Propensity Analysis in Stata
Revision: 1.1

Mark Lunt

October 14, 2014

Contents

List of Figures

List of Listings

1. Introduction

Propensity scores can be very useful in the analysis of observational studies. They enable us to balance a large number of covariates between two groups (referred to as exposed and unexposed in this tutorial) by balancing a single variable, the propensity score. There are three ways to use the propensity score to do this balancing: matching, stratification and weighting. We will explore all three ways in this tutorial.

Propensity models depend on the potential outcomes model popularized by Don Rubin[1]. In this model, we assume every subject has two potential outcomes: one if they were treated, the other if they are not treated. The aim is to compare treated subjects to untreated subjects *with the same potential outcomes*: this ensures that the difference between treated and untreated subjects is due to the treatment, since the outcomes in both groups would have been the same had the treated subjects not received treatment. Rosenbaum and Rubin [2] have shown that subjects with the same propensity score have, on average, the same potential outcomes, so comparing treated and untreated subjects with the same propensity score gives an unbiased estimate of the effect of treatment.

1.1. The data used in this tutorial

We will use simulated data for this tutorial, since that way we can know what the correct answer is, and compare the results we get with different methods with the correct answer. The outcome we are interested in is the variable `y`, which is normally distributed. The treatment variable, `t`, has the effect of reducing `y` by 1. However, there are three confounding variables, `x1`, `x2` and `x3`: an increase in any of these variables increases the probability of receiving treatment, and also increases the outcome `y`. So you can think of `y` as being a measure of disease severity, with those with the highest disease severity being more likely to receive treatment. This data can be loaded into stata with the commands

```
global datadir http://personalpages.manchester.ac.uk/staff/mark.lunt
use "$datadir/pg_example.dta"
```

1.2. Additional programs required

I have written some ado-files which make analysis with propensity scores a little easier, and which we will use throughout this tutorial. They can be downloaded by entering the following command in stata:

```
net from http://personalpages.manchester.ac.uk/staff/mark.lunt
```

then clicking on "`propensity`" and finally clicking on "`click here to install`". We will also use the `pbalchk` command which can be installed in the same way.

2. Checking Balance

Before we start analysing the data, it will be useful to see how big a problem we have. We will therefore compare all of the confounders between the treated and untreated. One way to do this is with the `tabstat` command: Listing 1 shows how we can get the mean and standard deviation for each variable in the treated and untreated.

Listing 1 Mean and standard deviation of confounders in treated and untreated subjects

```
. tabstat x*, by(t) statistics(mean sd) columns(statistics)

Summary for variables: x1 x2 x3
     by categories of: t

       t |      mean        sd
---------+--------------------
       0 | -.4172519  .8412321
         |  6.118413  1.690448
         |  39.39428  7.560781
---------+--------------------
       1 |  .4777602   .943132
         |  7.992944  1.845908
         |  45.37631  11.72876
---------+--------------------
   Total |  .0087739  .9968548
         |   7.01069  1.998652
         |  42.24173  10.21349
------------------------------
```

We can see that there is a difference of about 1 in `x1`, 2 in `x2` and 6 in `x3`. However, since we don't know the units in which these variables are measured in, we don't know if the difference in `x3` is more important than the difference in `x1` or not. We could do a significance test, but that is very sample-size dependent, and does not tell us how big any differences between treated and untreated are. We are better looking at "standardised differences": the difference in terms of standard deviations.

We can get this data easily from the `pbalchk` program, the syntax of which is

`pbalchk` *`treatvar testvars`*

where *`treatvar`* is the treatment variable (in our case `t`) and *`testvars`* are the potential confounders, in our case `x1`, `x2` and `x3`. The result from `pbalchk` is shown in Listing 2.

Listing 2 Checking balance of confounders between treated and untreated

```
. pbalchk t x1 x2 x3

              Mean in treated   Mean in Untreated   Standardised diff.
------------------------------------------------------------------------
          x1 |            0.48               -0.42                -0.949
          x2 |            7.99                6.12                -1.016
          x3 |           45.38               39.39                -0.510
------------------------------------------------------------------------


Warning: Significant imbalance exists in the following variables:
x1 x2 x3
```

The output shows us that the treated and untreated differ by about 1 SD in `x1` and `x2`, and by 0.5 SD in `x3`. So the treated and untreated are more similar in `x3` than they are in `x1` or `x2`.

3. Calculating Propensity Scores

3.1. Using Logistic Regression

We use logistic regression to calculate the propensity scores. The stata commands to do this are

```
logistic t x1 x2 x3
predict propensity
```

We can now look at the distributions of the propensity score in the treated and the untreated with the command

```
graph tw kdensity propensity if t == 0 || ///
         kdensity propensity if t == 1
```

The output of this command is shown in Figure 1. You can see that propensity scores tend to be higher in the treated than the untreated, but because of the limits of 0 and 1 on the propensity score, both distributions are skewed.

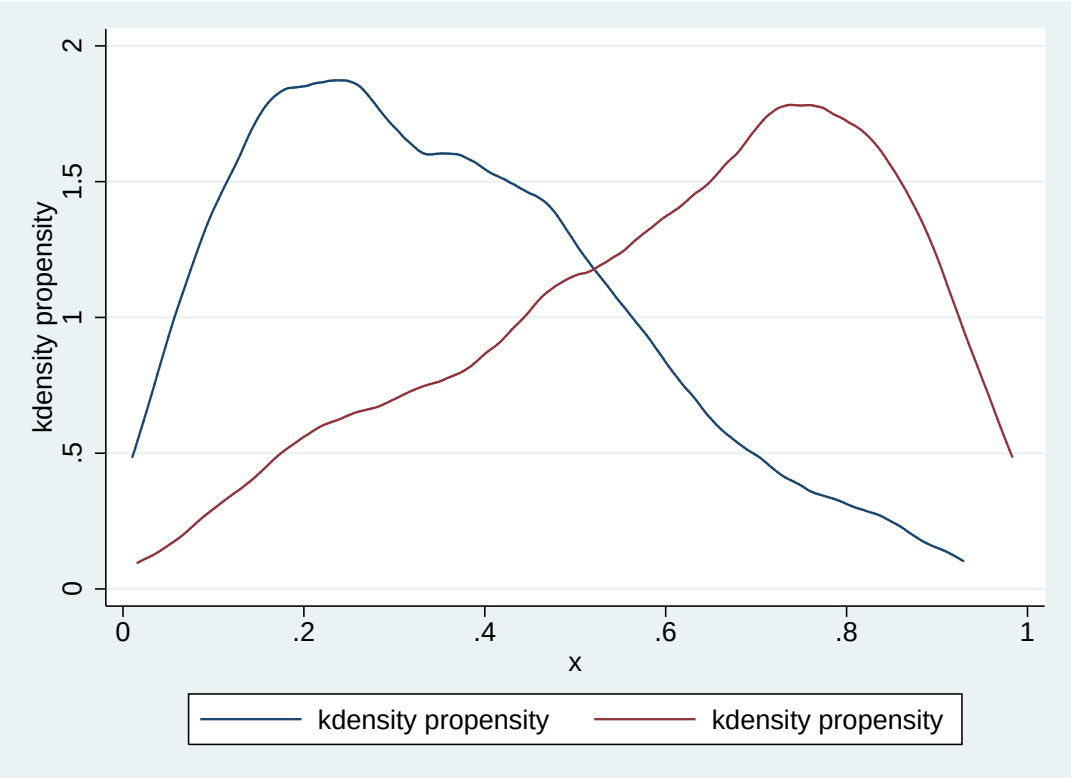


Figure 1: Distributions of Propensity Score

For this reason, it is sometimes recommended to work with the log of the odds of the propensity score (sometimes called the linear predictor), rather than the propensity score itself, since it tends to be more normally distributed. We can obtain and graph this with the commands

```
predict lp, xb
graph tw kdensity lp if t == 0 || kdensity lp if t == 1
```

The result is shown in Figure 2: a much more normal distribution in both subgroups.

