violations

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Friday 5th April, 2024 19:50

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

endogeneity

[*] (elements of) research design: causality

[*] more diagnostics

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

endogeneity

[*] (elements of) research design: causality

[*] more diagnostics

changes from before

◇ dropped autocorrelation–assuming you use cross-sec data

∘ not time series, not panel

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

endogeneity

[*] (elements of) research design: causality

[*] more diagnostics

violations

- ◇ so far we have just talked about the regressions that satisfy assumptions
- ◇ but what happens when assumptions are violated?
- ○ typically, they are!
- ◇ and what you can do about it ?

practical considerations

- ◇ usually have heteroskedasticity in crosssectional data
- ◇ (and autocorrelation in time-series data) [skipped]
- ◇ (and both in panel data) [skipped]
- ◇ “unobserved heterogeneity” = LOVB
- ◇ outliers/leverage
- ◇ normality of residuals
- ◇ you should *always* test all of them
- ◦ (except autocorr in unclustered cross-sectional data and normality in datasets>1k)
- ◇ when you report reg results, it is expected and assumed you took care of all assumptions

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

endogeneity

[*] (elements of) research design: causality

[*] more diagnostics

we discussed collinearity earlier

- ◇ if perfect, then you cannnot estimate std err
- ◦ stata will just drop a perfectly collinear var
- ◦ with dummies–if you incl all cat–it is so called "dummy trap"
- ◇ otherwhise, collinearity does not violate any assumption
- ◇ just makes std err bigger
- ◇ it is just like "micronumerosity"
- ◇ typically, do nothing

outline

misc

intuition

collinearity again

heteroskedasticity

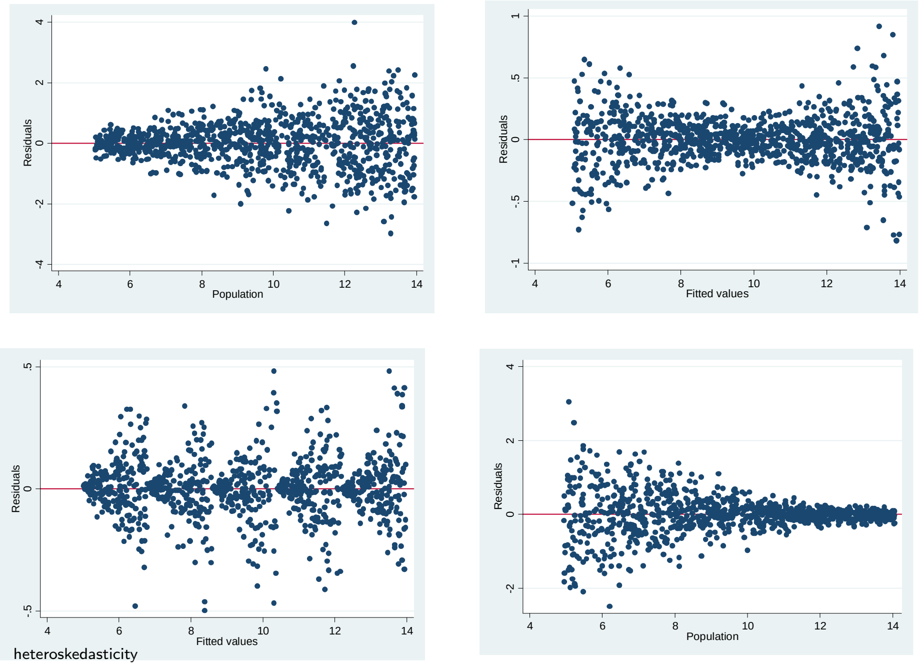
normality of residuals

endogeneity

[*] (elements of) research design: causality

[*] more diagnostics

examples


Residuals
Population
4
2
0
-2
-4
4
6
8
10
12
14
Residuals
Fitted values
1
.5
0
-.5
-1
4
6
8
10
12
14
Residuals
Fitted values
.5
0
-.5
4
6
8
10
12
14
Residuals
Population
4
2
0
-2
4
6
8
10
12
14

violation

- ◇ again, heteroskedascity=pattern in residuals
- ◇ the variance of Y conditional on X varies from one observation to another
  - ◦ eg it may depend on the values of X
- ◇ if true:
  - ◦ $\hat{\beta}_j$ still unbiased
  - ◦ $s_{\hat{\beta}_j}$ is not as accurate as reported by software
  - ◦ not BLUE because not efficient

diagnosis

- ◇ eyeball
- ◇ test
- ○ there are many tests... eg Breush-Pagan

solutions

- calculate robust se
- transform variables (*if* theoretically justifiable)
  - heteroskedasticity might indicate you are working in the wrong metric
  - a popular transformation that often works is log
  - log is popular for skewed distributions like income...
- dofile: het

outline

only worry if you have small sample

- don't have to worry about this at all if sample is big
- if sample is small, after running regress
- can predict residuals `predict resid,r`
- do a histogram and plot them
- if they look very unnormal, don't be too trusting in significance
- try to get more data!

