violations

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Monday 20th April, 2015 17:35

outline

misc

intuition

collinearity again

heteroskedasticity

autocorrelation

normality of residuals

more diagnostics

elements of research design: causality [bonus]

endogeneity [bonus]

outline

misc

intuition

collinearity again

heteroskedasticity

autocorrelation

normality of residuals

more diagnostics

elements of research design: causality [bonus]

endogeneity [bonus]

ps6

- added some comments to ps6.pdf–let's have a look!
- work on paper! there will be one more big ps:
- final draft of final project

outline

misc

intuition

collinearity again

heteroskedasticity

autocorrelation

normality of residuals

more diagnostics

elements of research design: causality [bonus]

endogeneity [bonus]

violations

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

practical considerations

- you will usually have heteroskedasticity in crosssectional data
- (and autocorrelation in time-series data)
- (and both in panel data)
- 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 assumptions

outline

we discussed collinearity earlier

- if perfect, then you cannnot estimate std err
  - stata will just drop a variable
  - 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"

outline

misc

intuition

collinearity again

heteroskedasticity

autocorrelation

normality of residuals

more diagnostics

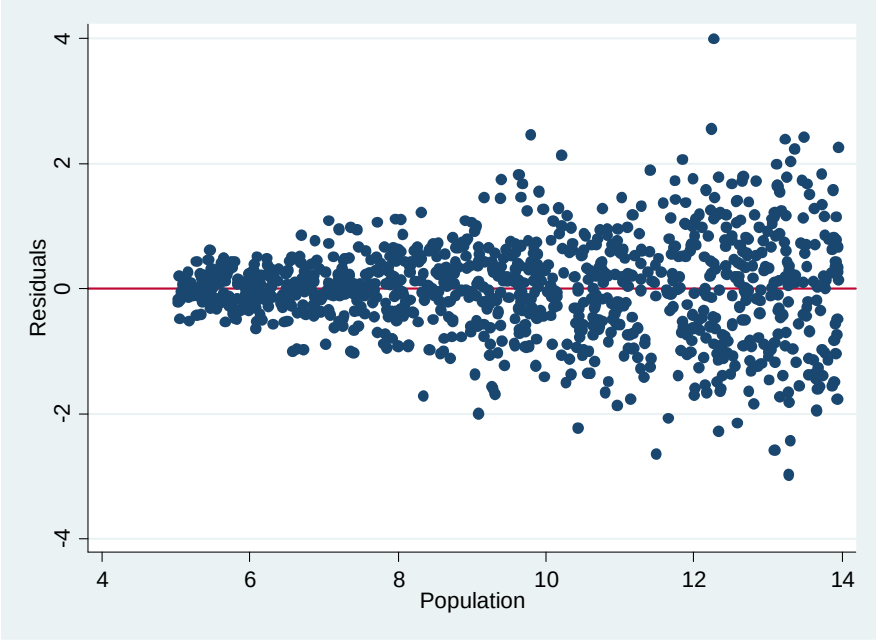
elements of research design: causality [bonus]

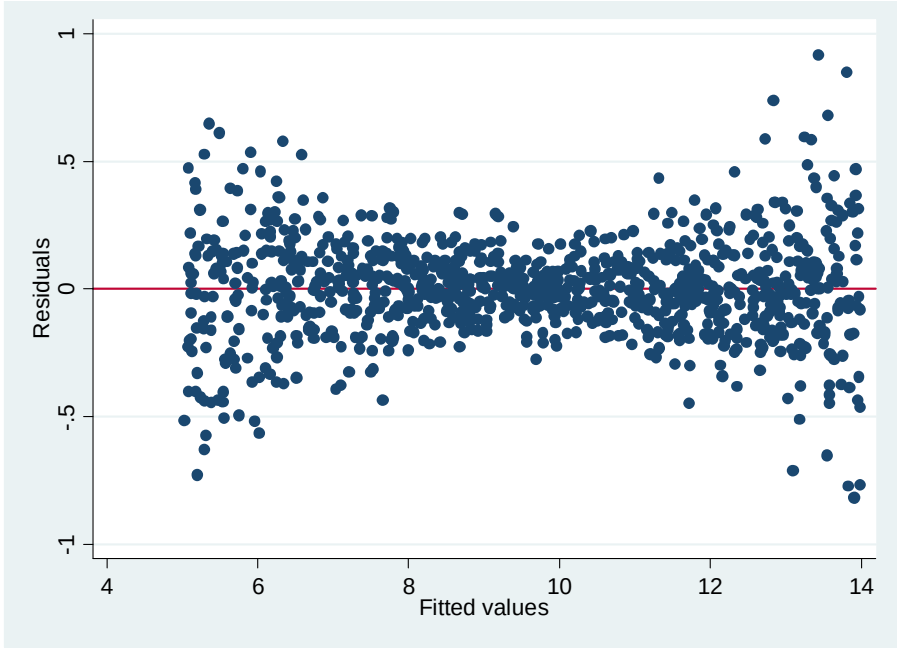
endogeneity [bonus]

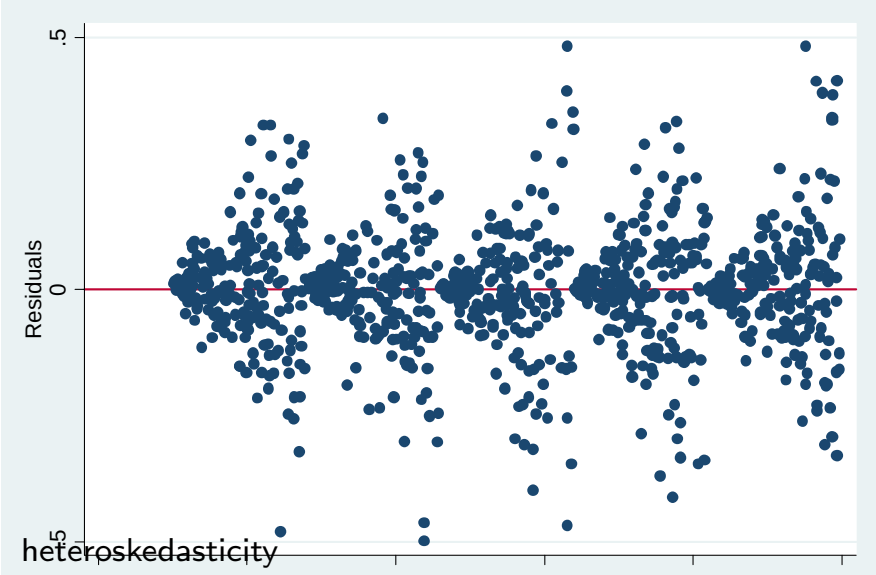
what is it ?

