Propensity Score Matching

Version 1.0 (February 2019)

Felix Bittmann (mail@statabook.com)

Table of Contents

1 Introduction

This paper will give a short introduction to applied propensity score matching (PSM). Main concepts about Stata, data handling and the foundations of causal analysis won't be discussed here, so refer to the book for more information. The focus lies on the direct application using Stata 15.

The main research question of this paper is to estimate the (causal) effect of smoking (of the mother) on the birth weight of her child. We will use freely available data on the internet to answer that question. Based on theoretical considerations we assume that smoking has negative effects on children and will result in a reduced birthweight. Usually, this argumentation should be grounded in a review of the literature. We can visualize our hypothesis like so:

Smoking → Birthweight

The arrow indicates that smoking during pregnancy should affect birthweight causally. However, to account for spurious correlations, we must consider all factors that simultaneously influence the notion to smoke and the birth weight. For example, education is such a factor as lowly educated women might be less informed about the negative consequences of smoking but will also have children with lower birthweight as they probably know less about adequate behaviour during pregnancy, which might result in a lower probability to visit prenatal screenings and checks. Therefore, if we do not account for the influence of education, we probably cannot identify the causal effect of smoking.

2 Description

Every causal analysis must start with a thorough description of the main variables of interest as a foundation. This is useful to inspect the basic relations between variables and identify problems and errors in the data. We will use tables and graphics for description. As a first step, we download the dataset directly into Stata.

```
webuse cattaneo2.dta, clear
```

As usual, it is recommended using a do-file to save all commands which are shown here. This makes documentation comfortable and allows for easy and fast correction of errors. A complete do-file can be found with this paper. The first command here downloads the dataset from the internet[1], the option *clear* will delete any other data loaded in Stata right now. We now inspect all variables contained in the data:

```
describe
```

1 http://www.stata-press.com/data/r13/cattaneo2.dta (2019-01-23).

```
Contains data from http://www.stata-press.com/data/r15/cattaneo2.dta
  obs:         4,642                          Excerpt from Cattaneo (2010) Journal of Econometrics 155: 138-154
 vars:            23                          14 Jan 2016 09:49
 size:       116,050
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
bweight         int     %9.0g                 infant birthweight (grams)
mmarried        byte    %10.0g     mmarried   1 if mother married
mhisp           byte    %9.0g                 1 if mother hispanic
fhisp           byte    %9.0g                 1 if father hispanic
foreign         byte    %9.0g                 1 if mother born abroad
alcohol         byte    %9.0g                 1 if alcohol consumed during pregnancy
deadkids        byte    %9.0g                 previous births where newborn died
mage            byte    %9.0g                 mother's age
medu            byte    %9.0g                 mother's education attainment
fage            byte    %9.0g                 father's age
fedu            byte    %9.0g                 father's education attainment
nprenatal       byte    %9.0g                 number of prenatal care visits
monthslb        int     %9.0g                 months since last birth
order           byte    %9.0g                 order of birth of the infant
msmoke          byte    %27.0g     smoke2     cigarettes smoked during pregnancy
mbsmoke         byte    %9.0g      mbsmoke    1 if mother smoked
mrace           byte    %9.0g                 1 if mother is white
frace           byte    %9.0g                 1 if father is white
prenatal        byte    %9.0g                 trimester of first prenatal care visit
birthmonth      byte    %9.0g                 month of birth
lbweight        byte    %9.0g                 1 if low birthweight baby
fbaby           byte    %9.0g      YesNo      1 if first baby
prenatal1       byte    %9.0g      YesNo      1 if first prenatal visit in 1 trimester
-------------------------------------------------------------------------------
Sorted by:
```

We learn that there are 23 variables and 4,642 observations (cases) in the dataset. The variable label gives us a first impression of the information contained. We locate the central dependent variable (bweight), which is the variable we want to explain. The central independent (explanatory) variable (mbsmoke) is a binary one that indicates whether a woman smokes or not. We should now select potential further explanatory variables that are important to estimate the causal effect of smoking. As explained before, the education of the mother seems central (medu). The following table will display all selected control variables. Note that this selection should be justified by the theoretical framework and grounded in previous empirical research (if available).

