Applied Econometrics for Researchers

Workshop on Differences-in-Differences (DiD), December 9th 2022

Data:

The following workshop is based on the study by Yakovlev (2018) on the alcohol consumption habits of Russian men.1 The dataset to be used in the following exercises is VodkaData.dta. It contains data on a sample of Russian men (18-65 years old) for the period 1994-2014. It includes information on individual's age, income, education, marital status, drinking and smoking behavior, among other details. We will be studying the effects of a federal tax on vodka consumption.

Background of the policy:

Russia implemented a policy in 2011 that aimed at reducing alcohol abuse with a strong increase in the federal excise tax on vodka. In this workshop, we will follow individuals' vodka consumption patterns over time and analyze some of the impacts of this policy.2

The variable *y2011* is a dummy variable equal to 1 for all years equal or larger than 2011; it is equal to zero for all years up to (and including) 2010.

Exercise 1.

a. Focus your attention on the variable *logvodka*, which captures the volume of consumption of vodka by the individuals in the sample. Implement a t-test in order to compare the consumption of vodka before and after the policy change. What do you conclude?

b. The dataset includes the variable *hdrinker50*, a dummy variable equal to 1 for heavy drinkers (those who belong to the yearly top 50% in total alcohol intake), and 0 for all others.3 Do you confirm the hypothesis that heavy drinkers have consumed more vodka than non-heavy drinkers during the period covered in the dataset? How large is the difference in relative terms?

c. Policy makers were hoping that the federal tax introduced in 2011 would reduce the gap in vodka consumption between heavy drinkers and non-heavy drinkers.

    i. By conducting two separate t-tests (for pre- and post-2011), what do you conclude regarding the difference in consumption between heavy drinkers and others?

1 Yakovlev, E. (2018), "Demand for alcohol consumption in Russia and its implication for mortality", *American Economic Journal: Applied Economics*, 10(1): 106-149.

2 Please note that this "policy change" (the introduction of the tax on vodka) affected everyone, so there is no natural "treated" and a "control" group in this case. However, we can explore whether this tax affected different groups of individuals differently – which is the focus of the exercises.

3 Please note that although we refer to non-heavy drinkers as a counterfactual, this group includes both low(er)-volume drinkers and individuals who did not consume any vodka in the reference year (non-drinkers).

ii. Still based on t-tests, how much did each group reduce their vodka intake after the introduction of the policy?

d. On average, how much did the price of vodka increase after 2011?

Exercise 2.

a. Using the command "*egen*" and the function "*mean*", construct two variables that measure the average consumption of vodka per year – one for heavy drinkers and another for non-heavy drinkers. (Hint: consult the slides of the lecture for an example.)

b. Plot these two variables over time, such that the two variables constructed in a) are represented in the y-axis and the years are represented in the x-axis (Hint: explore the command "*twoway*", with the option "*connected*"). Add a vertical line in the x-axis for the year 2011. Based on the plot that you obtain, what do you conclude regarding the evolution of alcohol consumption over time, the parallel trends assumption, and the impact of the policy for each group of drinkers?

(Note: we will return to the topic of parallel trends in Exercise 3)

c. Estimate a simple OLS regression (without any control variables) that allows you to estimate the differences between the two groups, before and after the policy change, as well as the difference in these two differences (i.e. the difference-in-differences). Cluster the standard errors at the individual level (using the variable *idind*) and explain why you should do so. Interpret the three coefficients of interest.

d. Repeat the same model as in c), but now include the following set of control variables: *logincome age age_2 smokes married college curwrk stroke*. Briefly interpret their coefficients. What is the estimated difference-in-differences once these controls are included? Why do you think that the estimates of interest are slightly different once we include these control variables?

Exercise 3.

a. Repeat the model estimated in questions 2c) and 2d) using a panel regression with individual fixed effects, i.e. using within-individual variation to estimate the key coefficients of interest.

a.i. What is the estimated impact of the federal tax on the vodka consumption of non-heavy-drinkers in either case? How do you interpret the different estimates obtained in models with and without control variables?

a.ii. What is the estimated impact of the federal tax on the vodka consumption of heavy-drinkers in either case? Did the policy have a significantly different effect for heavy- and non-heavy-drinkers?

b. Look at the following regression output. It covers the period 2008-2014, i.e. 3 year before and 3 years after the policy change. The variable *years_pre* measure the number of years prior to the policy change (years_pre = 1, 2, 3 in 2010, 2009, and

