EVALUATION GUIDELINES - Take-home examination

DRE 70061
Panel Data/Microeconometrics

Department of Economics

Start date:	10.10.2019	Time 09:00
Finish date:	11.10.2019	Time 15:00

For more information about formalities, see examination paper.

Question 1 (counts 60 %)

Download the paper “Economic Shocks and Crime: Evidence from the Brazilian Trade Liberalization” and the data and programs that are available with the paper.

a) Give a short description (one page) of the empirical strategy applied in the paper, with an emphasis on what happens before Section IV. In particular, explain how the approach is related to a difference-in-differences approach, and in particular, how Figure 3 is used to validate the approach.

- Here we expect just an explanation of the empirical approach: Trade liberalization affected different regions differently, because of different sector composition. The analysis amounts to a comparison of crime rates in regions hit hard (in terms of labor market outcomes) by trade liberalization and other regions. If this treatment had been binary, this would have been a difference-in-differences strategy, itis rather similar when the treatment is continuous. The key assumption is that the regions that were hard hit by trade liberalization would not have different trends in crime rates from other regions in the absence of the trade liberalization. This is in a sense testable by looking backwards in time, and Figure 3 does not show any clear breach of this assumption.

b) In Tables 2 and 3 standard errors are adjusted for 91 meso-region clusters. Explain what this means and show how standard errors would have been without using this clustering.

- Here, we expect students to be able to replicate the analysis in Table 2 and 3 and replace the standard errors with what we get without applying the clustering option in STATA.
- The reason for clustering here is spatial correlation. Standard errors without clustering are only valid if there is no problem with spatial correlation. (The problem with spatial correlation is that it is likely that two adjacent regions have the same sectoral composition and the same development in crime, and that it is therefore not obvious that they should be treated as two independent observations).
- The clustered standard errors are quite a bit higher than the others, suggesting that spatial correlations are indeed a problem.

Table 2: The first two lines of numbers replicates Table 2 in the paper. The last two lines are without the clustering (with/without heteroskedasticity-robust standard errors, it makes most sense to do the heteroskedasticity-robust ones)

Model	(1)	(2)	(3)	(4)	(5)
RTCr	-1.976	.2,444	-3.838	-3.769	-3.853
standard error	(0.822)	(0.723)	(1.426)	(1.365)	(1.403)
new standard error	(0.850)	(0.518)	(0.591)	(0.564)	(0.584)
new standard error (robust)	(0.707)	(0.836)	(1.112)	(1.070)	(1.096)

c) In Tables 2 and 3, results from a two-stage least squares (TSLS) estimation procedure are reported, along with various OLS results. In this TSLS procedure, a control variable is instrumented because endogeneity in the control variable may contaminate other results. Explain the difference between instrumenting a control variable and using the instrument directly in place of the control variable. Illustrate this replicating the TSLS results from Tables 2 and 3 and comparing with the alternative suggested approach.

- This is a hard question. Expect some credit even if not precisely answered.
- It is not required that control variables are exogenous, what is required for a causal interpretation of the effect of trade liberalization on crime, is that the variation in trade liberalization is exogenous, conditional on the control variables. Therefore, it does not really make sense to instrument a control variable.
- It may be that the problem with the control variable is that it is a “bad control” and not only that it is endogenous. It is still not clear that instrumenting the variable is a solution to a bad control problem. Fortunately, none of these things seem to matter much for the results.

d) In the analyses reported in Tables 2 and 3, results are not very powerful because the authors only use data for a single year. Redo Tables 2 and 3 using data for all the years 1996-2003 and 2004-2010 for the two analyses. How do you handle autocorrelation?

- It takes some coding to produce the relevant variables.
- There seems to be two straightforward ways of dealing with the autocorrelation. One of them is to cluster the standard errors. Since different years from the same region will be within the same mesoregion, this should be handled by the clustering that is

already in the analysis. The other way is to construct one crime rate for each period (1996-2003) and (2004-2010) and to run the regression on this aggregated crime rate. I did version 2, in the table results from the early period are reported. Original estimate from paper, new estimate using all the early data and the new (meso-region-clustered) standard error.

Model	(1)	(2)	(3)	(4)	(5)
RTCr (original model)	-1.976	-2.444	-3.838	-3.769	-3.853
more data	-1.463	-2.006	-3.852	-3.771	-3.875
new standard error	(0.676)	(0.478)	(0.514)	(0.470)	(0.523)

e) Based on the analysis in d), apply quantile regression to see if the mean regressions reported in Tables 2 and 3 hide distributional effects. Specifically, do quantile regression for the three quartiles and compare these with mean regression results. Interpret the results.

