Statistics 222, Education 351A Autumn 2020

Statistical Methods for Longitudinal Research

Autumn 2020 Remote Asynchronous Instruction

David Rogosa Sequoia 224, rag{AT}stanford{DOT}edu
Course web page: http://rogosateaching.com/stat222//

To see full course materials from Autumn 2018 go here

Course Welcome and Logistics (first day stuff, to be posted in August, call it Week0)
Lecture slides, week 0 (pdf) **Audio companion, week 0**
For recreation of in-classroom experience, linked below are youtube versions of the music I play
before starting lecture and after lecture concludes. Some may wish to reverse that ordering.

```
Registrar's information
STATS 222 (Same as EDUC 351A): Statistical Methods for Longitudinal Research   Units: 2
Grading Basis: Letter or Credit/No Credit

Course Description:
 STATS 222: Statistical Methods for Longitudinal Research (EDUC 351A)
Research designs and statistical procedures for time-ordered (repeated-measures) data.
The analysis of longitudinal panel data is central to empirical research on learning, development, aging, and the effects of interventions.
Topics include: measurement of change, growth curve models, analysis of durations including survival analysis,
experimental and non-experimental group comparisons, reciprocal effects, stability.
See http://rogosateaching.com/stat222/. Prerequisite: intermediate statistical methods
Terms: Aut | Units: 2 | Grading: Letter or Credit/No Credit
Instructors: Rogosa, D. (PI)
```

Preliminary Course Outline

Week 1. Course Overview, Longitudinal Research; Analyses of Individual Histories and Growth Trajectories
Week 2. Introduction to Data Analysis Methods for assessing Individual Change for Collections of Growth Curves (mixed-effects models)
Week 3. Analysis of Collections of growth curves: linear, generalized linear and non-linear mixed-effects models
Week 4. Special case of time-1, time-2 data; Traditional measurement of change for individuals and group comparisons
Week 5. Assessing Group Growth and Comparing Treatments: Traditional Repeated Measures Analysis of Variance and Linear Mixed-effects Models
Week 6. Comparing group growth continued: Power calculations, Cohort Designs, Cross-over Designs, Methods for missing data, Observational studies.
Week 7. Analysis of Durations: Introduction to Survival Analysis and Event History Analysis
Weeks 8-9. Further topics in analysis of durations: Diagnostics and model modification; Interval censoring, Time-dependence, Recurrent Events, Frailty Models, Behavioral Observations and Series of Events (renewal processes)
Dead Week. Assorted Special Topics (enrichment) and Overflow (weeks 1-8): Assessments of Stability (including Tracking), Reciprocal Effects, (mis)Applications of Structural Equation Models, Longitudinal Network Analysis

Texts and Resources for Course Content

1. Garrett M. Fitzmaurice Nan M. Laird James H. Ware Applied Longitudinal Analysis (Wiley Series in Probability and Statistics; 2nd ed 2011)
Text Website second edition website Text lecture slides
2. Judith D. Singer and John B. Willett . Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence New York: Oxford University Press, March, 2003.
Text web page Text data examples at UCLA IDRE Powerpoint presentations good gentle intro to modelling collections of growth curves (and survival analysis) is Willett and Singer (1998)
3. Douglas M. Bates. lme4: Mixed-effects modeling with R February 17, 2010 Springer (chapters). A merged version of Bates book: lme4: Mixed-effects modeling with R January 11, 2010 has been refound
Manual for R-package lme4 and mlmRev, Bates-Pinheiro book datasets.
Additional Doug Bates materials. Collection of all Doug Bates lme4 talks Mixed models in R using the lme4 package Part 2: Longitudinal data, modeling interactions Douglas Bates 8th International Amsterdam Conference on Multilevel Analysis 2011-03-16 another version
Original Bates-Pinheiro text (2000). Mixed-Effects Models in S and S-PLUS (Stanford access). Appendix C has non-linear regression models.
Fitting linear mixed-effects models using lme4, *Journal of Statistical Software* Douglas Bates Martin Machler Ben Bolker. Technical topics: Mixed models in R using the lme4 package Part 4: Theory of linear mixed models
4. A handbook of statistical analyses using R (second edition). Brian Everitt, Torsten Hothorn CRC Press, Index of book chapters Stanford access
Longitudinal chapters: Chap11 Chap12 Chap13. Data sets etc Package 'HSAUR2' August 2014, Title A Handbook of Statistical Analyses Using R (2nd Edition)
There is now a third edition of HSAUR, but full text not yet available in crcnetbase.com. CRAN HSAUR3 page with Vignettes (chapter pieces) and data in reference manual
5. Peter Diggle , Patrick Heagerty, Kung-Yee Liang , Scott Zeger. Analysis of Longitudinal Data 2nd Ed, 2002
Amazon page Peter Diggle home page Book data sets
A Short Course in Longitudinal Data Analysis Peter J Diggle, Nicola Reeve, Michelle Stanton (School of Health and Medicine, Lancaster University), June 2011 earlier version associated exercises: Lab 1 Lab2 Lab3
6. Longitudinal and Panel Data: Analysis and Applications for the Social Sciences by Edward W. Frees (2004). Full book available and book data and

programs (mostly SAS).
7. Growth Curve Analysis and Visualization Using R. Daniel Mirman Chapman and Hall/CRC 2014 Print ISBN: 978-1-4665-8432-7 Stanford Access Mirman web page (including data links).
8. *Longitudinal Data Analysis* Edited by Geert Verbeke , Marie Davidian , Garrett Fitzmaurice , and Geert Molenberghs Chapman and Hall/CRC 2008. online supplement for LDA book .
9. Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. Springer Series in Statistics. New-York: Springer. Extended presentation: Introduction to Longitudinal Data Analysis A shorter exposition: Methods for Analyzing Continuous, Discrete, and Incomplete Longitudinal Data
10. Survival analysis Rupert G. Miller. Available as Stanford Tech Report
11. Event History Analysis with R (Stanford access). Goran Brostrom CRC Press 2012. R-package `eha`
12. John D. Kalbfleisch , Ross L. Prentice The Statistical Analysis of Failure Time Data 2nd Ed
Amazon page online from Wiley
13. Klein J, Moeschberger M (2003). Survival Analysis, 2nd edition. New York: Springer.
14. Therneau TM, Grambsch PM (2000). Modeling Survival Data: Extending the Cox Model. New York: Springer.
15. Advanced survival analysis topics.
Interval-Censored Time-to-Event Data Methods and Applications Chapman and Hall/CRC 2012 (esp Chap 14--glrt).
Recurrent Events: Chapter 9 of Kalbfleisch and Prentice (2nd edition), "Modeling and Analysis of Recurrent Event Data".
Cook, R. J. and Lawless, J. F. (2007). The Statistical Analysis of Recurrent Events. (Stanford access) Springer, New. York.
Joint Models for Longitudinal and Time-to-Event Data. With Applications in R. Dimitris Rizopoulos. Chapman and Hall/CRC 2012(Stanford access) Book website

Additional Specialized Resources
Harvey Goldstein. The Design and Analysis of Longitudinal Studies: Their Role in the Measurment of Change (1979). Elsevier
Amazon page Goldstein Chap 6 Repeated measures data Multilevel Statistical Models by Harvey Goldstein with data sets
David Roxbee Cox, Peter A. W. Lewis The statistical analysis of series of events. Chapman and Hall, 1966
Google books Poisson process computing program
David J Bartholomew. Stochastic Models for Social Processes, Chichester 3rd edition: John Wiley and Sons.
David J Bartholomew web page

Grading, Exams, and Credit Units
Stat222/Ed351A is listed as Letter or Credit/No Credit grading for 2-units
For Autumn 2020 grading for the 2-units will be based on a 'take home'(i.e. do at home) Problem Set.
Each week are posted a few exercises for that week's content--towards the end of the qtr I will identify a subset of those exercises to be turned in.
Those selected problems will constitute the graded Problem Set.
Also as you will see, for each week's content a number of Review Questions with Solutions are posted.

Course Problem Set 2020 to be posted

Cumulative Collection of Course Handounts 2020 to be posted

Statistical computing
Class presentation will be in, and students are encouraged to use, R (occasionally, some references to SAS and Mathematica).
Current version of R is R version 4.0.2 (Taking Off Again) released 2020-06-22.
For references and software: The R Project for Statistical Computing Closest download mirrors in the past, UCLA and Berkeley, seem no longer avaliable, pick your fave anywhere in the world.
The CRAN Task View: Statistics for the Social Sciences provides an overview of some relevant R packages. Also the new CRAN Task View: Psychometric Models and Methods and CRAN Task View: Survival Analysis and CRAN Task View: Computational Econometrics.
A good R-primer on various applications (repeated measures and lots else). Notes on the use of R for psychology experiments and questionnaires Jonathan Baron, Yuelin Li. Another version
A Stat209 text, Data analysis and graphics using R (2007) J. Maindonald and J. Braun, Cambridge 2nd edition 2007. 3rd edition 2010 has available a short version in CRAN .
According to Peter Diggle: "The best resource for R that I have found is Karl Broman's Introduction to R page."

Course Content: Files, Readings, Examples

Week 1. First class: Longitudinal Research Overview, Analysis of Individual Trajectories.

Lecture slides, week 1 (pdf)
Audio companion, week 1
parta partb partc

Lecture Topics
A. Longitudinal research overview
B. Examples, illustrations for longitudinal research overview, taken from course resources above:
Laird,Ware (#1) slides 1-16; Diggle (#5) slides 4-14, 22-28 Verbeke (#9) slides from Ch 2 and Sec3.3
C. Data Analysis Examples of Model Fitting for Individual Trajectories and Histories.
Motto: Individual trajectories are the proper starting point for longitudinal data analysis
ascii version of class handout annotated version pdf version with plots datasets
Starting up R-addendum: installing packages and obtaining data (sleepstudy in lme4)

Additional materials for the trajectory examples
For Count Data (glm) example. Link functions for generalized linear mixed models (GLMMs), Bates slides (pdf pages 11-18)

Week 6. Comparing Group Growth, continued. Observational Studies, Cohort Designs.

Lecture Topics

1. Observational Studies: Group Comparisons in Longitudinal Observational (non-experimental, "quasi"-experimental) Designs

A. Regression adjustments in quasi-experiments.

Technical resource: Weisberg, H. I. Statistical adjustments and uncontrolled studies. Psychological Bulletin, 1979, 86, 1149-1164. class handout

B. Lord's paradox; pre-post group comparisons. Lord notes

Publications: Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304-305.

Wainer, H. (1991). Adjusting for differential base rates: Lord's Paradox again. Psychological Bulletin, 109, 147-151.

C. Economist's differences in differences (or diffs in diffs with matching) for observational studies. class slide

A very popular subject these days. Pretty good Wiki page LSE slides

Austin Nichols slides. Causal inference with observational data A brief review of quasi-experimental methods July 2009

Angrist Ch 5, MHE. Card and Krueger (1994) data, minumimum wage ex

paper On the Use of Linear Fixed Effects Regression Models for Causal Inference (sec 3.2)

R-package `did`

D. Interrupted time-series.

Intros: Interrupted Time Series Quasi-Experiments Gene V Glass Arizona State University.

Assessing impact of stewardship: the why, when, and how of interrupted time series

Time Series Analysis with R section 4.6 December 2012 Handbook of Statistics 30(1):661-712

ITSpages for closing time example Class example: Closing time (glm kludge)

Rogosa R-session

Original publication (ozone data):

Box, G. E. P. and G. C. Tiao. 1975. Intervention Analysis with Applications to Economic and Environmental Problems." Journal of the American Statistical Association. 70:70-79. SAS example for ozone data

Applications:

Did fertility go up after the Oklahoma City bombing? An analysis of births in metropolitan counties in Oklahoma, 1990-1999. Demography, 2005.

Box-tiao time series models for impact assessment Evaluation Quarterly 1979

Interrupted time-series analysis and its application to behavioral data Donald P. Hartmann, John M. Gottman, Richard R. Jones, William Gardner, Alan E. Kazdin, and Russell S. Vaught J Appl Behav Anal. 1980 Winter; 13(4): 543-559.

Segmented regression analysis of interrupted time series studies in medication use research. By: Wagner, A. K.; Soumerai, S. B.; Zhang, F.; Ross-Degnan, D.. Journal of Clinical Pharmacy & Therapeutics, Aug2002, Vol. 27 Issue 4, p299-309,

R-packages:

`tscount`, vignette `BayesSingleSub`: Computation of Bayes factors for interrupted time-series designs

New resource, `Package Wats` Oklahoma City Fertility analyses

E. Longitudinal Treatments (exposures)

Propensity Scores for Repeated Treatments A Tutorial for the iptw Function in the TWANG Package

Using cobalt with Longitudinal Treatments

F. Value-added analysis

Value-added does New York City. New York schools release 'value added' teacher rankings

from the unions: THIS IS NO WAY TO RATE A TEACHER

Value-Added Models to Evaluate Teachers: A Cry For Help H Wainer, Chance, 2011.

American Statistical Association Statement on Using Value-Added Models for Educational Assessment

2. Cohort effects. Cohort-sequential, Accelerated longitudinal designs.

Robinson, K., Schmidt, T. and Teti, D. M. (2008) Issues in the Use of Longitudinal and Cross-Sectional Designs, in Handbook of Research Methods in Developmental Science (ed D. M. Teti), Blackwell Publishing Ltd, Oxford, UK

3. Econometric Approaches to Longitudinal Panel Data.

vignette for Panel Data Econometrics in R: The plm Package Yves Croissant Giovanni Millo (esp. section 7. "plm versus nlme/lme4").

R-package `plm` Class handout Maybe more in Week 10.

WEEK 6 Review Questions

1. Interrupted Time Series example, redux

Create a version of the its 'closing time' example presented in class (example linked above) with the 50 months before intervention having mean fatality = 1 and after intervention mean fatality = 2.

Carry out the glm approximation to the time series analysis.

Solution for Review Question 1

2. Observational Studies: Lord's Paradox.

Part 1. Lord's paradox example

a. construct a two-group pre-post example with 20 observations in each group that mimics the description in Lord (1967):

statistician 1 (difference scores) obtains 0 group effect

statistician 2 (analysis of covariance) obtains large group effect for the group higher on the pre-existing differences in pretest

b. construct second example for which

statistician 1 (difference scores) obtains large group effect

statistician 2 (analysis of covariance) obtains 0 group effect

c. construct a third example (if possible) for which

statistician 1 (difference scores) obtains large postive group effect

statistician 2 (analysis of covariance) obtains large negative group effect

Part 2. Group Comparisons by repeated measures analysis of variance or lmer

For the examples in part 1, (a and c), carry out the group comparison (i.e. is there differential change?) for the artificial data using a repeated measures anova (one within, one between factor) or lmer equivalent.

Demonstrate the equivalence from Brogan-Kutner paper that testing the groupXtime interaction term is equivalent to a t-test between groups on individual improvement (i.e. a statistician 1 analysis).

Solution to Question 2 RQ2 solution

3. Observational Studies: Regression Adjustments. *more in week 10*
The display from lecture of the regression adjustments also has a numerical example (page 2 of pdf). Recreate the results shown for the Anderson et al Head Start example. Also for lecture materials, Regression Adjustments with Non-equivalent groups Week 6, show the Belson adjustment procedure (using control group slope) is equivalent to evaluating the vertical distance between the within-group regression fits at the mean of the treatment group. written out proof.