Construct	Variable Name	Reason
Education	medu	Uneducated women might smoke more and care less about general health recommendations
Alcohol consumption	alcohol	Drinking women might also smoke more and alcohol affects the fetus negatively
Marriage	mmarried	Married women might smoke less but also have more social support to care for them and their health during pregnancy
Previous dead children	deadkids	Women with previous births that resulted in the death of the child might smoke more due to depression and also have a worse general health
Age	mage	Younger women might smoke more but are also fitter for pregnancy than older women
Ethnic	mrace	Ethnic and birthweight might be correlated but also ethnic and smoking, possible due to different cultural views on drugs and smoking
Number of prenatal care visits	nprenatal	Women with a high awareness of health might have more check ups as they care for the child but also smoke less as they are aware of the negative consequences of smoking

Note that this selection is not written in stone but rather ad-hoc. There might be more good reasons to include more or different variables. However, the overall workflow of the method will not be different.

Next, we want to inspect all the variables that we want to use in the analysis. We use the *fre* command, so make sure that you have this installed.

```
ssc install fre, replace
fre mbsmoke bweight medu alcohol mmarried deadkids mage mrace nprenatal
```

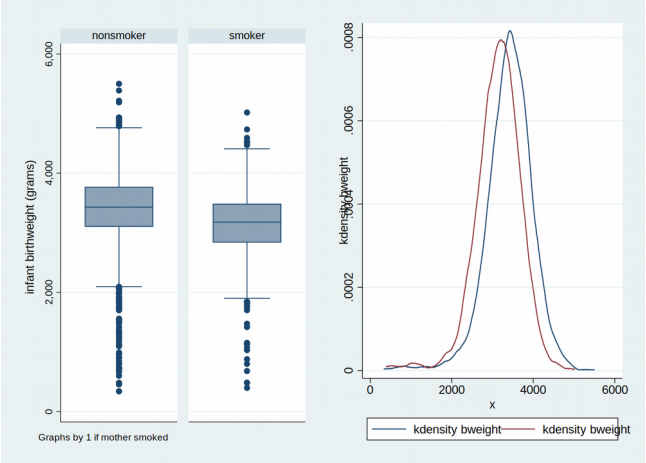
You will receive a quite long list of output that is not shown here. This gives you a first impression of the data and makes coding errors obvious (for example, when there are negative numbers included, which are commonly used to indicate missing values). Yet it seems, everything is fine here. To get more information about the metric (continuous) variables, we use the *summarize* command.

```
summarize bweight medu mage nprenatal
```

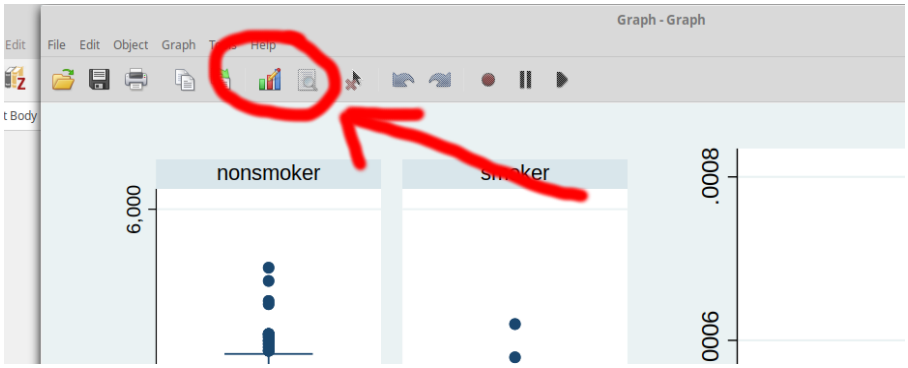
Variable	Obs	Mean	Std. Dev.	Min	Max
bweight	4,642	3361.68	578.8196	340	5500
medu	4,642	12.68957	2.520661	0	17
mage	4,642	26.50452	5.619026	13	45
nprenatal	4,642	10.75808	3.681084	0	40