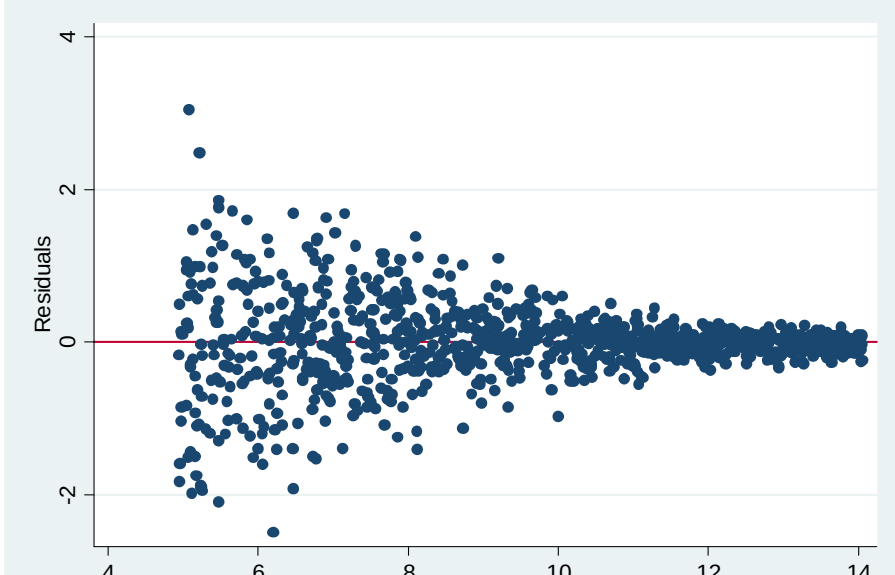
- Homoskedastic: $var(u_i) = \sigma^2$ note there is no subscript on *sigma*
- also, note that there is a subscript on u_i – this means that every obs has error variance

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


Residuals
-2
0
2
4

violation; $var(u_i) = \sigma_i^2$

- $u_i \sim N(0, \sigma_i^2)$
- subscript i means that the disturbance term has greater variability for some observations than others
- Another way to put it: the variance of Y conditional on X varies from one observation to another. For example, 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

implications

- $u_i \sim N(0, \sigma_i^2)$
- $var(\hat{\beta}_2) = \frac{\sum \sigma_i^2 x_i^2}{(\sum x_i^2)^2}$
- if $\sigma_i^2 = \sigma^2 \rightarrow var(\hat{\beta}_2) = \frac{\sigma^2 \sum x_i^2}{(\sum x_i^2)^2} = \frac{\sigma^2}{(\sum x_i^2)}$
- $s^2 = \frac{\sum e_i^2}{n-k}$
- $s_{\hat{\beta}_2} = \frac{s}{\sqrt{\sum x_i^2}}$
- hence, the reported se are wrong (because we assume σ)
  - t tests, F tests, confidence intervals are wrong, too

diagnosis

- eyeball
- test
- there are many tests... e.g. Breush-Pagan

solutions

- you can do weighted least squares
  - Weighted Least Squares (WLS) is a special case of Generalized Least Squares (GLS) - transform the variables until they satisfy Gauss- Markov assumptions, then use OLS on the transformed variables.
- calculate robust se $var(\hat{\beta}_2) = \frac{\sum \sigma_i^2 x_i^2}{(\sum x_i^2)^2}$
- 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

autocorrelation

normality of residuals

more diagnostics

elements of research design: causality [bonus]

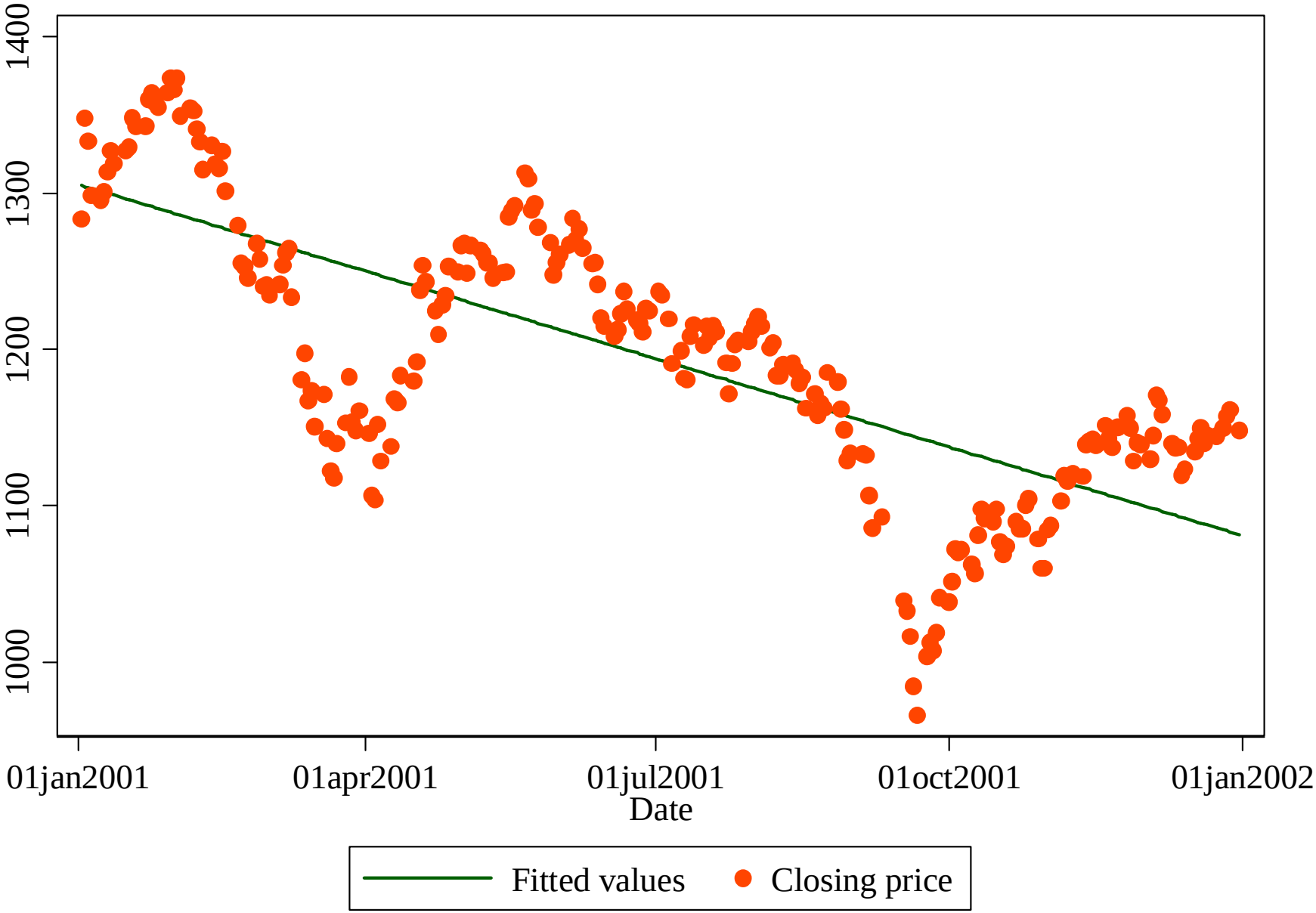
endogeneity [bonus]

what is it?