4. Time 1 Time 2 observational data, Differences in Differences analysis.
We reuse some time-1, time-2 observational data generated to illustrate Lord's paradox (RQ2) -- gender differences in weight gain. (The 'paradox' is solved by Holland, Wainer, Rubin using potential outcomes.) The set up for these artificial data is females gain, males no change

```
  corr .7 within gender, equal vars time1 time 2 within gender
means
                M               F
X (t1)        170             120
Y (t2)        170             130
```

comparison of "gains" 170 - 170 - (130 - 120) = -10 negative effect males (females gain more).
ancova: 170 - 130 - .7*(170 - 120) = 5 positive male effect
So: does being male cause a student to gain weight or lose weight? Illustrate forms of diffs-in-diffs analyses.
wide form for these data long form for these data

Solution for Review Question 4

WEEK 6 Exercises

1. Smoking and Lung Function. Data and description available here (from ALA main page Datasets/ Vlagtwedde-Vlaardingen Study). From these (rather meager) data, what do you make of the effect of being a (self-selected) smoker vs non-smoker throughout the 19 years of this study.
2. Longitudinal Observational Study: Wages for High School Drop-outs. Data obtained from the National Longitudinal Survey on Youth can be used to look-at the labor-market experiences of high school drop-outs. The subset of these data we will use is available at UCLA-- it's a csv file here's the appropriate read.table statement.

```
read.table("http://stats.idre.ucla.edu/stat/r/examples/alda/data/wages_pp.txt", header=T, sep=",")

'data.frame':   6402 obs. of  15 variables:
 $ id           : int  31 31 31 31 31 31 31 31 36 36 ...
 $ lnw          : num  1.49 1.43 1.47 1.75 1.93 ...
 $ exper        : num  0.015 0.715 1.734 2.773 3.927 ...
 $ ged          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ postexp      : num  0.015 0.715 1.734 2.773 3.927 ...
 $ black        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ hispanic     : int  1 1 1 1 1 1 1 1 0 0 ...
 $ hgc          : int  8 8 8 8 8 8 8 8 9 9 ...
 $ hgc.9        : int  -1 -1 -1 -1 -1 -1 -1 -1 0 0 ...
 $ uerate       : num  3.21 3.21 3.21 3.29 2.9 ...
 $ ue.7         : num  -3.79 -3.79 -3.79 -3.71 -4.11 ...
 $ ue.centert1  : num  0 0 0 0.08 -0.32 ...
 $ ue.mean      : num  3.21 3.21 3.21 3.21 3.21 ...
 $ ue.person.cen: num  0 0 0 0.08 -0.32 ...
 $ ue1          : num  3.21 3.21 3.21 3.21 3.21 ...
```

Variables we will use are `id`, log-wage (hourly) `lnw` for each observation, `exper` time in labor force to the nearest day (in years), `black` (isblack = 1), `hcg` (highest grade completed; note hgc.9 is hgc - 9)
a. How many individuals in this data set? Give a five-number summary of the number of observations per person. How many of the individuals in these data have black = 1?
b. SFYS descriptive analyses. We are interested in wages (measured by lnw) as a function of experience (lnw ~ exper). Show a five-number summary of the gradient (slope; i.e. change in log-wage for unit change in exper) and level (here fit for exper = 0, initial status) for the set of individuals. Then stratify on black = 1 vs black = 0 (combining the white and hispanic high school drop-outs). Also show side-by-side boxplots for gradient and initial level stratifying on `black`. What do these displays indicate. Also show a plot of the lnw ~ exper fits seperately for black = 1 and black = 0. What do these analyses and displays indicate?
c. Use a formal mixed-effects model analysis to obtain random and fixed effects for the `lnw~ exper` individual level model. Obtain a point estimate and confidence interval for the variance of gradients. Does the bootstrap CI differ from the profile CI?
d. Are there differences in the lnw ~ exper relation for students black = 1 vs black = 0? Show by estimates and confidence intervals from mixed-effects models.
e. Does inclusion of `hgc` information confirm or alter your indications in part d?
3. Observational study: Gender differences in Vocabulary learning data-- see Week 3 problem 2-- from test results on file in the Records Office of the Laboratory School of the University of Chicago. Source D R Bock, MSMBR. The data consist of scores, obtained from a cohort of pupils at the eigth through eleventh gade level on alternative forms of the vocabulary section of the Cooperative Reading Tests." There are 64 students in all, 36 male, 28 female (ordered) each with four equally spaced observations (test scores). Wide form of these data are in BOCKwide.dat and I kindly also made a long-form version BOCKlong.dat .
For this problem consider gender differences in Vocabulary growth. Obtain the means (over persons) and plot the group growth curves, separately by gender. Does there appear to be curvature (i.e. deceleration in vocabulary skill growth) for both males and females? Construct an lmer model with the individual growth curve a quadratic function of grade (year), most convenient to use uncorrelated predictors `grade - mean(grade)` and `(grade - mean(grade))^2`. In the level II model allow each of the three parameters of the individual quadratic curves to differ by gender. Fit the lmer model and interpret the fixed and random effects you obtain. Compare the results with a lmer model in which the individual trajectories are straight-line. Use the `anova` model comparison functionality in R (e.g. `anova(modLin, modQuad)` to test whether the quadratic function for individual growth produces a better model fit.
4. Observational Studies: Regression Adjustments.
The class handout on regression adjustments shown in class and linked in RQ3 above contained summary statistics for the Head Start data considered in Anderson et al (1980) *Statistical Methods for Comparative Studies*. I constructed a corresponding data set located at W6prob3dat
Try out the various regression adjustments described on the handout for these pretest-posttest data. (Handout shows some approximate estimates). Also

show the result for the basic diffs-in-diffs estimator from Week 6.

Week 7. Analysis of Durations: Introduction to Survival Analysis (aka *event history*) Methods

Gmail Search mail

Compose

Inbox 254
Starred
Snoozed
Sent
Drafts
More

David

1 of 1,485

[dbds-affiliated-faculty] [dbds-all-faculty] [dbds-events] [qsu_research_methods_seminar] QSU Research Methods Seminar: Charles McCulloch, PhD - Tuesday November 6 at 4pm - 1070 Arastradero Road, Room 109, Palo Alto

Inbox ×

Ni Deng <nideng@stanford.edu> 2:45 PM (2 minutes ago)
to Ni

QSU Research Methods Seminar

Tuesday, November 6 at 4:00pm

Location: 1070 Arastradero Road, #109, Palo Alto, CA 94304

or join the seminar remotely via: https://stanford.zoom.us/j/858345929

Charles McCulloch, PhD
Professor and Head of Biostatistics Division
Department of Epidemiology & Biostatistics
University of California San Francisco

"Improving the prediction of extreme clusters in multilevel data"

Predicted random effects are widely used to evaluate the performance of and rank clusters such as patients and hospitals using longitudinal and multilevel data. For example, the Center for Medicare and Medicaid Services uses these approaches to estimate hospital-specific performance metrics such as mortality or re-admission rates. Using predicted random effects from mixed regression models has been shown to outperform using standard regression to estimate hospital specific effects. However, predicted random effects are often used to identify extreme values such as poorly performing hospitals and their performance has not been systematically evaluated in this context. After introducing the ideas of mixed models and traditional methods of best prediction, I motivate new methods to predict extreme clusters and evaluate their performance. I show that methods that assume distributions with heavier tails than the normal distribution can produce predicted values with smaller mean square or absolute error of prediction than standard best predicted values when interest focuses on extreme clusters. The methods are illustrated using data on length of stay in hospitals.

****

Refreshments will be served. Please note that parking at 1070 Arastradero is free!
We welcome any suggestions or questions about the seminar series - please contact Ni Deng (nideng@stanford.edu).

We look forward to seeing you!

If you are unable to join us in person, please watch the seminar remotely at https://stanford.zoom.us/j/858345929

** To get to 1070 Arastradero:*

Driving: enter the driveway and take the first left into the parking lot and park (PARKING IS FREE!). Enter through the front door - you will be in the lobby. Take the door on the right and follow it toward the back of the building. Room 109 will be on your left before you reach the glass door leading outside.
Shuttle: 1070 Arastradero is served by the Marguerite Shuttle, which picks up at the medical center on Pasteur Drive.

No recent chats
Start a new one

Examples

```
(r5 <- GHrule(5, asMatrix=FALSE))
## second, fourth, sixth, eighth and tenth central moments of the
## standard Gaussian density
with(r5, sapply(seq(2, 10, 2), function(p) sum(w * z^p)))
```

---

glmer — *Fitting Generalized Linear Mixed-Effects Models*

Description

Fit a generalized linear mixed-effects model (GLMM). Both fixed effects and random effects are specified via the model `formula`.

Usage

```
glmer(formula, data = NULL, family = gaussian, control = glmerControl(),
      start = NULL, verbose = 0L, nAGQ = 1L, subset, weights, na.action,
      offset, contrasts = NULL, mustart, etastart,
      devFunOnly = FALSE, ...)
```

Arguments

`formula` a two-sided linear formula object describing both the fixed-effects and random-effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. Random-effects terms are distinguished by vertical bars (`"|"`) separating expressions for design matrices from grouping factors.

`data` an optional data frame containing the variables named in `formula`. By default the variables are taken from the environment from which `lmer` is called. While `data` is optional, the package authors *strongly* recommend its use, especially when later applying methods such as `update` and `drop1` to the fitted model (*such methods are not guaranteed to work properly if* `data` *is omitted*). If `data` is omitted, variables will be taken from the environment of `formula` (if specified as a formula) or from the parent frame (if specified as a character vector).

`family` a GLM family, see `glm` and `family`.

`control` a list (of correct class, resulting from `lmerControl()` or `glmerControl()` respectively) containing control parameters, including the nonlinear optimizer to be used and parameters to be passed through to the nonlinear optimizer, see the `*lmerControl` documentation for details.

`start` a named list of starting values for the parameters in the model, or a numeric vector. A numeric `start` argument will be used as the starting value of `theta`. If `start` is a list, the `theta` element (a numeric vector) is used as the starting value for the first optimization step (default=1 for diagonal elements and 0 for off-diagonal elements of the lower Cholesky factor); the fitted value of `theta` from the first step, plus `start[["fixef"]]`, are used as starting values for the

second optimization step. If `start` has both `fixef` and `theta` elements, the first optimization step is skipped. For more details or finer control of optimization, see `modular`.

`verbose` integer scalar. If `> 0` verbose output is generated during the optimization of the parameter estimates. If `> 1` verbose output is generated during the individual penalized iteratively reweighted least squares (PIRLS) steps.

`nAGQ` integer scalar - the number of points per axis for evaluating the adaptive Gauss-Hermite approximation to the log-likelihood. Defaults to 1, corresponding to the Laplace approximation. Values greater than 1 produce greater accuracy in the evaluation of the log-likelihood at the expense of speed. A value of zero uses a faster but less exact form of parameter estimation for GLMMs by optimizing the random effects and the fixed-effects coefficients in the penalized iteratively reweighted least squares step. (See Details.)

`subset` an optional expression indicating the subset of the rows of `data` that should be used in the fit. This can be a logical vector, or a numeric vector indicating which observation numbers are to be included, or a character vector of the row names to be included. All observations are included by default.

`weights` an optional vector of 'prior weights' to be used in the fitting process. Should be `NULL` or a numeric vector.

`na.action` a function that indicates what should happen when the data contain `NA`s. The default action (`na.omit`, inherited from the 'factory fresh' value of `getOption("na.action")`) strips any observations with any missing values in any variables.

`offset` this can be used to specify an *a priori* known component to be included in the linear predictor during fitting. This should be `NULL` or a numeric vector of length equal to the number of cases. One or more `offset` terms can be included in the formula instead or as well, and if more than one is specified their sum is used. See `model.offset`.

`contrasts` an optional list. See the `contrasts.arg` of `model.matrix.default`.

`mustart` optional starting values on the scale of the conditional mean, as in `glm`; see there for details.

`etastart` optional starting values on the scale of the unbounded predictor as in `glm`; see there for details.

`devFunOnly` logical - return only the deviance evaluation function. Note that because the deviance function operates on variables stored in its environment, it may not return *exactly* the same values on subsequent calls (but the results should always be within machine tolerance).

`...` other potential arguments. A `method` argument was used in earlier versions of the package. Its functionality has been replaced by the `nAGQ` argument.

Details

Fit a generalized linear mixed model, which incorporates both fixed-effects parameters and random effects in a linear predictor, via maximum likelihood. The linear predictor is related to the conditional mean of the response through the inverse link function defined in the GLM `family`.

The expression for the likelihood of a mixed-effects model is an integral over the random effects space. For a linear mixed-effects model (LMM), as fit by `lmer`, this integral can be evaluated

exactly. For a GLMM the integral must be approximated. The most reliable approximation for GLMMs is adaptive Gauss-Hermite quadrature, at present implemented only for models with a single scalar random effect. The `nAGQ` argument controls the number of nodes in the quadrature formula. A model with a single, scalar random-effects term could reasonably use up to 25 quadrature points per scalar integral.

Value

An object of class `merMod` (more specifically, an object of *subclass* `glmerMod`) for which many methods are available (e.g. `methods(class="merMod")`)

See Also

`lmer` (for details on formulas and parameterization); `glm` for Generalized Linear Models (*without* random effects). `nlmer` for nonlinear mixed-effects models.

`glmer.nb` to fit negative binomial GLMMs.

Examples

```
## generalized linear mixed model
library(lattice)
xyplot(incidence/size ~ period|herd, cbpp, type=c('g','p','l'),
       layout=c(3,5), index.cond = function(x,y)max(y))
(gm1 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd),
              data = cbpp, family = binomial))
## using nAGQ=0 only gets close to the optimum
(gm1a <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd),
               cbpp, binomial, nAGQ = 0))
## using  nAGQ = 9  provides a better evaluation of the deviance
## Currently the internal calculations use the sum of deviance residuals,
## which is not directly comparable with the nAGQ=0 or nAGQ=1 result.
(gm1a <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd),
               cbpp, binomial, nAGQ = 9))

## GLMM with individual-level variability (accounting for overdispersion)
## For this data set the model is the same as one allowing for a period:herd
## interaction, which the plot indicates could be needed.
cbpp$obs <- 1:nrow(cbpp)
(gm2 <- glmer(cbind(incidence, size - incidence) ~ period +
    (1 | herd) +  (1|obs),
              family = binomial, data = cbpp))
anova(gm1,gm2)

## glmer and glm log-likelihoods are consistent
gm1Devfun <- update(gm1,devFunOnly=TRUE)
gm0 <- glm(cbind(incidence, size - incidence) ~ period,
           family = binomial, data = cbpp)
## evaluate GLMM deviance at RE variance=theta=0, beta=(GLM coeffs)
gm1Dev0 <- gm1Devfun(c(0,coef(gm0)))
## compare
stopifnot(all.equal(gm1Dev0,c(-2*logLik(gm0))))
## the toenail oncholysis data from Backer et al 1998
```

```
Stat222 Oct'18 numerical convergence (ep3)

> ep3 <- glmer(seizure.rate ~  treatment + offset(per) + base +(period|subject),   data = epilepsy, family = "poisson")
Warning message:
In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge with max|grad| = 0.00460772 (tol = 0.001, component 1)
> summary(ep3)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: seizure.rate ~ treatment + offset(per) + base + (period | subject)
   Data: epilepsy

     AIC      BIC   logLik deviance df.resid
  1271.4   1316.5   -622.7   1245.4      223

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.23598 -0.45284 -0.04729  0.41114  1.74768

Random effects:
 Groups  Name        Variance Std.Dev. Corr
 subject (Intercept) 0.2813   0.5304
         period.L    0.1193   0.3454   -0.03
         period.Q    0.1414   0.3760   -0.43 -0.59
         period.C    0.1183   0.3439   -0.30 -0.18  0.10
Number of obs: 236, groups:  subject, 59

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.297014   0.138704   2.141   0.0322 *
treatmentProgabide -0.328191   0.145136  -2.261   0.0237 *
base                0.025385   0.002555   9.937   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) trtmnP
trtmntPrgbd -0.513
base        -0.660 -0.018
convergence code: 0
Model failed to converge with max|grad| = 0.00460772 (tol = 0.001, component 1)

> ep3 <- glmer(seizure.rate ~  treatment + offset(per) + base +(period|subject),   data = epilepsy, family = "poisson", nAGQ = 9)
Error in updateGlmerDevfun(devfun, glmod$reTrms, nAGQ = nAGQ) :
  nAGQ > 1 is only available for models with a single, scalar random-effects term

# the nAGQ switch is only avaible at present for single random factor-- here (1|subject) for random effects like ep2, ep4

> ep2 <- glmer(seizure.rate ~  treatment + offset(per) + base +(1|subject),   data = epilepsy, family = "poisson", nAGQ = 20)
```

```
#no warning
> ep4 <- glmer(seizure.rate ~ base + age + treatment + offset(per) +(1|subject), data = epilepsy, family = "poisson")
Warning message:
In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge with max|grad| = 0.00116699 (tol = 0.001, component 1)
> ep4 <- glmer(seizure.rate ~ base + age + treatment + offset(per) +(1|subject), data = epilepsy, family = "poisson", nAGQ = 20)
# no warning if you up nAGQ, no noticable increase in computation time

## doing (as.numeric(period)|subject) as random effect doesn't help

## what does help ep3 is downgrading the required numerical accuracy, but you can see (.328 to .322, 2%) change in treatment effect est
# no convergence warning for

> ep3 <- glmer(seizure.rate ~  treatment + offset(per) + base +(period|subject),   data = epilepsy, family = "poisson", nAGQ = 0)
> summary(ep3)
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature, nAGQ = 0) ['glmerMod']
 Family: poisson  ( log )
Formula: seizure.rate ~ treatment + offset(per) + base + (period | subject)
   Data: epilepsy

     AIC      BIC   logLik deviance df.resid
  1271.8   1316.8   -622.9   1245.8      223

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.25823 -0.47510 -0.05909  0.38963  1.74261

Random effects:
 Groups  Name        Variance Std.Dev. Corr
 subject (Intercept) 0.2831   0.5321
         period.L    0.1182   0.3439   -0.03
         period.Q    0.1405   0.3748   -0.44 -0.60
         period.C    0.1178   0.3432   -0.31 -0.18  0.10
Number of obs: 236, groups:  subject, 59

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.349899   0.130787   2.675  0.00747 **
treatmentProgabide -0.321793   0.142556  -2.257  0.02399 *
base                0.024872   0.002383  10.436  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) trtmnP
trtmntPrgbd -0.535
base        -0.630 -0.030
```

