened to the standard error of your ATE? (**5 points**)

12. What is the assumption necessary to interpret the matching estimator in #11 as causal? Do you believe it holds in this case? Why or why not? (**5 points**)

See attached do file for solutions.

```
------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\corcorsp\Dropbox\_TEACHING\Regression II\Problem sets\Problem s
> et 2 - Matching and weighting 1\PS2.txt
  log type:  text
 opened on:   9 Sep 2025, 13:42:57

.
.
. // ************************************************************
. //
. // Problem set 2
. // Last updated: September 9, 2025
. //
. // ************************************************************
.
. use https://stats.idre.ucla.edu/stat/stata/examples/methods_matter/chapter12/cathol
> ic, clear

.
. // *********************************
. // Question 1 - descriptives
. // *********************************
.
.         summ, sep(0)

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
          id |      5,671     4626664     2700654     124902    7979086
      read12 |      5,671    51.00126    9.476733      29.15      68.09
      math12 |      5,671    51.05124    9.502415      29.88      71.37
      hsgrad |      5,671    .9169459    .2759884          0          1
       inpse |      5,671    .7092224    .4541612          0          1
    catholic |      5,671    .1043908    .3057938          0          1
       read8 |      5,671    51.54138    9.695829      32.05      70.55
       math8 |      5,671    51.48952    9.683425      34.48       77.2
      female |      5,671    .5200141    .4996433          0          1
        race |      5,671    3.532887    .9537466          1          5
       white |      5,671    .6892964    .4628225          0          1
       black |      5,671    .0975137    .2966821          0          1
        hisp |      5,671    .1162053    .3204992          0          1
         api |      5,671    .0585435    .2347889          0          1
     nativam |      5,671    .0384412    .1922758          0          1
     parmar8 |      5,671    5.344384    1.576191          1          6
     faminc8 |      5,671    9.526186    2.217688          1         12
     fathed8 |      5,671    3.606948    2.267043          1          8
     mothed8 |      5,671    3.380356    2.141246          1          8
     fhowfar |      5,671    4.818198    1.105028          1          6
     mhowfar |      5,671    4.858226    1.074148          1          6
      fight8 |      5,671    .2191853    .5005381          0          2
       nohw8 |      5,671     .143361    .3504715          0          1
    disrupt8 |      5,671    .1795098    .3838125          0          1
   riskdrop8 |      5,671    .6236995    .9031568          0          5

.
.         // There are 5,671 observations, where 10.4% attended a Catholic HS. 91.7%
.         // graduated HS on time and 70.9% enrolled in post-secondary education
.         // after HS. The mean (sd) for 12th grade math and 12th grade reading are
.         // 51.1 (9.5) and 51.0 (9.5).
.
```

```
.
. // ********************************
. // Question 2 - create some new vars
. // ********************************
.
.         *ssc install fre
.         fre faminc8

faminc8 -- total annual family income in 8th grade
-------------------------------------------------------------------------------
                          |      Freq.    Percent      Valid       Cum.
--------------------------+----------------------------------------------------
Valid   1  none           |         18       0.32       0.32       0.32
        2  <$1000         |         42       0.74       0.74       1.06
        3  $1000-$2999    |         84       1.48       1.48       2.54
        4  $3000-$4999    |         85       1.50       1.50       4.04
        5  $5000-$7499    |        144       2.54       2.54       6.58
        6  7500-$9999     |        175       3.09       3.09       9.66
        7  $10000-$14999  |        447       7.88       7.88      17.55
        8  $15000-$19999  |        441       7.78       7.78      25.32
        9  $20000-$24999  |        655      11.55      11.55      36.87
        10 $25000-$34999  |       1267      22.34      22.34      59.21
        11 35000-$49999   |       1419      25.02      25.02      84.24
        12 50000-$74999   |        894      15.76      15.76     100.00
        Total             |       5671     100.00     100.00
-------------------------------------------------------------------------------

.
.         // "continuous" version of family income
.         gen faminc8b=0 if faminc8==1
(5,653 missing values generated)

.         replace faminc8b = (0+1000)/2 if faminc8==2
(42 real changes made)

.         replace faminc8b = (1000+2999)/2 if faminc8==3
(84 real changes made)

.         replace faminc8b = (3000+4999)/2 if faminc8==4
(85 real changes made)

.         replace faminc8b = (5000+7499)/2 if faminc8==5
(144 real changes made)

.         replace faminc8b = (7500+9999)/2 if faminc8==6
(175 real changes made)

.         replace faminc8b = (10000+14999)/2 if faminc8==7
(447 real changes made)

.         replace faminc8b = (15000+19999)/2 if faminc8==8
(441 real changes made)

.         replace faminc8b = (20000+24999)/2 if faminc8==9
(655 real changes made)

.         replace faminc8b = (25000+34999)/2 if faminc8==10
(1,267 real changes made)

.         replace faminc8b = (35000+49999)/2 if faminc8==11
(1,419 real changes made)
```

```
.         replace faminc8b = (50000+74999)/2 if faminc8==12
(894 real changes made)

.         label var faminc8b "family income in 8th grade (dollars)"

.
.         // 3-category version of family income (following Murnane and Willett)
.         gen faminc8c = 1 if faminc8<=8
(4,235 missing values generated)

.         replace faminc8c = 2 if faminc8>=9 & faminc8<=10
(1,922 real changes made)

.         replace faminc8c = 3 if faminc8>=11 & faminc8~=.
(2,313 real changes made)

.         label var faminc8c "family income in 8th grade (three categories)"

.
.         // 4-category version of 8th grade math scores
.         egen math8b=cut(math8), at(30,38,44,51,80) icodes

.         replace math8b=math8b+1
(5,671 real changes made)

.         label var math8b "8th grade math score (four categories)

.
.         // father's highest education
.         codebook fathed8

-------------------------------------------------------------------------------
fathed8                                          father^s highest level of education