Next, we want to have some information about the basic relations within the data. How are smoking and birthweight related? We start with two graphical ways using kernel-density plots and boxplots. We will combine the two in one final graph, which is better for publication.

```
graph box bweight, by(mbsmoke) name(boxplot, replace)
twoway (kdensity bweight if mbsmoke == 0) ///
    (kdensity bweight if mbsmoke == 1), name(kernel, replace)
graph combine boxplot kernel
```



This gives a clear first impression. Smoking mothers have lighter babies on average as is shown by the thick lines in the boxplots (the median) or by the fact that the weight distribution is shifted to the left in the kernel-density plot (red plot). There are many further options available to make these graphs nicer and more informative. For example, the labelling of the two plots in the kernel-density plot is absolutely not informative. By starting the graph editor you can make these changes.



Another option to test statistically whether the two variables are associated is to use the t-test. This test compares means between groups, which are defined by our binary explanatory variable (smokers VS non-smokers). We run this test by typing

```
ttest bweight, by(mbsmoke)
```

```
Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
nonsmoke |   3,778    3412.912    9.284683    570.6871    3394.708    3431.115
  smoker |     864     3137.66    19.08197    560.8931    3100.207    3175.112
---------+--------------------------------------------------------------------
combined |   4,642    3361.68    8.495534    578.8196    3345.025    3378.335
---------+--------------------------------------------------------------------
    diff |            275.2519    21.4528                233.1942    317.3096
------------------------------------------------------------------------------
    diff = mean(nonsmoke) - mean(smoker)                          t =  12.8306
Ho: diff = 0                     degrees of freedom =     4640

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

We notice that the overall difference is quite large (275.2519). To check whether this is statistically different, we should locate the desired alternative hypothesis. Our null hypothesis is that the group means are not different from each other (*Ho: diff = 0*). We decide to use a one-sided test as we have the strong assumption that smoking affects birthweight negatively. If we did not have this hypothesis and just assumed it could be a positive or negative association, we would use the two-sided test (*Ha: diff != 0*). Therefore, we locate the correct alternative hypothesis (*Ha: diff > 0*), which means: our difference is positive and nonsmokers should have a larger birthweight on average. This interpretation depends on the coding of your smoking variable so be careful and take some time and think to get the logic right. The line *diff = mean(nonsmoke) – mean(smoker)* helps a lot. The result is highly significant and the displayed result is smaller than 0.01 (*Pr(T > t)*).

Consequently, we know that the group means are statistically different from each other and the sampling error cannot be the source of this difference. However, this does not prove that smoking *causes* low birth weight as we did not account for any spurious correlations!

3 Matching

After the description, we can start with the matching. We will use Ben Jann's Stata program *kmatch* which implements a fast and robust form of kernel-matching[2]. I would prefer this over Stata's internal *teffects* (included since version 13). We download and install this by typing

```
ssc install kmatch, replace
```

We can now use it. The syntax for our model is as follows:

```
kmatch ps mbsmoke c.medu i.alcohol i.mmarried i.deadkids ///
   c.mage i.mrace c.nprenatal (bweight)
```

First, we type the name of the command (kmatch) and then *ps* to tell Stata to use the propensity-score algorithm (others are included, see the help files). Then the binary grouping variable is typed (mbsmoke). After that, we type all control variables in arbitrary order. Metric variables get the prefix *c.*, all others (nominal or ordinal) receive the prefix *i.* This is a Stata trick so we do not have to create dummies manually. Then follows in parentheses the dependent variable (bweight). Note that you can also include more than one dependent variable if so desired.