- I skip the IV estimation here. As indicated in d), it is hard to understand the point with the IV estimation anyway. I also skip the weights. It is not straightforward to apply weights to quantile regression in STATA.

Model	(1)	(3)	(4)
RTCr (original model)	-1.463 (0.676)	-2.717 (0.763)	-2.811 (0.667)
Quantile regression (median)	-1.507 (0.789)	-2.031 (0.780)	-3.130 (0.673)
Quantile regression (first quartile)	-3.066 (0.766)	-2.486 (0.730)	-2.535 (0.834)
Quantile regression (third quartile)	-0.083 (1.053)	-1.748 (1.328)	-3.293 (1.028)

We can clearly see that the weights matter quite a bit for the effects. That was not the main point here. The main point is that the quantile regressions show clearly stronger effects at the first quartile than the median and the third quartile. That is, the effects

of the trade liberalization on crime rates seem to be stronger in regions that had low crime rates to start with. Note that this relates to model (1). In models (3) and (4), the same does not hold. Now, is this a consequence of the fact that we control for fixed effects, or is it because we estimate something different when we control for fixed effects (this is dealt with in problem f)?

f) Describe and apply alternatives to quantile regression - such as unconditional quantile regression - in the case with control variables. What is the difference in interpretation between standard and unconditional quantile regression?

- Standard quantile regression with control variables describes the effect on the quantiles on the conditional distribution. With state fixed effects, this means that the lower quartile regression describes what happens to the lower quartile crime rate within each state. It does not address what happens to the lower quartile crime rates globally. We can estimate what happens to lower quartile crime rates globally - while still controlling for state fixed effects using unconditional quantile regression.
- Unconditional quantile regression gives the corresponding table below. With unconditional quantile regression, there seems to be a somewhat lower effect of trade liberalization on crime in regions in the top of the crime rate distribution. (We have not dealt with how to test if the difference in effects is statistically significant).

Model	(1)	(3)	(4)
RTCr (original model)	-1.463 (0.676)	-2.717 (0.763)	-2.811 (0.667)
Quantile regression (median)	-2.171 (0.759)	-2.472 (0.893)	-2.551 (0.886)
Quantile regression (first quartile)	-2.589 (0.699)	-2.280 (0.888)	-2.350 (0.845)
Quantile regression (third quartile)	-0.141 (0.969)	-1.388 (1.200)	-1.477 (1.168)

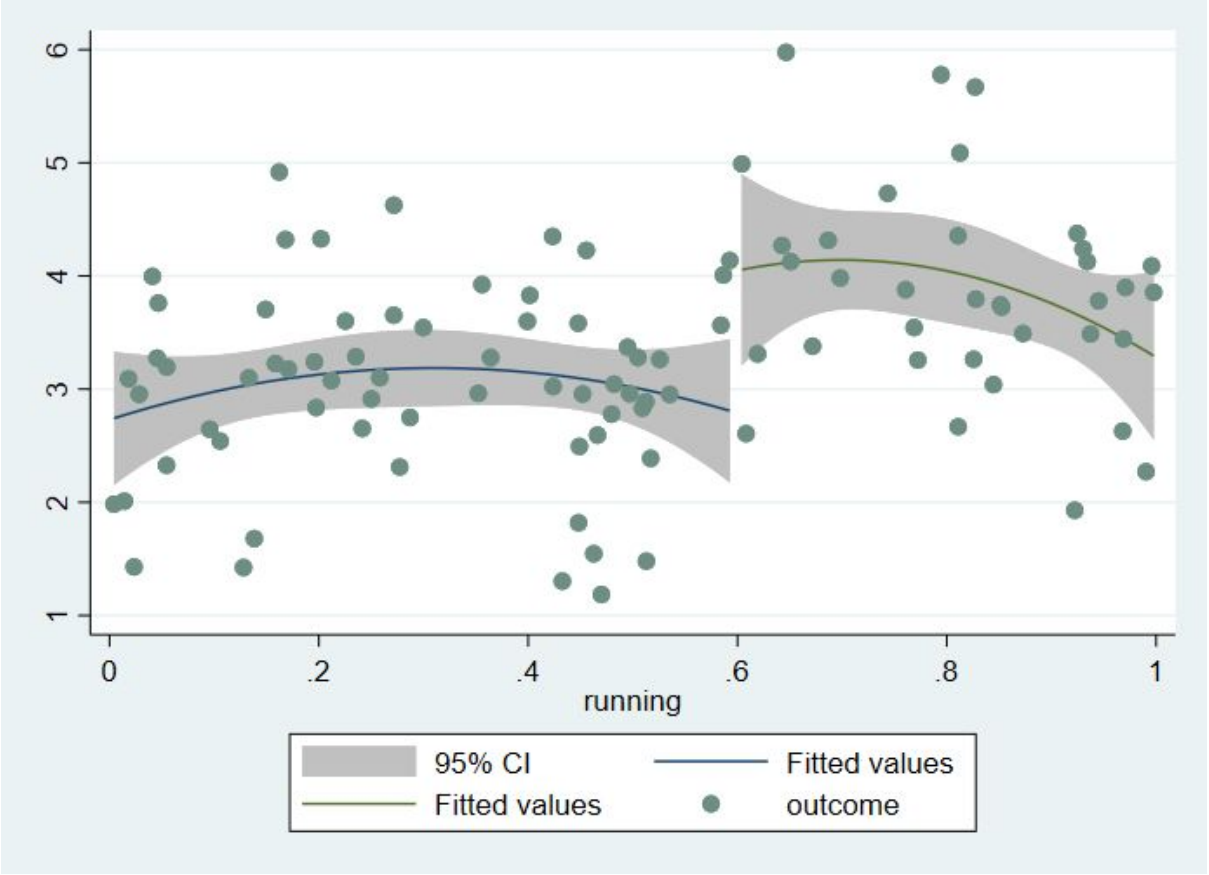
Question 2 (counts 20%)