-------------------------------------------------------------------------------

                  Type: Numeric (byte)
                 Label: farcat

                 Range: [1,8]                         Units: 1
         Unique values: 8                         Missing .: 0/5,671

            Tabulation: Freq.   Numeric  Label
                          873         1  not finish hs
                        1,778         2  hs grad
                          660         3  junior coll
                          443         4  coll <4
                          743         5  coll grad
                          346         6  masters
                          141         7  doctorate
                          687         8  dont know

.         gen fathed1 = fathed8==1 /* hs dropout */

.         gen fathed2 = fathed8==2 /* hs grad */

.         gen fathed3 = (fathed8>=3 & fathed8<=4) /* some college */

.         gen fathed4 = (fathed8>=5 & fathed8<=8) /* 4yr college or more */

.         label var fathed1 "father's highest ed: hs dropout"
```

```
.         label var fathed2 "father's highest ed: hs grad"

.         label var fathed3 "father's highest ed: some college"

.         label var fathed4 "father's highest ed: 4yr college or more"

.
.         // mother's highest education
.         codebook mothed8

-------------------------------------------------------------------------------
mothed8                                          mother^s highest level of education
-------------------------------------------------------------------------------

                  Type: Numeric (byte)
                 Label: farcat

                 Range: [1,8]                         Units: 1
         Unique values: 8                         Missing .: 0/5,671

            Tabulation: Freq.   Numeric  Label
                          815         1  not finish hs
                        2,091         2  hs grad
                          686         3  junior coll
                          468         4  coll <4
                          655         5  coll grad
                          299         6  masters
                           82         7  doctorate
                          575         8  dont know

.         gen mothed1 = mothed8==1 /* hs dropout */

.         gen mothed2 = mothed8==2 /* hs grad */

.         gen mothed3 = (mothed8>=3 & mothed8<=4) /* some college */

.         gen mothed4 = (mothed8>=5 & mothed8<=8) /* 4yr college or more */

.         label var mothed1 "mother's highest ed: hs dropout"

.         label var mothed2 "mother's highest ed: hs grad"

.         label var mothed3 "mother's highest ed: some college"

.         label var mothed4 "mother's highest ed: 4yr college or more"

.
.         forvalues j=1/4 {
  2.            replace fathed`j'=. if fathed8==.
  3.            replace mothed`j'=. if mothed8==.
  4.            }
(0 real changes made)
(0 real changes made)
(0 real changes made)
(0 real changes made)
(0 real changes made)
(0 real changes made)
(0 real changes made)
(0 real changes made)
```

```
.
.         // highest education of two parents
.         egen pared8=rowmax(fathed8 mothed8)

.         gen pared1 = pared8==1 /* hs dropout */

.         gen pared2 = pared8==2 /* hs grad */

.         gen pared3 = (pared8>=3 & pared8<=4) /* some college */

.         gen pared4 = (pared8>=5 & pared8<=8) /* 4yr college or more */

.         label var pared1 "parent's highest ed: hs dropout"

.         label var pared2 "parent's highest ed: hs grad"

.         label var pared3 "parent's highest ed: some college"

.         label var pared4 "parent's highest ed: 4yr college or more"

.         label var pared8 "parent's highest education"

.
.
. // ********************************
. // Question 3 - replicate Table 12.1
. // ********************************
.
.         // use 3-category strata of family income
.         tabulate faminc8c catholic, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

    family |
 income in |
 8th grade |
    (three |
categories | attended catholic hs?
         ) |        no        yes |     Total
-----------+----------------------+----------
         1 |     1,365         71 |     1,436
           |     95.06       4.94 |    100.00
-----------+----------------------+----------
         2 |     1,745        177 |     1,922
           |     90.79       9.21 |    100.00
-----------+----------------------+----------
         3 |     1,969        344 |     2,313
           |     85.13      14.87 |    100.00
-----------+----------------------+----------
     Total |     5,079        592 |     5,671
           |     89.56      10.44 |    100.00

.
.         // These counts correspond exactly to those in Table 12.1.
.
```

```
.         // Mean income by catholic enrollment, by strata
.         // Note: the dtable and collect commands in Stata 17+ offers other
.         //    alternatives to the below approach to formatting results
.         forvalues j=1/3 {
  2.            qui estpost ttest faminc8b if faminc8c==`j', by(catholic)
  3.            esttab, cell((mu_2(fmt(%12.0fc) label("Catholic")) mu_1(fmt(%12.0fc)
> ///
>               label("Public")) b(fmt(%12.0fc) label("Diff")) t(fmt(%12.3fc) ///
>                   label("t-statistic") star))) nonumb ///
>                   title(Mean income by Catholic enrollment - strata `j')
  4.            }

Mean income by Catholic enrollment - strata 1
------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------
faminc8b           12,774       11,251       -1,523       -2.333*
------------------------------------------------------------------------
N                    1436
------------------------------------------------------------------------

Mean income by Catholic enrollment - strata 2
------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------
faminc8b           28,008       27,386         -622       -2.219*
------------------------------------------------------------------------
N                    1922
------------------------------------------------------------------------

Mean income by Catholic enrollment - strata 3
------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------
faminc8b           50,988       50,097         -891       -1.565
------------------------------------------------------------------------
N                    2313
------------------------------------------------------------------------

.
.         // The tables above show the mean (continuous) income for Catholic and