- violation of $E[u_i, u_j] = 0$, with $i \neq j$
- correlations in the disturbance terms of observation in time series
  - this year is a lot like last year
  - specifically, even after controlling for x, "nearby" years have similar disturbances
- correlation in space, nearby neighborhoods (houses, blocks, countries, etc.) have similar disturbance terms

, robust cluster(space)

S&P 500, 2001


1400
1300
1200
1100
1000
01jan2001
01apr2001
01jul2001
01oct2001
01jan2002
Date
Fitted values
Closing price

causes

- random factors correlated in time or space
- inertia
- incorrect functional form
- lagged responses
- data manipulations
- transformations e.g. first differences
- non-stationarity

consequences

- OLS estimates still unbiased
- standard errors are incorrect
  - usually underestimate the true standard errors
  - therefore, overestimate t/F statistics
  - increase probability of Type I error (ouch!)
- estimates are no longer efficient (not BLUE)
- estimate of σ^2 biased, therefore R^2 unreliable

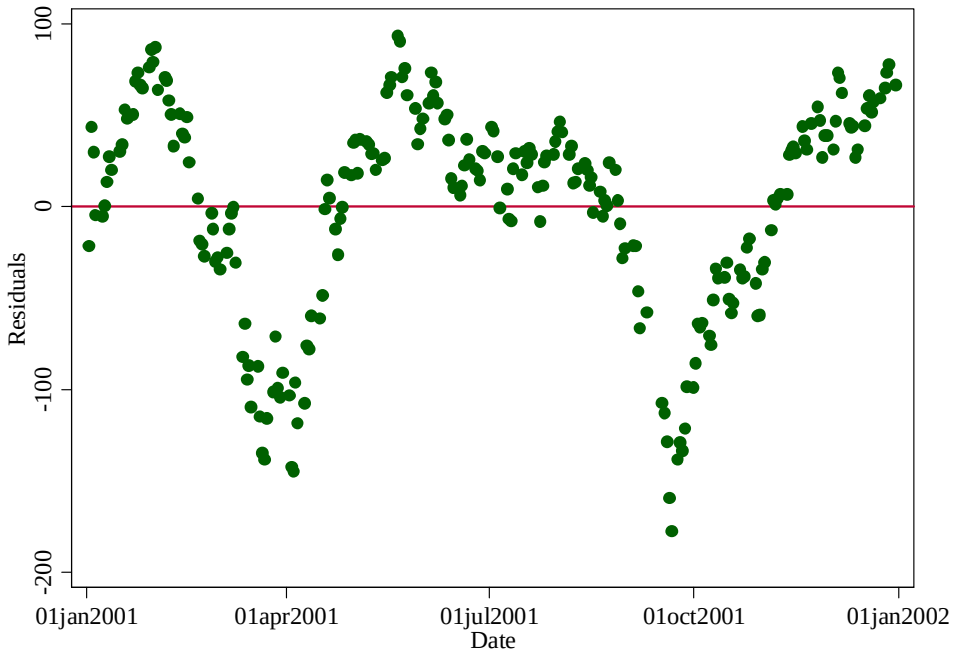
the simplest case, AR(1)

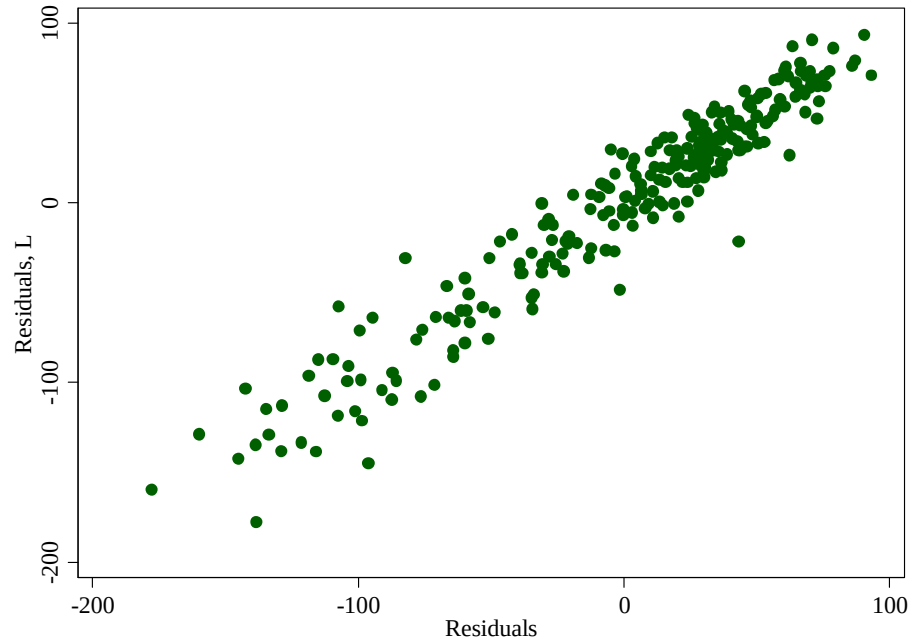
- AR(1) is First Order Autoregressive Process
- $Y_t = \beta_1 + \beta_2 X_t + u_t \quad t = time$
- $u_t = \rho u_{t-1} + \epsilon_t \quad -1 < \rho < +1$
  - ρ is correlation with previous u
  - ϵ is well-behaved residual
    $E[\epsilon_t] = 0;\ var[\epsilon] = \sigma^2;\ cov[\epsilon_t, \epsilon_{t+s}] = 0,\ s \neq 0$
- if $\rho = 0$ then we have OLS; otherwise problems

AR(1) chain of correlations

- u_t is correlated with every other u_{t+s} by $\rho^{|s|}$
- $u_t = \rho u_{t-1} + \epsilon_t$
- $u_t = \rho(\rho u_{t-2} + \epsilon_{t-1}) + \epsilon_t$
- $u_t = \rho^2 u_{t-2} + \rho\epsilon_{t-1} + \epsilon_t$
- $u_t = \rho^2(\rho u_{t-3} + \epsilon_{t-2} + \rho\epsilon_{t-1} + \epsilon_t$
- $u_t = \rho^3 u_{t-3} + \rho\epsilon_{t-2} + \rho\epsilon_{t-1} + \epsilon_t$
- $u_t = \ldots$