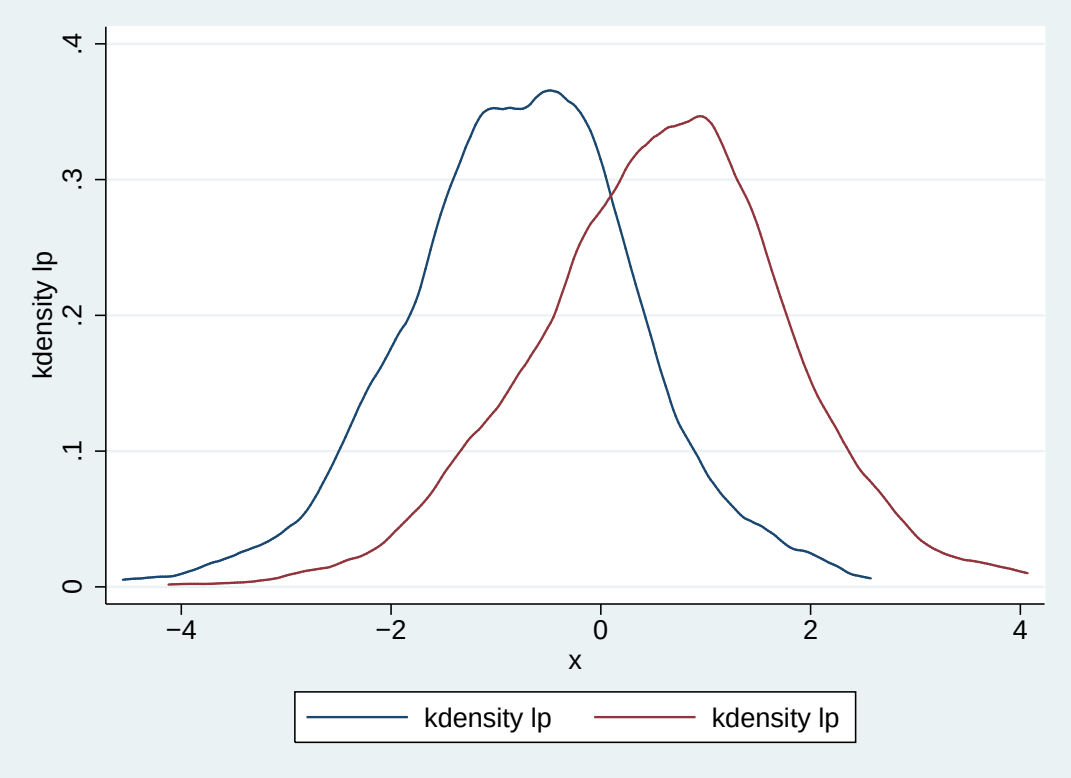


Figure 2: Distributions of Log Odds of Propensity Score

3.2. Diagnostics for the propensity score

Having calculated a propensity score, we need to check that it is correct. If there are important confounders that we have not measured, the propensity score will not work, and there is no way really of testing that. However, if we have got the functional form of our regression equation wrong, a Hosmer-Lemeshow test will show that. Listing 3 shows the command to do this and the resultant output.

Listing 3 Hosmer-Lemeshow test on initial propensity model

```
. estat gof, group(10) table

Logistic model for t, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.1397 |    26 |  17.5 |   174 | 182.5 |   200 |
  |     2 | 0.2243 |    43 |  37.1 |   157 | 162.9 |   200 |
  |     3 | 0.3021 |    51 |  51.9 |   149 | 148.1 |   200 |
  |     4 | 0.3890 |    58 |  68.7 |   142 | 131.3 |   200 |
  |     5 | 0.4650 |    69 |  85.0 |   131 | 115.0 |   200 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.5436 |    98 | 101.2 |   102 |  98.8 |   200 |
  |     7 | 0.6417 |   110 | 118.3 |    90 |  81.7 |   200 |
  |     8 | 0.7336 |   153 | 137.4 |    47 |  62.6 |   200 |
  |     9 | 0.8277 |   168 | 156.2 |    32 |  43.8 |   200 |
  |    10 | 0.9832 |   176 | 178.5 |    24 |  21.5 |   200 |
  +--------------------------------------------------------+

      number of observations =      2000
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =        25.14
                Prob > chi2 =         0.0015
```

You can see that the test is significant, showing that the logistic regression model does not fit our data well. This suggests that either there is a non-linearity in the relationships between the confounders and the log odds of being treated, or there is an interaction between two of the confounders. We can find out which by generating and testing all 3 squared terms (`x1*x1`, `x2*x2` and `x3*x3`) and all 3 interaction terms (`x1*x2`, `x1*x3`, `x2*x3`). Some code to achieve this with not too much typing is given in Listing 4

Listing 4 Testing interactions in propensity model

```
foreach var of varlist x1 x2 x3 {
  foreach var2 of varlist x1 x2 x3 {
    capture drop temp
    gen temp = `var' * `var2'
    logit t x1 x2 x3 temp
    di "Testing `var' * `var2'"
    estat gof, table group(10)
  }
}
```

If you run this code, you will see that the best fit is achieved when `x3*x3` is added to the model (Listing 5):