```
(computing bandwidth for treated ... done)
(computing bandwidth for untreated ... done)

Propensity-score kernel matching                Number of obs     =      4,642
                                                Kernel            =      epan
Treatment : mbsmoke = 1
Covariates: medu i.alcohol i.mmarried i.deadkids mage i.mrace nprenatal
PS model  : logit (pr)

Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	827	37	864	3673	105	3778	.00192
Untreated	3597	181	3778	813	51	864	.0013
Combined	4424	218	4642	4486	156	4642	.

```
Treatment-effects estimation
```

bweight	Coef.
ATE	-204.4944

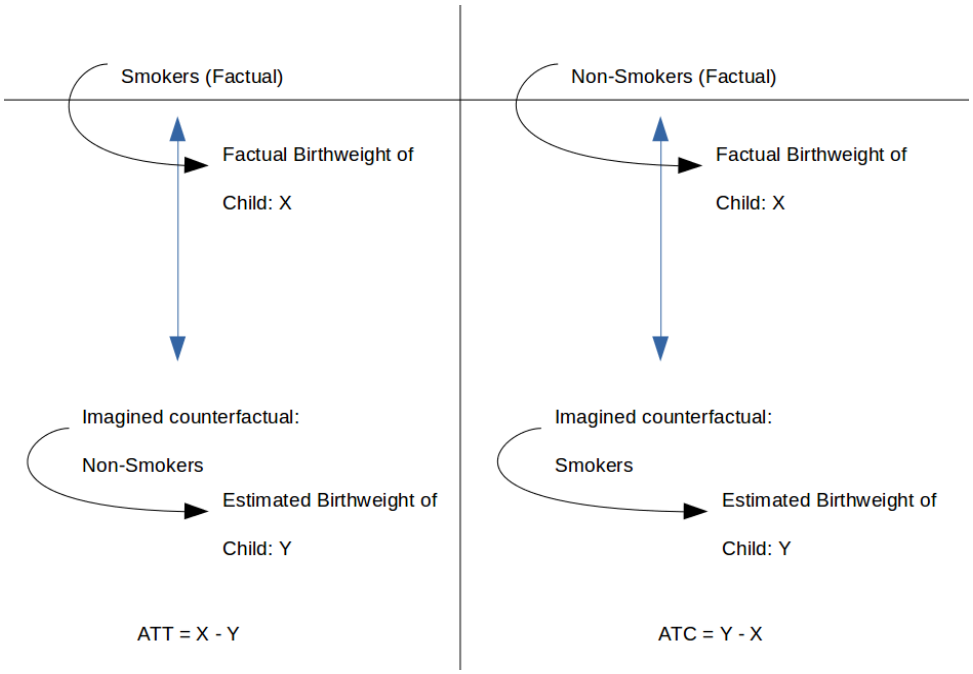
```
.
end of do-file
```

2 https://ideas.repec.org/c/boc/bocode/s458346.html

The ATE is -204. The interpretation is that women who smoke have babies that are, on average, about 204 grams lighter than babies from mothers who do not smoke. To better understand this, we can also receive other common statistics. Type

```
kmatch ps mbsmoke c.medu i.alcohol i.mmarried i.deadkids ///
    c.mage i.mrace c.nprenatal (bweight), ate att atc
```

What does this mean? Starting with the ATC (average treatment effect for the control): there are women in the dataset who are nonsmokers (thus, they are the control group). Suppose, these women would be smokers (the counterfactual, the imagined change in smoking status), then their babies would be -203.8 grams lighter. The other statistic, the ATT, is the average treatment effect for the treated, thus, the women who are smokers in reality. Again, we imagine these women had a counterfactual, so they were non-smokers. We compare the differences in birthweight and see that the difference would be -207.3 grams.