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

endogeneity

[*] (elements of) research design: causality

[*] more diagnostics

closely related to design!

- ◇ if you have bad design, you'll have endogeneity
- ◇ curiously, economists are obsessed with it
- ◇ but other fields aren't
- ◇ a superb and readable reference is Sorensen (2012)

  `http://people.bu.edu/tsimcoe/code/Endog-PDW.pdf`

what is it

- ◇ technically, if x and error term are correlated
- ◇ so there is some Z that predicts Y and correlates with X
- ○ (see also discussion of Z in res des sec)
- ◇ so it can be just LOVB, or unobserved heterogeneity
- ◇ unobserved heterogeneity: see Rumsfeld's unknown unknowns in res des sec

simultaneity and self-selection

- ◇ but usually by endogenity we mean bigger problems
- ◇ simultaneity and self-selection
- ◇ and they are bigger problems because no amount of control vars helps!
- ◇ simultaneity not only $X \rightarrow Y$ but also $Y \rightarrow X$
- ○ could do Granger causality or IV
- ◇ but best do an experiment, or natural experiment
- ◇ think deeply about the relationship between X and Y
- ◇ one of the best ways to think deeply, i think, is to use INUS condition (res des sec)

the bottom line

- the bottom line is that in experiment U/As are assigned to levels of X at random
- think about whether that is the case in your study (after controlling for other Xs)
- or at least if that's the case to large degree
- you want to think about selectivity and self-selection early in the process: at the research design stage
- think about **source of variability** in X
  - or data generating process as pol sci would put it

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

endogeneity

[*] (elements of) research design: causality

[*] more diagnostics

research design

- ◇ whether you have good or bad research design does not violate assumptions
- ◇ but it is critical for ability to argue causality
- ◇ causality is acheived with design, not with statistics (incl regression)!!
  - ◦ sure trying to get closer to it with multiple regressions, but cannot really get there with much confidence
  - ◦ indeed multiple regression results themselves (without design or at least much thought given to it), are about as good as an educated guess!!

research design is a class itself

- research design is about designing your research
- i will just mention few things that will be important for this class
- a quick, useful and applied reference is http://www.socialresearchmethods.net/kb/design.php
- a more in-depth treatment is Lawrence B. Mohr, Impact Analysis for Program Evaluation books.google.com/books?isbn=0803959362
- also see http://knowledge.sagepub.com/view/researchdesign/SAGE.xml
    - guess have to be on campus to access it for free

causality

- ◇ much of research design is about causality
- ○ want to show X $\rightarrow$ Y
- ◇ correlation is necessary for causality
- ○ (in rare cases suppressor var makes it unnecessary, eg (Mazur, 2011))
- ◇ but not sufficient
- ◇ `http://www.tylervigen.com/`

INUS condition (Mackie, 1980)

- ◇ a useful way of thinking about causality: Insufficient but Non-redundant part of Unnecessary but Sufficient Condition
- ◇ many, if not most causes are INUS conditions
- ◇ eg a cigarette as a cause of forrest fire
- ◦ it's Insufficient, because by itself it is not enough, eg you also need oxygen, dry leaves, etc
- ◦ it is contributing to fire, hence Non-redundant
- ◇ and along with other stuff (oxygen, dry leaves etc) it constitutes Unnecessary but Sufficient Condition
- ◦ it's not necessary for fire, it can be lightening, etc
- ◦ but it's sufficient – it's enough to start the fire

basic concepts

- ◇ Y: a dependent variable, outcome
- ◇ X: an independent variable, predictor
- ◦ (T: (treatment), like X)
- ◇ Z: some other variable
- ◇ want to show $X \rightarrow Y$ (X affects (causes) Y)
- ◦ and not the other way round ($Y \rightarrow X$)
- ◦ and not $Z \rightarrow Y$; eg X(CO_2),Y(temp), Z(sun temp)
- ◦ it is difficult to argue !
- ◦ after all, there are unknown unknowns (Z's that we are unaware of)

The Problem: Unknown Unknowns

- there are known knowns; there are things we know that we know
- there are known unknowns; that is to say, there are things that we now know we don't know
- but there are also unknown unknowns–there are things we do not know we don't know
- (Donald Rumsfeld)
- how do we deal with unknown unknowns?
- do an experiment!