Listing 5 Goodness of fit of improved propensity model

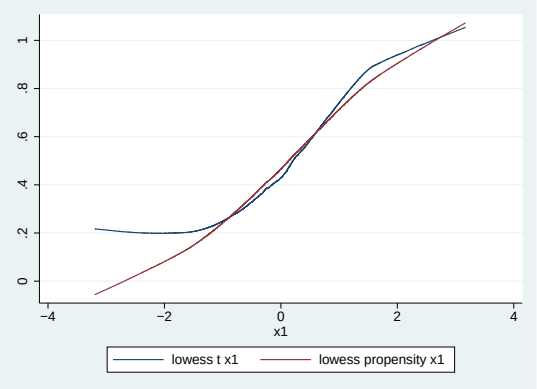
```
. gen x32 = x3 * x3

. logit t x1 x2 x3 x32

Logistic regression                                     Number of obs   =       2000
                                                        LR chi2(4)      =     877.82
                                                        Prob > chi2     =     0.0000
Log likelihood = -945.07838                             Pseudo R2       =     0.3171

------------------------------------------------------------------------------
           t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .9236019   .1041029     8.87   0.000     .7195641     1.12764
          x2 |   .4648093   .0440131    10.56   0.000     .3785452    .5510734
          x3 |  -.8005345   .0544132   -14.71   0.000    -.9071825   -.6938866
         x32 |   .0095544    .000661    14.46   0.000      .008259    .0108498
       _cons |     12.597   1.088999    11.57   0.000      10.4626     14.7314
------------------------------------------------------------------------------

. estat gof, table group(10)

Logistic model for t, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0968 |    12 |  11.4 |   188 | 188.6 |   200 |
  |     2 | 0.1659 |    25 |  26.1 |   175 | 173.9 |   200 |
  |     3 | 0.2414 |    41 |  40.6 |   159 | 159.4 |   200 |
  |     4 | 0.3316 |    59 |  56.9 |   141 | 143.1 |   200 |
  |     5 | 0.4316 |    76 |  75.7 |   124 | 124.3 |   200 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.5519 |    91 |  98.1 |   109 | 101.9 |   200 |
  |     7 | 0.6777 |   128 | 122.4 |    72 |  77.6 |   200 |
  |     8 | 0.8153 |   148 | 148.6 |    52 |  51.4 |   200 |
  |     9 | 0.9438 |   176 | 175.9 |    24 |  24.1 |   200 |
  |    10 | 1.0000 |   196 | 196.3 |     4 |   3.7 |   200 |
  +--------------------------------------------------------+

       number of observations =      2000
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =      1.90
                Prob > chi2 =      0.9839
```

The fit of this model is now very good, so we will save the propensity score and linear predictor from this model to use later.

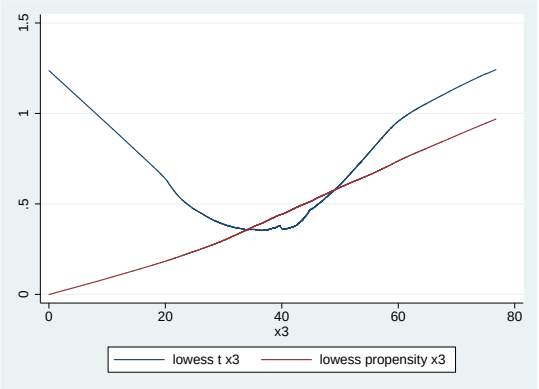
```
predict prop2
predict lp2, xb
```

There are (at least) two other approaches we could have used to determine which variable was causing the poor fit in the propensity model. One is to look at the observed and expected proportion of subjects treated at each level of the confounder. An easy way to do this is to use lowess smoothed plots: the lines corresponding to the observed and predicted proportions should be parallel if we have got the model correct. The commands to produce these plots for `x1` and `x3` are

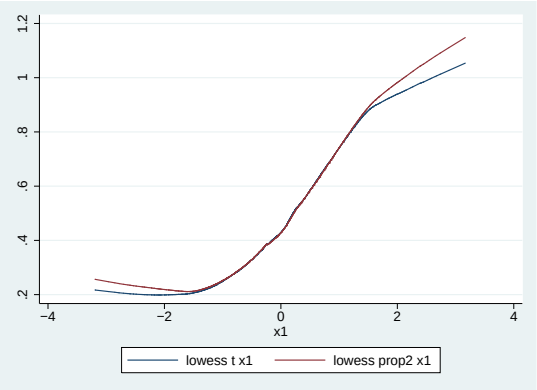
```
gr tw lowess t x1 || lowess propensity x1
```



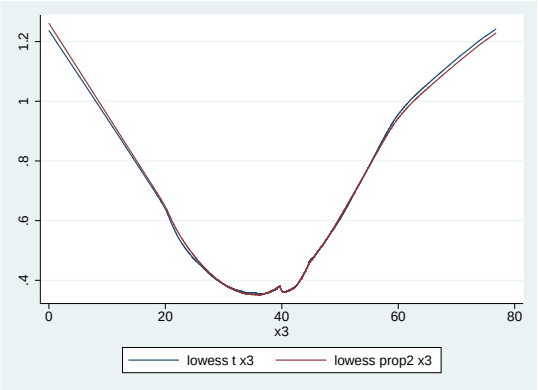
(a) `x1`, Initial propensity model



(b) `x3`, Initial propensity model



(c) `x1`, Improved propensity model



(d) `x3`, Improved propensity model

Figure 3: Comparison of observed and predicted associations between confounders and treatment

```
gr tw lowess t x3 || lowess propensity x3
gr tw lowess t x1 || lowess prop2 x1
gr tw lowess t x3 || lowess prop2 x3
```

and the plots themselves are given in Figure 3.

It is clear that we had the wrong functional form for `x3` in our initial propensity score, but that it was correct in the second one (`prop2`).

The other alternative is to fit different functions of the confounders in the propensity model and see if the fit of the model improves. You can either generate your own variables to do this:

```
gen x12 = x1^2
gen x13 = x1^3
```

to generate the square and cube of `x1`, or use stata's `fracpoly` command, which fits various

functions *of the first variable in the list of predictors* to test if a non-linear association is present. The results of using `fracpoly` to test `x1` and `x3` are given in Listing 6

Although fitting a non-linear function of `x1` improved the fit of the model, this was due to its correlation with `x3`: allowing `x3` to have a non-linear association produced a dramatically better improvement in the fit of the model.

3.3. Using the propensity score

We mentioned above that there are three ways to use the propensity score: matching, stratification and weighting.

3.3.1. Stratification

The simplest method is stratification: we divide our subjects into strata based on the propensity score, and look at the effect of treatment within strata. Listing 7 shows how to generate quintiles of the propensity score, and gives a cross-tabulation of treatment by quintile.

As you can see, there are some treated and some untreated subjects in every quintile of the propensity score, so it is possible to assess the effect of treatment in each quintile. In the lowest quintile, over 90% of subjects do not receive treatment, whilst in the highest quintile, over 90% *do* receive treatment.

Quintiles are commonly used for adjustment, since they are expected to remove 90% of the confounding [3]. However, the smaller the strata are, the better they will balance the covariates and the more confounding they will remove. To illustrate this, we will also created deciles of the propensity score (Listing 8):