OBSERVATIONAL DESIGNS

Stat 209 [redacted]

ANCOVA CNRL equations

precursor: t-test $\boxed{Y = \beta_0 + \beta_1 G + e}$ $\hat{\beta}_1 / se(\hat{\beta}_1)$

pooled t-test

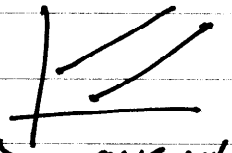
$G = 0, 1$ group membership

$$\boxed{Y = \gamma_0 + \gamma_1 G + \gamma_2 X + e}$$

ancova

$$\boxed{\hat{\gamma}_1 = \bar{Y}_1 - \bar{Y}_0 - \hat{\delta}_p (\bar{X}_1 - \bar{X}_0)}$$

constant treatment effect



$\hat{\gamma}_2 = \hat{\delta}_p$ ave w/in group slopes

---

more general model (CNRL)

$$Y = \beta_1 + \beta_2 G + \beta_3 X + \beta_4 XG$$

"interaction" term

Difference Scores, a good thing ~~week 9~~

$Y_2 - Y_1 = D$ outcome

② correlates of change r_{DW} W close group exogenous var

not $D \sim W + Y_1$ $\Rightarrow$ $Y_2 \sim W + Y_1$

③ pre post experiments $G = 0, 1$ (control, treat)

r_{DG} or t.test $(D \sim G)$

(B-K) equiv to repeated measures anova

do with lmList, lmer

④ Lord's paradox observational study

$Y_2 \sim Y_1 + G$ ancova

vs $D \sim G$ t-test on improvement gain

⑤ Diffs in diffs econ observational studies

diff 1 $D = Y_2 - Y_1$

diff 2 $D \sim G$

obs version of B-K matching?

1970's

Regression Adjustment in Quasi-experiments

outcome Y
group $G = 1, 0$

Head Start
Bryk, Weisburg
in Oakes

$$Y = \beta_0 + \beta_1 G + u$$

t-test.
but G, u not indep

adjustment (premeasure) X (pre-test?)

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0 - \hat{\beta}(\bar{X}_1 - \bar{X}_0)$$

$\hat{\beta} = 0$ t-test no adj

$\hat{\beta} = 1$ gain scores if X "pretest"

$\hat{\beta} = \hat{\beta}_{YX \cdot G}$ standard ancova (under adjusts, overadjusts)

$\hat{\beta} = \hat{\beta}_{YX \cdot G} / rel(X)$ adjustment for measurement error in slope

$\hat{\beta} = \hat{\beta}_{YX \cdot G} / r_{YX \cdot G} = \frac{S_{Y \cdot G}}{S_{X \cdot G}}$ validity correction Campbell-Erlebacher st. change scores

$\hat{\beta} = \hat{\beta}_0$ control slope
Belson equiv to $D(\bar{X}_1)$ adjustment

week 9

STANDARDIZED CHANGE SCORES

$G = 1\ T$
$\quad\ \ 0\ C$

Y post test
X pretest

---

change in treat — change in control

$$\bar{Y}_1 - \bar{X}_1 - \bar{Y}_0 - \bar{X}_0$$

or $\bar{Y}_1 - \bar{Y}_0 - (\bar{X}_1 - \bar{X}_0)$

Standardize Y and X (w in pooled sd at each time)

$$\frac{\bar{Y}_1 - \bar{Y}_0}{S_{Y \cdot G}} - \frac{\bar{X}_1 - \bar{X}_0}{S_{X \cdot G}}$$

$$\bar{Y}_1 - \bar{Y}_0 - \left(\frac{S_{Y \cdot G}}{S_{X \cdot G}}\right)\left[\bar{X}_1 - \bar{X}_0\right]$$

same as validity correction.

PROBLEM 4 Belson estimate

In Week 5 (see handout) and HW5 prob 7 we considered the Belson estimate for observational studies
Belson, W. A. (1956), “A technique for studying the effects of a television broadcast,” Applied Statistics, 5, 195–202.

AKA Peters-Belson, this estimate is widely used at present in health outcomes research and in wage-discrimination studies.
A representative statement (Medical Care Volume 42, Number 8, August 2004)

"The PB approach has been used in wage discrimination studies and race (sex) discrimination cases13 to predict the experience a minority (female) individual would have had if they were white (male). The conventional regression approach, which includes a dummy variable to identify race/ ethnicity, assumes a common amount (degree) of disparity for all minority group members regardless of their individual characteristics. In contrast, the PB method produces estimates of disparity for each minority group member by incorporating their individual characteristics. Our study explores how the PB approach can be similarly used to understand disparities in public health outcomes as illustrated from studying cancer screening."

In "Agnostic Notes On Regression Adjustments To Experimental Data: Reexamining Freedman’s Critique By Winston Lin UC Berkeley in The Annals of Applied Statistics, 2013, Vol. 7, No. 1, 295–318 p.299 states
" estimator, ATEinteract, can be computed as the estimated coefficient on Ti in the OLS regression of Yi on Ti, zi, and Ti(zi - z*). In the context of observational studies, Imbens and Wooldridge [(2009), pp. 28–30] give a theoretical analysis of ATEinteract, and a related method is known as the Peters–Belson or Oaxaca–Blinder estimator ... "

[notational notes for above.
Yi is outcome for individual i ;
Ti is the group indicator for individual i, as Ti = 1 for individuals in the treatment group, Ti = 0 for control
zi is the value of the covariate for individual i
z* my notation (was zoverbar in paper) covariate mean.
You may specify over what population or subpopulation this mean is taken over.

Can you show that ATEinteract is the Belson estimator as defined in class?
What assumption/specification do you need to make for ATEinterect to match the Belson estimate?

HW5 Belson Stat 209 week 5

refer to comparing regressions handout
CNRL Math notes

---

start with $D(\bar{X}_1)$ diff of regressions at treatment group mean on X

for easy notation make:

$\hat{\delta}_1$ the Y on X slope in treatment group

$\hat{\delta}_0$ the Y on X slope in control group

[in CNRL model in handout $\hat{\delta}_0 = \hat{\beta}_3$, $\hat{\delta}_1 = \hat{\beta}_3 + \hat{\beta}_4$]

$$D(\bar{X}_1) = \hat{\beta}_2 + \hat{\beta}_4 \bar{X}_1$$ from CNRL handout

$$= (\bar{Y}_1 - \hat{\delta}_1 \bar{X}_1) - (\bar{Y}_0 - \hat{\delta}_0 \bar{X}_0)$$ $\hat{\beta}_2$ part

$$+ (\hat{\delta}_1 - \hat{\delta}_0)\bar{X}_1$$ $\hat{\beta}_4$ part

cancel terms regroup

$$D(\bar{X}_1) = (\bar{Y}_1 - \bar{Y}_0) - \hat{\delta}_0 (\bar{X}_1 - \bar{X}_0)$$

as indicated in class handout use control group slope in ancova style adjustment.

---

small note of $\hat{\beta}_2$ part above

for OLS fit $\bar{Y} = a + b\bar{X}$ so $a = \bar{Y} - b\bar{X}$ term of control and treatment groups above.

numerical illustrations

Anderson et al (1980) Ch. 12 Table 12.1 Head Start Data

	Pre	Post	$r_{prepost}$	n
Innovative curriculum	17.1 (6.1)	23.3 (4.6)	.67	157
Standard Head Start	14.6 (6.2)	18.9 (5.8)	.78	669

pre diff 2.5 post diff 4.4

Inference?

$\hat{\beta}$ options

t-test	$\hat{\beta} = 0$	$\hat{\alpha} = 4.4$
gain	$\hat{\beta} = 1$	$\hat{\alpha} = 1.9$
ancova	$\hat{\beta} = \beta_{Y1X \cdot G} = .76$	$\hat{\alpha} = 2.5$
C-E	$\hat{\beta} = \frac{5.6}{6.2} = .9$	$\hat{\alpha} = 2.1$
Belson	$\hat{\beta} = .73$	$\hat{\alpha} = 2.57$

Analytic results

Weisberg (1979) Ancova bias positive or negative

Potential outcomes setup: w_i outcome if T $(Q=1)$

$\alpha_i = w_i - z_i$ treatment effect. z_i outcome if C $(Q=0)$

observable $y_i = z_i + Q\alpha$ (set $\alpha_i = \alpha$) $P = E(Q)$

for non-random assignment $(\rho_{zQ} \neq 0)$

$$\mu_{y_1} - \mu_{y_0} = \alpha + (\mu_{z_1} - \mu_{z_0})$$ ← selection bias

Can ancova with covariate X reduce or eliminate bias?

result. bias from ancova $\delta = (\mu_{z_1} - \mu_{z_0})\left(\frac{\rho_{zQ \cdot X}}{\rho_{zQ}}\right)\left(\frac{\sqrt{1-\rho_{zX}^2}}{\sqrt{1-\rho_{XQ}^2}}\right)$

H.I.W uses $\pi = \delta/(\mu_{z_1} - \mu_{z_0})$ "proportion bias"

Psychological Bulletin
1979, Vol. 86, No. 5, 1149–1164

Statistical Adjustments and Uncontrolled Studies

Herbert I. Weisberg
The Huron Institute, Cambridge, Massachusetts

Many evaluations of social interventions are based on uncontrolled assignments of individuals to treatment groups. Statistical adjustments are often used to compensate for naturally occurring differences between groups. There is much confusion and controversy about the adequacy of these statistical methods. A variety of interrelated problems have been identified, including measurement error, unequal growth rates across groups, and regression artifacts. In this article it is shown that these problems can all be subsumed under a general conceptual framework, as particular examples of model misspecification. This perspective is helpful in revealing clearly the nature of the problems posed by lack of experimental control. The important case of linear adjustment (analysis of covariance) is given special attention. An expression is derived for the proportion of bias remaining after adjustment, in terms of easily interpretable parameters. Implications of these results for research and evaluation design are considered.

To evaluate the effectiveness of a social intervention, the performance of a group receiving the "treatment" must be compared with a standard representing the expected performance in the absence of intervention. The fundamental problem in research design is to find a valid standard of comparison. Randomization is generally accepted as the ideal approach. That is, we use a random mechanism to assign individuals to either a treatment group or an untreated control group. Random selection virtually guarantees (at least for large samples) that the control group's performance will correspond to that of the treatment group without the intervention. So a straightforward comparison of mean outcomes for the two groups will provide an unbiased estimate of the treatment's effect.

Often, however, it is impossible to exercise experimental control. Complex social forces unknown to the investigator determine which individuals wind up in each of the groups. With such uncontrolled selection designs, the straightforward difference of group means may be a biased estimate of the effect. In these situations a variety of statistical methods have been proposed to compensate for this bias and thus provide an unbiased estimate. The analysis of covariance (ANCOVA) is perhaps most widely used for this purpose.

Recently there has been a great deal of concern about the adequacy of ANCOVA and other statistical adjustment procedures. Several investigators have shown that under models representing uncontrolled selection, the ANCOVA may either overadjust or underadjust (Bryk & Weisberg, 1977; Cain, 1975; Cochran & Rubin, 1973; Cronbach, Rogosa, Floden, & Price, Note 1). The estimates generated may in some instances be seriously misleading. It is even possible for the remaining bias after adjustment to be larger in absolute value than the initial bias without any adjustment.

Confusion over the adequacy of statistical adjustments is part of a larger debate about the usefulness of designs based on uncontrolled

This work was supported by Grant NIE-G-76-0090 from the National Institute of Education, U.S. Department of Health, Education and Welfare. However, points of view or opinions stated do not represent official NIE position or policy.

The author gratefully acknowledges the contribution of Anthony S. Bryk to the development of the ideas expressed in this article and his many helpful comments on earlier drafts. Sincere thanks also to Walt Haney, David Rogosa, and the referees for their valuable suggestions, many of which have been incorporated in the final version.

Requests for reprints should be sent to Herbert I. Weisberg, The Huron Institute, 123 Mount Auburn Street, Cambridge, Massachusetts 02138.

Weisberg Ancova Results

STAT 209
week 5

Potential Outcomes
person i

W_i outcome if T $(Q=1)$
Z_i outcome if C $(Q=0)$

Recast week 2
Holland
treatment/control
difference
$\alpha_i = W_i - Z_i$
(set $\alpha_i = \alpha$)

---

observable $Y_i = Z_i + Q_i \alpha$
$P = E(Q)$ prop in T

Non-random assignment $\rho_{ZQ} \neq 0$

observables
$\mu_{Y_1} - \mu_{Y_0}$
$= \alpha + (\mu_{Z_1} - \mu_{Z_0})$ selection bias

point-biserial ρ_{ZQ}
$$\mu_{Z_1} - \mu_{Z_0} = \frac{\sigma_Z}{\sqrt{P(1-P)}} \rho_{ZQ}$$

recall week 2 FACE results BIAS $= E(Y_c | S=t) - E(Y_c | S=c)$

Ancova with covariate X, saves the day?