detection: eyeball: e_t vs time and e_t vs e_{t-1}



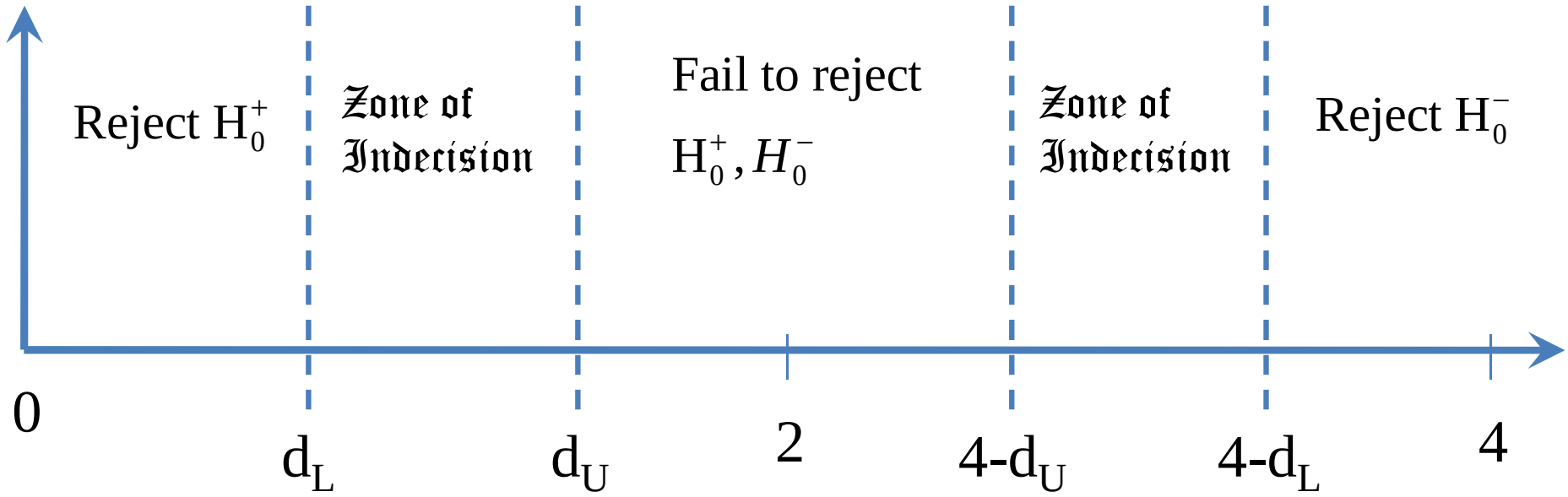


detection: durbin-watson

- $d = \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=2}^{N} e_t^2} = \frac{\sum_{t=2}^{N} e_t^2 + \sum_{t=2}^{N} e_{t-1}^2 - 2\sum_{t=2}^{N} e_t e_{t-1}^2}{\sum_{t=2}^{N} e_t^2} \approx$
  $\frac{2\sum_{t=2}^{N} e_t^2 - 2\sum_{t=2}^{N} e_t e_{t-1}^2}{\sum_{t=2}^{N} e_t^2} \approx 2\left(1 - \frac{\sum_{t=2}^{N} e_t e_{t-1}^2}{\sum_{t=2}^{N} e_t^2}\right)$
- $\hat{\rho} = \frac{\sum_{t=2}^{N} e_t e_{t-1}^2}{\sum_{t=2}^{N} e_t^2}$
- $d = 2(1 - \hat{\rho})$
- if $\rho = 0 \rightarrow E[d] = 2$
- if $\rho = 1 \rightarrow E[d] = 0$
- if $\rho = -1 \rightarrow E[d] = 4$

interpretation

- H_0^+ : no $+$ autocorrelation
- H_0^- : no $-$ autocorrelation



- help estat dwatson
- table: `http://web.stanford.edu/~clint/bench/dwcrit.htm`
- dofile:autocorr

other tests

- Geary Runs Test help runtest
- Bruesch-Godfrey help estat bgodfrey
- (i teach dw because it is in the curiculum; in practice rather use other tests)

solution

- Prais-Winsten and Cochrane-Orcutt help prais
- Newey-West help newey
- dofile:autocorr

[*] stata time series

- `fmwww.bc.edu/ec-p/wp598.pdf`

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

we'll do causality/endogeneity next week

- it is recently fashinable to do it

[*] Nick's modeldiag

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

ucla diagnostics

- `http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm`
- most useful:
  - scatter dfbeta ...
  - lvr2plot, ml()
  - avplot(s)
- thase are the thing that you should always do in your research

bonus

- ucla scroll to 1.5 transforming variables `http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter1/statareg1.htm`
- help regress postestimation

outline

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)

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, e.g. (Mazur, 2011))
- but not sufficient
- `http://www.tylervigen.com/`

INUS condition (Mackie and 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
- e.g. a cigarette as a cause of forrest fire
  - it's Insufficient, because by itself it is not enough, e.g. 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 $Z \rightarrow Y$; e.g. 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
- e.g. 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
  e.g. 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
  - e.g. 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

- draw a scatter, fit line of some Y and some X
  say X is banana production in Honduras, Y is deaths on US highways
- you think that X causes Y, but actually it is Z
- another example: global warming...
  - we have it–we can measure temperature
  - but what's the cause: CO_2 or Sun activity
- another way to say it: correlation is not causation

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

- you may try to find some other X that measures the same or similar concept and that cannot be caused by Y
- e.g. 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`

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
- e.g. earthquake (any weather, e.g. storm); terrorist attack; policy change (less random)

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...
- and try to abolish your story / find an alternative explanation
- if you cannot find any, then your story is right ...
  - until disproved
  - just use regression and "control" for other variables [elaborate later in semester]
- there are some designs that improve our inference greatly over having no design at all (ex post facto, observational)
- e.g. get data from `http://www.statepolicyindex.com/` and

PRE, POST

- look over time (PRE, POST)(BEFORE, AFTER)
  e.g. you can trace unemployment over time in Camden
  draw interrupted time series
- and, say, you can find that it increased during Reagan administration...
- still, you cannot argue causality right away !
- there may be lots of alternative explanations, e.g. shift away from manufacturing during the same time, etc etc

T, C (treatment, control)

- ◇ and you can look across groups or space (T, C)
- e.g. you can compare crime in Camden and Newark, while there was some intervention in Camden only, say new policing approach
- draw 4 boxplots: Camden, Newark: before after
- remeber blood pressure in stata?
- `http://www.ats.ucla.edu/Stat/stata/library/GraphExamples/code/grbox1.htm`
- or see actual paper–scroll down to box plots in the middle
- `http://www.bmj.com/content/341/bmj.c3215`

few basic designs

- see p.62, 63 http://books.google.com/books?id=GBxhOT8btfYC&printsec=frontcover
  - also: http://knowledge.sagepub.com/view/researchdesign/n353.xml#n353
- ex post facto: $X_1 Y_1$ (T,Y at the same time)
- (one group) pre-post or before-after: $Y_1 X_2 Y_3$
- (two group) comparative change: $\frac{Y_{E1} X_2 Y_{E3}}{Y_{C1} \quad Y_{C3}}$
- interrupted time series: $Y_1 Y_{...} Y_{10} X_{11} Y_{12} Y_{...} Y_{13}$
  - can also have interrupted time series with a control group
  - $\frac{Y_{E1} Y_{...} Y_{E10} X_{11} Y_{E12} Y_{...} Y_{E13}}{Y_{C1} Y_{...} Y_{C10} \quad Y_{C12} Y_{...} Y_{C13}}$
- subscripts measure time; E:experimental/treatment; C:control
- control group: http://knowledge.sagepub.com/view/researchdesign/n76.xml#n76

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
- e.g. smoking→cancer was found out using ex post facto
- and then confirm using better designs
- e.g. correlate happiness and income
- e.g. correlate crime and poverty
- e.g. correlate car sales and pollution
- `http://knowledge.sagepub.com/view/researchdesign/n145.xml`
- `http://knowledge.sagepub.com/view/researchdesign/n271.xml#n271`

before-after: $Y_1 X_2 Y_3$

- measured Y, then do X, and then measured Y again
- e.g. measured readership at the library (Y_1), buy some cool stats books (X_2); measured readership again (Y_3)
- e.g. measured crime rate (Y_1), put more police on the streets (X_2); measured crime again (Y_3)
- e.g. measured soup consumption (Y_1), changed soup (X_2); measured soup consumption again (Y_3)