Listing 6 Using `fracpoly` to test for non-linearity

```
. fracpoly logit t x1 x2 x3

Logistic regression                               Number of obs   =       2000
                                                  LR chi2(4)      =     604.09
                                                  Prob > chi2     =     0.0000
Log likelihood = -1081.9432                       Pseudo R2       =     0.2182

------------------------------------------------------------------------------
           t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Ix1__1 |  -1.285783   .3309939    -3.88   0.000    -1.934519   -.6370464
      Ix1__2 |   .3267525   .0537353     6.08   0.000     .2214332    .4320718
      Ix2__1 |   .4126267   .0403556    10.22   0.000     .3335312    .4917222
      Ix3__1 |   -.015576   .0072658    -2.14   0.032    -.0298167   -.0013354
       _cons |  -.3419402   .0633793    -5.40   0.000    -.4661613   -.2177191
------------------------------------------------------------------------------
Deviance:  2163.89. Best powers of x1 among 44 models fit: 1 2.

. estat gof, group(10)

Logistic model for t, goodness-of-fit test

       number of observations =      2000
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =        15.72
                  Prob > chi2 =         0.0466

. fracpoly logit t x3 x2 x1

Logistic regression                               Number of obs   =       2000
                                                  LR chi2(4)      =     879.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -944.14098                       Pseudo R2       =     0.3178

------------------------------------------------------------------------------
           t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Ix3__1 |  -4.433401   .3000949   -14.77   0.000    -5.021576   -3.845226
      Ix3__2 |   .0816554   .0056721    14.40   0.000     .0705382    .0927725
      Ix2__1 |   .4624106   .0439802    10.51   0.000      .376211    .5486102
      Ix1__1 |   .9306171   .1042095     8.93   0.000     .7263703    1.134864
       _cons |  -.9380883   .0760635   -12.33   0.000     -1.08717   -.7890067
------------------------------------------------------------------------------
Deviance:  1888.28. Best powers of x3 among 44 models fit: 1 3.

. estat gof, group(10)

Logistic model for t, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =      2000
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         2.47
                  Prob > chi2 =         0.9632
```

Listing 7 Stratifying into quintiles of propensity score

```
. xtile pq = lp2, n(5)

. tab pq t, ro

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

         5 |
 quantiles |           t
    of lp2 |         0          1 |     Total
-----------+----------------------+----------
         1 |       363         37 |       400
           |     90.75       9.25 |    100.00
-----------+----------------------+----------
         2 |       300        100 |       400
           |     75.00      25.00 |    100.00
-----------+----------------------+----------
         3 |       233        167 |       400
           |     58.25      41.75 |    100.00
-----------+----------------------+----------
         4 |       124        276 |       400
           |     31.00      69.00 |    100.00
-----------+----------------------+----------
         5 |        28        372 |       400
           |      7.00      93.00 |    100.00
-----------+----------------------+----------
     Total |     1,048        952 |     2,000
           |     52.40      47.60 |    100.00
```

Listing 8 Stratifying into deciles of propensity score

```
. xtile p_d = lp2, n(10)

. tab p_d t, ro
```

10 quantiles of lp2	t 0	1	Total
1	188	12	200
	94.00	6.00	100.00
2	175	25	200
	87.50	12.50	100.00
3	159	41	200
	79.50	20.50	100.00
4	141	59	200
	70.50	29.50	100.00
5	124	76	200
	62.00	38.00	100.00
6	109	91	200
	54.50	45.50	100.00
7	72	128	200
	36.00	64.00	100.00
8	52	148	200
	26.00	74.00	100.00
9	24	176	200
	12.00	88.00	100.00
10	4	196	200
	2.00	98.00	100.00
Total	1,048	952	2,000
	52.40	47.60	100.00

We still have both treated and untreated subjects in every stratum, so we can use this stratification in our analysis.

3.3.2. Weighting

The second way to use the propensity score is to reweight the data. I will spare you the gory details, I will just say that by reweighting, we ensure that the distribution of confounders is the same in the treated and untreated subjects, so they are no longer confounders. Sato and Matsuyama [4] have written a good, comprehensible introduction to how weighting works if you are interested.

In practice, we commonly use two kinds of weights: inverse probability of treatment (IPT) weights and SMR weights. IPT weights change the distribution of confounders in both the treated and untreated subjects so that they are the same as the distribution in the entire sample. The IPT weighted analysis therefore compares what we would expect to see if everyone received treatment to what we would expect to see if no-one received treatment. SMR weights

do not change the distribution in the treated, and change the distribution in the untreated to match it. They therefore compare what happened to the treated subjects with what would have happened to them if they had remained untreated. If we assume that treatment has the same effect on everyone (as in this simulated data), these two analyses are estimating the same thing. However, if we assume that subjects who will benefit more from treatment are more likely to be given it, a very reasonable assumption, the SMR effect will be greated than the IPT effect.

The program `propwt` can be used to create both IPT and SMR weights. For full details of the syntax, type

```
help propwt
```

into stata, but for now, simply enter the command

```
propwt t prop2, ipt smr
```

This will create two new variables, `smr_wt` and `ipt_wt` which we will use later.

3.3.3. Matching

There are a huge number of ways of performing matching, which I am not going to discuss here: Rosenbaum has written some accessible papers on the subject ([5, 6]). I am simply going to show one method of obtaining matches.

We are going to use greedy matching. That is, we will compare every treated subject to every untreated subject, and find the closest match we can. These subjects will be paired off, then we will compare the remaining subjects and pick the best match. This procedure will continue until there are no more possible pairings. The program `gmatch` will do this for us, using the command

```
gmatch t lp2, set(set1) diff(diff1)
```

This stores a case-control pair identifier in `set1`, and the magnitude of the difference between the case and the control in `diff1`.

However, there is a problem here, as we can see if we look at the distribution of the differences within each set (Figure 4). At first, we had a lot of very good matches, but at a certain point, the matches became very poor. Since we insisted on finding a match for every treated subject, and we don't have enough suitable matches, we end up accepting unsuitable matches.

We can improve on this matching by insisting that matched pairs cannot differ by more than a fixed amount referred to as a *caliper*. This will limit the number of matches we can make, but ensure that the matches are good. Zooming in on the left hand end of the histogram in Figure 4 with the command

```
histogram diff1 if diff1 < 0.2
```

suggests that a caliper of 0.1 would be appropriate, so lets use that:

```
gmatch t lp2, set(set2) diff(diff2) cal(0.1)
```

The differences are now all relatively small, as shown in Figure 5.