.         // public HS students, separately by strata. Note the t-test direction is
.         // reversed, so a negative number means Catholic school students had
.         // *higher* values. (For some reason the "reverse" option is not working
.         // here. The results are not exactly comparable to Table 12.1
.         // since M&W used the categorical income values, not continuous.
.
.         // Now, mean 12th grade math by catholic enrollment, by strata
.         forvalues j=1/3 {
  2.            qui estpost ttest math12 if faminc8c==`j', by(catholic)
  3.            esttab, cell((mu_2(fmt(%5.2fc) label("Catholic")) mu_1(fmt(%5.2fc) //
> /
>               label("Public")) b(fmt(%5.2fc) label("Diff")) t(fmt(%12.3fc) ///
>                   label("t-statistic") star))) nonumb ///
>                   title(Mean 12th grade math by Catholic enrollment - strata `j')
  4.            }

Mean 12th grade math by Catholic enrollment - strata 1
------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------
math12              50.54        46.77        -3.76       -3.480***
------------------------------------------------------------------------
N                    1436
------------------------------------------------------------------------
```

```
Mean 12th grade math by Catholic enrollment - strata 2
------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------
math12              53.86        50.34        -3.52       -4.823***
------------------------------------------------------------------------
N                    1922
------------------------------------------------------------------------

Mean 12th grade math by Catholic enrollment - strata 3
------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------
math12              55.72        53.60        -2.12       -4.027***
------------------------------------------------------------------------
N                    2313
------------------------------------------------------------------------

.
.         // The tables above show the mean 12th grade math score for Catholic and
.         // public HS students, separately by strata. Note the t-test direction is
.         // reversed, so a negative number means Catholic school students had
.         // *higher* values. The results replicate Table 12.1 exactly.
.
.         // One way to get ATE and ATT manually (there are others)
.         preserve

.         gen math12p = math12 if catholic==0
(592 missing values generated)

.         gen math12c = math12 if catholic==1
(5,079 missing values generated)

.         collapse (mean) math12p math12c (count) n=math12 np=math12p nc=math12c, ///
>                 by(faminc8c)

.         gen te = math12c - math12p

.         // ATE = weight by # of observations in faminc8c cell
.         summ te [weight=n]
(analytic weights assumed)

    Variable |     Obs      Weight        Mean   Std. dev.       Min        Max
-------------+--------------------------------------------------------------------
          te |       3        5671     3.00864   .9125871   2.117859   3.762051

.         // ATT = weight by # of catholic observations in faminc8c cell
.         summ te [weight=nc]
(analytic weights assumed)

    Variable |     Obs      Weight        Mean   Std. dev.       Min        Max
-------------+--------------------------------------------------------------------
          te |       3         592    2.733595   .8924686   2.117859   3.762051

.         restore

.
.         // The ATE is 3.01 and the ATT is 2.73. These replicate Table 12.1 exactly.
```

```
.
.         // Simple difference in means - compare to ATE and ATT above
.         ttest math12, by(catholic) rev

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
     yes |     592    54.53951    .3478334    8.463153    53.85637    55.22265
      no |   5,079    50.64465    .1337825    9.534295    50.38238    50.90692
---------+--------------------------------------------------------------------
Combined |   5,671    51.05124     .126184    9.502415    50.80387    51.29861
---------+--------------------------------------------------------------------
    diff |             3.89486    .4094621                3.092157    4.697562
------------------------------------------------------------------------------
    diff = mean(yes) - mean(no)                                   t =   9.5121
H0: diff = 0                                     Degrees of freedom =     5669

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

.
.         // The simple difference in means is 3.89, which is larger than the two
.         // estimates above. This is not surprising, as we expect there to be
.         // positive selection bias.
.
.
. // *********************************
. // Question 4 - replicate Table 12.2
. // *********************************
.
.         // Mean 12th grade math by catholic enrollment, by strata
.         // Note: the tables command in Stata 17+ offers better alternatives to the
.         //   below approach
.         forvalues j=1/3 {
  2.            forvalues k=1/4 {
  3.               qui estpost ttest math12 if faminc8c==`j' & math8b==`k', by(catholic)
  4.                  display "Income group `j' baseline math group `k'"
  5.                  esttab, cell((mu_2(fmt(%5.2fc) label("Catholic")) mu_1(fmt(%5.2fc)
>  ///
>                       label("Public")) b(fmt(%5.2fc) label("Diff")) t(fmt(%12.3fc) /
> //
>                           label("t-statistic") star))) nonumb ///
>                           title(Mean 12th grade math by Catholic enrollment - income
>  `j' math `k')
  6.            }
  7.            }
Income group 1 baseline math group 1

Mean 12th grade math by Catholic enrollment - income 1 math 1
--------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
--------------------------------------------------------------------
math12              42.57        36.81        -5.76            .
--------------------------------------------------------------------
N                     143
--------------------------------------------------------------------
Income group 1 baseline math group 2

Mean 12th grade math by Catholic enrollment - income 1 math 2
--------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
--------------------------------------------------------------------
math12              41.70        40.99        -0.71       -0.621
--------------------------------------------------------------------
N                     454
--------------------------------------------------------------------
Income group 1 baseline math group 3
```