The ATE is the weighted average of these two values. In summary, these results clearly indicate that women who smoke have lighter babies. As we control for certain confounders we could argue that this is a causal effect and smoking is the cause of this difference. But as there are no standard errors, we do not know if this difference is actually statistically significant (albeit large). We can get these values using a bootstrap method, which is a form of resampling. The computer repeatedly draws a random sample of the women and calculates the effects. Then the average is shown. This is a valid and useful statistical technique. We can get the results by first setting a seed (so we all get *exactly* the same numbers as the random number generated in Stata is set to the same state) and then typing the command again with the *reps* option:

```
set seed 12345
```

```
kmatch ps mbsmoke c.medu i.alcohol i.mmarried i.deadkids ///
    c.mage i.mrace c.nprenatal (bweight), att vce(boot, reps(500))
```

This might take a few minutes to run. Usually the higher the number of replications the more stable the results. 500 is a good start and for a final result for publication you could increase this up to 5000 as you only have to do this once.

```
Propensity-score kernel matching                Number of obs     =      4,642
                                                Replications      =        500
                                                Kernel            =      epan
Treatment : mbsmoke = 1
Covariates: medu i.alcohol i.mmarried i.deadkids mage i.mrace nprenatal
PS model  : logit (pr)

Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	**827**	**37**	**864**	**3673**	**105**	**3778**	**.00192**

```
Treatment-effects estimation
```

bweight	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
ATT	**-207.3363**	**27.61005**	**-7.51**	**0.000**	**-261.451**	**-153.2216**

The results indicate that this statistic is highly significant (the p-value is smaller than 0.000). Usually, we would say that any result below 0.05 is significant, which means we can reject the null hypothesis which assumes that our observed coefficient is equal to zero.

Finally, I want to introduce an aspect that can further improve your model, which is exact matching. Imagine we have one variable in our model that we view as central as it highly influences both the outcome as well as the treatment variable. We can specify to match exactly for this variable so only matches are computed that have *preciesly* the same value for this variable. A candidate could be the marriage variable. This means that married women are only compared to other married women and never to women who are not married. This is conceptually a good idea but puts constraints on the model, which can be a problem if your sample or one group of the exact variable is small. This should not be a problem here as about only 30% of women are not married. We specify the command as follows:

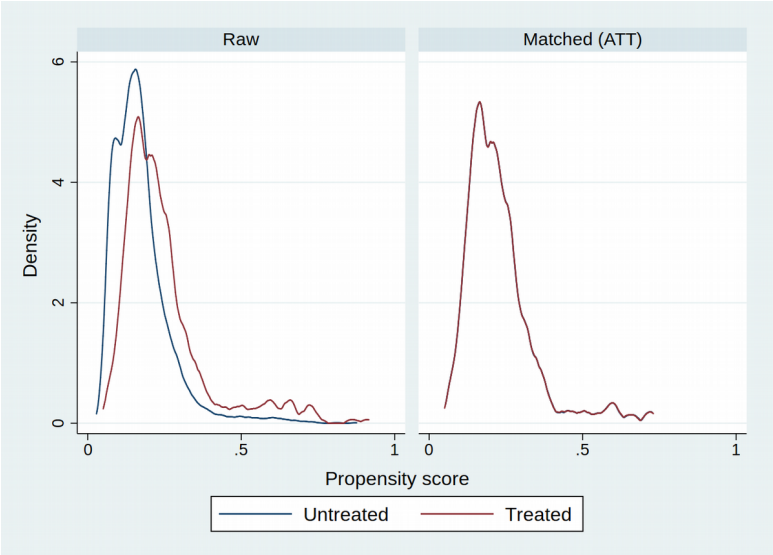
```
kmatch ps mbsmoke c.medu i.alcohol i.deadkids ///
    c.mage i.mrace c.nprenatal (bweight), att vce(boot, reps(500)) ///
    ematch(mmarried)
```

Usually, this variable should only have a few distinct categories. You cannot use factor-variable-notation in the *ematch* option. Note that you then should remove this variable form the other controls. We see that the result changed only slightly (the ATT is now -203).

4 Diagnostics

After the models were estimated it is crucial to check if your model assumptions actually hold. If we ignore these aspects this could lead to serious problems and wrong inferences. The following diagnostics are always based on the last model run. The first assumption is about common support. To get this we type

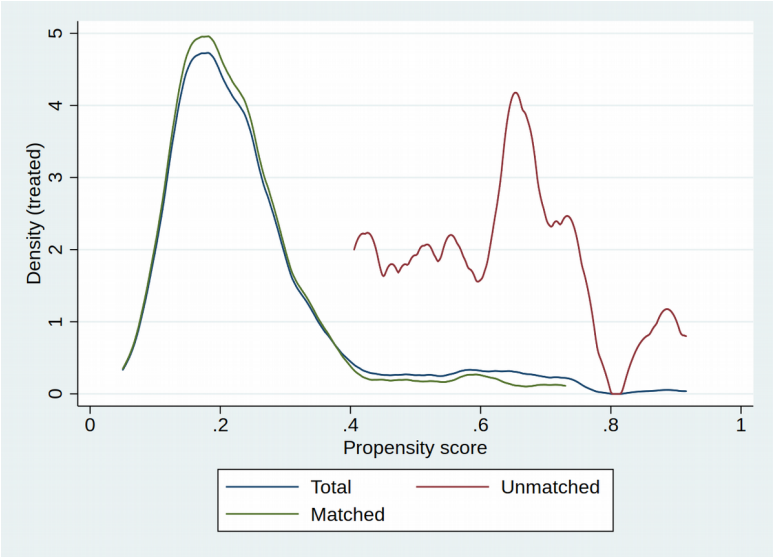
```
kmatch density
```



We see that the matched sample (on the right) as only one line, which is good as there are no large deviations. After matching was applied, the common support is very good, so we know both groups are similar on average. Common support ensures that persons with the same X-values have a positive probability of being both treatment and non-treatment.

The second aspect is about the cases that cannot be matched as no good match could be found. These cases are dropped from the analysis. If too many cases are lost and these cases are somewhat special, you might end up with a sample that is no longer representative of the sample you started with. This is a bad thing so we should check it by typing