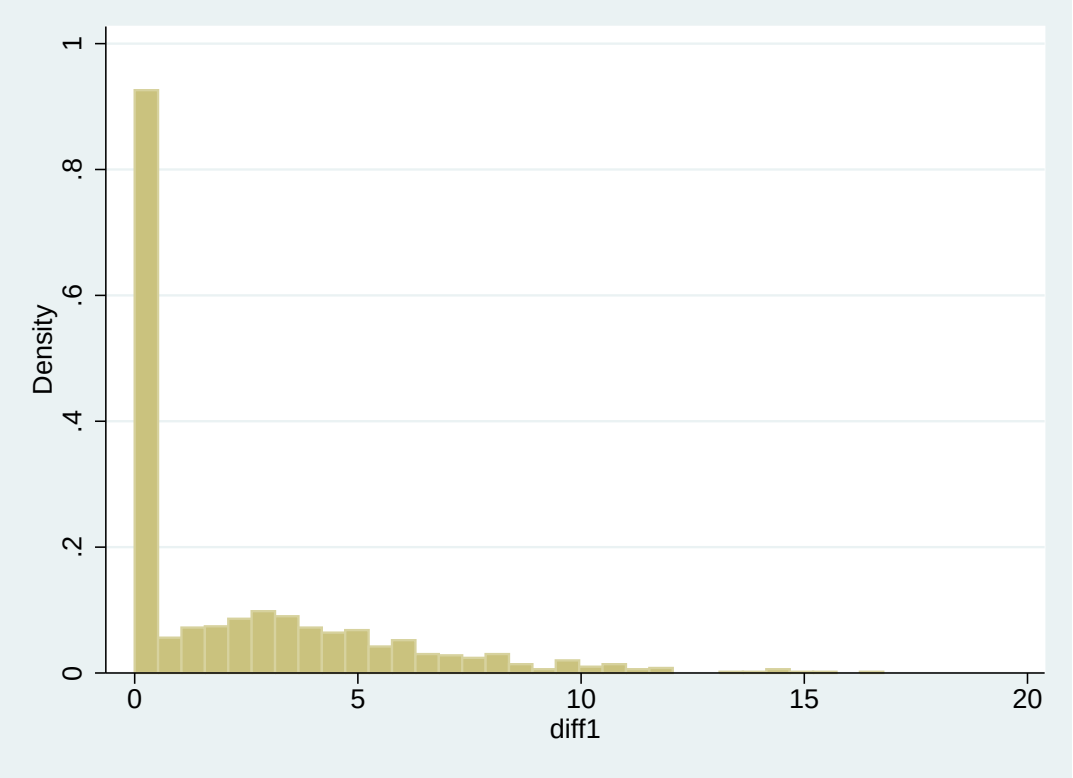


Figure 4: Distribution of case-control differences after naïve matching

4. Rechecking Balance

4.1. Stratification

We can check the balance after stratification by giving the `strata` option to `pbalchk`, as shown in Listing 9.

Listing 9 Checking balance of confounders between treated and untreated after stratifying into quintiles

```
. pbalchk t x1 x2 x3, strata(pq)

i.pq              _Spq_1-5            (naturally coded; _Spq_1 omitted)
               Mean in treated   Mean in Untreated   Standardised diff.
------------------------------------------------------------------------
          x1 |            0.48                0.41               -0.069
          x2 |            7.99                7.87               -0.067
          x3 |           45.38               44.80               -0.049
------------------------------------------------------------------------
```

The differences are now much smaller than before: all less than 0.1 SD. This suggests that the balancing has been successful. However, we do get better balance if we use 10 strata rather than 5, as seen in Listing 10.

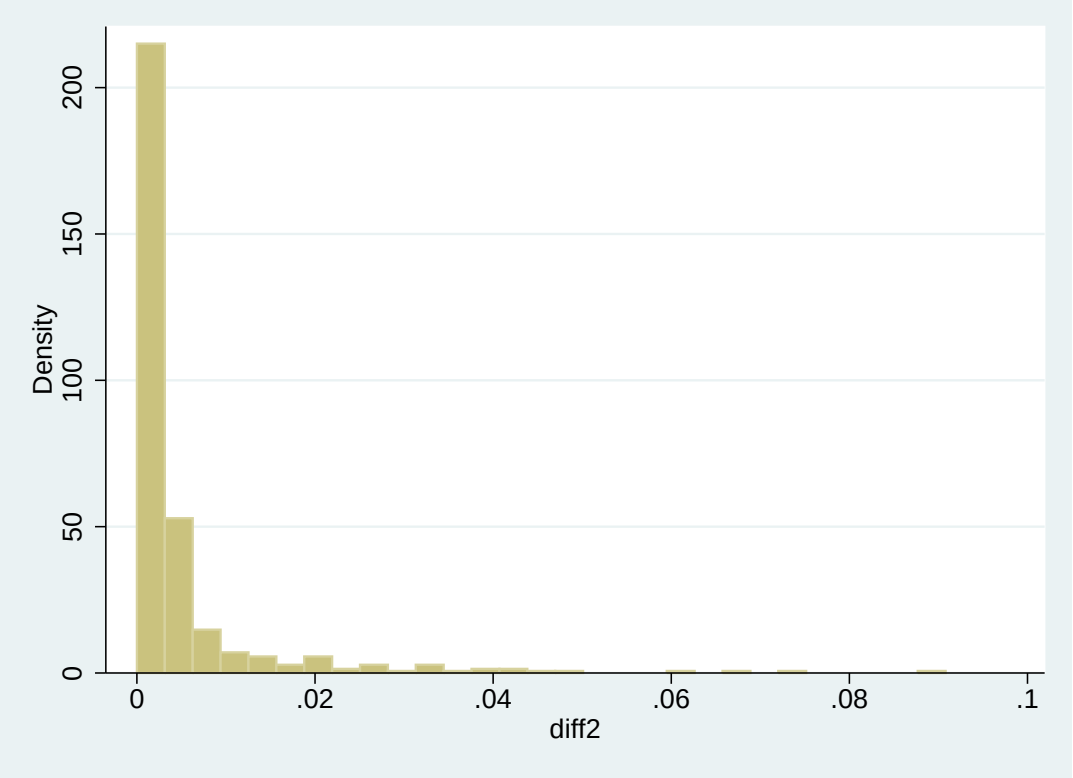


Figure 5: Distribution of case-control differences after matching with a caliper of 0.1

Listing 10 Checking balance of confounders between treated and untreated after stratifying into deciles

```
. pbalchk t x1 x2 x3, strata(p_d)

i.p_d             _Sp_d_1-10          (naturally coded; _Sp_d_1 omitted)
               Mean in treated   Mean in Untreated   Standardised diff.
------------------------------------------------------------------------
          x1 |            0.48                0.46                -0.020
          x2 |            7.99                7.96                -0.017
          x3 |           45.38               45.20                -0.015
------------------------------------------------------------------------
```

The balance is now markedly better on all three variables.

4.2. Weighting

To check the balance between treated and untreated after weighting, we use the `wt()` option to `pbalchk`. Both sets of weights also improve the balance of all three covariates markedly: they have done better than quintiles but not as well as deciles (Listing 11).

Listing 11 Checking balance of confounders between treated and untreated after weighting

```
. pbalchk t x1 x2 x3, wt(ipt_wt)

               Mean in treated   Mean in Untreated   Standardised diff.
------------------------------------------------------------------------
          x1 |            0.01                0.02                0.005
          x2 |            7.08                7.02               -0.034
          x3 |           42.25               41.81               -0.037
------------------------------------------------------------------------


. pbalchk t x1 x2 x3, wt(smr_wt)

               Mean in treated   Mean in Untreated   Standardised diff.
------------------------------------------------------------------------
          x1 |            0.48                0.52                0.044
          x2 |            7.99                8.06                0.038
          x3 |           45.38               44.59               -0.067
------------------------------------------------------------------------
```

Note that whilst the distribution of all three confounders is similar in the treated and untreated subjects with both weighting schemes, the actual means differ between the weightings: x2 is about 7 with IPT weights and about 8 with SMR weights, for example.

4.3. Matching

As we saw earlier, our first attempt at matching produced some very poor matches, so we would not expect that to do a good job of balancing the confounders. This is shown to be true in Listing 12 (note that only subjects who were matched will have a non-missing value for `set1`, so unmatched subjects are excluded but the clause `if set1 < .`).