(18) $E(Y|X,Q) = \mu + \beta X + (\delta + \alpha) Q$
observables

[note: matching works, eqs 9-12]

can only estimate $\delta + \alpha$, want α

bias from ancova
(19)
$$\delta = \rho_{ZQ \cdot X} \frac{\sigma_{Z \cdot X}}{\sigma_{Q \cdot X}} \quad (19)$$
$$= \rho_{ZQ \cdot X} \frac{\sigma_Z}{\sqrt{P(1-P)}} \frac{\sqrt{1-\rho_{XZ}^2}}{\sqrt{1-\rho_{XQ}^2}}$$

proportion of bias
$$\pi = \frac{\delta}{\mu_{Z_1} - \mu_{Z_0}} = \frac{\rho_{ZQ \cdot X}}{\rho_{ZQ}} \frac{\sqrt{1-\rho_{XZ}^2}}{\sqrt{1-\rho_{XQ}^2}}$$

---

ancova can overadjust or increase bias

WEISBERG

Table 1
Range of π for Different Combinations of ρ_{ZQ}, ρ_{XZ}, ρ_{ZX}

Basic situation	Case	Sign (ρ_{ZQ})	Sign (ρ_{XQ})	Sign (ρ_{ZX})	π
1	1	+	+	+	$-\infty$ to $+1$
	2	+	−	−	$-\infty$ to $+1$
	3	−	+	−	$-\infty$ to $+1$
	4	−	−	+	$-\infty$ to $+1$
2	5	−	−	−	1 to $+\infty$
	6	−	+	+	1 to $+\infty$
	7	+	−	+	1 to $+\infty$
	8	+	+	−	1 to $+\infty$

Case 1 right direction, but can overadjust

Case 2 adjustment in wrong direction

Lord's Paradox (1967 - forever?)

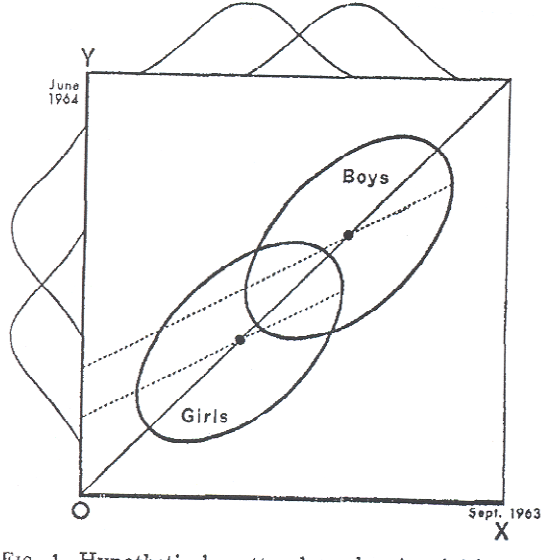


FIG. 1. Hypothetical scatterplots showing initial and final weight for boys and for girls.

A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex difference in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded.

At the end of the school year, the data are independently examined by two statisticians. Both statisticians divide the students according to sex. The first statistician examines the mean weight of the girls at the beginning of the year and at the end of the year and finds these to be identical. On further investigation, he finds that the frequency distribution of weight for the girls at the end of the year is actually the same as it was at the beginning.

He finds the same to be true for the boys. Although the weight of individual boys and girls has usually changed during the course of the year, perhaps by a considerable amount, the group of girls considered as a whole has not changed in weight, nor has the group of boys. A sort of dynamic equilibrium has been maintained during the year.

The whole situation is shown by the solid lines in the diagram. Here the two ellipses

Psychological Bulletin
1967, Vol. 68, No. 5, 304–305

A PARADOX IN THE INTERPRETATION OF GROUP COMPARISONS

FREDERIC M. LORD

Educational Testing Service

Attention is called to a basic source of confusion in the interpretation of certain types of group comparison data.

It is common practice in behavioral research, and in other areas, to apply the analysis of covariance in the investigation of preexisting natural groups. The research worker is usually interested in some criterion variable (y) and would like to make allowances for the fact that his groups are not matched on some important independent variable or "control" variable (x). The situation is such that observed differences in the dependent variable might logically be caused by differences in the independent variable, and the research worker wishes to rule out this possibility.

It is widely recognized that ideally the research worker should assign cases or individuals at random to the groups that are to be studied by analysis of covariance. In behavioral research and in many other areas, such random assignment is usually difficult or impossible—as, for example, in a comparison of the educational achievements of different racial groups. The research worker usually uses analysis of covariance regardless, or he may try to resort to a simple and direct interpretation of group means.

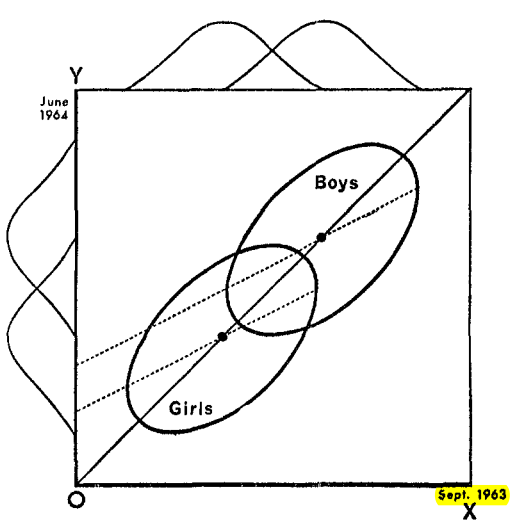


FIG. 1. Hypothetical scatterplots showing initial and final weight for boys and for girls.

The present note points out a type of problem that arises in interpreting data on preexisting groups. The difficulty can most easily be pointed out with the help of a hypothetical illustrative example.

A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex difference in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded.

At the end of the school year, the data are independently examined by two statisticians. Both statisticians divide the students according to sex. The first statistician examines the mean weight of the girls at the beginning of the year and at the end of the year and finds these to be identical. On further investigation, he finds that the frequency distribution of weight for the girls at the end of the year is actually the same as it was at the beginning.

He finds the same to be true for the boys. Although the weight of individual boys and girls has usually changed during the course of the year, perhaps by a considerable amount, the group of girls considered as a whole has not changed in weight, nor has the group of boys. A sort of dynamic equilibrium has been maintained during the year.

The whole situation is shown by the solid lines in the diagram. Here the two ellipses represent separate scatterplots for the boys and the girls. The frequency distributions of

initial weight are indicated at the top of the diagram and the identical distributions of final weight are indicated on the left side. People falling on the solid 45° line through the origin are people whose initial and final weight are identical. The fact that the center of each ellipse lies on this 45° line represents the fact that there is no mean gain for either sex.

The first statistician concludes that as far as these data are concerned, there is no evidence of any interesting effect of the school diet (or of anything else) on student weight. In particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change.

The second statistician, working independently, decides to do an analysis of covariance. After some necessary preliminaries, he determines that the slope of the regression line of final weight on initial weight is essentially the same for the two sexes. This is fortunate since it makes possible a fruitful comparison of the intercepts of the regression lines. (The two regression lines are shown in the diagram as dotted lines. The figure is accurately drawn, so that these regression lines have the appropriate mathematical relationships to the ellipses and to the 45° line through the origin.) He finds that the difference between the intercepts is statistically highly significant.

The second statistician concludes, as is customary in such cases, that the boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes. When pressed to explain the meaning of this conclusion in more precise terms, he points out the following: If one selects on the basis of initial weight a subgroup of boys and a subgroup of girls having identical frequency distributions of initial weight, the relative position of the regression lines shows that the subgroup of boys is going to gain substantially more during the year than the subgroup of girls.

The college dietician is having some difficulty reconciling the conclusions of the two statisticians. The first statistician asserts that there is no evidence of any trend or change during the year for either boys or girls, and consequently, a fortiori, no evidence of a differential change between the sexes. The data clearly support the first statistician since the distribution of weight has not changed for either sex.

The second statistician insists that wherever boys and girls start with the same initial weight, it is visually (as well as statistically) obvious from the scatterplot that the subgroup of boys gains more than the subgroup of girls.

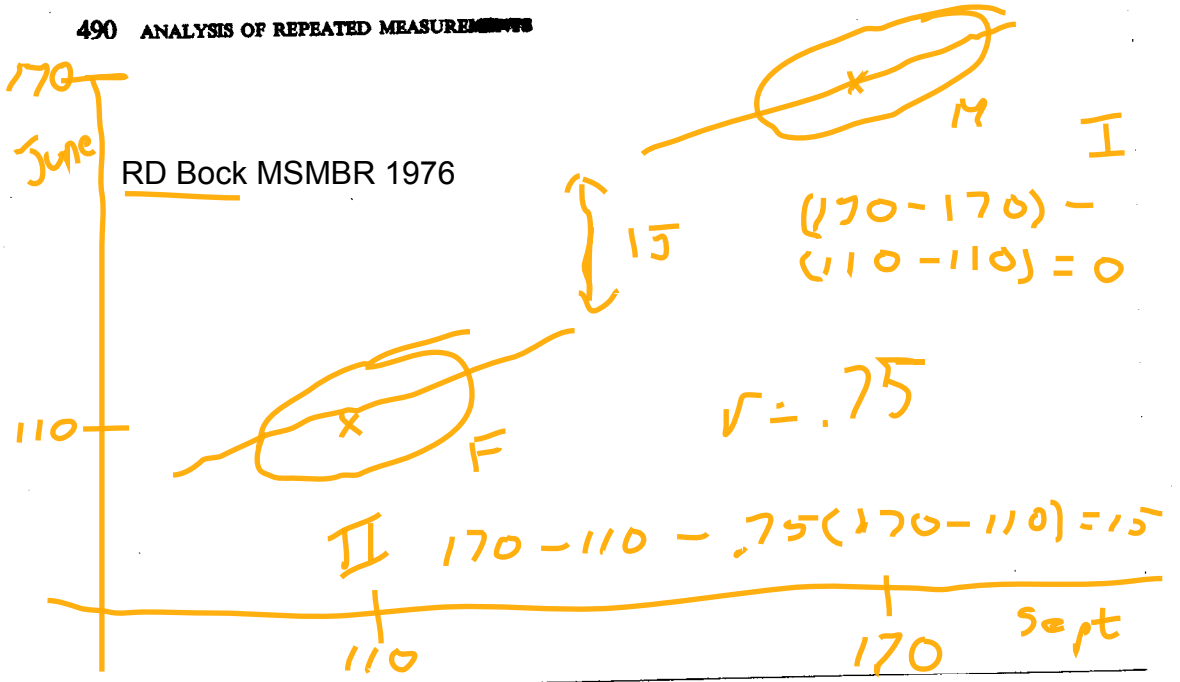
It seems to the present writer that if the dietician had only one statistician, she would reach very different conclusions depending on whether this were the first statistician or the second. On the other hand, granted the usual linearity assumptions of the analysis of covariance, the conclusions of each statistician are visibly correct.

This paradox seems to impose a difficult interpretative task on those who wish to make similar studies of preformed groups. It seems likely that confused interpretations may arise from such studies.

What is the "explanation" of the paradox? There are as many different explanations as there are explainers.

In the writer's opinion, the explanation is that with the data usually available for such studies, there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups. The researcher wants to know how the groups would have compared if there had been no preexisting uncontrolled differences. The usual research study of this type is attempting to answer a question that simply cannot be answered in any rigorous way on the basis of available data.

(Received May 10, 1965)



Suppose a large university obtains measurements, at the beginning and end of the school year, of the weight of each student who takes his meals in the university dining halls. When the resulting data are classified by sex of student, their scatter plot takes the form shown schematically in Fig. 7.3-1. The 45° line represents equality of weights in September and June. The ellipses of concentration represent the presumably bivariate normal distribution of weight on the two occasions.

Suppose two statisticians analyze these data for differences in weight gain of men versus women. The first statistician analyzes simple gain scores and concludes that "as far as these data are concerned, there is no evidence of any interesting effect of the school diet (or of anything else) on student weight; in particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change."

The second statistician, on the other hand, decides to do an analysis of covariance.

'After some necessary preliminaries, he determines that the slope of the regression line of final weight on initial weight is essentially the same for the two sexes. This is fortunate since it makes possible a fruitful comparison of the intercepts of the regression lines . . . . He finds that the difference between the intercepts is statistically highly significant. The second statistician concludes . . . that the [men] showed significantly more gain in weight than the [women] when proper allowance is made for differences in initial weight between the two sexes.'*

As they are stated, the conclusions of the two statisticians are contradictory, and some form of paradox seems implied. On closer inspection, however, it is seen that these alternative methods of analyzing the data are actually directed toward

* From Lord [1967].

picture from Bock 1975 MSMBR

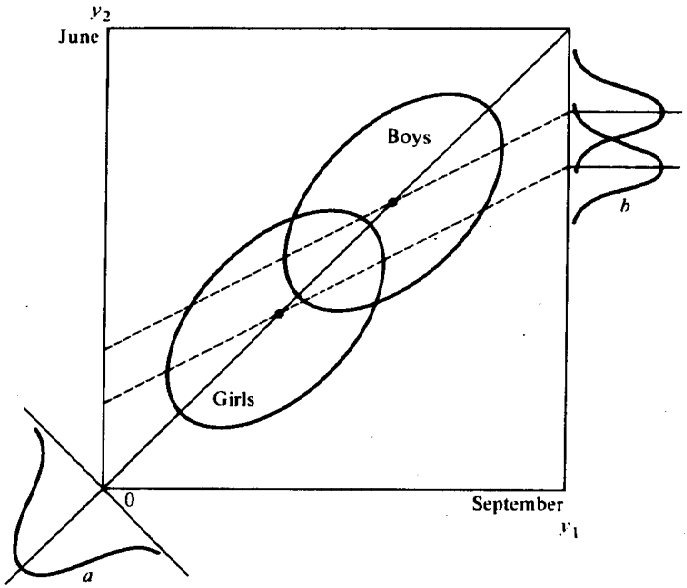


FIGURE 7.3-1
Distribution of gains (*a*) unconditional and (*b*) conditional on initial score.

different inferential problems. Moreover, each method provides the correct solution of the problem to which it is relevant.

These inferential problems may be described briefly as *unconditional* and *conditional*, respectively. The first statistician correctly analyzes gain scores to answer the unconditional question, "Is there a difference in the average gain in weight of the populations?" For the data of Fig. 7.3-1, the answer to this question is, "No, there is no difference in average gain represented by the two sexes."

At the same time, the second statistician correctly employs analysis of covariance to answer the conditional question, "Is a man expected to show a greater weight gain than a woman, given that they are initially of the same weight?" For the data in Fig. 7.3-1, the answer to this question is, "Yes, the man will be expected to gain more, for if he is initially of the same weight as the woman, he is either underweight and will be expected to gain, or the woman is overweight and will be expected to lose." Because the regression lines are parallel, this expectation is independent of the given initial weight.

The conditional inference refers only to the distribution of final weight given initial weight, and says nothing about the average weight gain in the populations from which the given subjects are drawn. The paradox in the original statement of the conclusions stems from the ambiguous phrase "when proper allowance is made for the differences in initial weight between the two sexes." The paradox vanishes

from Rubin et al JEBS

2.2 A Classic Example of Poorly Formulated Causal Assessment–Lord's Paradox

Lord's Paradox is a classic example to illustrate the importance of defining appropriate comparisons and stating clearly any assumptions underlying estimates implied to be "causal". This example originally arose in a similar educational setting, in discussion of gain scores vs. covariance (regression) adjustment (Lord 1967). Lord described the following "paradox:" A university is interested in estimating the effect of the university diet on student's weight, and is particularly interested in any differential effect on males and females. Simple descriptions are given of the data at the beginning and end of the year. For both males and females, the distribution of weights is the same at the beginning and end of the year (the average female weight is the same, the female variance is the same, the average male weight is the same, the male variance is the same, the correlation between September and June weight is 0.8 for both males and females, etc.). Lord then posits two statisticians. Statistician 1 uses gain scores (comparing the change in weight from September to June between males and females) and claims that since on average neither males nor females gained or lost weight during the year, then there is no differential effect of the diet on males or females. Statistician 2 computes a covariance adjusted difference of the two group means and sees that, for males and females of the same initial weight, the males weigh more at the end of the year. He thus concludes that there is a differential effect of the diet for males and females, with males gaining more weight on average. For a graphical representation of the analyses of statisticians 1 and 2, see Bock (1975). Lord's primary question concerned which of these statisticians was correct.