Mean 12th grade math by Catholic enrollment - income 1 math 3

	Catholic	Public	Diff	t-statistic
math12	48.65	47.12	-1.53	-0.949
N	398			

Income group 1 baseline math group 4

Mean 12th grade math by Catholic enrollment - income 1 math 4

	Catholic	Public	Diff	t-statistic
math12	56.59	56.12	-0.47	-0.412
N	441			

Income group 2 baseline math group 1

Mean 12th grade math by Catholic enrollment - income 2 math 1

	Catholic	Public	Diff	t-statistic
math12	39.77	37.94	-1.83	-0.455
N	98			

Income group 2 baseline math group 2

Mean 12th grade math by Catholic enrollment - income 2 math 2

	Catholic	Public	Diff	t-statistic
math12	44.56	41.92	-2.64	-2.520*
N	423			

Income group 2 baseline math group 3

Mean 12th grade math by Catholic enrollment - income 2 math 3

	Catholic	Public	Diff	t-statistic
math12	50.14	47.95	-2.19	-2.570*
N	518			

Income group 2 baseline math group 4

Mean 12th grade math by Catholic enrollment - income 2 math 4

	Catholic	Public	Diff	t-statistic
math12	59.42	57.42	-2.00	-2.740**
N	883			

Income group 3 baseline math group 1

```
Mean 12th grade math by Catholic enrollment - income 3 math 1
------------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------------
math12              40.40        39.79        -0.62       -0.224
------------------------------------------------------------------------------
N                     63
------------------------------------------------------------------------------
Income group 3 baseline math group 2

Mean 12th grade math by Catholic enrollment - income 3 math 2
------------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------------
math12              44.23        42.75        -1.48       -1.490
------------------------------------------------------------------------------
N                    359
------------------------------------------------------------------------------
Income group 3 baseline math group 3

Mean 12th grade math by Catholic enrollment - income 3 math 3
------------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------------
math12              50.71        49.18        -1.53       -2.139*
------------------------------------------------------------------------------
N                    505
------------------------------------------------------------------------------
Income group 3 baseline math group 4

Mean 12th grade math by Catholic enrollment - income 3 math 4
------------------------------------------------------------------------------

                 Catholic       Public         Diff  t-statistic
------------------------------------------------------------------------------
math12              59.66        58.93        -0.72       -1.587
------------------------------------------------------------------------------
N                   1386
------------------------------------------------------------------------------

.
.         // The tables above show the mean 12th grade math score for Catholic and
.         // public HS students, separately by strata. Note the t-test direction is
.         // reversed, so a negative number means Catholic school students had
.         // *higher* values. The results replicate Table 12.2 exactly.
.
.         // One way to get ATE and ATT manually (there are others)
.         preserve

.         gen math12p = math12 if catholic==0
(592 missing values generated)

.         gen math12c = math12 if catholic==1
(5,079 missing values generated)

.         collapse (mean) math12p math12c (count) n=math12 np=math12p nc=math12c, ///
>                  by(faminc8c math8b)
```

```
.         gen te = math12c - math12p

.         // ATE = weight by # of observations in faminc8c cell
.         summ te [weight=n]
(analytic weights assumed)

    Variable |     Obs      Weight        Mean   Std. dev.       Min        Max
-------------+--------------------------------------------------------------------
          te |      12        5671    1.499595   1.005945   .4710312   5.764858

.         // ATT = weight by # of catholic observations in faminc8c cell
.         summ te [weight=nc]
(analytic weights assumed)

    Variable |     Obs      Weight        Mean   Std. dev.       Min        Max
-------------+--------------------------------------------------------------------
          te |      12         592    1.313056   .7175966   .4710312   5.764858

.         restore

.
.         // The ATE is 1.50 and the ATT is 1.31. These replicate Table 12.2 exactly.
>
.
.
. // ********************************
. // Questions 5-7 exact matching
. // ********************************
.
.         // exact matching on 3-category income strata
.         teffects nnmatch (math12 faminc8c) (catholic), ematch(faminc8c) ate

Treatment-effects estimation                   Number of obs      =      5,671
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =         71
Distance metric: Mahalanobis                               max =       1969
------------------------------------------------------------------------------
             |               AI robust
      math12 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATE          |
    catholic |
(yes vs no)  |   3.008641   .4010225     7.50   0.000     2.222651     3.79463
------------------------------------------------------------------------------

.         teffects nnmatch (math12 faminc8c) (catholic), ematch(faminc8c) atet

Treatment-effects estimation                   Number of obs      =      5,671
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =         71
Distance metric: Mahalanobis                               max =       1969
------------------------------------------------------------------------------
             |               AI robust
      math12 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
    catholic |
(yes vs no)  |   2.733596    .369277     7.40   0.000     2.009826    3.457365
------------------------------------------------------------------------------
```

```
.     tebalance summarize faminc8b math8 read8
(refitting the model using the generate() option)

Covariate balance summary

                          Raw      Matched
------------------------------------------
Number of obs =         5,671        1,184
Treated obs   =           592          592
Control obs   =         5,079          592
------------------------------------------

-------------------------------------------------------------
                |Standardized differences          Variance ratio
                |        Raw     Matched           Raw    Matched
----------------+--------------------------------------------
       faminc8b |   .4572743    .0543268      .8885848   .9955878
          math8 |   .2606657    .1418083      .8200688   .7999527
          read8 |   .4356571    .3273512      .9223486   .9083642
-------------------------------------------------------------

.
.         // Q5: the ATE and ATT estimates are exactly the same as those found in #3.
.         // Algebraically, taking each observation and matching to its (treated or
.         // untreated) counterpart with the same income strata (and taking all ties)
.         // and then differencing these is the same as differencing the mean
.         // outcomes within strata. There was a minimum of 71 exact matches and a
.         // maximum of 1969. Q6: the exact matching on income strata significantly
.         // improved balance by (continuous) income--which is to be expected--but
.         // considerable imbalance on 8th grade tests scores remains. Balancing
.         // by income improved the balance on 8th grade scores a bit, but since
.         // this wasn't part of the matching algorithm, it's unsurprising that
.         // it remains unbalanced. Note when checking balance in the distribution
.         // of the covariates, one should look at both the means and ratio of the
.         // variances. For income, the variance ratio is ~1.
.
.         // exact matching on 3-category income strata AND 4-category baseline
.         // achievement strata. Note: vce(iid) is needed since there are fewer than
.         // 2 matches in some cases
.         teffects nnmatch (math12 faminc8c math8b) (catholic), ///
>                 ematch(faminc8c math8b) vce(iid) ate

Treatment-effects estimation                   Number of obs      =      5,671
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =       1159
------------------------------------------------------------------------------
      math12 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATE          |
    catholic |
(yes vs no)  |   1.499595    .316312     4.74   0.000     .8796352    2.119555
------------------------------------------------------------------------------

.         teffects nnmatch (math12 faminc8c math8b) (catholic), ///
>                 ematch(faminc8c math8b) vce(iid) atet

Treatment-effects estimation                   Number of obs      =      5,671
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =         57
Distance metric: Mahalanobis                               max =       1159
------------------------------------------------------------------------------
      math12 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
    catholic |
(yes vs no)  |   1.313056   .2519908     5.21   0.000     .8191632    1.806949
------------------------------------------------------------------------------
```