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

- e.g. 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

interrupted time series: $Y_1 Y_{\ldots} Y_{10} X_{11} Y_{11} Y_{\ldots} Y_{20}$

- e.g. H_0 : the new anti-unemployment program in Camden decreased unemployment
- get data about unemployment in Camden from 1990 to 2010 ($Y_1 Y_{\ldots} Y_{10}$) AND ($Y_{11} Y_{\ldots} Y_{20}$)
- say the unemployment program began in 2001 (X_{11})
- produce a time series plot (mark a vertical line in 2001: intervention/treatment)
- if there was a change in trend after 2001, we conclude that the program worked

interrupted time series: $Y_1 Y_{\ldots} Y_{10} X_{11} Y_{11} Y_{\ldots} Y_{20}$

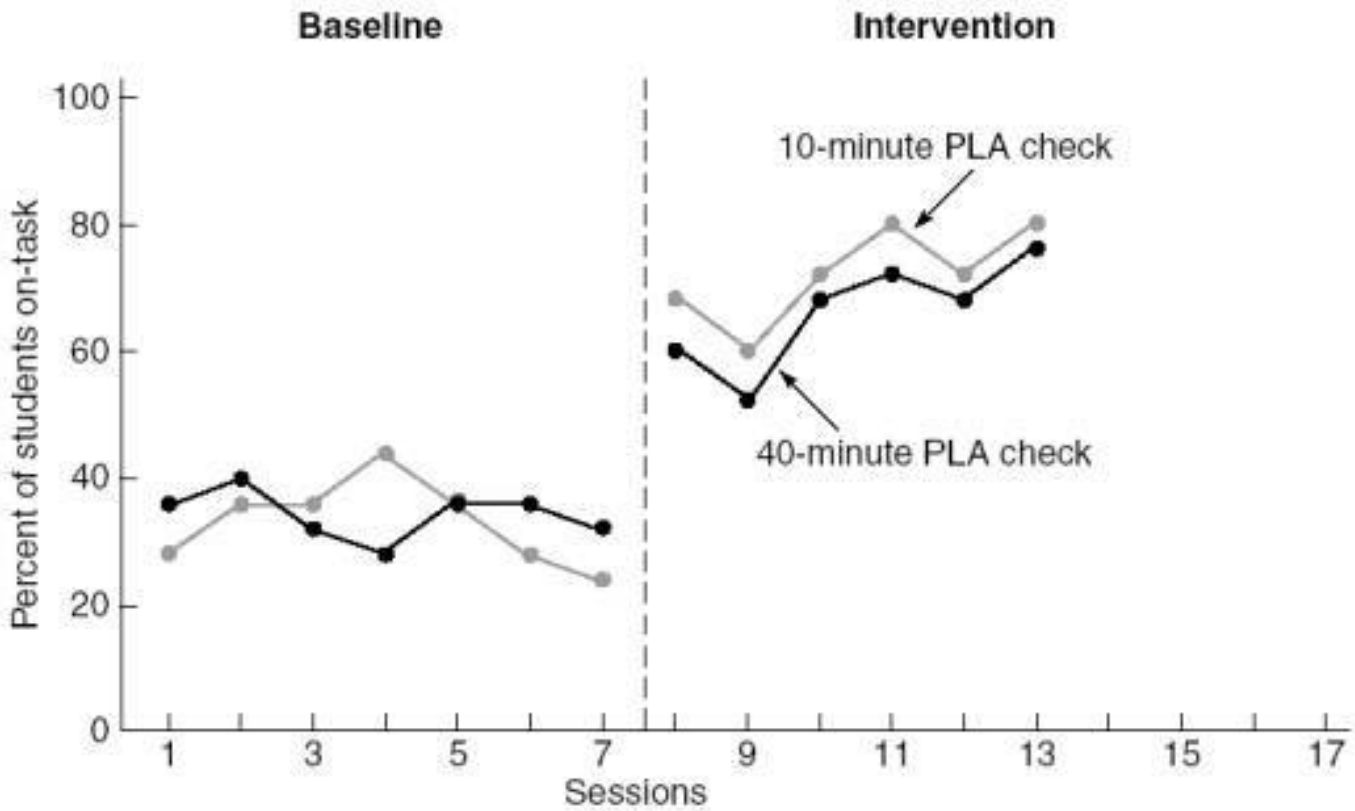
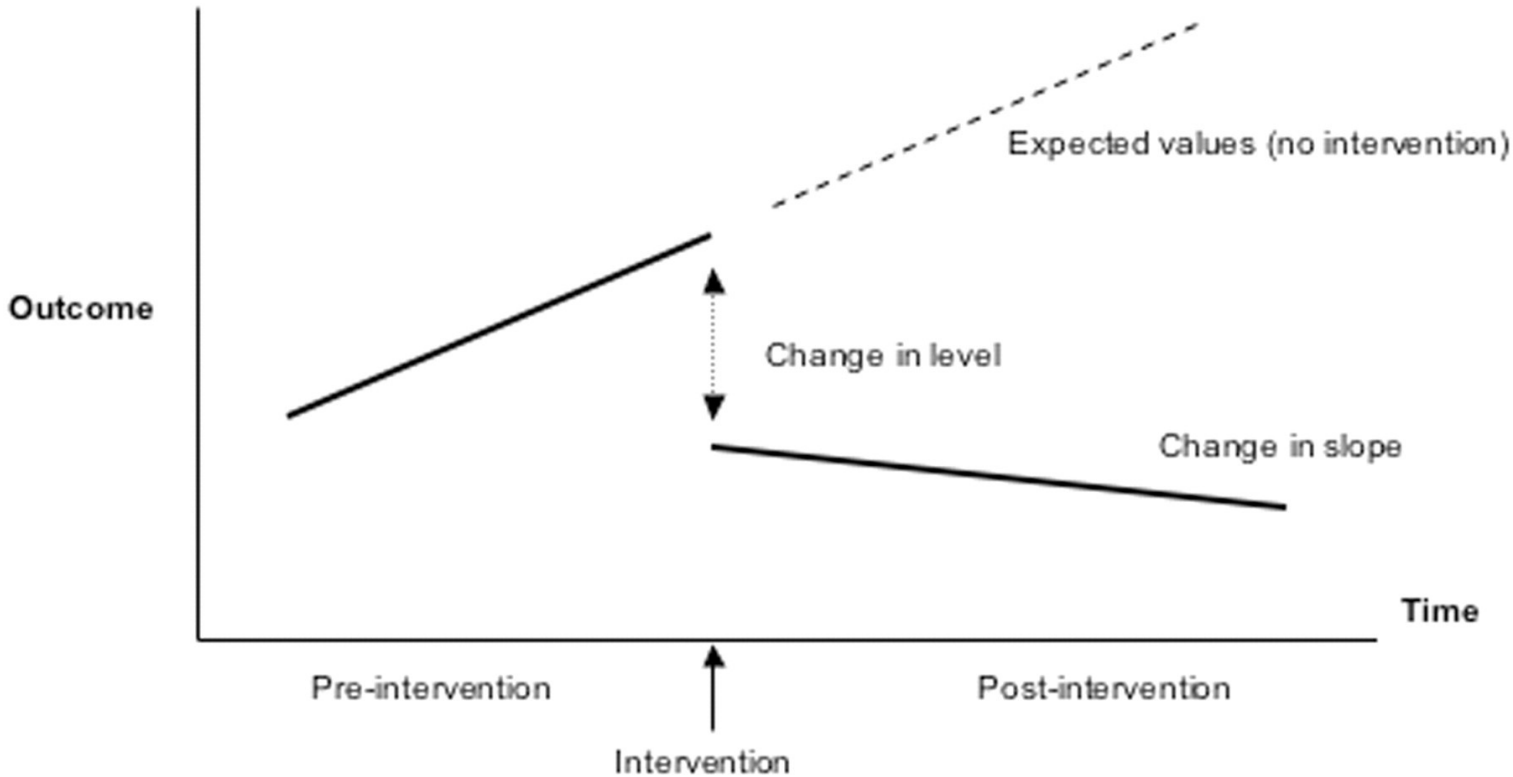


FIGURE 10.5 Percentage of students who are on-task at 10 minutes and 40 minutes into the class period. The figure presented here depicts the results of one of five classrooms investigated by Mayer et al. Only one classroom is presented here to illustrate a time-series design, whereas Mayer et al. used five classrooms and a multiple-baseline design. PLA refers to planned activity.