```
kmatch cdensity
```



Here we see that the total sample (blue) and the matched sample (green) are very close, so we can assume that the final sample that is used in the matching is very similar on average to your original sample.

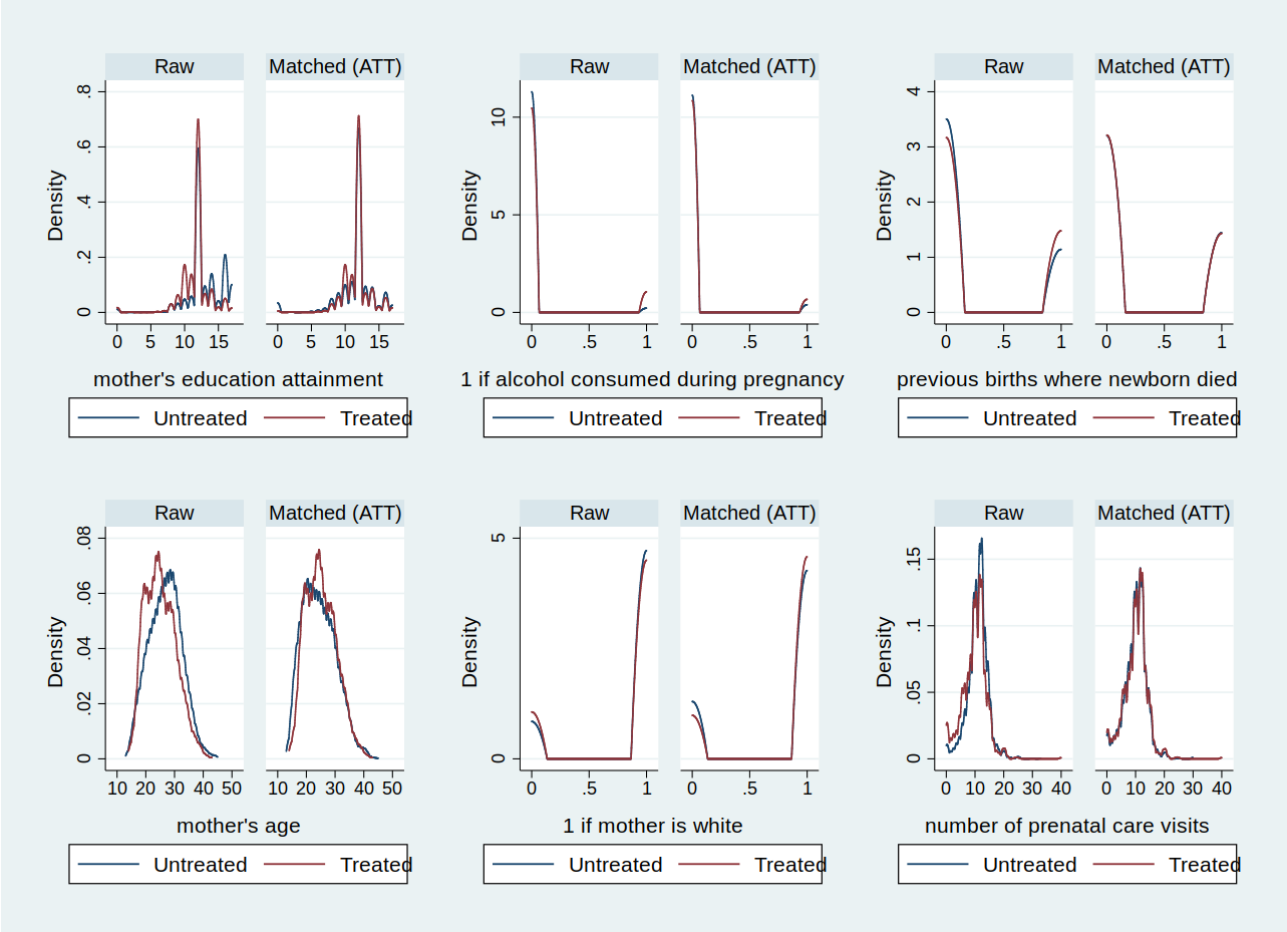
These tests are more global so you have an average result for all variables. This is good but not good enough as sometimes, in large models, some variables might cause trouble. Therefore, we should check all the variables separately. Type