2.3 Lord's Paradox Resolved

Holland and Rubin (1983) explain the apparent paradox by noting that either statistician can be correct, depending on the assumptions made. In the hypothetical scenario, all students receive the new diet; no students receive the undefined "control" diet, whatever it is (no diet? the "old" university diet? the diet the students ate before attending university?). Thus, the only potential outcome that is even observed is that under the treatment (university diet). The potential outcomes under the control diet are completely missing.

If it is assumed that under the control diet each student's weight in June would be the same as their weight in September, then statistician 1 is correct. Statistician 2 is correct under the assumption that weight gain under the control diet is a linear function of the student's weight in September with a common slope but varying intercept for males and females. This simple example is very instructive regarding the importance of thinking carefully about what is being estimated and what is the quantity of interest. It is easy to focus on estimation methods without thinking about the underlying problem–what the technical methods are trying to estimate. Many statisticians and educational researchers were perplexed by Lord's paradox–valid causal inference does not come easily or naturally, except in randomized experiments, and even there only with no complications such as those discussed in Sections 2.6 and 2.7.

see handout for details

2.4 Post-test Scores versus Gain Scores

Despite the debate about whether post-test scores or gain scores should be used as outcomes, the RCM perspective makes it clear that the same causal effect is the estimand whether using either post-test scores or gain scores because test score before treatment assignment is a covariate, unaffected by treatment

Distillation of Lord's Paradox (and Neyman-Rubin-Holland) via Rubin, Wainer, Mavis

Stat 209 [redacted]

p. 2

Wainer '91

	Treatment S	Sub-Population G	Outcome Y_t	Outcome Y_c	Concomitant Variable X_t	Concomitant Variable X_c
Population U of units: 1	t or c	1 or 2				
2	All diet t	male 1 fem 2 gender	June wt		Sept wt	
3						
.						
.						
.						
N						

Figure 1. A framework for causal inference (From "On Lord's Paradox" [p. 5] by P. W Holland & D. B. Rubin, 1983, in H. Wainer & S. [illegible] Principals of modern psychological measurement.

$$\boxed{\begin{array}{l}\text{Average Causal Effect ACE}\\ E(Y_t - Y_c) = E(Y_t) - E(Y_c)\\ \text{observe in comparative studies}\\ E(Y_t \mid S=t) \quad E(Y_c \mid S=c)\end{array}}$$

observables potential outcomes

In whole pop $E(Y_t) = E(Y_t \mid S=t)P(S=t) + E(Y_t \mid S=c)P(S=c)$

$E(Y_c) = E(Y_c \mid S=c)P(S=c) + E(Y_c \mid S=t)P(S=t)$

in Lord's paradox (also Wainer MCAT, not heart rate) c doesn't occur. [no "control" diet]

$$\boxed{\begin{array}{l}\text{Gender Effect} \quad \text{male} \qquad\qquad \text{female}\\ \Delta = E(Y_t - Y_c \mid G=1) - E(Y_t - Y_c \mid G=2)\end{array}}$$

$= E(Y_t \mid G=1) - E(Y_t \mid G=2) - E(Y_c \mid G=1) + E(Y_c \mid G=2)$

obs obs

response function $Y_c = X$ under "control" diet same wt in June

$\Delta = E(Y_t \mid G=1) - E(Y_t \mid G=2)$

$- E(X \mid G=1) + E(X \mid G=2)$

diff of gains gives Δ

Rubin's Model for Causal Inference

The structure used to unravel this mystery involves Rubin's model (Rubin, 1974, 1977, 1978, 1980; Holland, 1986a, 1986b) for the analysis of causal effects. This model allows absolute explicitness about certain distinctions and elements that are often left implicit in other accounts. This model is not meant to find the cause of an effect; rather it tells how to measure the effect of a cause. This purpose is made explicit in Equation 1.

The basic elements of the model are as follows:

1. A population of units, U
2. An "experimental manipulation," with levels t or c, and its associated indicator variable, S
3. A subpopulation indicator variable, G
4. An outcome variable, Y
5. A concomitant variable, X.

This [illegible]

$$\boxed{\begin{array}{l}\text{if instead response function}\\ Y_c = a + bX \quad \text{wt in June}\\ \qquad\qquad \text{linear funct of Sept.}\end{array}}$$

$\Delta = E(Y_t \mid G=1) - E(Y_t \mid G=2)$

$-\left(E(a+bX \mid G=1) - E(a+bX \mid G=2)\right)$

or $-b\left(E(X \mid G=1) - E(X \mid G=2)\right)$

ancova works for Δ! (MCAT)

DD Estimation: Early Examples

1849: London's worst cholera epidemic claims 14,137 lives

- Two companies supplied water to much of London: the Lambeth Waterworks Co. and the Southwark and Vauxhall Water Co.
  - Both got their water from the Thames
- John Snow believed cholera was spread by contaminated water

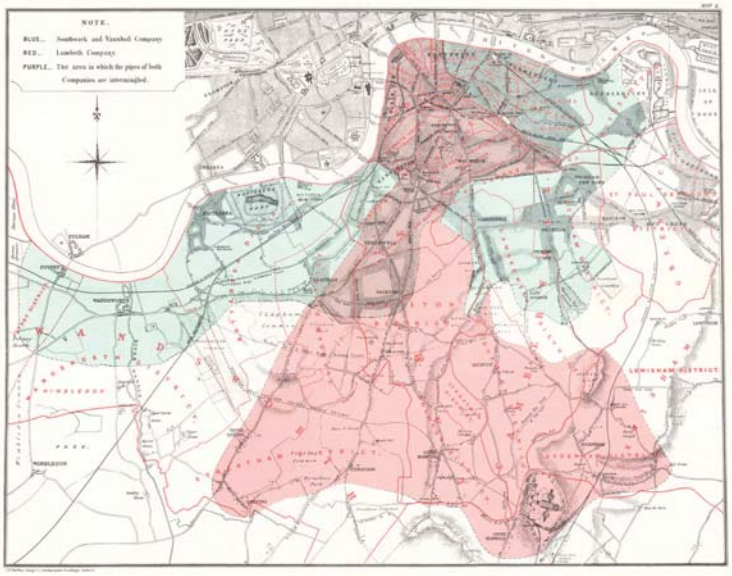
1852: Lambeth Waterworks moved their intake upriver

- Everyone knew that the Thames was dirty below central London

1853: London has another cholera outbreak

- Are Lambeth Waterworks customers less likely to get sick?

DD Estimation: Early Examples



Source: John Snow Archive and Research Companion

DD Estimation: Early Examples

John Snow's Grand Experiment:

- Mortality data showed that very few cholera deaths were reported in areas of London that were **only** supplied by the Lambeth Waterworks
- Snow hired John Whiting to visit the homes of the deceased to determine which company (if any) supplied their drinking water
- Using Whiting's data, Snow calculated the death rate
  - Southwark and Vauxhall: 71 cholera deaths/10,000 homes
  - Lambeth: 5 cholera deaths/10,000 homes
- **Southwark and Vauxhall responsible for 286 of 334 deaths**
  - Southwark and Vauxhall moved their intake upriver in 1855

DD Estimation: Early Examples

In the 1840s, observers of Vienna's maternity hospital noted that death rates from postpartum infections were higher in one wing than the other

- Division 1 patients were attended by doctors and trainee doctors
- Division 2 patients were attended by midwives and trainee midwives

Ignaz Semmelweis noted that the difference emerged in 1841, when the hospital moved to an "anatomical" training program involving cadavers

- Doctors received new training; midwives never handled cadavers
- Did the transference of "cadaveric particles" explain the death rate?

Semmelweis proposed an intervention: hand-washing with chlorine

- Policy implemented in May of 1847

False Counterfactuals

Before vs. After Comparisons:

- **Compares:** same individuals/communities before and after program
- **Drawback:** does not control for time trends

Participant vs. Non-Participant Comparisons:

- **Compares:** participants to those not in the program
- **Drawback:** selection — why didn't non-participants participate?

Two Wrongs Sometimes Make a Right

Difference-in-difference (or "diff-in-diff" or "DD") estimation combines the (flawed) pre vs. post and participant vs. non-participant approaches

- This can sometimes overcome the twin problems of [1] selection bias (on fixed traits) and [2] time trends in the outcome of interest
- The basic idea is to observe the (self-selected) treatment group and a (self-selected) comparison group before and after the program

The diff-in-diff estimator is:

$$DD = \bar{Y}_{post}^{treatment} - \bar{Y}_{pre}^{treatment} - \left(\bar{Y}_{post}^{comparison} - \bar{Y}_{pre}^{comparison}\right)$$

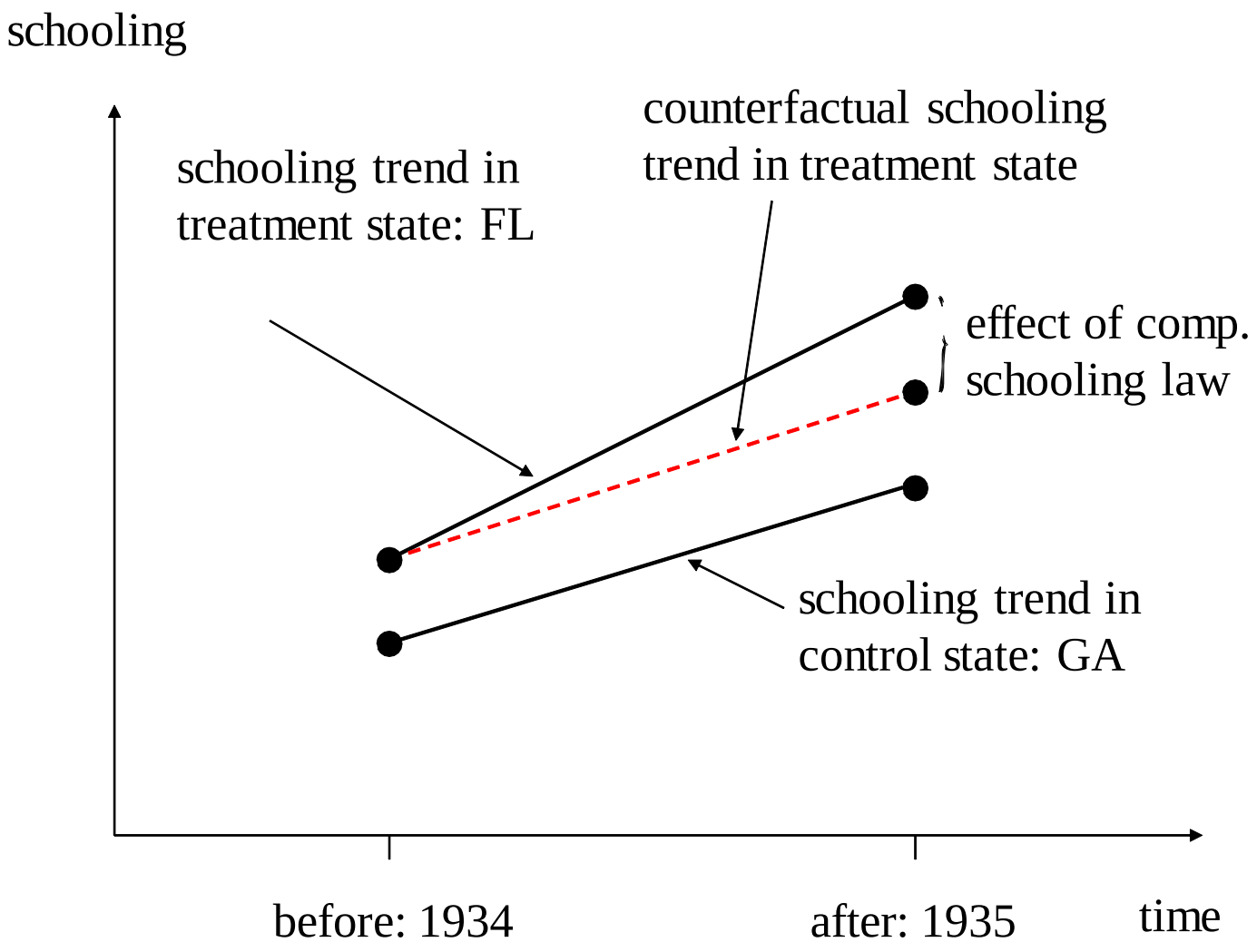
Example: Minimum Wage and Employment

- Do higher minimum wages decrease employment?
- Card and Krueger (1994) consider impact of New Jersey's 1992 minimum wage increase from $4.25 to $5.05 per hour
- Compare employment in 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise
- Survey data on wages and employment from two waves:
  - Wave 1: March 1992, one month before the minimum wage increase
  - Wave 2: December 1992, eight month after increase

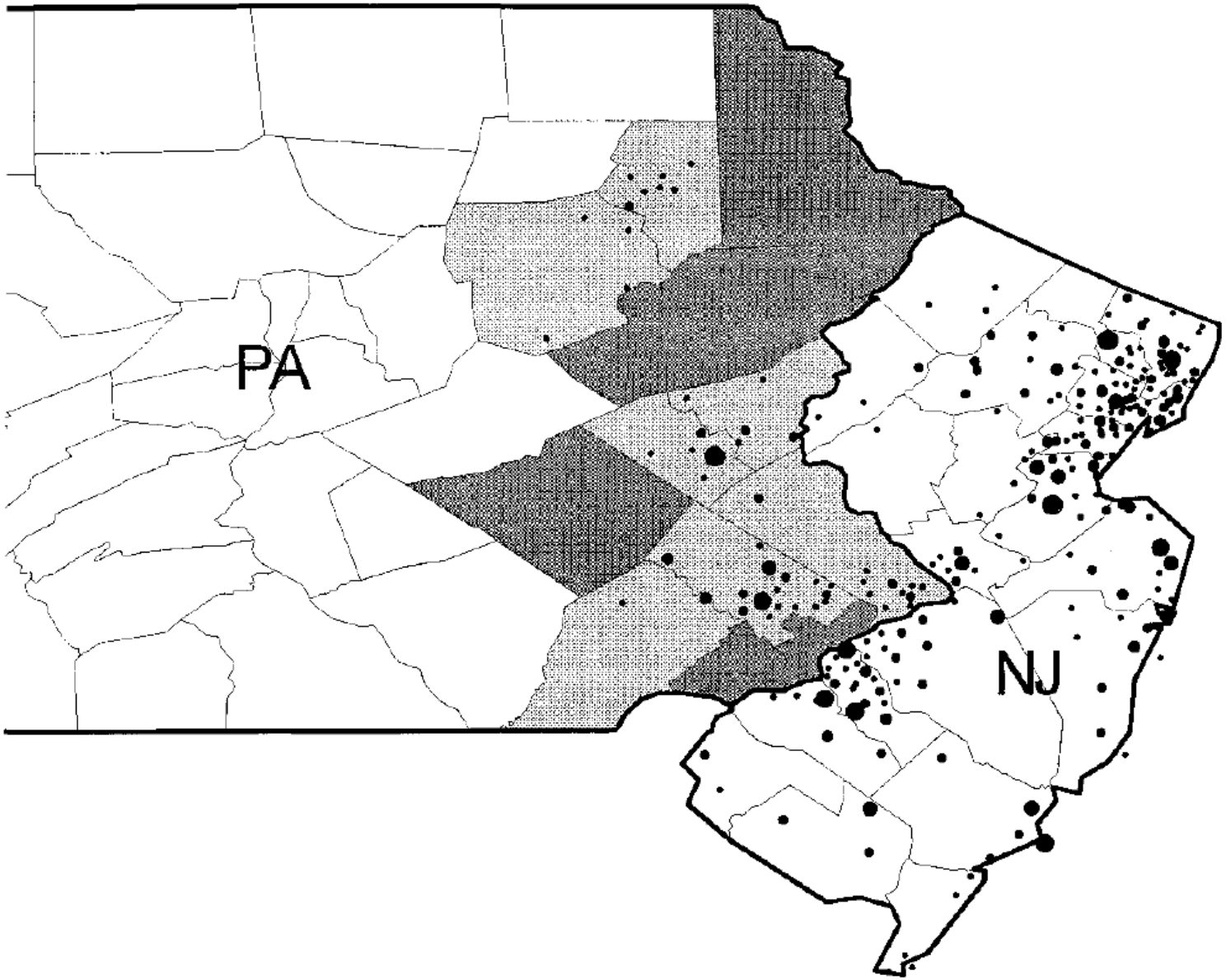
Compulsory schooling laws

- An example for a differences-in-differences setup would be the effect of compulsory schooling laws on schooling obtained in the US. These laws are set at the state level, and different states change the compulsory schooling laws at different times. For example, Florida raised its compulsory schooling requirement from 5 to 7 grades in 1935. Neighboring Georgia required 6 grades both before and after 1935.
- We can think of FL as the treatment state and GA as the control state.
- 1934 is a control period and 1935 is the treatment period.