a) Generate data for a regression discontinuity analysis. Specifically, generate a running variable taking on values on the unit interval. Assign treatment based on whether the running variable is above 0.6. Let expected outcomes be third order polynomials, with different coefficients at either side of the threshold. Introduce noise in observed outcomes. Present standard RDD plots.

- The code could be something like this:

```
clear
set obs 100

gen running = runiform()
gen treatment = running>0.6
gen runningt = (running-0.6)*treatment

gen running2 = running^2
gen running3 = running^3
gen runningt2 = runningt^2
gen runningt3 = runningt^3

gen outcomeu = 3 + rnormal()
gen outcomet = 4 -0.6-10*0.6^2+10*0.6^3 + running + 10*running^2 -
10*running^3+rnormal()
* treatment effect is 1.

gen outcome = treatment*outcomet + (1-treatment)*outcomeu

twoway (qfitci outcome running if running<0.6) (qfitci outcome running if
running >0.6) (scatter outcome running)
```

This produces a graph like:



b) Generate such data sets repeatedly. Choose parameter values for the expected outcomes and demonstrate how local linear models can outperform global linear models and global quadratic models. Is there a trade-off between bias and variance.

This could be achived by the following code:

```
capture program drop runrdd
program runrdd
clear
set obs 10000

gen running = runiform()
gen treatment = running>0.6

gen runningt = (running-0.6)*treatment

gen running2 = running^2
gen running3 = running^3
gen runningt2 = runningt^2
gen runningt3 = runningt^3

gen outcomeu = 3 + rnormal()*0.1
gen outcomet = 4 -0.6-10*0.6^2-10*0.6^3 + running + 10*running^2 +
10*running^3+rnormal()*0.1
```

```
* treatment effect is 1.

gen outcome = treatment*outcomet + (1-treatment)*outcomeu

gen local1 = running>0.55 & running < 0.65
gen local2 = running>0.4 & running < 0.8

reg outcome treatment running runningt running2 runningt2 running3 runningt3
scalar beta1 = _b[treatment]
reg outcome treatment running runningt
scalar beta2 = _b[treatment]
reg outcome treatment running runningt running2 runningt2
scalar beta3 = _b[treatment]
reg outcome treatment running runningt if local2
scalar beta4 = _b[treatment]
reg outcome treatment running runningt if local1
scalar beta5 = _b[treatment]

end
```

This program computes five estimates of the treatment effect. beta1 is based on a global cubic model with different slopes at either side of threshold, beta2 and beta3 is based on corresponding (misspecified) linear and quadratic models. Beta4 and beta5 are based on corresponding linear models, but bandwidths concentrates around the threshold. Beta4 has running variables ranging from 0.4 to 0.8, while beta5 has running variables ranging from 0.55 to 0.65.