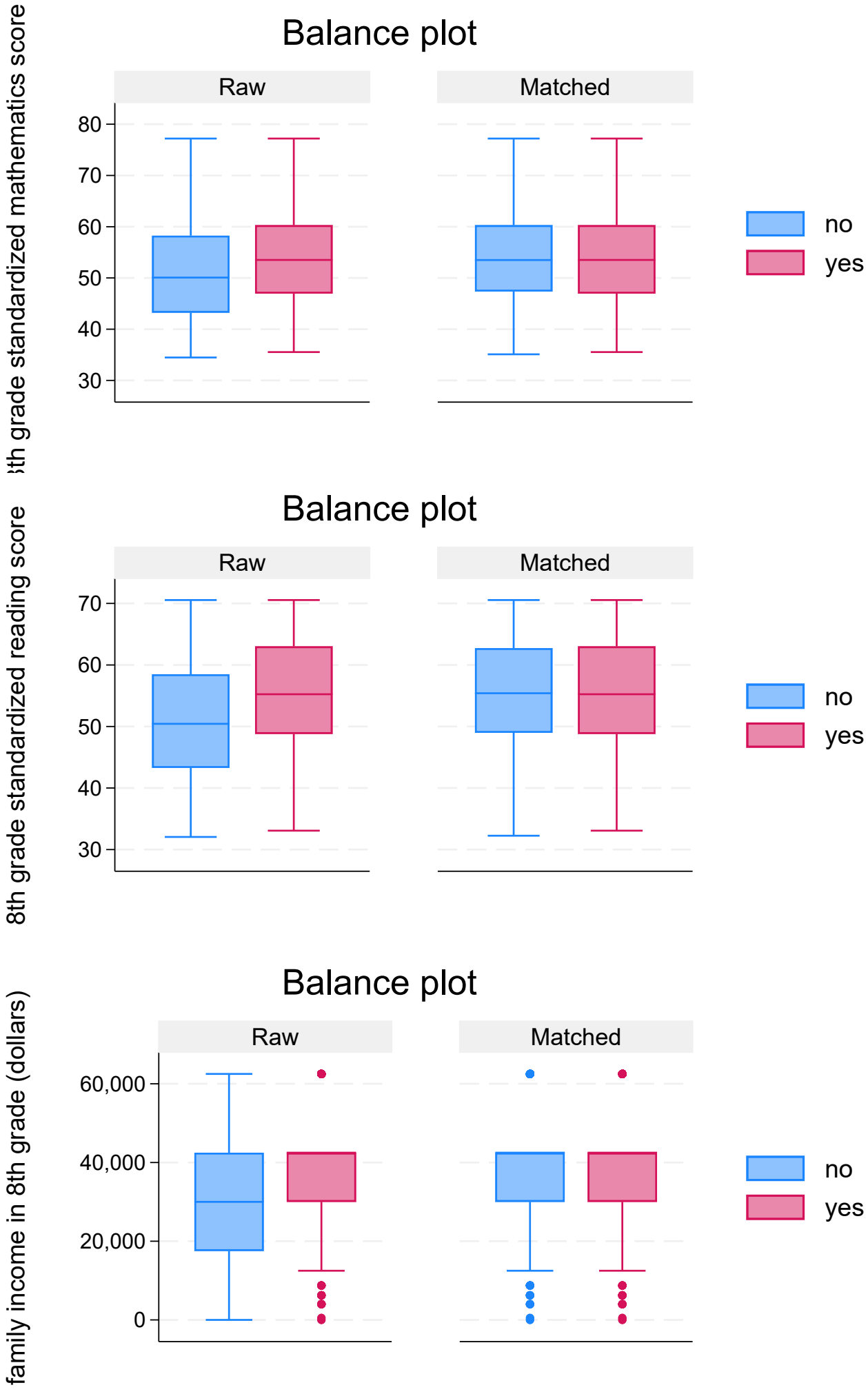
```
.     tebalance summarize faminc8b math8 read8
(refitting the model using the generate() option)

Covariate balance summary

                          Raw      Matched
----------------------------------------
Number of obs =         5,671        1,184
Treated obs   =           592          592
Control obs   =         5,079          592
----------------------------------------

------------------------------------------------------------------
                |Standardized differences          Variance ratio
                |        Raw     Matched           Raw    Matched
----------------+-------------------------------------------------
       faminc8b |   .4572743    .0450045      .8885848   1.000473
          math8 |   .2606657   -.0413211      .8200688   .8691157
          read8 |   .4356571    .2065909      .9223486    .948264
------------------------------------------------------------------

.
.         // Q7: the ATE and ATT estimates are exactly the same as those found in #4
.         // (for the same reason as noted in Q5). Here (for the ATE) there was a
.         // minimum number of matches of 1 and a maximum of 1159. (The min was 57
.         // for the ATT). The balance has improved even more on family income and
.         // 8th grade math, although 8th grade reading remains quite unbalanced
.         // (by about 0.2 sd).
.
.
. // ***********************************
. // Questions 8-9 - nearest neighbor match
. // ***********************************
.
.         teffects nnmatch (hsgrad math8 read8 faminc8b pared8) (catholic), atet

Treatment-effects estimation                   Number of obs      =      5,671
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =          1
------------------------------------------------------------------------------
             |              AI robust
      hsgrad | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
    catholic |
(yes vs no)  |    .035473   .0125184     2.83   0.005     .0109373    .0600087
------------------------------------------------------------------------------

.         teffects nnmatch (inpse math8 read8 faminc8b pared8) (catholic), atet

Treatment-effects estimation                   Number of obs      =      5,671
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =          1
------------------------------------------------------------------------------
             |              AI robust
       inpse | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
    catholic |
(yes vs no)  |   .0726351   .0221456     3.28   0.001     .0292306    .1160396
------------------------------------------------------------------------------
```