Listing 12 Checking balance of confounders between treated and untreated after naïve matching

```
. pbalchk t x1 x2 x3 if set1 < .

               Mean in treated   Mean in Untreated   Standardised diff.
------------------------------------------------------------------------
          x1 |            0.48               -0.30               -0.829
          x2 |            7.99                6.36               -0.884
          x3 |           45.38               39.64               -0.489
------------------------------------------------------------------------


Warning: Significant imbalance exists in the following variables:
x1 x2 x3
```

However, the restricted matching did balance the confounders much better (Listing 13):

Listing 13 Checking balance of confounders between treated and untreated after matching within caliper

```
. pbalchk t x1 x2 x3 if set2 != .

              Mean in treated   Mean in Untreated   Standardised diff.
------------------------------------------------------------------------
          x1 |            0.01              -0.01               -0.025
          x2 |            7.06               7.03               -0.020
          x3 |           40.86              40.63               -0.027
------------------------------------------------------------------------
```

5. Assessing the effect of treatment

5.1. Naïvely

If we were to compare the outcome between the two treatment groups without allowing for the presence of confounding, we would get a biased estimate of the effect of treatment. This analysis is given in Listing 14.

Listing 14 Naïve analysis of the effect of treatment

```
. regress y t

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  1,  1998) =  140.71
       Model |  172.114155     1  172.114155           Prob > F      =  0.0000
    Residual |   2443.8897  1998  1.22316802           R-squared     =  0.0658
-------------+------------------------------           Adj R-squared =  0.0653
       Total |  2616.00386  1999  1.30865626           Root MSE      =   1.106

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.5873868   .0495175   -11.86   0.000    -.6844982   -.4902754
       _cons |   .0448315   .0341635     1.31   0.190    -.0221684    .1118313
------------------------------------------------------------------------------
```

The observed effect of treatment (-0.59) is markedly less than the true effect (-1), and indeed the true effect does not lie in the 95% confidence interval.

5.2. Stratification

In order to look at the effect of treatment within strata of the propensity score, we add indicator variables for the strata to the regression equation, as shown in Listing 15

Listing 15 Analysis of the effect of treatment, stratifying by propensity score in 5 strata

```
. xi: regress y t i.pq
i.pq              _Ipq_1-5            (naturally coded; _Ipq_1 omitted)

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  5,  1994) =   58.17
       Model |   333.00699     5   66.601398           Prob > F      =  0.0000
    Residual |  2282.99687  1994  1.14493323           R-squared     =  0.1273
-------------+------------------------------           Adj R-squared =  0.1251
       Total |  2616.00386  1999  1.30865626           Root MSE      =    1.07

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.9521784    .060055   -15.86   0.000    -1.069955   -.8344013
      _Ipq_2 |  -.1253259   .0762505    -1.64   0.100    -.2748648     .024213
      _Ipq_3 |   .0110136   .0781384     0.14   0.888    -.1422279    .1642552
      _Ipq_4 |   .2584805   .0837391     3.09   0.002     .0942551    .4227059
      _Ipq_5 |   .8192614   .0908535     9.02   0.000     .6410836    .9974391
       _cons |   .0257863   .0537884     0.48   0.632     -.079701    .1312737
------------------------------------------------------------------------------
```

The estimated treatment effect (-0.95) is now much closer to the true value (-1). Most of the confounding was removed by stratification, but not all. Received wisdom is that 5 strata will remove about 90% of confounding, and that seems to be the case here. Adjusting using 10 strata is more successful, as we would expect from the fact that it balanced the covariates better, as seen in Listing 16

Listing 16 Analysis of the effect of treatment, stratifying by propensity score in 10 strata

```
. xi: regress y t i.p_d
i.p_d             _Ip_d_1-10          (naturally coded; _Ip_d_1 omitted)

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F( 10,  1989) =   40.74
       Model |  444.719762    10  44.4719762           Prob > F      =  0.0000
    Residual |  2171.28409  1989   1.0916461           R-squared     =  0.1700
-------------+------------------------------           Adj R-squared =  0.1658
       Total |  2616.00386  1999  1.30865626           Root MSE      =  1.0448

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.9921445   .0589945   -16.82   0.000    -1.107842    -.876447
     _Ip_d_2 |  -.1188487   .1045522    -1.14   0.256     -.323892    .0861947
     _Ip_d_3 |  -.1933323   .1048315    -1.84   0.065    -.3989233    .0122587
     _Ip_d_4 |  -.1635789   .1053976    -1.55   0.121    -.3702803    .0431224
     _Ip_d_5 |  -.1454589   .1061737    -1.37   0.171    -.3536822    .0627644
     _Ip_d_6 |   .0746155    .107049     0.70   0.486    -.1353244    .2845554
     _Ip_d_7 |    .155955    .109942     1.42   0.156    -.0596586    .3715686
     _Ip_d_8 |   .2899168   .1119186     2.59   0.010     .0704268    .5094068
     _Ip_d_9 |   .2832083   .1151375     2.46   0.014     .0574055    .5090111
    _Ip_d_10 |   1.303409    .117738    11.07   0.000     1.072506    1.534312
       _cons |   .0889075   .0739646     1.20   0.229    -.0561487    .2339637
------------------------------------------------------------------------------
```

The estimated treatment effect, -0.99, is now almost exactly equal to the true value of -1.

5.3. Weighting

In order to use the IPT and SMR weights in an analysis, we need to specify them as part of the stata command. The syntax is to add `[pw=wt_var]` to the command before any options, where *wt_var* is the name of the variable to use for the weighting. The results of using both IPT and SMR weights are given in Listing 17

Listing 17 Analysis of the effect of treatment, using weighting

```
. xi: regress y t [pw=ipt_wt]
(sum of wgt is   1.9748e+03)

Linear regression                                      Number of obs =    2000
                                                       F(  1,  1998) =  149.07
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1382
                                                       Root MSE      =   1.072

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.8589639   .0703525   -12.21   0.000    -.9969359   -.7209919
       _cons |   .1351195   .0539146     2.51   0.012     .0293848    .2408541
------------------------------------------------------------------------------

. xi: regress y t [pw=smr_wt]
(sum of wgt is   9.3002e+02)

Linear regression                                      Number of obs =    2000
                                                       F(  1,  1998) =   52.00
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1124
                                                       Root MSE      =  1.0984

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.7816477   .1083953    -7.21   0.000    -.9942274   -.5690681
       _cons |   .2390924   .1016333     2.35   0.019     .0397741    .4384107
------------------------------------------------------------------------------
```

Although markedly better than the naïve estimate, these estimates are not as good as those obtained through stratifying. One reason for this can be seen in Figure 2. There are treated subjects who have higher propensities than any untreated subject, and hence we have no suitable comparison subjects for them. Equally, there are untreated subjects with lower propensities than any treated subject (but not as many). Strictly, when doing propensity analysis, we should restrict our analysis to subjects on the "common support", that is restricted to the range of propensity scores at which we observe both treated and untreated subjects. Listing 18 shows how to calculate the common support and restrict the analysis to it.