```
kmatch summarize
```

	Raw			Matched(ATT)		
Means	Treated	Untrea~d	StdDif	Treated	Untrea~d	StdDif
medu	11.6389	12.9299	-.547436	11.7791	11.5993	.076236
1.alcohol	.091435	.018793	.322273	.058252	.034206	.106678
1.deadkids	.318287	.245897	.161322	.308252	.310623	-.005282
mage	25.1667	26.8105	-.300179	25.1626	24.483	.124109
1.mrace	.809028	.847803	-.102945	.822816	.766577	.149308
nprenatal	9.86227	10.9629	-.283799	10.0243	9.96418	.015495

	Raw			Matched(ATT)		
Variances	Treated	Untrea~d	Ratio	Treated	Untrea~d	Ratio
medu	4.69911	6.4232	.731585	3.49429	7.55803	.462328
1.alcohol	.083171	.018445	4.50921	.054926	.033077	1.66056
1.deadkids	.217232	.185481	1.17118	.213492	.214396	.995781
mage	28.1043	31.8714	.881803	27.9978	33.1336	.844997
1.mrace	.154681	.129067	1.19845	.145967	.179154	.814758
nprenatal	17.7041	12.3794	1.43013	16.7625	14.5843	1.14935

We see means and variances for all models. Note that variables that are used in the *ematch* option are not included as these have perfect values by design and cannot have a bad influence on these results here. Basically, you want the values of StdDif in the Matched columns to be as close to zero as possible (for means; for variances, they should approach 1). This means that, after matching, treatment and control are very similar to each other with respect to all variables. Also, the differences should also be smaller than in the *Raw* columns. We see that this is the case, except for mrace, where the value is a little further from zero than before matching. This could be due to the unequal distribution of this variable (only 16% of all women are not white). Therefore, finding matches for these non-white women is more difficult. Also, we see that the variances for medu are a bit worse. As there is no rule of thumb when a model becomes problematic, it is left to you to decide if you trust these values. At least make sure to report them in your paper. If you prefer a graphical representation type



The idea here is to plot before and after values. After matching, the two lines should always be more similar and closer to each other than before the matching.

This concludes the introduction to matching. By following these simple rules you can build quite complex models. Make sure to read the official documentation (*help kmatch*) to learn more as many more options are available. Also, check out the following presentation.

https://www.stata.com/meeting/germany17/slides/Germany17_Jann.pdf

5 Rosenbaum Bounds

Matching relies on observational data, therefore, we cannot test how good we can approximate causal effects in our findings. However, we can see how robust our results are with respect to omitted variables. These are variables that are not included in the data as they are not measured (for example, psychological constructs like intelligence or motivation) which might introduce spurious correlations, which can undermine our findings. The idea of Rosenbaum bounds is to simulate how strong these unobserved factors must be to undermine the effects we found with matching. Therefore, they are a type of sensitivity analysis. First, we run our model again, but this time save potential outcome differences for each treated case.