Identification in the differences-in-differences model



Location of Restaurants



Observational Longitudinal Data
t1, t2 Diffs in Diffs

time 1 $\quad \bar{Y}_{1t} - \bar{Y}_{1c} =$ selection bias (weighing w/s, z)

time 2 $\quad \bar{Y}_{2t} - \bar{Y}_{2c} =$ (ATE) effect + assume same selection bias

ATE effect $= \bar{Y}_{2t} - \bar{Y}_{2c} - (\bar{Y}_{1t} - \bar{Y}_{1c})$

$$\boxed{= \bar{Y}_{2t} - \bar{Y}_{1t} - (\bar{Y}_{2c} - \bar{Y}_{1c})}$$

diff in diffs

+ use matching at time 1 to reduce selection bias

$\boxed{\text{i.e. mimic RK experiment}}$

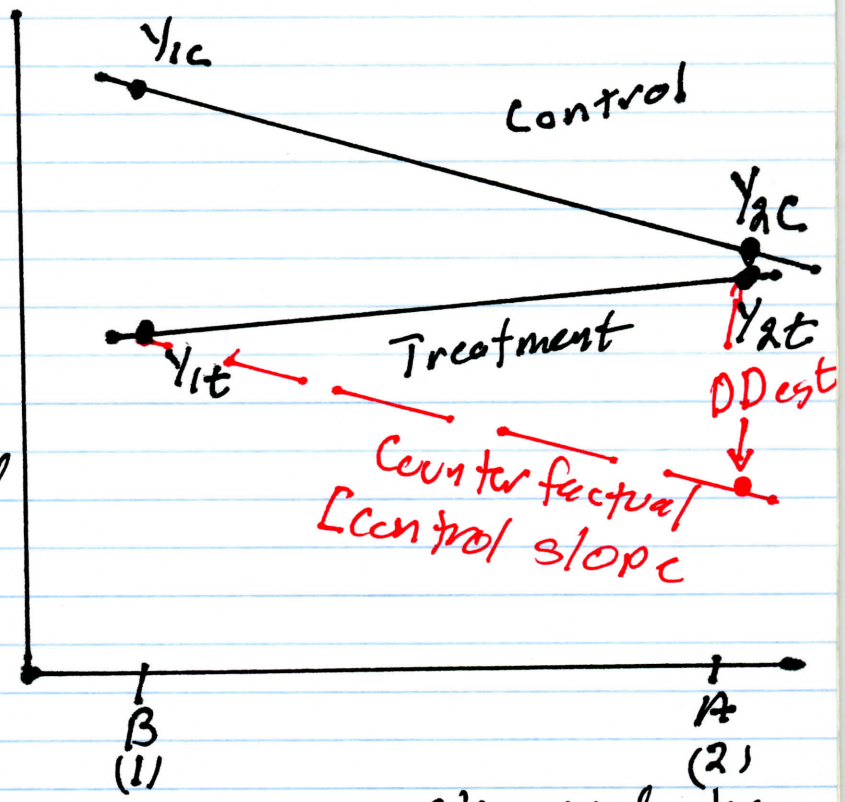
STAT 222 WEEK 6

Diffs in Diffs - Counterfactual Trend Interpretation

DD estimate

$y_{2t} - y^*_{2t}$ counterfactual

time 2 treatment if had control trend



DD estimate

counterfactual y_{2t}

$y_{2t} - (y_{1t} + (y_{2c} - y_{1c}))$

observed change

T $y_{2t} - y_{1t}$

C $y_{2c} - y_{1c}$

$= (y_{2t} - y_{1t}) - (y_{2c} - y_{1c})$

Diffs in improvement again

"ass|u|me"

same estimate - harder to laugh at (understand) assumptions

Imbens/Wooldridge, Lecture Notes 10, Summer '07

What's New in Econometrics? **NBER, Summer 2007**

Lecture 10, Tuesday, July 31st, 4.30-5.30 pm

Difference-in-Differences Estimation

These notes provide an overview of standard difference-in-differences methods that have been used to study numerous policy questions. We consider some recent advances in Hansen (2007a,b) on issues of inference, focusing on what can be learned with various group/time period dimensions and serial independence in group-level shocks. Both the repeated cross sections and panel data cases are considered. We discuss recent work by Athey and Imbens (2006) on nonparametric approaches to difference-in-differences, and Abadie, Diamond, and Hainmueller (2007) on constructing synthetic control groups.

1. Review of the Basic Methodology

Since the work by Ashenfelter and Card (1985), the use of difference-in-differences methods has become very widespread. The simplest set up is one where outcomes are observed for two groups for two time periods. One of the groups is exposed to a treatment in the second period but not in the first period. The second group is not exposed to the treatment during either period. In the case where the same units within a group are observed in each time period, the average gain in the second (control) group is substracted from the average gain in the first (treatment) group. This removes biases in second period comparisons between the treatment and control group that could be the result from permanent differences between those groups, as well as biases from comparisons over time in the treatment group that could be the result of trends. We will treat the panel data case in Section 4.

With repeated cross sections, we can write the model for a generic member of any of groups as

$$y = \beta_0 + \beta_1 dB + \delta_0 d2 + \delta_1 d2 \cdot dB + u \tag{1.1}$$

where y is the outcome of interest, $d2$ is a dummy variable for the second time period. The dummy variable dB captures possible differences between the treatment and control groups prior to the policy change. The time period dummy, $d2$, captures aggregate factors that would cause changes in y even in the absense of a policy change. The coefficient of interest, δ_1, multiplies the interaction term, $d2 \cdot dB$, which is the same as a dummy variable equal to one for those observations in the treatment group in the second period. The difference-in-differences estimate is

$$\hat{\delta}_1 = (\bar{y}_{B,2} - \bar{y}_{B,1}) - (\bar{y}_{A,2} - \bar{y}_{A,1}). \tag{1.2}$$

7. Addressing Unobserved Heterogeneity: *diff-in-diff* matching

In some cases, the conditional independence assumption is clearly not met because units are selected into an intervention on the basis of unmeasured characteristics that are expected to influence outcomes. The example from section II, where the more motivated teachers self-select into the program, clearly illustrates the point. Since motivation is typically not observable to the researcher, it cannot be introduced in the model, and thus, the matching estimator will be unable to isolate the impact of the treatment from the effect of the motivation. In fact, in the case of self-selection, it is usually reasonable to think that unobserved variables (like ability, intelligence, motivation, risk aversion) may critically determine the participation model. Unfortunately, we know from previous sections that the usual matching estimator may be seriously biased in case of *selection-on-unobservables*.

However, if pretreatment data are available, the strong Conditional Independence Assumption may be relaxed. More precisely, under the assumption that unobserved variables are *time-invariant* (that is, their value does not change with time), the effect can be cancelled out by taking the difference in outcomes before and after the program.

The implementation of the *difference-in-differences* (or *diff-in-diff*) *matching estimator* is very similar to the cross-sectional version, except that outcome is measured in changes (between the pretreatment and post-treatment periods) instead of in levels. For treated cases, the dependent variable is the difference between outcomes in a period following participation and prior to participation, and for comparison cases, the outcome difference is calculated over the same periods. Even if participating units differ in important ways from those in the comparison group, so long as such differences are stable over time in their influence on outcomes, this specification can eliminate bias resulting from differences between participants and nonparticipants. Letting *t* and *t'* represent the pretreatment and post-treatment periods, respectively, the outcome for individual *i* will be:

$$\Delta Y_i = Y_{it'} - Y_{it}$$

Note how this specification allows us to relax the CIA (*Conditional Independence Assumption*): the counterfactual outcome of the treated individuals is allowed to differ from the observed outcome of the untreated, as long as their *trend* is the same. In technical terms:

$$E\left(Y_{0t'} - Y_{0t} \mid D = 1, X\right) = E\left(Y_{0t'} - Y_{0t} \mid D = 0, X\right) for X \in S$$

where S is defined as the overlapping support among the treatment and comparison groups. In other words, even if treated units differ in important ways from comparison units, as long as such differences are stable over time in their impact on outcomes, this specification can eliminate bias resulting from differences between treated and untreated units (i.e., it allows for unobserved heterogeneity).

The diff-in-diff matching estimator is simply implemented by calculating the propensity score on the baseline year and applying the steps described above to the differenced outcome.

It is worth noting that it may still be important to control for unit characteristics (X) that do not change over time. For example, if individuals with higher levels of education experience greater growth over time in earnings, it may be necessary to match individuals with the same levels of education. Only if the change in outcomes is not associated with a particular characteristic is it appropriate to omit that measure.

Despite the benefits of difference-in-differences estimates, depending on the processes underlying the dynamics of program participation and outcomes, estimates may have biases that are not present in cross-sectional matching. If prior outcomes incorporate transitory shocks that differ for treatment and comparison units, since difference-in-differences estimation interprets such shocks as representing stable differences, estimates will contain a transitory component that does not represent the true program effect. More generally, the difference-in-differences estimates need to be understood as one of several estimates that rely on different assumptions.

Finally, another source of heterogeneity in effects may arise from different dosages of the treatment, which are neglected in models that record treatment (or participation) with a binary variable. If individuals or other units of analysis receive different levels of treatment that are influential in determining outcomes, an alternative matching technique, the generalized propensity score (GPS), can be applied to estimate the effects of different lengths of exposure to treatment on outcomes. Appendix 1 provides additional details on this matching estimation technique.

Health Serv Outcomes Res Method
DOI 10.1007/s10742-006-0016-x

Propensity score and difference-in-difference methods: a study of second-generation antidepressant use in patients with bipolar disorder

Alex Z. Fu · William H. Dow · Gordon G. Liu

Received: 3 April 2006 / Accepted: 11 December 2006

Abstract This article compared standard regression (logistic), propensity score weighting, propensity score matching, and difference-in-difference (DID) methods in determining the impact of second-generation antidepressant (AD) use on mania-related visits among adult patients with bipolar disorder. Using a large managed care claims database, a logistic regression was developed as a standard approach to predict the likelihood of having mania-related visits after receiving various types of treatments (AD monotherapy, mood stabilizer (MS) monotherapy, and AD-MS combination therapy) controlling for individual baseline characteristics. The propensity score method predicted the propensity to be with one treatment type versus another in the first-stage. Both weighting and greedy matching approaches were applied in the second-stage outcome model. For the DID method, a logistic regression was applied to predict the differential likelihood of having mania-related visits in post-baseline versus baseline periods on different treatments. Both full sample and propensity score-matched sample were applied for the DID method. Except DID with full sample, the results from all other methods suggested no higher likelihood of mania-related visits for second-generation AD-related therapies compared to MS monotherapy. We concluded that standard regression, propensity scoring, and DID methods may produce inconsistent outcomes in a logistic regression framework, when patient baseline characteristics are different between comparison groups and/or not all potential confounders can be correctly measured and

A. Z. Fu (✉)
Department of Quantitative Health Sciences, Cleveland Clinic, 9500 Euclid Avenue/Wb-4, Cleveland, OH 44195, USA
e-mail: fuz@ccf.org

W. H. Dow
Division of Health Policy and Management, University of California, Berkeley, CA, USA
e-mail: wdow@berkeley.edu

G. G. Liu
Department of Health Economics and Management, Peking University Guanghua School of Management, Beijing, China
e-mail: gordon@gsm.pku.edu.cn

Causal inference with observational data

A brief review of quasi-experimental methods

Austin Nichols

July 30, 2009

linked

DD

The simplest panel method is identical to a design used in many RCT's, the difference in differences (DD) method.

	Pre	Post
Treatment	y_1	y_2
Control	y_3	y_4

ATE

$$(y_2 - y_1) - (y_4 - y_3)$$

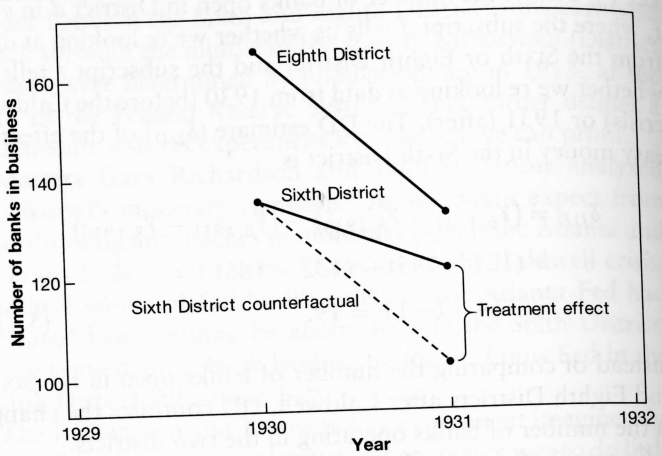
The average treatment effect (ATE) estimate is the difference in differences. For example, the estimate might be the test score gain from 8th grade to 12th grade for those attending charter schools less the test score gain for those in regular public schools. This assumes that the kids in charters, who had to apply to get in, would not have had the same gains in a regular school, i.e. that there was no selection into treatment.

Natural Experiments

The usual "good" diff-in-diff approach relies on a natural experiment, i.e. there was some change in policy or the environment expected to affect treatment for one group more than another, and the two groups should not otherwise have different experiences. For this to work well, the natural experiment should be exogenous itself (i.e. it should not be the case that the policy change is a reaction to behavior) and unlikely to induce people to "game the system" and change their behavior in unpredictable ways (e.g. the differentially treated group jealously overcompensates).

For example, in some US states in 1996, immigrants became ineligible for food stamps, but 17 states offered a substitute program for those in the country before 1996. As of July 2002, anyone in the country five years was eligible for food stamps and most of those in the country 4.9 years were not. One could compute a difference in mean outcomes (say, prevalence of obesity) across recent and less recent immigrants, across calendar years 1995 and 1996, across affected and unaffected states. Using 2002, you could compute a difference across the population of immigrants in the country 4 years or 5 years. See Kaushal (2007) for a related approach.