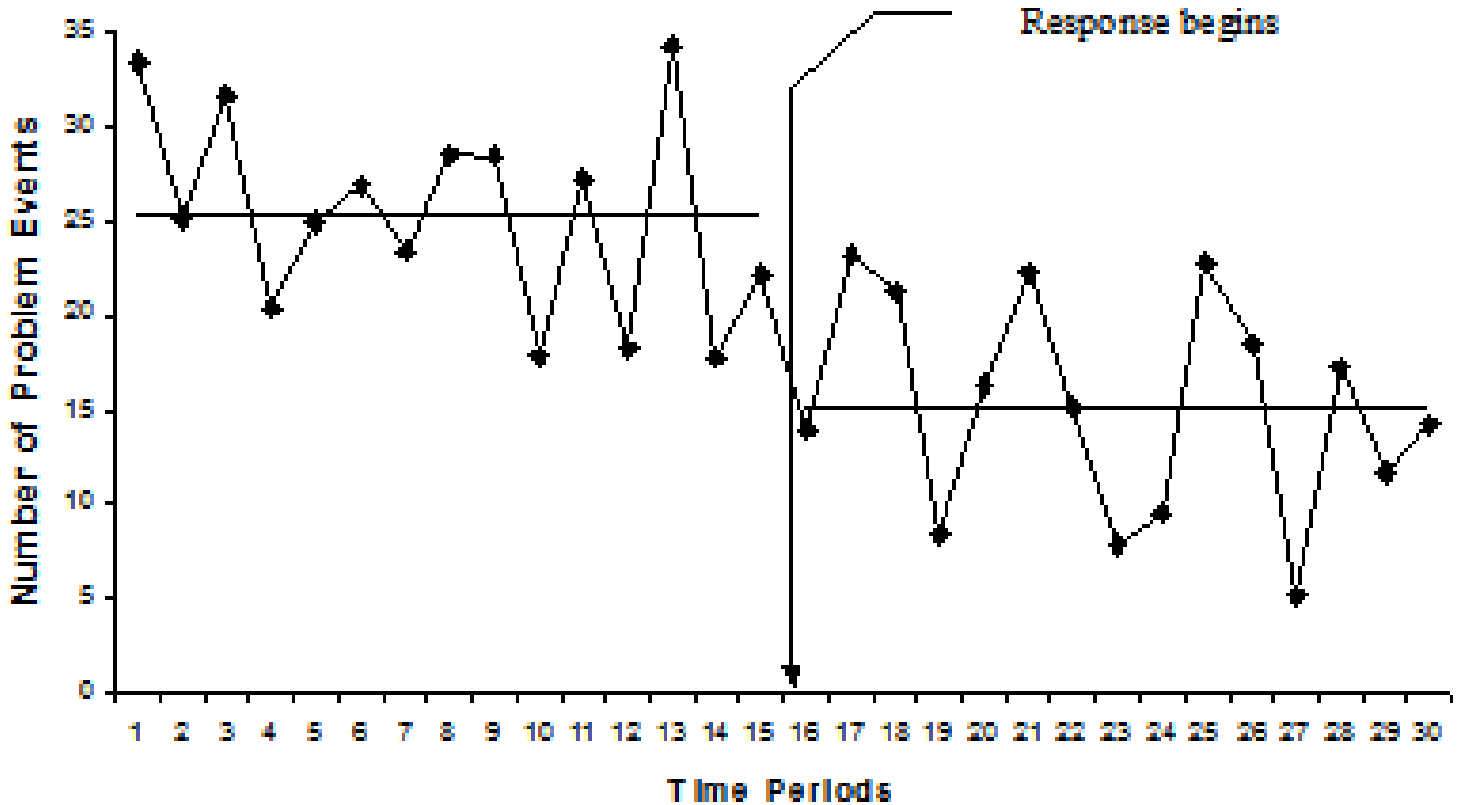
Adapted from G. R. Mayer, L. K. Mitchell, T. Clementi, E. Clement-Robertson, & R. Myatt (1993). "A dropout prevention program for at-risk high school students: Emphasizing consulting to promote positive classroom climates." *Education and Treatment of Children, 16,* 135–146. Reprinted by permission.

interrupted time series: $Y_1 Y_{\ldots} Y_{10} X_{11} Y_{11} Y_{\ldots} Y_{20}$



- ◇
- ◇ in general look at the trend

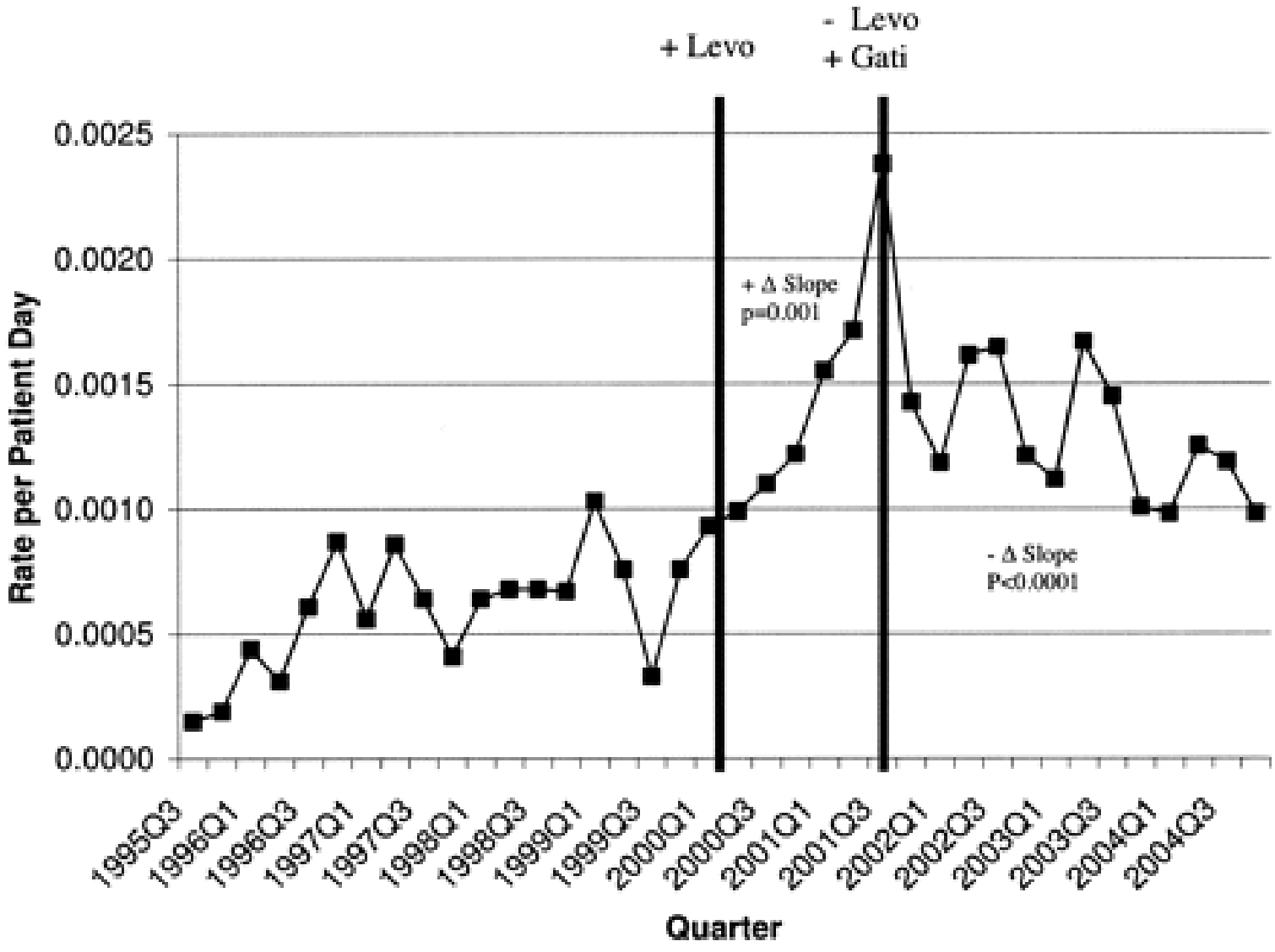
interrupted time series: $Y_1 Y_{\ldots} Y_{10} X_{11} Y_{11} Y_{\ldots} Y_{20}$