I run the simulation by:

```
do runrdd
simulate b1=beta1 b2=beta2 b3=beta3 b4=beta4 b5=beta5, reps(1000): runrdd
summarize
```

I get the following output:

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          b1 |      1,000    1.000062    .0082697   .9749515   1.024408
          b2 |      1,000    .1255132    .0157303   .0676311   .1715318
          b3 |      1,000    1.032152    .0064178   1.011298   1.052541
          b4 |      1,000    .7974576    .0083594   .7732221   .8252712
          b5 |      1,000    .9880015    .0131056   .9474736   1.030287
```

We see that beta1 works the way it is supposed to (no misspecification here), while beta2 and beta3 are biased, beta3 only slightly so. We also see that beta4 and particularly beta5 improves on beta2. Beta5 with the smallest bandwidth seems almost as good as the global cubic model.

Question 3 (counts 20 percent)

a) Construct a data set with three groups: always-takers, never-takers and compliers. Generate potential outcomes as treated and untreated for the two groups. Let the expected potential outcomes differ between groups. Randomize a binary instrument that affects treatment and assign treatment and outcomes (based on expected potential outcomes and noise). Present summary statistics from the data set.

- Some code like this would suffice:

```
clear

set obs 1000

gen grouper = runiform()
gen alwaystaker = grouper < 0.3
gen nevertaker = grouper> 0.7
gen complier = 1  -alwaystaker-nevertaker

gen instrgen = runiform()
gen instrument = instrgen>0.7

gen potoutcome_treated = rnormal() + 1 + 0.3 * alwaystaker + 0.1 * nevertaker
gen potoutcome_untreated = rnormal() + 0 - 0.2 * alwaystaker + 1.1 *
nevertaker

gen treatment = alwaystaker + complier*instrument

gen outcome = treatment * potoutcome_treated  +
(1-treatment)*potoutcome_untreated

sum outcome
sum outcome if treatment==1
sum outcome if treatment==0
```

b) Demonstrate that two stage least squares, indirect least squares, the indirect estimation method in Imbens and Rubin (1997) and the procedure for estimating LATE based on “Abadie’s kappa” presented in “Mostly harmless econometrics” give identical results.

- This following code validates that you get exactly the same estimate using the four different numbers:

```
* Task 1: Wald estimator / ILS
```

```
sum outcome
sum outcome if instrument==1
scalar outcome_on = r(mean)
sum outcome if instrument==0
scalar outcome_off= r(mean)

sum treatment
sum treatment if instrument==1
scalar treatment_on = r(mean)
sum treatment if instrument==0
scalar treatment_off= r(mean)

display (outcome_on-outcome_off)/(treatment_on-treatment_off)


* Task 2: Two-stage least squares

ivregress 2sls outcome (treatment=instrument)

* Task 3: Imbens-Rubin

* Generating some scalars based on sample means

sum outcome if treatment==1 & instrument==1
scalar treatment_on_on = r(mean)
sum outcome if treatment==1 & instrument==0
scalar treatment_on_off = r(mean)

sum treatment if instrument==0
scalar alwalwalw=r(mean)
sum treatment if instrument==1
scalar nevnevnev=1-r(mean)
scalar comcomcom = 1- nevnevnev - alwalwalw

* Estimate of pot outcome as treated for compliers

scalar pot_out_tre_comp = (alwalwalw+comcomcom)/comcomcom * treatment_on_on -
alwalwalw/comcomcom * treatment_on_off
display pot_out_tre_comp

sum outcome if treatment==0 & instrument==1
scalar treatment_off_on = r(mean)
sum outcome if treatment==0 & instrument==0
scalar treatment_off_off = r(mean)

display treatment_off_on
display treatment_off_off


* Estimate of pot outcome as untreated for compliers
```

```
scalar pot_out_untre_comp = (nevnevnev+comcomcom)/comcomcom *
treatment_off_off - nevnevnev/comcomcom * treatment_off_on
display pot_out_untre_comp

* Difference in pot. outcomes for compliers = LATE

display pot_out_tre_comp - pot_out_untre_comp

* Task 4: Abadie kappa computations:

sum instrument
scalar pz = r(mean)

gen kappa = (1 - treatment*(1-instrument)/(1-pz) -
(1-treatment)*instrument/pz)

reg outcome treatment [iw=kappa]
```

Some code for Problem 1

```
* Problem 1B Without "cluster(mesoreg)"

* Unweighted
reg delta_CR_1991_2000 rtc_kume_main, robust

* Weighted by population
reg delta_CR_1991_2000 rtc_kume_main [aw = weight_1991_2000], robust

* Weighted by population + state fixed effects
reg delta_CR_1991_2000 rtc_kume_main state1-state27 [aw = weight_1991_2000],
robust

* Weighted by population + state fixed effects + pre-trend
reg delta_CR_1991_2000 rtc_kume_main delta_CR_1980_1991 state1-state27 [aw =
weight_1991_2000], robust

ivregress 2sls delta_CR_1991_2000 rtc_kume_main (delta_CR_1980_1991 =
trend_iv_1980_1990) state1-state27 [aw = weight_1991_2000], robust
```

Problem 1c (just comparing the 2sls with a regression using the instrument from the 2sls as a control variable and not an instrument for an endogeneous control