```
kmatch ps mbsmoke c.medu i.alcohol i.deadkids ///
    c.mage i.mrace c.nprenatal (bweight), att ///
    ematch(mmarried) dy(outcomes) replace
```

The option *dy(outcomes)* saves these differences in a new variable called outcomes. Now we can install the ado that allows us to estimate the bounds. Note that this command works for metric outcomes.[3]

```
ssc install rbounds, replace
```

Next, we run the command with the variable we just created.

```
rbounds outcomes, gamma(1 (0.25) 3)
```

3 If your outcome is binary, type *findit mhbounds*.

```
Rosenbaum bounds for out (N = 824 matched pairs)

Gamma           sig+      sig-    t-hat+    t-hat-       CI+       CI-
----------------------------------------------------------------------
    1              0         0  -189.966  -189.966  -226.337  -153.183
 1.25              0   9.8e-13  -241.289   -137.98  -278.287  -101.322
  1.5              0   5.3e-07   -283.48  -96.3894  -321.057  -58.9289
 1.75              0   .000996  -318.911  -61.1078  -357.772  -23.1496
    2              0   .059854  -349.671  -31.1668  -389.556   8.09554
 2.25              0   .417712  -376.975  -4.02625  -417.739   35.6186
  2.5              0    .84075  -401.193   19.6118  -442.812   59.8861
 2.75              0   .981698  -422.977   40.7339  -466.013   81.9369
    3              0   .999005  -442.942   60.0097  -487.545   101.933

* gamma  - log odds of differential assignment due to unobserved factors
  sig+   - upper bound significance level
  sig-   - lower bound significance level
  t-hat+ - upper bound Hodges-Lehmann point estimate
  t-hat- - lower bound Hodges-Lehmann point estimate
  CI+    - upper bound confidence interval (a=  .95)
  CI-    - lower bound confidence interval (a=  .95)
```

It calculates results for gamma values between 1 and 3, in steps of 0.25. Gamma means here the strength of the unobserved factor that influences the chance to be part of the treatment group. What we do is check sig- until the value is larger than 0.05 and therefore above the normal level of significance. The critical value seems to be around 1.75 as above the value crosses the threshold. We conclude from this that even if there was an unobserved factor that influences the chance to be part of the treatment in contrast to being part of the control in a relation 1.75 to 1, we would still find a significant result. The confidence limits give us an approximation about the range of the size of the effect we would still find then (-357 to -23, measured in grams). This seems like a quite robust result as this unobserved factor would need to influence our treatment status quite strongly. However, it is up to you to assess the result in the presence of unobserved variables. Rosenbaum bounds can help you but the interpretation is not written in stone.