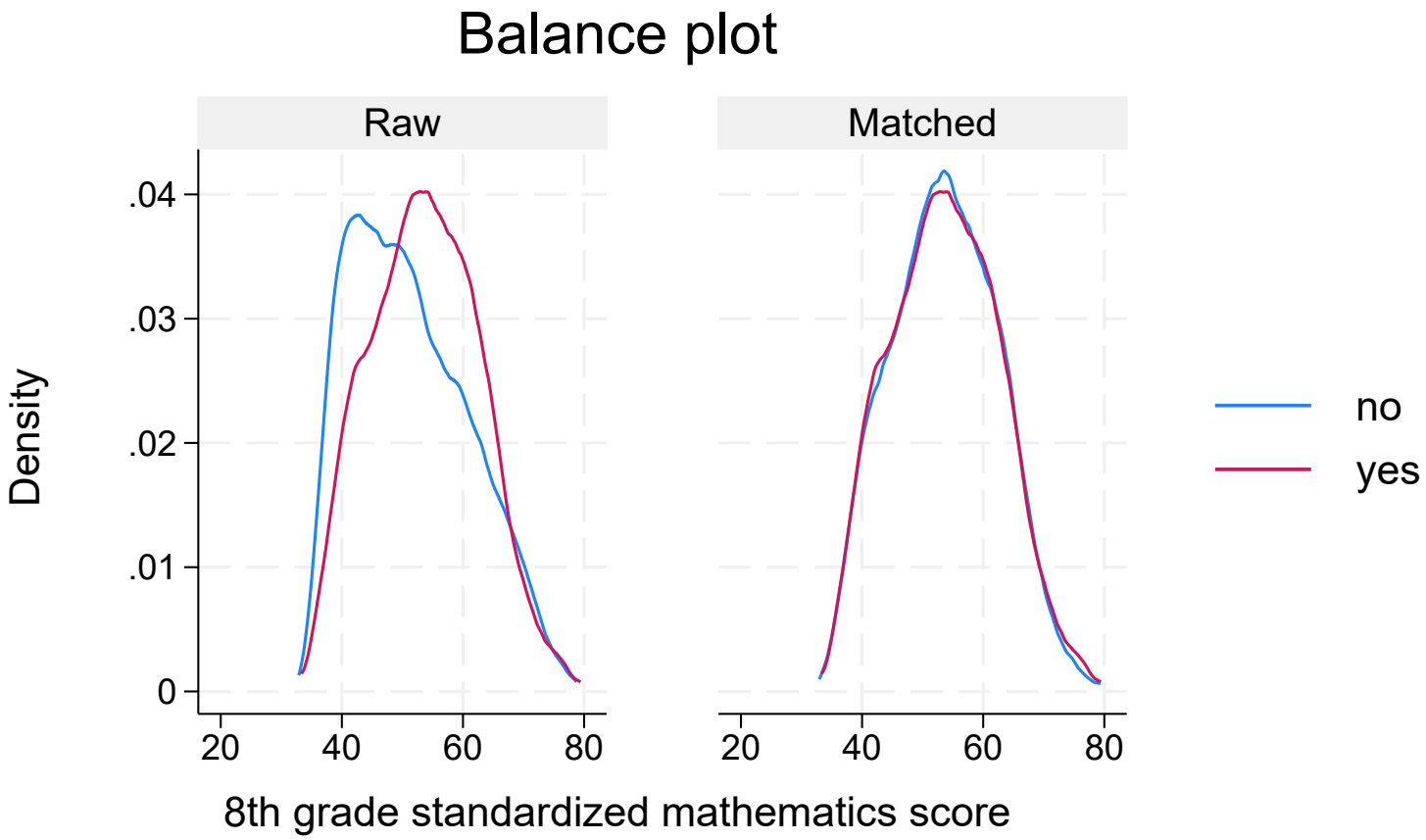
```
.
.         // Q8: The ATT for HS graduation is 0.0354 and for post-secondary
.         // enrollment is 0.0726. In other words, we estimate that Catholic HS
.         // grads are 3.5 ppts more likely to graduate from HS and 7.3 ppts more
.         // likely to enroll in post-secondary education. Only 1 nearest neighbor
.         // was used. (There were seemingly no ties).
.
.         tebalance summarize
(refitting the model using the generate() option)

Covariate balance summary

                          Raw      Matched
-----------------------------------------
Number of obs =         5,671        1,184
Treated obs   =           592          592
Control obs   =         5,079          592
-----------------------------------------

------------------------------------------------------------------
                |Standardized differences          Variance ratio
                |        Raw     Matched           Raw    Matched
----------------+-------------------------------------------------
          math8 |   .2606657    .0056571      .8200688   1.038155
          read8 |   .4356571    .0049383      .9223486   1.037133
       faminc8b |   .4572743   -.0035875      .8885848   1.011686
         pared8 |   .1701967   -.0057155      .8080069   .9853939
------------------------------------------------------------------

.         tebalance box math8, name(q8a, replace)
(refitting the model using the generate() option)

.         tebalance box read8, name(q8b, replace)
(refitting the model using the generate() option)

.         tebalance box faminc8b,name(q8c, replace)
(refitting the model using the generate() option)

.         tebalance density math8, name(q8d, replace)
(refitting the model using the generate() option)

.         tebalance density read8, name(q8e, replace)
(refitting the model using the generate() option)

.
.         graph combine q8a q8b q8c, col(1) xsize(4) ysize(6)

.         graph export q8a.pdf, as(pdf) replace
file q8a.pdf saved as PDF format

.         graph combine q8d q8e , col(1) xsize(4) ysize(5)

.         graph export q8b.pdf, as(pdf) replace
file q8b.pdf saved as PDF format

.
.         // Q9: the plots are attached. The distributions visually appear quite
.         // balanced for math and reading achievement. The means and variances
.         // are all quite comparable for the four matching variables.
.
```

```
.
. // ******************************************
. // Question 10 - force exact match on parents ed
. // ******************************************
.
.         teffects nnmatch (hsgrad math8 read8 faminc8b pared8) (catholic), ///
>                 atet ematch(pared8)

Treatment-effects estimation                    Number of obs      =      5,671
Estimator      : nearest-neighbor matching      Matches: requested =          1
Outcome model  : matching                                   min =          1
Distance metric: Mahalanobis                                max =          1
------------------------------------------------------------------------------
             |               AI robust
      hsgrad | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
    catholic |
(yes vs no)  |   .0371622   .0126988     2.93   0.003     .0122729    .0620514
------------------------------------------------------------------------------

.         teffects nnmatch (inpse math8 read8 faminc8b pared8) (catholic), ///
>                 atet ematch(pared8)

Treatment-effects estimation                    Number of obs      =      5,671
Estimator      : nearest-neighbor matching      Matches: requested =          1
Outcome model  : matching                                   min =          1
Distance metric: Mahalanobis                                max =          1
------------------------------------------------------------------------------
             |               AI robust
       inpse | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
    catholic |
(yes vs no)  |   .0760135   .0223403     3.40   0.001     .0322273    .1197997
------------------------------------------------------------------------------

.         tebalance summarize
(refitting the model using the generate() option)

Covariate balance summary

                          Raw      Matched
-----------------------------------------
Number of obs =         5,671        1,184
Treated obs   =           592          592
Control obs   =         5,079          592
-----------------------------------------

-------------------------------------------------------------------
                |Standardized differences          Variance ratio
                |        Raw     Matched           Raw    Matched
----------------+--------------------------------------------------