The Problem put another way: Counterfactual

- ◇ it all boils down to comparing
  what happened to what would have happened had the treatment not happened
- ◇ eg we got a new teacher and now kids perform better on SAT
- ◦ to know whether the teacher caused better performance we would need to know what would have happened to SAT scores without this teacher (scores might have gone up due to Z),
- ◦ and compare it to what actually happened

The Problem put another way: Counterfactual

- ◇ the problem is that we do not observe counterfactual (we can try to infer it though)
- ◇ counterfactual is the effect of all knowns/unknowns (incl. unknown unknowns)
- ◇ how do we deal with lack of counterfactual
- ◇ do an experiment!
- ◇ (or if you cannot, try to estimate it somehow)

the gold standard [ask IRB appr]

◇ the experimental design give few examples

◇ only with experimental design you can confidently argue causality

◇ and it is because randomization takes care of the known and unknown predictors of the outcome (draw a picture of 2 groups of people)

○ in other words, it establishes a counterfactual

◇ but wait !

○ most of the time we cannot have an experimental design because it is unethical and politically impossible
eg we cannot randomly assign kids to bad school or to smoking http://www.socialresearchmethods.net/kb/desexper.php

internal validity

- internal validity is about causality
- you have internal validity if you can claim that X causes Y
  - eg some drug X causes some disease Y to disappear
  - http://knowledge.sagepub.com/view/researchdesign/n43.xml#n43
  - http://knowledge.sagepub.com/view/researchdesign/n192.xml#n192

threats to internal validity

- ◇ history, maturation, regression to the mean
  - ○ something else happened that caused Y
  - ○ things develop over time in a certain way
- ◇ selection bias, self selection
  - ○ does smoking causes cancer ?
  - ○ maybe less healthy people select to smoke ?
- ◇ http://knowledge.sagepub.com/view/researchdesign/n192.xml#n192

spurious correlation

- ◇ you think that X causes Y, but actually it is Z
- ◇ global warming:
  - ◦ we have it–we can measure temperature
  - ◦ but what's the cause: CO_2 or Sun activity?

reverse causality

- a closely related topic to spurious correlation is reverse causality
- here, instead of some other Z that causes Y instead of X
- we have Y causing X, as opposed to X causing Y...
- what do we do ?

reverse causality OR chicken-egg dilemma

- ◇ you may try to find some other X that measures the same or similar concept and that cannot be caused by Y
- ◇ eg instead of education $\rightarrow$ wage; do father's education$\rightarrow$ wage (your wage can reverse cause your education, but not your father's education)
- ◇ find some exogenous (external) shock: policing$\leftrightarrow$crime
- ◇ but terror attack/alert $\rightarrow$policing$\rightarrow$crime; we know that policing$\rightarrow$crime; not the other way round
  - ○ https://www.law.upenn.edu/fac/jklick/48JLE267.pdf
- ◇ or dating happiness–which comes first? happy folks more likely to be dated!

natural experiment

- ◇ again most of the time you cannot have an experiment
- ◇ but there are natural experiments or exogenous shocks
- ◇ exogenous meaning that they are caused externally (like an experimenter's randomization) and somewhat randomly (at least with relation to a problem at hand
- ○ eg earthquake (any weather, eg storm); terrorist attack; policy change (less random)
- ◇ in model simply have dummy for U/As affected storm, policy etc

causality without experiment?