FIGURE 5.1
Bank failures in the Sixth and Eighth Federal Reserve Districts

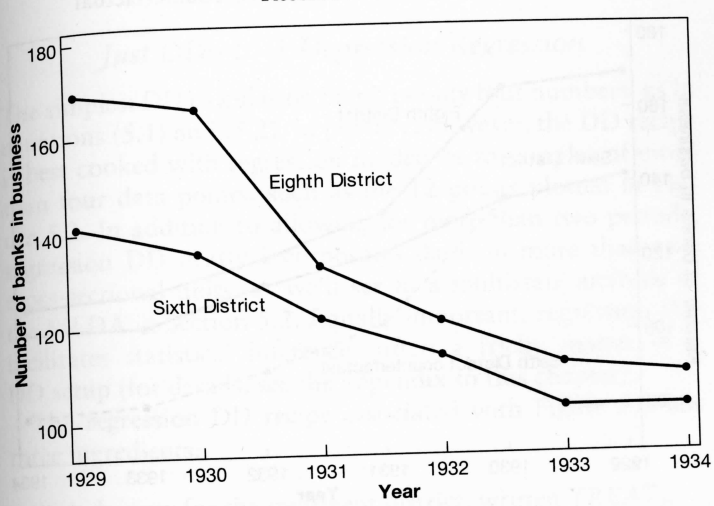


Notes: This figure shows the number of banks in operation in Mississippi in the Sixth and Eighth Federal Reserve Districts in 1930 and 1931. The dashed line depicts the counterfactual evolution of the number of banks in the Sixth District if the same number of banks had failed in that district in this period as did in the Eighth.

The DD tool amounts to a comparison of changes or trends across districts. The dotted line in Figure 5.1 is the counterfactual outcome—our imagined Y_{0i} in the notation of Chapter 1—that lies at the heart of the DD research design: this line tells us what would have happened in the Sixth District had everything evolved as it did in the Eighth. The fact that the solid line for the Sixth District declines much more gradually than this counterfactual line is evidence for the effectiveness of easy money. The 19 bank failures uncovered by our DD calculation is the difference between what really happened and what would have happened had bank activity in the two districts unfolded in parallel.

The DD counterfactual comes from a strong but easily stated assumption: *common trends*. In the Mississippi experiment, DD presumes that, absent any policy differences, the Eighth District trend is what we should have expected to see in the

FIGURE 5.2
Trends in bank failures in the Sixth and Eighth Federal Reserve Districts



Note: This figure shows the number of banks in operation in Mississippi in the Sixth and Eighth Federal Reserve Districts between 1929 and 1934.

Sixth. Although strong, the common trends assumption seems like a reasonable starting point, one that takes account of pretreatment differences in levels. With more data, the assumption can also be probed, tested, and relaxed.

Figure 5.2 provides evidence on the common trends assumption for Mississippi's Federal Reserve Districts. The evidence comes in the form of a longer time series on bank activity. Before 1931, the Great Depression had not yet hit Mississippi hard. Regional Fed policies in the two districts were also similar in this more relaxed period. The fact that bank failures moved almost in parallel in the two districts between 1929 and 1930, with the number of banks declining slightly in both districts, is therefore consistent with the common trends hypothesis for untreated periods. Figure 5.3 adds the Sixth District counterfactual implied by extrapolating Eighth District trends to the Sixth District for years after 1930. The gap

of union wage effects, Card (1996) uses external information from a separate validation survey to adjust panel data estimates for measurement error in reported union status. But data from multiple reports and repeated measures of the sort used by Ashenfelter and Krueger (1994) and Card (1996) are unusual. At a minimum, therefore, it's important to avoid overly strong claims when interpreting fixed effects estimates (never bad advice for an applied econometrician in any case).

5.2 Differences-in-Differences: Pre and Post, Treatment and Control

The fixed effects strategy requires panel data, that is, repeated observations on the same individuals (or firms, or whatever the unit of observation might be). Often, however, the regressor of interest varies only at a more aggregate or group level, such as state or cohort. For example, state policies regarding health care benefits for pregnant workers may change over time but are fixed across workers within states. The source of OVB when evaluating these policies must therefore be unobserved variables at the state and year level. In some cases, group-level omitted variables can be captured by group-level fixed effects, an approach that leads to the differences-in-differences (DD) identification strategy.

The DD idea was probably pioneered by physician John Snow (1855), who studied cholera epidemics in London in the mid-nineteenth century. Snow wanted to establish that cholera is transmitted by contaminated drinking water (as opposed to "bad air," the prevailing theory at the time). To show this, Snow compared changes in death rates from cholera in districts serviced by two water companies, the Southwark and Vauxhall Company and the Lambeth Company. In 1849 both companies obtained their water supply from the dirty Thames in central London. In 1852, however, the Lambeth Company moved its water works upriver to an area relatively free of sewage. Death rates in districts supplied by Lambeth fell sharply in comparison to the change in death rates in districts supplied by Southwark and Vauxhall.

SHOE LEATHER, Freedman

Snai

"To turn this grand experiment to account, all that was required was to learn the supply of water to each individual house where a fatal attack of cholera might occur. [pp. 74–75.]"

Snow's data are shown in Table 1. The denominator data—the number of houses served by each water company—were available from parliamentary records. For the numerator data, however, a house-to-house canvass was needed to determine the source of the water supply at the address of each cholera fatality. (The "bills of mortality" showed the address, but not the water source.) The death rate from the Southwark and Vauxhall water is about 9 times the death rate for the Lambeth water. This is compelling evidence.

Snow argued that the data could be analyzed as if they had resulted from an experiment of nature: there was no difference between the customers of the two water companies, except for the water. His sample was not only large but representative; therefore, it was possible to generalize to a larger population. Finally, Snow was careful to avoid the "ecological fallacy:" relationships that hold for groups may not hold for individuals (Robinson, 1950). It is the design of the study and the magnitude of the effect that compel conviction, not the elaboration of technique.

TABLE 1. Death rate from cholera by source of water. Rate per 10,000 houses. London, epidemic of 1853–54. Snow's Table IX.

	No. of Houses	Cholera Deaths	Rate per 10,000
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

More evidence was to come from other countries. In New York, the epidemics of 1832 and 1849 were handled according to the theories of the time. The population was exhorted to temperance and calm, since anger could increase the humor "choler" (bile), and imbalances in the humors of the body lead to disease. Pure water was brought in to wash the streets and reduce miasmas. In 1866, however, the epidemic was handled by a different method—rigorous isolation of cholera cases, with disinfection of their dejecta by lime or fire. The fatality rate was much reduced.

At the end of the 19th century, there was a burst of activity in microbiology. In 1878, Pasteur published *La théorie des germes et ses applications à la médecine et à la chirurgie.* Around that time, Pasteur and Koch isolated the anthrax bacillus and developed techniques for vaccination. The tuberculosis bacillus was next. In 1883, there was a cholera epidemic in Egypt, and Koch isolated the vibrio; he was perhaps anticipated by Filipo Pacini. There was an epidemic in Hamburg in 1892. The city fathers turned to Max von Pettenkofer, a leading figure in the German hygiene movement of the time. He did not believe Snow's theory, holding instead that cholera was caused by poison in the ground. Hamburg was a center of the slaughterhouse industry, and von Pettenkofer had the carcasses of dead animals dug up and hauled away, in order to reduce pollution of the ground. The epidemic continued its ravages, which ended only when the city lost faith in von Pettenkofer and turned in desperation to Koch.

The approach developed by Louis and Snow found many applications; I will mention only two examples. Semmelweiss (1867) discovered the cause of puerperal fever. Around 1914, Goldberger

To make matters more concrete, let us return to an example from economics. Suppose we are interested in the effect of the minimum wage on employment, a classic question in labor economics. In a competitive labor market, increases in the minimum wage move us up a downward-sloping labor demand curve. Higher minimums therefore reduce employment, perhaps hurting the very workers minimum wage policies were designed to help. Card and Krueger (1994) use a dramatic change in the New Jersey state minimum wage to see if this is true.[6]

On April 1, 1992, New Jersey raised the state minimum from \$4.25 to \$5.05. Card and Krueger collected data on employment at fast food restaurants in New Jersey in February 1992 and again in November 1992. These restaurants (Burger King, Wendy's, and so on) are big minimum wage employers. Card and Krueger also collected data from the same type of restaurants in eastern Pennsylvania, just across the Delaware River. The minimum wage in Pennsylvania stayed at \$4.25 throughout this period. They used their data set to compute differences-in-differences (DD) estimates of the effects of the New Jersey minimum wage increase. That is, they compared the February-to-November change in employment in New Jersey to the change in employment in Pennsylvania over the same period.

DD is a version of fixed effects estimation using aggregate data. To see this, let Y_{1ist} be fast food employment at restaurant i in state s and period t if there is a high state minimum wage, and let Y_{0ist} be fast food employment at restaurant i in state s and period t if there is a low state minimum wage. These are potential outcomes; in practice, we only get to see one or the other. For example, we see Y_{1ist} in New Jersey in November 1992. The heart of the DD setup is an additive structure for potential outcomes in the no-treatment state. Specifically, we assume that

$$E[Y_{0ist}|s,t] = \gamma_s + \lambda_t, \tag{5.2.1}$$

[6]The DD idea was first used to study the effects of minimum wages by Obenauer and von der Nienburg (1915), writing for the U.S. Bureau of Labor Statistics.

where s denotes state (New Jersey or Pennsylvania) and t denotes period (February, before the minimum wage increase, or November, after the increase). This equation says that in the absence of a minimum wage change, employment is determined by the sum of a time-invariant state effect and a year effect that is common across states. The additive state effect plays the role of the unobserved individual effect in section 5.1.

Let D_{st} be a dummy for high-minimum-wage states and periods. Assuming that $E[Y_{1ist} - Y_{0ist}|s,t]$ is a constant, denoted δ, observed employment, Y_{ist}, can be written:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist}, \tag{5.2.2}$$

where $E(\varepsilon_{ist}|s,t) = 0$. From here, we get

$$\begin{aligned} &E[Y_{ist}|s = PA, t = Nov] - E[Y_{ist}|s = PA, t = Feb] \\ &\quad = \lambda_{Nov} - \lambda_{Feb} \end{aligned}$$

and

$$\begin{aligned} &E[Y_{ist}|s = NJ, t = Nov] - E[Y_{ist}|s = NJ, t = Feb] \\ &\quad = \lambda_{Nov} - \lambda_{Feb} + \delta. \end{aligned}$$

The population difference-in-differences,

$$\begin{aligned} &\{E[Y_{ist}|s = NJ, t = Nov] - E[Y_{ist}|s = NJ, t = Feb]\} \\ &\quad - \{E[Y_{ist}|s = PA, t = Nov] - E[Y_{ist}|s = PA, t = Feb]\} = \delta, \end{aligned}$$

is the causal effect of interest. This is easily estimated using the sample analog of the population means.

Table 5.2.1 (based on table 3 in Card and Krueger, 1994) shows average employment at fast food restaurants in New Jersey and Pennsylvania before and after the change in the New Jersey minimum wage. There are four cells in the first two rows and columns, while the margins show state differences in each period, the changes over time in each state, and the difference-in-differences. Employment in Pennsylvania restaurants is somewhat higher than in New Jersey in February but falls by November. Employment in New Jersey, in contrast, increases slightly. These two changes produce

TABLE 5.2.1
Average employment in fast food restaurants before and after the New Jersey minimum wage increase

Variable	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (.94)	21.03 (.52)	−.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	.59 (.54)	2.76 (1.36)

Notes: Adapted from Card and Krueger (1994), table 3. The table reports average full-time-equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all restaurants with data on employment. Employment at six closed restaurants is set to zero. Employment at four temporarily closed restaurants is treated as missing. Standard errors are reported in parentheses.

a positive difference-in-differences, the opposite of what we might expect if a higher minimum wage pushed businesses up the labor demand curve.

How convincing is this evidence against the standard labor demand story? The key identifying assumption here is that employment *trends* would be the same in both states in the absence of treatment. Treatment induces a deviation from this common trend, as illustrated in figure 5.2.1. Although the treatment and control states can differ, this difference is meant to be captured by the state fixed effect, which plays the same role as the unobserved individual effect in (5.1.3).[7]

[7]The common trends assumption can be applied to transformed data, for example,

$$E[\ln Y_{0ist}|s,t] = \gamma_s + \lambda_t.$$

Note, however, that common trends in logs rule out common trends in levels and vice versa. Athey and Imbens (2006) introduce a semiparametric DD estimator that allows for common trends after an unspecified transformation of the dependent variable. Poterba, Venti, and Wise (1995) and Meyer, Viscusi, and Durbin (1995) discuss DD-type models for quantiles.

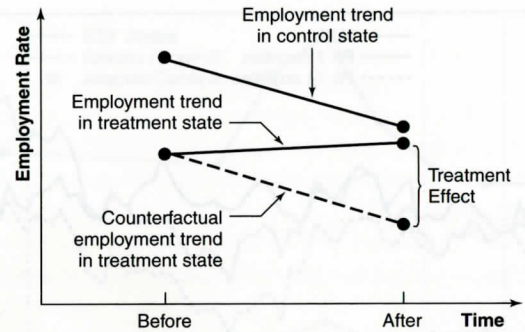


Figure 5.2.1 Causal effects in the DD model.

The common trends assumption can be investigated using data on multiple periods. In an update of their original minimum wage study, Card and Krueger (2000) obtained administrative payroll data for restaurants in New Jersey and a number of Pennsylvania counties. These data are shown here in figure 5.2.2, similar to figure 2 in their follow-up study. The vertical lines indicate the dates when the original Card and Krueger surveys were conducted, and the third vertical line indicates the October 1996 increase in the federal minimum wage to $4.75, which affected Pennsylvania but not New Jersey. These data give us an opportunity to look at a new minimum wage experiment.

As in the original Card and Krueger survey, the administrative data show a slight decline in employment from February to November 1992 in Pennsylvania, and little change in New Jersey over the same period. However, the data also reveal substantial year-to-year employment variation in other periods. These swings often seem to differ substantially in the two states. In particular, while employment levels in New Jersey and Pennsylvania were similar at the end of 1991, employment in Pennsylvania fell relative to employment in New Jersey over the next three years (especially in the 14-county group), mostly before the 1996 increase in the federal minimum wage. So Pennsylvania may not provide a very good measure of counterfactual employment rates in New Jersey in the absence of a minimum wage change.

Package ‘did’

February 18, 2020

Title Treatment Effects with Multiple Periods and Groups

Version 1.2.3

Description The standard Difference-in-Differences (DID) setup involves two periods and two groups -- a treated group and untreated group. Many applications of DID methods involve more than two periods and have individuals that are treated at different points in time. This package contains tools for computing average treatment effect parameters in Difference in Differences models with more than two periods and with variation in treatment timing using the methods developed in Callaway and Sant'Anna (2019) <https://ssrn.com/abstract=3148250>. The main parameters are group-time average treatment effects which are the average treatment effect for a particular group at a a particular time. These can be aggregated into a fewer number of treatment effect parameters, and the package deals with the cases where there is selective treatment timing, dynamic treatment effects, calendar time effects, or combinations of these. There are also functions for testing the Difference in Differences assumption, and plotting group-time average treatment effects.