          math8 |   .2606657   -.0006861      .8200688   1.041008
          read8 |   .4356571    .0052826      .9223486   1.039061
       faminc8b |   .4572743   -.0049718      .8885848   1.017085
         pared8 |   .1701967           0      .8080069          1
-------------------------------------------------------------------

.
```

```
.         // Q10: parents education is now exactly balanced, and the distributions
.         // of the other variables remain quite balanced (although perhaps a little
.         // less so than Q9).
.
.
. // ************************************
. // Question 11 - nearest neighbor match (5)
. // ************************************
.
.         teffects nnmatch (hsgrad math8 read8 faminc8b pared8) (catholic), atet ///
>                 nneighbor(5) ematch(pared8) biasadj(math8 read8 faminc8b)

Treatment-effects estimation                   Number of obs      =      5,671
Estimator      : nearest-neighbor matching     Matches: requested =          5
Outcome model  : matching                                  min =          5
Distance metric: Mahalanobis                               max =          5
------------------------------------------------------------------------------
             |               AI robust
      hsgrad | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
    catholic |
(yes vs no)  |   .0300248   .0085213     3.52   0.000     .0133234    .0467262
------------------------------------------------------------------------------

.         teffects nnmatch (inpse math8 read8 faminc8b pared8) (catholic), atet ///
>                 nneighbor(5) ematch(pared8) biasadj(math8 read8 faminc8b)

Treatment-effects estimation                   Number of obs      =      5,671
Estimator      : nearest-neighbor matching     Matches: requested =          5
Outcome model  : matching                                  min =          5
Distance metric: Mahalanobis                               max =          5
------------------------------------------------------------------------------
             |               AI robust
       inpse | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
    catholic |
(yes vs no)  |   .0688116    .016185     4.25   0.000     .0370895    .1005337
------------------------------------------------------------------------------

.         tebalance summarize
(refitting the model using the generate() option)

Covariate balance summary

                          Raw      Matched
-----------------------------------------
Number of obs =         5,671        1,184
Treated obs   =           592          592
Control obs   =         5,079          592
-----------------------------------------

-------------------------------------------------------------------
                |Standardized differences          Variance ratio
                |        Raw     Matched           Raw    Matched
----------------+--------------------------------------------------
          math8 |   .2606657   -.0075643      .8200688   1.030665
          read8 |   .4356571   -.0005527      .9223486   1.054308
       faminc8b |   .4572743   -.0050568      .8885848    1.02276
         pared8 |   .1701967           0      .8080069          1
-------------------------------------------------------------------
```

```
.
.         // The only difference here is number of neighbors (5) and the bias
.         // adjustment. The latter applies only to the ATT estimate. The balance
.         // here appears slightly less good as compared to Q10--due to the request
.         // for more neighbors--but the change is small. The standard errors for
.         // the ATTs are lower. This is as expected, since the sample size has
.         // increased with the number of neighbors.
.
.
. // ************************************
. // Question 12 - assumption
. // ************************************
.
.         // In matching estimators, the key assumption for causal inference is the
.         // conditional independence assumption. That is, conditional on X (the
.         // variables on which we matched) treatment assignment (here, Catholic HS)
.         // and potential outcomes are independent. It seems unlikely to hold in
.         // this case. Even if the covariates are well-balanced in the two samples
.         // being compared, there are likely *unobserved* covariates that are
.         // related to selection into Catholic HS *and* outcomes like graduation
.         // and post-secondary enrollment.
.
.
. // Close log and convert to PDF
. log close
      name:  <unnamed>
       log:  C:\Users\corcorsp\Dropbox\_TEACHING\Regression II\Problem sets\Problem s
> et 2 - Matching and weighting 1\PS2.txt
  log type:  text
 closed on:   9 Sep 2025, 13:44:23
-------------------------------------------------------------------------------------
```