- ◇ yes! well maybe, but you need to do lots of work...
- ◇ essentially you want to exclude alternative explanations
- ◇ so you act like a devil's advocate...
- ◇ try to abolish your story / find an alt explanation
- ◇ if you cannot find any, then your story is right ...
- ○ until disproved
- ○ just use regression and "control" for other vars
- ◇ there are some designs that improve our inference greatly over having no design at all (ex post facto, observational)

ex post facto: $X_1 Y_1$

- ◇ very common...it is *no* design
- ◇ non-experimental, cross-sectional, observational, correlational; you'll most likey do this
- ◇ we start investigation "after the fact"
- ◇ no time involved, don't know whether X precedes Y
- ◇ both, X and Y are observed at the same time examples?
- ◦ (but X must precede Y in order to be causal)
- ◇ practically impossible to argue causality here
- ◇ but cheap and big N, and good external validity

ex post facto: $X_1 Y_1$

- useful, many "causes" were discovered using observational studies
- eg smoking→cancer was found out using ex post facto
- and then confirm using better designs
- http://knowledge.sagepub.com/view/researchdesign/n145.xml
- http://knowledge.sagepub.com/view/researchdesign/n271.xml#n271

before-after (pre-post): blackboard: schematic

- ◇ measured Y, then do X, and then measured Y again
- ◇ eg measured readership at the library , buy some cool stats books ; measured readership again
- ◇ eg measured crime rate , put more police on the streets ; measured crime again
- ◇ eg measured soup consumption , changed soup ; measured soup consumption again
- ◇ anyone did pre/post? eg working at school?
- ○ tried new programs, new approaches?
- ○ or simply pre-post without T, say to identify highest and lowest gain students

(2 group) comparative change: $\frac{Y_{E1} X_2 Y_{E3}}{Y_{C1} \quad Y_{C3}}$

- ◇ eg H_0 : police with better guns fights crime better
- ◇ measured crime rate in 2010 in Camden (Y_{E1}) and Newark (Y_{C1})
  - ◦ in 2011 give super guns to police in Camden (X_2), (but not in Newark)
  - ◦ in 2012 measured crime rate Camden (Y_{E3}) and Newark (Y_{C3})
- ◇ if crime rate dropped more in Camden than in Newark, then we have evidence that the guns worked
- ◇ stata: see so called DID http://www.princeton.edu/~otorres/DID101.pdf

difference in difference (p.235 Wheelan, 2013)

◇ just 'before after' with a comparison group

◇ did sth to one group, and not to the other group

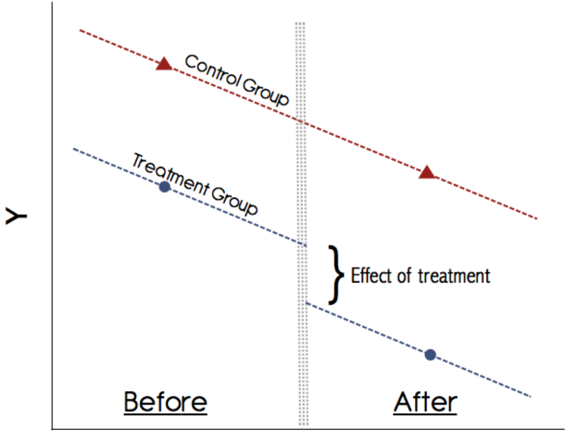
◦ over time (pre post) see if there is any difference