◇

◇ look at the trend: may be difficult to see response

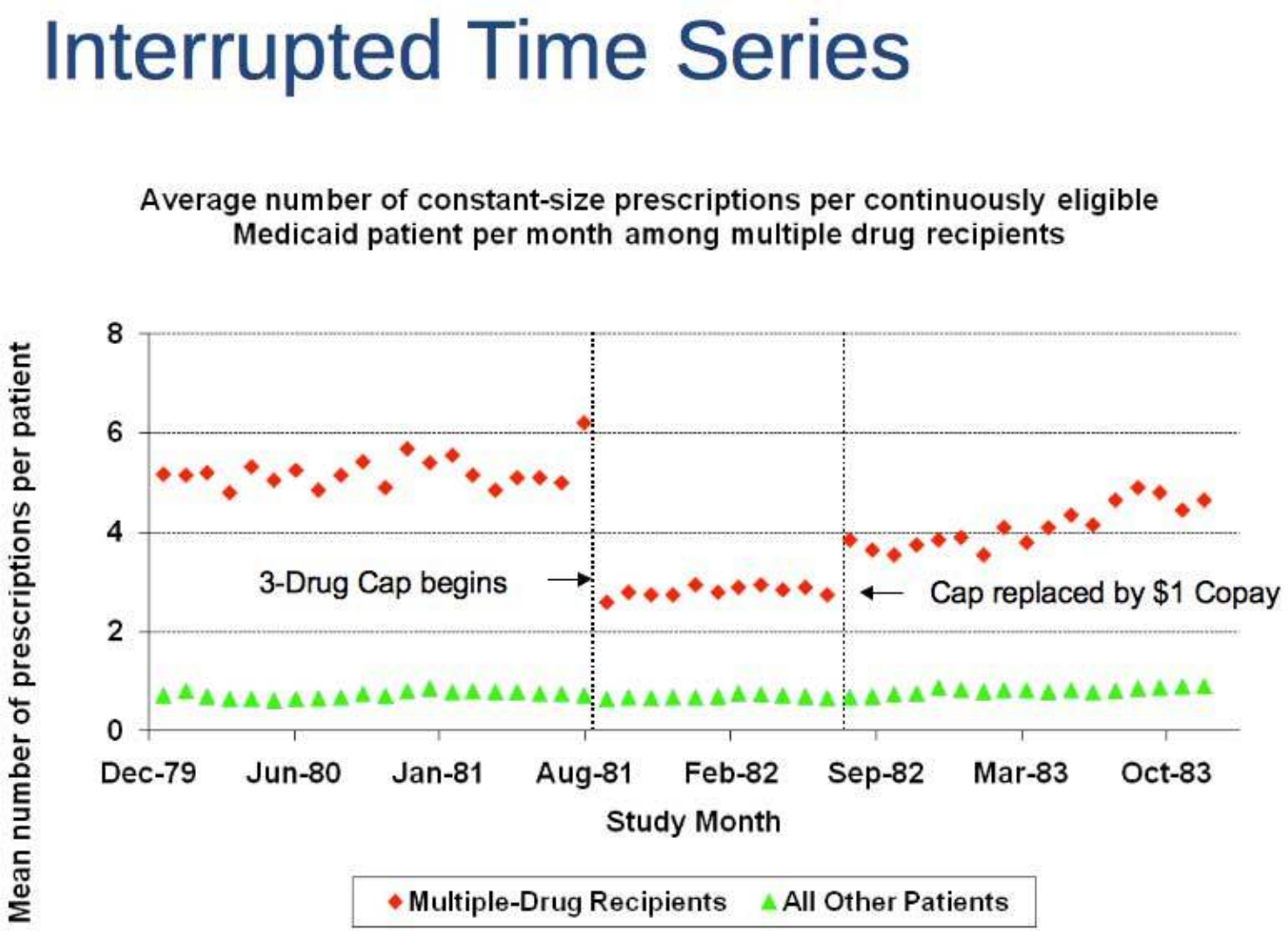
interrupted time series: $Y_1 Y_{\ldots} Y_{10} X_{11} Y_{11} Y_{\ldots} Y_{20}$