```
ivregress 2sls delta_CR_1991_2000 rtc_kume_main
(delta_CR_1980_1991 = trend_iv_1980_1990) ///
       state1-state27 [aw = weight_1991_2000], cluster(mesoreg)

regress delta_CR_1991_2000 rtc_kume_main trend_iv_1980_1990 ///
       state1-state27 [aw = weight_1991_2000], cluster(mesoreg)
```

Problem 1 d

```
*gen delta_CR_1991_early =
log((1/7)*(CR_1996+CR_1997+CR_1998+CR_1999+CR_2000+CR_2001+CR_2002+CR_2003))
- log(CR_1991)
*gen delta_CR_1991_later =
log((1/6)*(CR_2004+CR_2005+CR_2006+CR_2007+CR_2008+CR_2009+CR_2010)) -
log(CR_1991)

* Unweighted
reg delta_CR_1991_early rtc_kume_main

* Weighted by population
reg delta_CR_1991_early rtc_kume_main [aw = weight_1991_2000]

* Weighted by population + state fixed effects
reg delta_CR_1991_early rtc_kume_main state1-state27 [aw = weight_1991_2000]

* Weighted by population + state fixed effects + pre-trend
reg delta_CR_1991_early rtc_kume_main delta_CR_1980_1991 state1-state27 [aw =
weight_1991_2000]

ivregress 2sls delta_CR_1991_early rtc_kume_main (delta_CR_1980_1991 =
trend_iv_1980_1990) state1-state27 [aw = weight_1991_2000]


* Unweighted
reg delta_CR_1991_late rtc_kume_main

* Weighted by population
reg delta_CR_1991_late rtc_kume_main [aw = weight_1991_2000]

* Weighted by population + state fixed effects
reg delta_CR_1991_late rtc_kume_main state1-state27 [aw = weight_1991_2000]

* Weighted by population + state fixed effects + pre-trend
reg delta_CR_1991_late rtc_kume_main delta_CR_1980_1991 state1-state27 [aw =
weight_1991_2000]

ivregress 2sls delta_CR_1991_late rtc_kume_main (delta_CR_1980_1991 =
trend_iv_1980_1990) state1-state27 [aw = weight_1991_2000]
```

Problem 1e

```
reg delta_CR_1991_early rtc_kume_main

qreg delta_CR_1991_early rtc_kume_main
qreg delta_CR_1991_early rtc_kume_main, q(0.25)
qreg delta_CR_1991_early rtc_kume_main, q(0.75)
```

```
reg delta_CR_1991_early rtc_kume_main state1-state27

qreg delta_CR_1991_early rtc_kume_main state1-state27
qreg delta_CR_1991_early rtc_kume_main state1-state27, q(0.25)
qreg delta_CR_1991_early rtc_kume_main state1-state27, q(0.75)


reg delta_CR_1991_early rtc_kume_main delta_CR_1980_1991 state1-state27

qreg delta_CR_1991_early rtc_kume_main delta_CR_1980_1991 state1-state27
qreg delta_CR_1991_early rtc_kume_main delta_CR_1980_1991 state1-state27,
q(0.25)
qreg delta_CR_1991_early rtc_kume_main delta_CR_1980_1991 state1-state27,
q(0.75)
```

Problem 1 f

```
xtile deltacat = delta_CR_1991_early, nq(4)

gen delta1 = deltacat<=1
gen delta2 = deltacat<=2
gen delta3 = deltacat<=3

pctile quartile = delta_CR_1991_early, n(4)

kdensity delta_CR_1991_early, at(quartile) generate(xes densities)

gen rif1 = quartile[1] + (0.25-delta1) / densities[1]
gen rif2 = quartile[2] + (0.5-delta2) / densities[2]
gen rif3 = quartile[3] + (0.75-delta3) / densities[3]

reg rif1 rtc_kume_main, robust
reg rif2 rtc_kume_main, robust
reg rif3 rtc_kume_main, robust

reg rif1 rtc_kume_main state1-state27, robust
reg rif2 rtc_kume_main state1-state27, robust
reg rif3 rtc_kume_main state1-state27, robust

reg rif1 rtc_kume_main delta_CR_1980_1991 state1-state27, robust
reg rif2 rtc_kume_main delta_CR_1980_1991 state1-state27, robust
reg rif3 rtc_kume_main delta_CR_1980_1991 state1-state27, robust
```