◇ like we discussed earlier in res_des.pdf

◇ blackboard: fig: first from p236, and then from p237

◦ and pictures similar to those from res_des.pdf follow

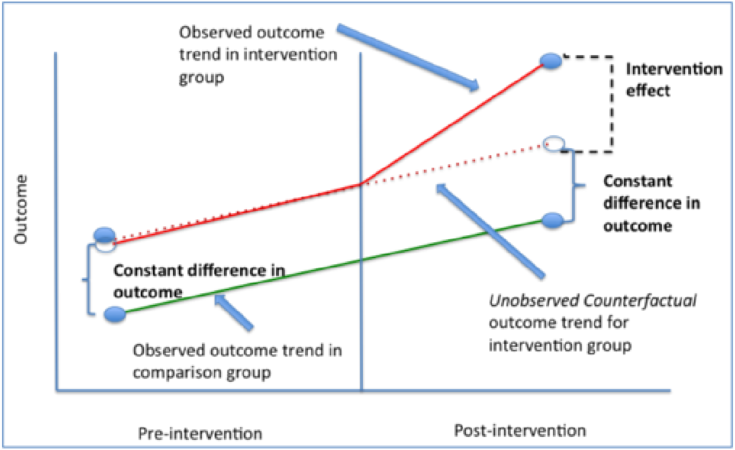
DID



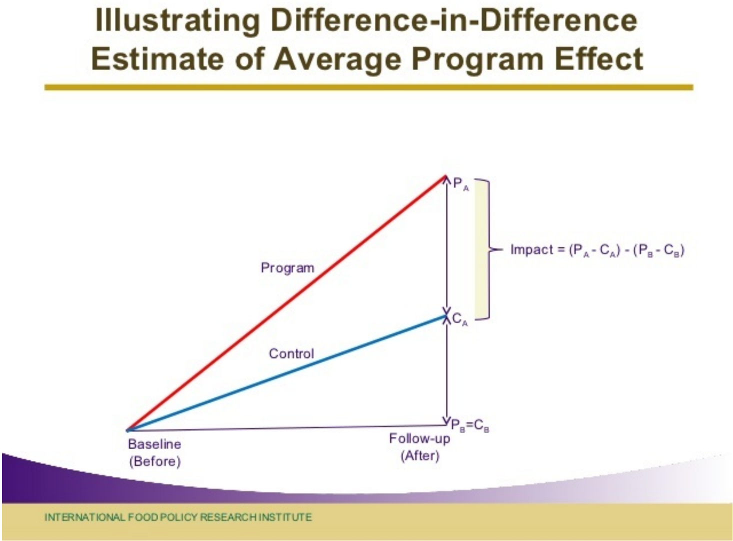
Y
Control Group
Treatment Group
Effect of treatment
Before
After

◇

DID


Observed outcome
trend in intervention
group
Intervention
effect
Outcome
Constant
difference in
outcome
Constant difference in
outcome
Unobserved Counterfactual
outcome trend for
intervention group
Observed outcome trend in
comparison group
Pre-intervention
Post-intervention

DID


Illustrating Difference-in-Difference
Estimate of Average Program Effect
P_A
Impact = $(P_A - C_A) - (P_B - C_B)$
Program
C_A
Control
$P_B = C_B$
Baseline
(Before)
Follow-up
(After)
INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

discontinuity analysis (p.238 Wheelan, 2013)

◇ can use when there is some rigid cutoff for something, say:

◦ remedial program for F grades

◦ prison sentence for a crime

◇ then compare those who just made it (C-, or a ticket)

◦ v those who didn't (F, prison)–but they were just above the cutoff

◇ the cool thing is that the two groups are similar, especially:

◦ not really any difference whatsoever with respect to cause of treatment!

◦ so the treatment is arbitrary (random), so we have experiment!

example

- ◇ new jersey state government workforce profile 2010
- ◇ http://www.nj.gov/csc/about/publications/workforce/pdf/wf2010.pdf
- ◇ p37: minorities in state govt over time
- ◇ how increase internal validity?
- ◇ compare to PA, DE, NY etc
- ◇ factor in minority population; applications
- ◇ do experiments! many already done! again, read lit!!
  - ◦ say people with black names apply for jobs
  - ◦ students with Asian names email professors
- ◇ and both, employers and professors discriminate against!

eg: tacit knowledge is the key!

◇ if you know sth about state govt

∘ you know that it is concentrated in Trenton

∘ (one student said so)

◇ hence, the key is population characteristics

∘ around Trenton!

next step

- ◇ if you are interested in program evaluation:
  - ○ quick http://www.socialresearchmethods.net/kb/evaluation.php
  - ○ in-depth, advanced: Mohr (1995), Shadish et al. (2002)

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

endogeneity

[*] (elements of) research design: causality

[*] more diagnostics

Nick's modeldiag

- http://www.stata-journal.com/sjpdf.html?articlenum=gr0009
- dofile:modeldiag

ucla diagnostics

- https://stats.idre.ucla.edu/stata/webbooks/reg/chapter2/stata-webbooksregressionwith-statachapter-2-regression-d
- most useful:
  - scatter dfbeta ...
  - lvr2plot, ml()
  - avplot(s)
- you should always do these in your research
- may also want to transform variables if needed: 1.5 transforming variables https://stats.idre.ucla.edu/stata/webbooks/reg/chapter1/regressionwith-statachapter-1-simple-and-multiple-regres
- and see help regress postestimation

Mackie, J. (1980): The cement of the universe, Clarendon Press Oxford.

Mazur, A. (2011): "Does increasing energy or electricity consumption improve quality of life in industrial nations?" Energy Policy, 39, 2568–2572.

Mohr, L. B. (1995): Impact Analysis for Program Evaluation, Sage, Beverly Hills CA, second edition ed.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002): Experimental and quasi-experimental designs for generalized causal inference, Wadsworth Cengage learning.

Sorensen, J. B. (2012): "Endogeneity is a fancy word for a simple problem," Unpublished.

Wheelan, C. (2013): Naked statistics: stripping the dread from the data, WW Norton & Company.