Listing 18 Analysis of the effect of treatment, using weighting, restricted to common support

```
. bys t: summ prop2

-------------------------------------------------------------------------------
-> t = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       prop2 |      1048    .2978684    .2170018   .0037738   .9868041

-------------------------------------------------------------------------------
-> t = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       prop2 |       952    .6720945    .2659372   .0280815   .9999992


. xi: regress y t [pw=ipt_wt] if prop2 <= 0.9868041 & prop2 >= 0.0280815
(sum of wgt is   1.9136e+03)

Linear regression                                      Number of obs =    1875
                                                       F(  1,  1873) =  166.85
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1630
                                                       Root MSE      =  1.0489

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.9269146   .0717594   -12.92   0.000    -1.067651   -.7861777
       _cons |   .1307708   .0545776     2.40   0.017     .0237315    .2378101
------------------------------------------------------------------------------

. xi: regress y t [pw=smr_wt] if prop2 <= 0.9868041 & prop2 >= 0.0280815
(sum of wgt is   8.8311e+02)

Linear regression                                      Number of obs =    1875
                                                       F(  1,  1873) =   72.27
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1582
                                                       Root MSE      =  1.0591

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.9206137    .108293    -8.50   0.000    -1.133001    -.708226
       _cons |   .2389587   .1016888     2.35   0.019     .0395234    .4383941
------------------------------------------------------------------------------
```

The largest propensity score in the untreated was 0.9868041 and the smallest in the treated was 0.0280815. Restricting our analysis to subjects within this range means that we lost 125 subjects, but the effect estimates are less biased.

5.4. Matching

We would not expect the initial matching we did to perform very well at removing bias, and this is borne out by Listing 19

Listing 19 Analysis of the effect of treatment, using naïve matching

```
. regress y t if set1 != .

      Source |       SS       df       MS              Number of obs =    1904
-------------+------------------------------           F(  1,  1902) =  122.70
       Model |  151.086553     1  151.086553           Prob > F      =  0.0000
    Residual |  2341.96921  1902  1.23131925           R-squared     =  0.0606
-------------+------------------------------           Adj R-squared =  0.0601
       Total |  2493.05577  1903  1.31006609           Root MSE      =  1.1096

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.5633904   .0508606   -11.08   0.000    -.6631389   -.4636419
       _cons |   .0208351   .0359639     0.58   0.562    -.0496978    .0913679
------------------------------------------------------------------------------
```

This is almost exactly the same as the naïve analysis. However, the restricted matching fares much better (Listing 20):

Listing 20 Analysis of the effect of treatment, using matching with a caliper

```
. regress y t if set2 != .

      Source |       SS       df       MS              Number of obs =     906
-------------+------------------------------           F(  1,   904) =  218.31
       Model |  240.813657     1  240.813657           Prob > F      =  0.0000
    Residual |  997.184884   904  1.10308062           R-squared     =  0.1945
-------------+------------------------------           Adj R-squared =  0.1936
       Total |  1237.99854   905  1.36795419           Root MSE      =  1.0503

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -1.031113   .0697862   -14.78   0.000    -1.168075   -.8941516
       _cons |   .1340728   .0493463     2.72   0.007     .0372261    .2309194
------------------------------------------------------------------------------
```

The estimated treatment effect is very close to the true value here. However, the analysis we have done is simply compared 453 treated subjects to 453 comparable untreated subjects: we have in fact ignored the matching. If we wish to use the fact that we have matched data, we need to use `xtreg` for the analysis. The matched analysis should not give a different estimate, but it should give a smaller standard error (Listing 21):

Listing 21 Matched analysis of the effect of treatment, using matching with caliper

```
. xtreg y t, i(set2) fe

Fixed-effects (within) regression               Number of obs      =       906
Group variable (i): set2                        Number of groups   =       453

R-sq:  within  = 0.3367                         Obs per group: min =         2
       between = 0.0000                                        avg =       2.0
       overall = 0.1945                                        max =         2

                                                F(1,452)           =    229.49
corr(u_i, Xb)  = 0.0000                         Prob > F           =    0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -1.031113    .068065   -15.15   0.000    -1.164877   -.8973503
       _cons |   .1340728   .0481292     2.79   0.006     .0394879    .2286576
-------------+----------------------------------------------------------------
     sigma_u |  .76053338
     sigma_e |  1.0243726
         rho |  .35534384   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(452, 452) =     1.10            Prob > F = 0.1501
```

The effect estimate is exactly the same, as expected, and the standard error is very slightly smaller, but the difference would not affect our inference from the result.

6. Trimming

We have seen that failing to exclude subjects outside the common support can lead to biased estimates. Even within the common support, there may be highly atypical subjects (untreated with very high propensity scores, or treated with very low propensity scores). This may be a sign of unmeasured confounding (there is a reason why these subjects were not treated as we would have expected, but we have not recorded it in our potential confounders), which can lead to biased estimates. It is therefore sometimes recommended to trim subjects with particularly large or small propensity scores[7].

One common way to do this is to calculate the x^{th} centile of the propensity score in the treated and the $100 - x^{th}$ centile in the untreated, and remove all subjects outside these limits. This can lead to far more than $x\%$ of subjects being removed, since there are likely to be more untreated than treated at the lower end of the propensity range and more treated than untreated at the higher end. In this scheme, limiting the analysis to the common support is equivalent to trimming at the 0^{th} centile.

There is a program `proptrim` which will create variables identifying subjects to be included in the analysis after trimming at various centiles. The basic command

```
proptrim t prop2
```

will create the variables `keep_0`, `keep_1` and `keep_5`, which take the value 1 for subjects to be included in the analysis and 0 for those to be excluded when trimming at the 0^{th}, 1^{st} and 5^{th} centile respectively, although other centiles can be requested.

The effect of trimming at the 5^{th} centile on the IPT weighted estimate is showing in Listing 22.

Listing 22 Analysis of the effect of treatment, using weighting, trimmed at the fifth centile

```
. xi: regress y t [pw=ipt_wt] if keep_5
(sum of wgt is   1.0301e+03)

Linear regression                                      Number of obs =    1027
                                                       F(  1,  1025) =  191.70
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1789
                                                       Root MSE      =  1.0495

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.9800994   .0707877   -13.85   0.000    -1.119005    -.841194
       _cons |   .0559337   .0461471     1.21   0.226    -.0346199    .1464873
------------------------------------------------------------------------------
```

Two things to note:

1. The estimate of the treatment effect is now very much closer to the true value than in the untrimmed analysis.

2. The standard error has not changed appreciably, despite using only about half as many subjects as the untrimmed analysis. This is because the variance of the weights has been reduced (subjects with particularly high weights have been excluded).

So the gain in precision has not been at the expense of a loss of efficiency.