2008 respectively). Likewise, *years_post* measure the number of years after the policy change (years_post = 1, 2, 3 in 2012, 2013, and 2014 respectively). The year 2011 (i.e. the year of the introduction of the federal tax) is used as the baseline.

```
xtreg logvodka i.hdrinker50##i.years_pre i.hdrinker50##i.years_post $controls if year >= 2008,
fe r

Fixed-effects (within) regression               Number of obs     =     35,337
Group variable: idind                           Number of groups  =     13,363

R-squared:                                      Obs per group:
     Within  = 0.4097                                         min =          1
     Between = 0.5855                                         avg =        2.6
     Overall = 0.5336                                         max =          7

                                                F(21,13362)       =     419.35
corr(u_i, Xb) = 0.1779                          Prob > F          =     0.0000

                                    (Std. err. adjusted for 13,363 clusters in idind)
------------------------------------------------------------------------------------
                     |               Robust
            logvodka | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
---------------------+--------------------------------------------------------------
         1.hdrinker50 |   3.653408    .063289    57.73   0.000     3.529353    3.777463
                      |
            years_pre |
                   1 |  -.0130925   .0624875    -0.21   0.834    -.1355768    .1093918
                   2 |   .0749833   .0946966     0.79   0.428    -.1106355    .2606021
                   3 |  -.0525739   .1310422    -0.40   0.688    -.3094353    .2042874
                      |
 hdrinker50#years_pre |
                 1 1 |   .0131028   .0896595     0.15   0.884    -.1626425    .1888482
                 1 2 |  -.2464012   .0834221    -2.95   0.003    -.4099203    -.082882
                 1 3 |  -.0170154    .084402    -0.20   0.840    -.1824553    .1484246
                      |
           years_post |
                   1 |  -.1260879   .0526852    -2.39   0.017    -.2293583   -.0228175
                   2 |  -.1729076   .0891398    -1.94   0.052    -.3476342    .0018191
                   3 |  -.2280786   .1268946    -1.80   0.072    -.4768099    .0206528
                      |
hdrinker50#years_post |
                 1 1 |   -.377689   .0728358    -5.19   0.000    -.5204575   -.2349205
                 1 2 |   -.384279   .0758887    -5.06   0.000    -.5330317   -.2355264
                 1 3 |  -.4588321    .081265    -5.65   0.000    -.6181231   -.2995411
                      |
            logincome |   .0181487   .0108974     1.67   0.096    -.0032118    .0395093
                  age |   .0050412   .0449289     0.11   0.911    -.0830257    .0931081
                age_2 |  -.0000512   .0002528    -0.20   0.839    -.0005468    .0004443
               smokes |   .1038342   .0464003     2.24   0.025     .0128831    .1947853
              married |  -.0645199    .058724    -1.10   0.272    -.1796273    .0505876
              college |  -.1102237   .0729204    -1.51   0.131     -.253158    .0327106
               curwrk |   .0001932   .0454676     0.00   0.997    -.0889298    .0893162
               stroke |  -.0964041   .1390985    -0.69   0.488    -.3690569    .1762488
                _cons |   .4957966   1.623898     0.31   0.760    -2.687273    3.678866
---------------------+--------------------------------------------------------------
              sigma_u |  1.556954
              sigma_e |  1.6174961
                  rho |  .48093527   (fraction of variance due to u_i)
------------------------------------------------------------------------------------
```

Based on the output above:

i. why are there several coefficients for *years_pre* and *years_post* instead of one coefficient for each?

ii. do you observe any significant pre-trends (i.e. in the years prior to the federal tax) in vodka consumption for non-heavy drinkers? Which coefficients do you consider to make this assessment?

iii. do you observe any significant pre-trends in vodka consumption for heavy drinkers? Which coefficients do you consider to make this assessment?

iv. How much had non-heavy drinkers changed their consumption of vodka in 2012 compared to 2011? Is this difference statistically significant?

v. How much had heavy-drinkers changed their consumption of vodka by 2014: a) compared to their own consumption in 2011? b) compared to how much non-heavy drinkers have changed their consumption from 2011 to 2014?

vi. Policymakers were, at first, concerned that the effect of this tax would be short-lived, especially for heavy-drinkers, who could quickly go back to their old drinking habits. Based on the output above, what do you conclude regarding the long-term effects of this tax change for heavy-drinkers?

Exercise 4.

a. Implement a difference-in-differences estimation using the command "*diff*" in Stata (note that you may need to install it before, if you have done so yet). Compare your estimate of the difference of the differences with that obtained in Exercise 2c). What do you conclude?

b. Extend this model with the same controls listed in Exercise 2d). Repeat the interpretations as in the previous question.

c. Finally, run the same model as in 4b) but with the option "*kernel*", which will allow you to use kernel propensity score matching and DiD at the same time. (Hint: you need to specify which variable identifies the individuals – *idind* in this dataset). What is the estimate for the DiD, and how has it changed compared to the estimates obtained before? **(Note: To accelerate the computation of this command, please reduce your analysis to the period 2008-2014 by adding the condition "if year >= 2008" in the code).**