◇

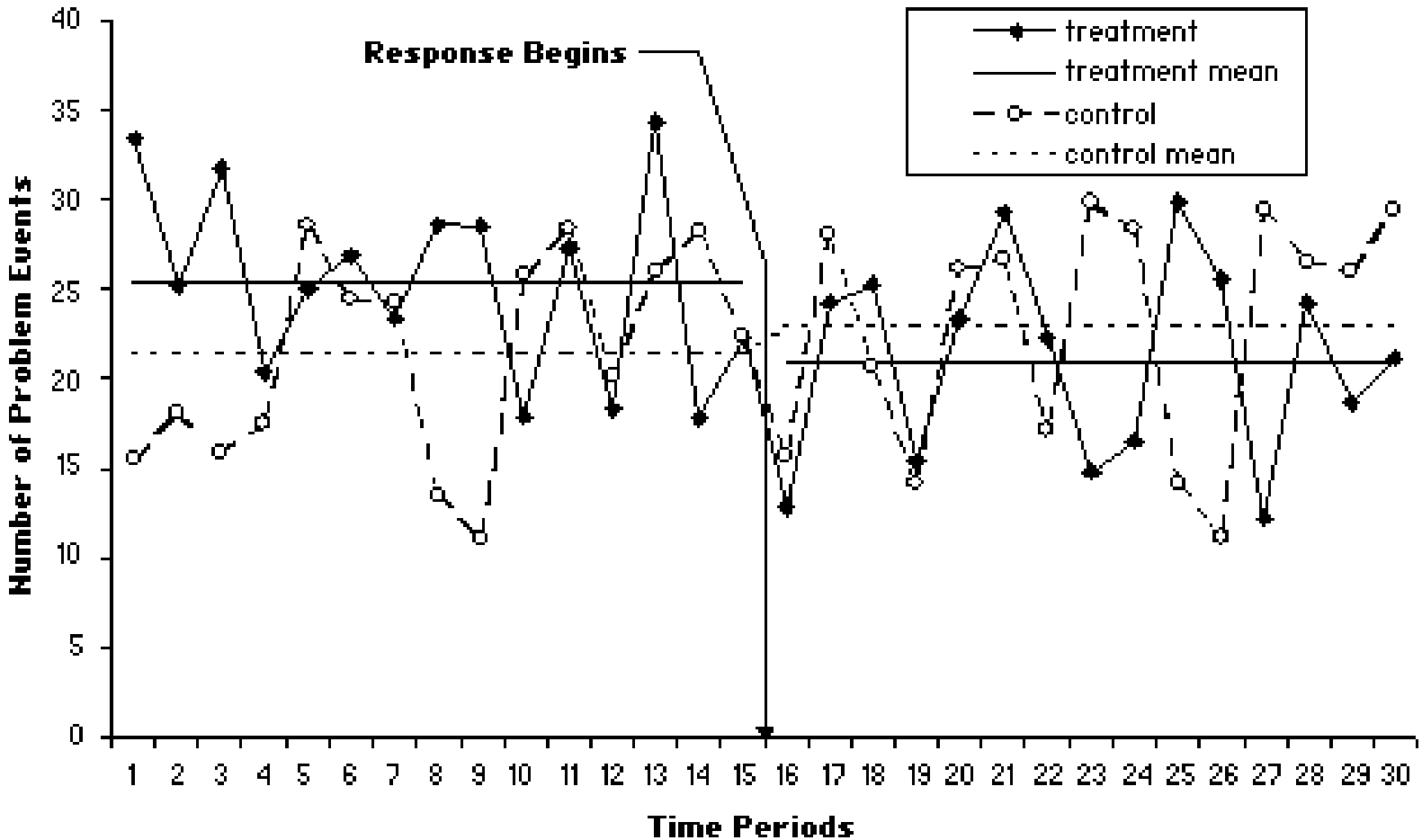
◇ more powerful: take away T→ effect dies

interrupted time series with a control



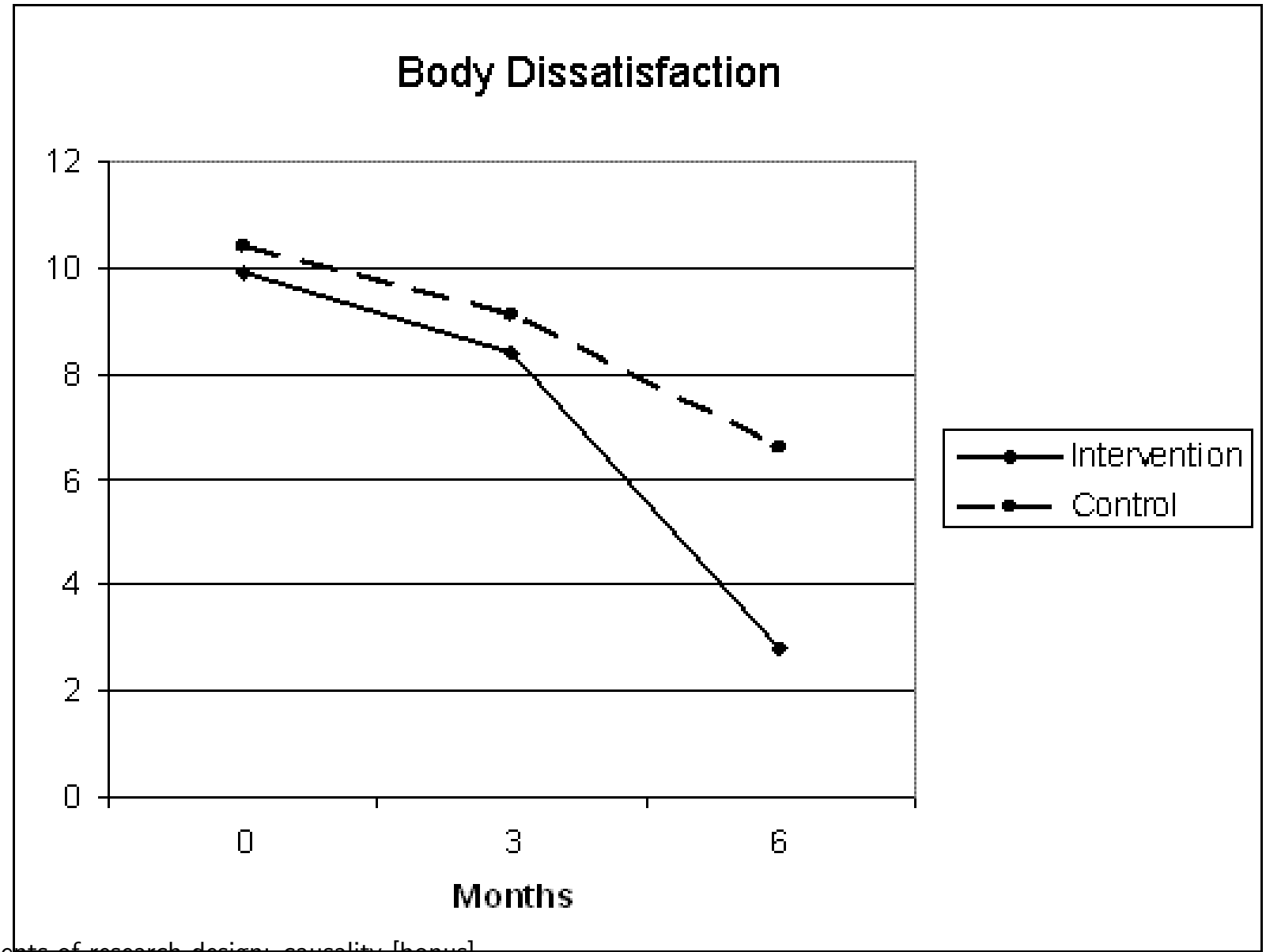
Adapted from: Soumerai et al, N Engl J Med 1987

interrupted time series with a control



◇

interrupted time series with a control



outline

closely related to design!

- if you have bad design, you'll have endogeneity
- curiously, it is something that economists are obsessed with
- but other fields aren't
- a good reference is Sorensen (2012)
  `http://people.bu.edu/tsimcoe/code/Endog-PDW.pdf`
- a. gujarati "a note on causality and exogeneity" ed5 p.657, ed4 p.701

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 previous research resign section
- so it can be just LOVB, or unobserved heterogeneity
- unobserved heterogeneity: see Rumsfeld's unknown unknowns in previous section; can use FE, RE, etc

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
- in any case as Jesper B Sorensen advocates (Sorensen, 2012) `http://people.bu.edu/tsimcoe/code/Endog-PDW.pdf`
- think deeply about the relationship between X and Y
- one of the best ways to think deeply, i think, is to use INUS condition

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

Mackie, J. and J. MacKie (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.

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