If we do a stratified analysis on the 5% trimmed data, we get a very similar result (Listing 23):

Listing 23 Analysis of the effect of treatment, using weighting, trimmed at the fifth centile

```
. xi: regress y t i.p_d if keep_5
i.p_d             _Ip_d_1-10          (naturally coded; _Ip_d_1 omitted)

      Source |       SS       df       MS              Number of obs =    1027
-------------+------------------------------           F(  6,  1020) =   34.63
       Model |  222.136794     6  37.0227991           Prob > F      =  0.0000
    Residual |  1090.41913  1020  1.06903837           R-squared     =  0.1692
-------------+------------------------------           Adj R-squared =  0.1644
       Total |  1312.55593  1026  1.27929428           Root MSE      =  1.0339

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.9816952   .0685561   -14.32   0.000    -1.116222   -.8471681
     _Ip_d_2 |  (dropped)
     _Ip_d_3 |  -.5581856   .1448573    -3.85   0.000     -.842438   -.2739332
     _Ip_d_4 |  -.4285173   .1358399    -3.15   0.002    -.6950748   -.1619598
     _Ip_d_5 |  -.4112855   .1348454    -3.05   0.002    -.6758915   -.1466795
     _Ip_d_6 |  -.1919948   .1341721    -1.43   0.153    -.4552796    .0712901
     _Ip_d_7 |  -.1125884   .1333471    -0.84   0.399    -.3742543    .1490775
     _Ip_d_8 |  (dropped)
     _Ip_d_9 |  (dropped)
    _Ip_d_10 |  (dropped)
       _cons |   .3507634   .1206996     2.91   0.004     .1139154    .5876113
------------------------------------------------------------------------------
```

The thing to note this time is that we have lost all of the lowest decile, along with the 3 highest deciles. Nonetheless, the estimate is similar to that we saw without trimming, although in this case there has been an increase in the standard error.

7. Alternative Analyses

So far, we have only seen how propensity scores can be used with the `regress` command. In fact, they can be used with many other commands as well, but not all commands can be used with all methods.

The good news is that the simplest method, stratification, works with any command. The thing to remember with this method is the more strata the better, provided you have at least one case and one control in every stratum.

Weighting can in principle be used with any regression command, but in practice not every command in stata will accept `pweights`, the name stata gives to the type of weight we need to use. You can find out if the command accepts `pweights` by typing `help` *`cmd`*, where *`cmd`* is the name of the command you want to use.

Matching needs to be used carefully. Matched data should normally be analysed using a method that takes into account the matching. As we have seen, with linear regression it makes no difference to the estimate, but it may reduce the standard error. However, with other forms of regression, an unmatched analysis might give an incorrect answer.

8. Conclusions

Stratification can work well provide there are enough strata. This requires large sample sizes: our sample of 2,000 could only just support deciles. Using a small number of strata will lead to residual confounding. Stratification is also the easiest method to use, so is recommended whenever the sample size is adequate.

With weighting, you have to be careful about which subjects you include. Trimming to the common support is essential, but trimming more may give added benefit.

The thing to watch for with matching is the caliper that you select for the matches. This needs to be sufficiently small that the estimate in the matched data is unbiased, but sufficiently large that enough subjects are included in the analysis to give and efficient estimate.

References

[1] Rubin DB Using propensity scores to help design observational studies: Application to the tobacco litigation Health Services & Outcomes Research Methodology Dec 2001; 2:169–188.

[2] Rosenbaum PR, Rubin DB The central role of the propensity score in observational studies for causal effects Biometrika 1983;70:41–55.

[3] Cochran WG The effectiveness of adjustment by subclassification in removing bias in observational studies Biometrics 1968;24:295–313.

[4] Sato T, Matsuyama Y Marginal structural models as a tool for standardization Epidemiology 2003;14:680–686.

[5] Rosenbaum PR Optimal matching for observational studies Journal of the American Statistical Association 1989;84:1024–1302.

[6] Rosenbaum PR, Rubin DB Constructing a control group using multivariate matched sampling methods that incorporate the propensity score The American Statistician 1985; 39:33–38.

[7] Cole SR, Hernán MA Constructing inverse probability weights for marginal structural models American Journal of Epidemiology 2008;168:656–664.

A. Complete do file for tutorial

```
set more off

// Getting data
global datadir http://personalpages.manchester.ac.uk/staff/mark.lunt
use "$datadir/pg_example.dta"

// Checking confounders
tabstat x*, by(t) statistics(mean sd) columns(statistics)
pbalchk t x1 x2 x3

// Intial propensity score
logistic t x1 x2 x3
predict propensity

graph tw kdensity propensity if t == 0 || ///
         kdensity propensity if t == 1

predict lp, xb
graph tw kdensity lp if t == 0 || kdensity lp if t == 1

estat gof, group(10) table

foreach var of varlist x1 x2 x3 {
  foreach var2 of varlist x1 x2 x3 {
    capture drop temp
    gen temp = `var' * `var2'
    logit t x1 x2 x3 temp
    di "Testing `var' * `var2'"
    estat gof, table group(10)
  }
}

// Improved propensity score
gen x32 = x3 * x3
logit t x1 x2 x3 x32
estat gof, table group(10)
predict prop2
predict lp2, xb

gr tw lowess t x1 || lowess propensity x1
gr tw lowess t x3 || lowess propensity x3
gr tw lowess t x1 || lowess prop2 x1
gr tw lowess t x3 || lowess prop2 x3

fracpoly logit t x1 x2 x3
estat gof, group(10)
fracpoly logit t x3 x2 x1
estat gof, group(10)

// Creating strata
```

```
xtile pq = lp2, n(5)
tab pq t, ro
xtile p_d = lp2, n(10)
tab p_d t, ro

// Creating weights
propwt t prop2, ipt smr

// Creating matched sets
gmatch t lp2, set(set1) diff(diff1)
histogram diff1 if diff1 < 0.2
gmatch t lp2, set(set2) diff(diff2) cal(0.1)
histogram diff2 if set2 < .

// Rechecking Balance
pbalchk t x1 x2 x3, strata(pq)
pbalchk t x1 x2 x3, strata(p_d)
pbalchk t x1 x2 x3, wt(ipt_wt)
pbalchk t x1 x2 x3, wt(smr_wt)
pbalchk t x1 x2 x3 if set1 < .
pbalchk t x1 x2 x3 if set2 != .

// Estimating treatment effect
regress y t

xi: regress y t i.pq
xi: regress y t i.p_d

regress y t [pw=ipt_wt]
regress y t [pw=smr_wt]
bys t: summ prop2
regress y t [pw=ipt_wt] if prop2 <= 0.9868041 & prop2 >= 0.0280815
regress y t [pw=smr_wt] if prop2 <= 0.9868041 & prop2 >= 0.0280815

regress y t if set1 != .
regress y t if set2 != .
xtreg y t, i(set2) fe

// Trimming
proptrim t prop2
regress y t [pw=ipt_wt] if keep_5
xi: regress y t i.p_d if keep_5
```

B. Do file used to generate example dataset

```
set obs 2000
gen x1 = invnorm(uniform())
gen x2 = sqrt(0.5)*x1 + sqrt(0.5) * invnorm(uniform())
gen x3 = sqrt(0.5)*x1 + sqrt(0.5) * invnorm(uniform())
gen l = x1 + x2 + x3^2
replace l = l - 1
```

```
gen p = exp(l) / (1 + exp(l))
gen t = uniform() < p
gen y = (x1^2 + x2 + x3) /4  - t + invnorm(uniform())
replace x2 = 2*x2 + 7
replace x3 = (x3 + 4.22 ) * 10
drop p
drop l
save P:/home/teaching/propensity_guide/pg_example.dta, replace
saveold P:/public_html/pg_example.dta, replace
```