Balance plot
Raw
Matched
80
70
60
50
40
30
8th grade standardized mathematics score
no
yes
Balance plot
Raw
Matched
70
60
50
40
30
8th grade standardized reading score
no
yes
Balance plot
Raw
Matched
60,000
40,000
20,000
0
family income in 8th grade (dollars)
no
yes


Balance plot
Raw
Matched
Density
.04
.03
.02
.01
0
20
40
60
80
8th grade standardized mathematics score
no
yes

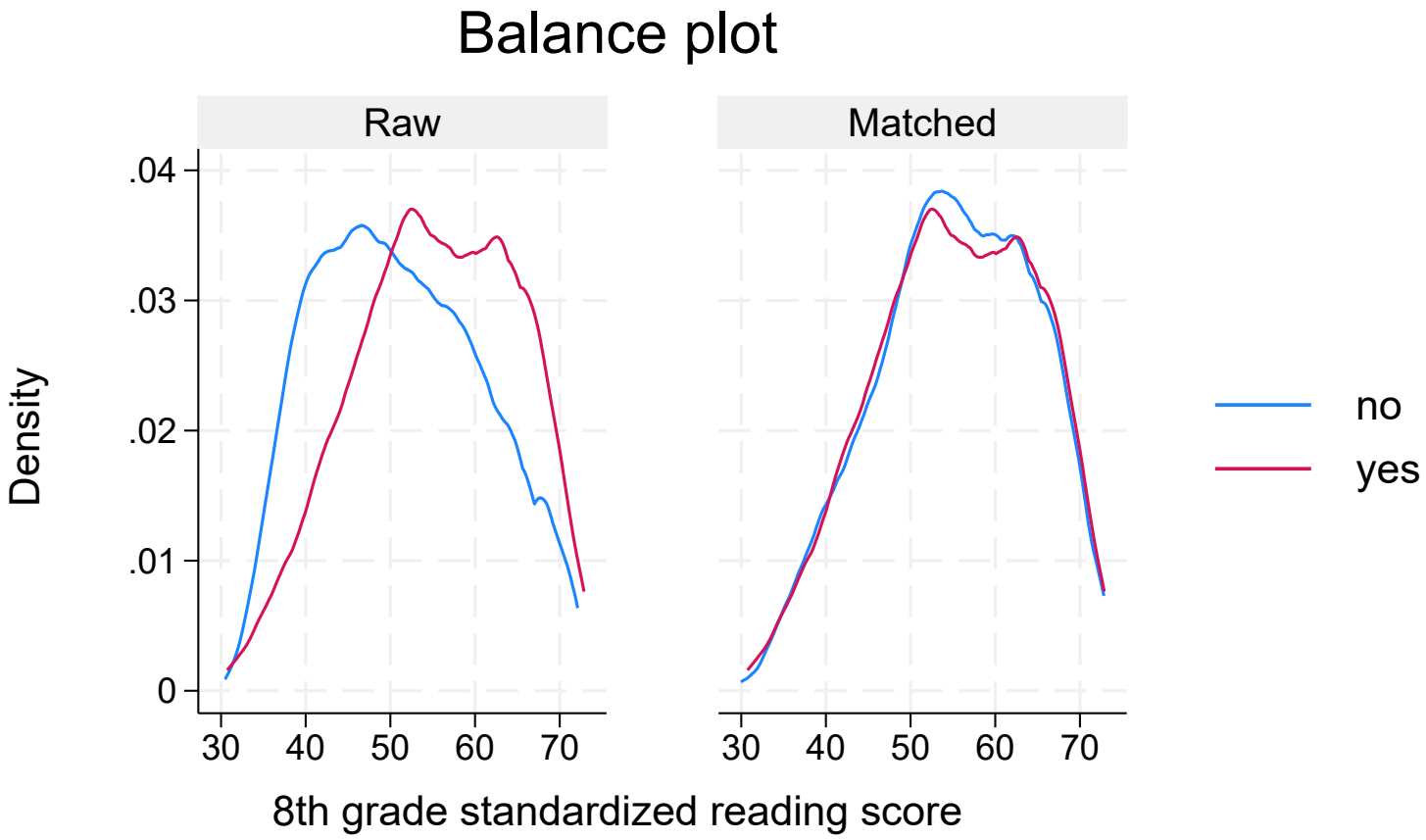

Balance plot
Raw
Matched
Density
.04
.03
.02
.01
0
30
40
50
60
70
8th grade standardized reading score
no
yes

Table 12.1 Descriptive statistics on annual family income, by stratum, overall and by type of high school attended, and average twelfth-grade mathematics achievement by income stratum and by high-school type (n = 5,671)

Stratum		Average Base-Year Annual Family Income *(1988 dollars, 15-point ordinal scale)*			Cell Frequencies		Average Mathematics Achievement (12th grade)		
Label	Income Range	Sample Variance	Sample Mean		Public	Catholic (% of stratum total)	Public	Catholic	Diff.
			Public	Catholic					
Hi_Inc	$35,000 to $74,999	0.24	11.38	11.42	1,969	344 (*14.87%*)	53.60	55.72	2.12***,†
Med_ Inc	$20,000 to $34,999	0.22	9.65	9.73	1,745	177 (*9.21%*)	50.34	53.86	3.52***,†
Lo_Inc	≤$19,999	3.06	6.33	6.77	1,365	71 (*4.94%*)	46.77	50.54	3.76***,†
							Weighted Average ATE		***3.01***
							Weighted Average ATT		***2.74***

~p <0.10; *p <0.05; **p <0.01; ***p <0.001
†One-sided test.

Table 12.2 Sample frequencies and average twelfth-grade mathematics achievement, by high-school type, within 12 strata defined by the crossing of stratified versions of base-year annual family income and mathematics achievement (n = 5,671)

Stratum		Cell Frequencies		Average Mathematics Achievement (12th Grade)		
Base-Year Family Income	Base-Year Mathematics Achievement	Public	Catholic	Public	Catholic	Diff.
Hi_Inc	*Hi_Ach*	1,159	227	58.93	59.66	0.72
	MHi_Ach	432	73	49.18	50.71	1.53*,†
	MLo_Ach	321	38	42.75	44.23	1.48
	Lo_Ach	57	6	39.79	40.40	0.62
Med_Inc	*Hi_Ach*	790	93	57.42	59.42	2.00**,†
	MHi_Ach	469	49	47.95	50.14	2.19**,†
	MLo_Ach	390	33	41.92	44.56	2.64*,†
	Lo_Ach	96	2	37.94	39.77	1.83
Lo_Inc	*Hi_Ach*	405	36	56.12	56.59	0.47
	MHi_Ach	385	13	47.12	48.65	1.53
	MLo_Ach	433	21	40.99	41.70	0.71
	Lo_Ach	142	1	36.81	42.57	5.76
				Weighted Average ATE		***1.50***
				Weighted Average ATT		***1.31***

~p <0.10; *p <0.05; **p <0.01; ***p <0.001
†One-sided test.