violations

Adam Okulicz-Kozaryn
adam.okulicz.kozaryn@gmail.com

this version: Saturday 6th April, 2024 11:47

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

[*] more diagnostics

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

[*] more diagnostics

outline

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

[*] 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

[*] 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

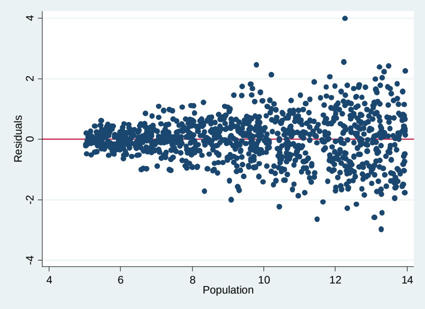
collinearity again

heteroskedasticity

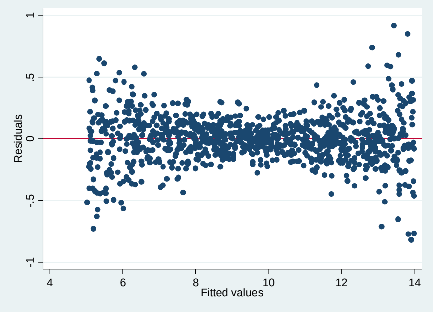
normality of residuals

[*] more diagnostics

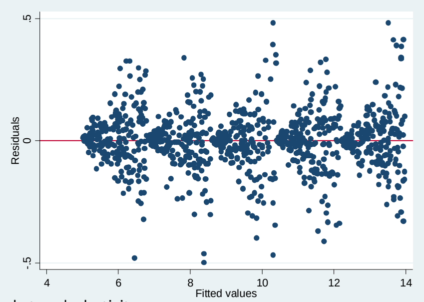
examples



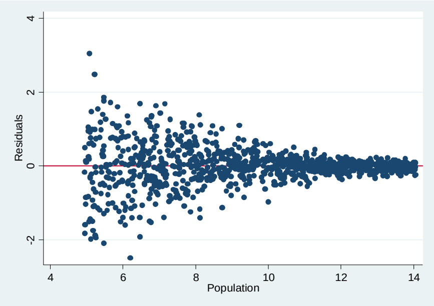
Residuals
4
2
0
-2
-4
4
6
8
10
12
14
Population



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



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



Residuals
4
2
0
-2
4
6
8
10
12
14
Population

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

misc

intuition

collinearity again

heteroskedasticity

normality of residuals

[*] more diagnostics

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

[*] 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.