Depends R (>= 2.10)

License GPL-2

Encoding UTF-8

LazyData true

Imports BMisc (>= 1.3.1), MASS, pbapply, stats, ggplot2, knitr, utils, gridExtra

RoxygenNote 7.0.2

VignetteBuilder knitr

Suggests rmarkdown

NeedsCompilation no

Author Brantly Callaway [aut, cre], Pedro H.C. Sant'Anna [aut]

Maintainer Brantly Callaway <bmcallaw@olemiss.edu>

Repository CRAN

Date/Publication 2020-02-18 00:00:02 UTC

Difference in differences (DID)
Estimation step-by-step

 # Estimating the DID estimator (using the multiplication method, no need to generate the interaction)

```
didreg1 = lm(y ~ treated*time, data = mydata)
summary(didreg1)
```

```
Call:
lm(formula = y ~ treated * time, data = mydata)

Residuals:
       Min         1Q     Median         3Q        Max 
-9.768e+09 -1.623e+09  1.167e+08  1.393e+09  6.807e+09 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)  
(Intercept)   3.581e+08  7.382e+08   0.485   0.6292  
treated       1.776e+09  1.128e+09   1.575   0.1200  
time          2.289e+09  9.530e+08   2.402   0.0191 *
treated:time -2.520e+09  1.456e+09  -1.731   0.0882 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.953e+09 on 66 degrees of freedom
Multiple R-squared:  0.08273, Adjusted R-squared:  0.04104 
F-statistic: 1.984 on 3 and 66 DF,  p-value: 0.1249
```

 # The coefficient for ‘treated#time’ is the differences-in-differences estimator (‘did’ in the previous example). The effect is significant at 10% with the treatment having a negative effect.

References

Introduction to econometrics, James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.

"Difference-in-Differences Estimation", Imbens/Wooldridge, Lecture Notes 10, summer 2007.

http://www.nber.org/WNE/lect_10_diffindiffs.pdf

"Lecture 3: Differences-in-Differences", Fabian Waldinger

http://www2.warwick.ac.uk/fac/soc/economics/staff/ffwaldinger/teaching/ec9a8/slides/lecture_3_-_did.pdf

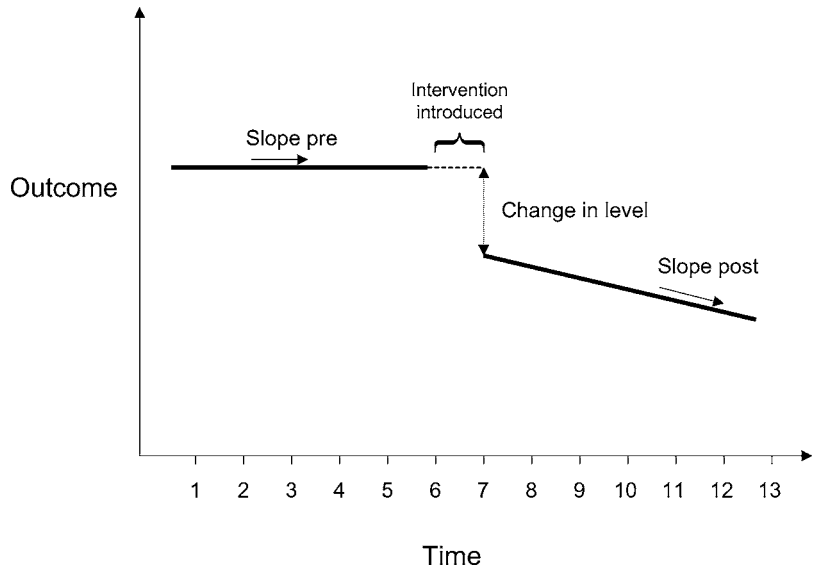


Figure 1. The effect sizes estimated by time series regression analysis of an interrupted time series design.

review (1–4;12–14;16;22;26;28–33;37;38;43–45;50;52;54;56;57;59;62;63;65–68;70; 72–75). Eighteen studies were published before 1990 and forty studies after 1990. The characteristics of the ITS studies are displayed in Table 1. Both reviews had similar numbers of preintervention data points; however, the average ratio of postintervention points to preintervention points indicated that the guideline studies tended to collect more postintervention points than preintervention points within each study. The time interval between data points varied, with "monthly" the most common in both reviews.

The quality criteria for the ITS studies are shown in Table 2. Thirty-eight (66%) studies did not rule out the threat that another event could have occurred at the same time as the intervention. Reporting of factors related to data collection, the primary outcome, and completeness of the data set were generally done in both reviews. No study provided a justification for the number of data points used or a rationale for the shape of the intervention effect.

Table 1. Characteristics of Included Interrupted Time Series Studies

	Mass media review	Guidelines review
Number of included studies	20	38
Median number of preintervention points	9	10
Median number of postintervention points	6	12
Ratio of postintervention points to preintervention points – mean (SD)	0.9 (0.8)	1.9 (2.0)
Time interval between points		
5 days	0	1
1 week	3	3
1 month	9	25
2 months	3	0
3 months	1	5
1 year	4	4

CALIFORNIA

California ballot measure blamed for shoplifting jump

Published May 14, 2016 0 0

Associated Press

ROCKLIN, Calif. – Perry Lutz says his struggle to survive as a small businessman became a lot harder after California voters reduced theft penalties 1½ years ago.

About a half-dozen times this year, shoplifters have stolen expensive drones or another of the remote-controlled toys he sells in HobbyTown USA, a small shop in Rocklin, northeast of Sacramento. "It's just pretty much open season," Lutz said. "They'll pick the $800 unit and just grab it and run out the door."

Anything below $950 keeps the crime a misdemeanor — and likely means the thieves face no pursuit and no punishment, say retailers and law enforcement officials. Large retailers including Safeway, Target, Rite Aid and CVS pharmacies say shoplifting increased at least 15 percent, and in some cases, doubled since voters approved Proposition 47 and ended the possibility of charging shoplifting as a felony with the potential for a prison sentence.

Shoplifting reports to the Los Angeles Police Department jumped by a quarter in the first year, according to statistics the department compiled for The Associated Press. The ballot measure also lowered penalties for forgery, fraud, petty theft and drug possession.

Public Policy Institute of California researcher Magnus Lofstrom noted a troubling increase in property crime inCalifornia's largest cities in the first half-year after Proposition 47 took effect. Preliminary FBI crime reports show a 12 percent jump in larceny-theft, which includes shoplifting, but he said it is too early to determine what, if any, increase is due to the ballot measure.

The increase in shoplifting reports set up a debate over how much criminals pay attention to penalties, and whether law enforcement is doing enough to adapt to the legal change.

Prosecutors, police and retailers, including California Retailers Association President Bill Dombrowski and CVS Health spokesman Mike DeAngelis, say the problem is organized retail theft rings whose members are well aware of the reduced penalties.

"The law didn't account for that," said Capt. John Romero, commander of the LAPD's commercial crimes division. "It did not give an exception for organized retail theft, so we're seeing these offenders benefiting and the retailers are paying the price."

Lenore Anderson, executive director of Californians for Safety and Justice, who led the drive to pass Proposition 47, said law enforcement still has plenty of tools, including using the state's general conspiracy law and proving that the same thief is responsible for multiple thefts that together top $950.

Shoplifting rings generally recruit society's most vulnerable — the homeless, low-end drug users, those living in the country illegally — to steal merchandise that can be sold for a discount on the streets or over the Internet, said Joseph LaRocca, a Los Angeles-based theft-prevention consultant and formerly the National Retail Federation's vice president of loss prevention.

While misdemeanors, in theory, can bring up to a year in county jail, Fresno Police Sgt. Mark Hudson said it's not worth it to

issue a citation or arrest a suspect who would likely be immediately released because of overcrowding.

"We've heard of cases where they're going into stores with a calculator so they can make sure that what they steal is worth less than $950," said Robin Shakely, Sacramento County assistant chief deputy district attorney.

Adam Gelb, director of the public safety performance project at The Pew Charitable Trusts, disputes those sorts of anecdotes.

"The vast majority of offenders just aren't fine-tuning their behavior that way," Gelb said.

His organization recently reported finding no effect on property crimes and larceny rates in 23 states that increased the threshold to charge thefts as felonies instead of misdemeanors between 2001 and 2011.California raised its threshold from $400 in 2010.

"It's hard to see how raising the level to $950 in California would touch off a property crime wave when raising it to $2,000 in South Carolina six years ago hasn't registered any impact at all," Gelb said.

The study did not include the effects of Proposition 47, but Gelb and other Pew researchers said there is no reason to believe adding shoplifting to the list would spark an increase in thefts.

California is among 17 states without an organized retail crime law that specifically targets shoplifting rings with tougher penalties, according to the Organized Retail Crime Resource Center. Results vary: Of the top five states for shoplifting last year, three — Florida, Pennsylvania and Texas — had such laws, while California and New York did not.

For his part, Lutz, the hobby shop owner, has provided police with surveillance videos, and even the license plate, make and model of the getaway vehicles.

"They go, 'Perry, our hands are tied because it's a misdemeanor,'" Lutz said. "It's not worth pursuing, it's just a waste of manpower."

Market Data

Trending in U.S.

1. Coast Guard searching for woman who fell off cruise ship in Gulf of Mexico
2. Elderly man's Quaker Oats contest submission rejected because… it's hand-written
3. Manhunt on for 1 of 2 Wyoming men accused of tying up mom, 4 teenage daughters in basement

Interrupted Time Series

Above, in the discussion of non-equivalent control group designs, it was suggested that pretest-posttest versions could be improved by having at least two pretests to establish linear tendencies apart from treatment. Cook and Campbell (1979) list six interrupted time series designs which extend this suggestion by having multiple pretests and posttests.

1. **Simple Interrupted Time Series Design**. This is the one-group pretest-posttest design augmented with multiple pretests and posttests. The trend found in multiple pretests can be compared to the trend found in multiple posttests to assess whether apparent post-treatment improvement may simply be an extrapolation of a maturation effect which was leading toward improvement anyway. There may be other confounding factors as well, such as failure to seasonally adjust data, confounding a seasonal effect with a treatment effect. In general, this design is liable to history-type challenges to validity: the possibility that other factors historically coterminous with the treatment actually led to the observed effect. Other threats to validity include selection bias, as due to non-random attrition of subjects in the posttest; instrumentation bias (the posttest is not equivalent to the pretest); and testing (there may be a learning effect from the pretest such that the observed effect is one a test artifact rather than a treatment effect).

2. **Interrupted Time Series with a Nonequivalent No-Treatment Comparison Group** : This is the two-group pretest-posttest design using an untreated control group, but with multiple pretests and posttests. By having a comparison group, even if nonequivalent (not randomized), the same threats to validity can occur, but they most occur in a more complex and hence more easily disproved way. For instance, if this design shows an improvement in the treatment but not comparison group, it may still be true that there is historical bias, but such bias must be unique to the treatment group for some reason, not including variables which also affect the comparison group. There could be seasonal bias, but only if the seasonal factors were thought to be uniquely associated with treatment, and so on.

3. **Interrupted Time Series with Nonequivalent Dependent Variables** : This is the nonequivalent dependent variables pretest-posttest design with multiple pretests and posttests. The object is to find dependent variables related to the one of thought to be influence by treatment, but where the related variables are not. Cook and Campbell (1979) give the example of influence of breathalyzer tests given by police when bars are open weekend nights, but not given at other times. The dependent variable of interest is accident rates on weekend nights. The related dependents are accident rates on weekday nights when bars are open, and accident rates at times when bars are not open. the expectation is that the treatment effect of breathalyzer testing will be significantly greater for weekend nights than at other times. Counter-explanations for lower accident rates (ex., safer cars, stricter court treatment of offenders) must explain not only the lower accident rate on weekend nights, but also the lack of effect at other times. Of course, confounding factors may well exist, but they must be unique to the dependent variable of interest.

4. **Interrupted Time Series with Removed Treatment** : This is the removed-treatment pretest-posttest design with multiple pretests and posttests, including ones in between the original treatment and its removal, and hence is a more powerful test. For instance, the threat of history is reduced because any historical forces coincident with treatment would also have increase after treatment and decrease after removal, an unlikely circumstance. Ideally removal of treatment does not occur until enough observations have been taken to rule out any seasonal or other cyclical effects.

5. **Interrupted Time Series with Multiple Replications** . This is simply the interrupted time series with removed treatment design, except that treatment and removal occur multiple times on a schedule. Circumstances rarely permit such a design, but it is stronger yet. By timing the replications randomly, the researcher is able to minimize contamination from cyclical factors. This design assumes one is dealing with a treatment effect which dissipates in a timely manner before the next replication, without carryover effects.

6. **Interrupted Time Series with Switching Replications** . This is a further refinement in which there are two groups, each serving as either the treatment or comparison group on an alternating basis, through multiple replications of treatment and removal.This requires an even higher level of control over subjects by the researcher but is a particularly strong design in ruling out threats to validity. It does not lend itself to studies where the treatment intervention has been gradual, or where treatment effect does not decay well.

Interrupted Time Series ARIMA

A common research questions in time series analysis is whether an outside event affected subsequent observations. For example, did the implementation of a new economic policy improve economic performance; did a new anti-crime law affect subsequent crime rates; and so on. In general, we would like to evaluate the impact of one or more discrete events on the values in the time series. This type of interrupted time series analysis is described in detail in McDowall, McCleary, Meidinger, & Hay (1980). McDowall, et. al., distinguish between three major types of impacts that are possible: (1) permanent abrupt, (2) permanent gradual, and (3) abrupt temporary.

linked in lecture topics

Interrupted Time Series Quasi-Experiments[1]

Gene V Glass
Arizona State University

Researchers seek to establish causal relationships by conducting experiments. The standard for causal proof is what Campbell and Stanly (1963) called the "true experiment." Often, circumstances will not permit meeting all the conditions of a true experiment. Then, a quasi-experiment is chosen. Among the various quasi-experimental designs is one that rivals the true experiment: the interrupted time-series design. It has become the standard method of causal analysis in applied behavioral research.

Just what is a "cause" is a matter of deep philosophical debate. Perhaps I can safely ignore that debate and appeal to your intuitive understanding that renders meaningful such statements as "The nail caused the tire to go flat" or "Owning a car causes teenagers' grades to drop." If every relationship were causal, the world would be a simple place; but most relationships are not. In schools where teachers make above-average salaries, pupils score above average on achievement tests. It is not safe, however, to say that increasing teachers' salaries will cause an increase in pupils' achievement. Business executives who take long, expensive vacations make higher salaries than executives who don't. But will taking the summer off and touring Europe increase your salary? Try it and find out.

Relationships: Causal and Spurious

Relationships can fail to be causal relationships in two principal ways: because of (a) third variables and (b) an ambiguous direction of influence. The third-variable situation occurs when two things are related because each is causally related to a third variable, not because of any causal link between each other. The teachers' salaries and pupil achievement example is probably an instance of the third-variable situation. In this case, the third variable might be the wealth of the community; rich communities pay teachers more and have pupils who score higher on achievement tests for a host of reasons connected to family wealth but not to teachers' pay. Teachers are professionals who want to be paid well and deserve to be; but I doubt that, once the negotiations are finished, a teacher tries any harder to teach the pupils because of a few hundred dollars on the salary schedule. So the relationship of teachers' salaries and pupil achievement—a relationship that is an empirical fact, incidentally—is due to common relationships to a third variable.

The business executive's vacation is an example of ambiguous direction of influence. A travel agent might publish a graph in an advertisement that shows this relationship. However, the simple fact of the relationship leaves quite ambiguous whether long vacations cause higher salaries (presumably through improving morale and vitality and the like) or higher salaries cause long, expensive vacations. The truth is obvious in this case, and it is quite the opposite of the direction of influence that the travel agents wants people to believe. But many other examples are less clear. Does enhanced motivation cause pupils to learn successfully in school, or is it mainly the other way around: success in learning causes an increase in motivation to learn? The truth is probably some of each in unknown amounts, which goes to show how ill-

---

[1] Reprinted from Jaeger, R. M. (1997). *Complementary methods for research in education. 2nd Edition*. Pp. 589-608. Washington D. C.: American Educational Research Association.

Investigation and Detective Work

In some time-series experiments, it frequently happens that subsections of a large data pool reveal intervention effects that are not apparent in the complete body of data. Consider the data in Figure 2, for example. If there is any effect of the British Road Safety Act of 1967 on the traffic fatality rate, it certainly isn't very apparent in the graph in Figure 2. But look at the graph in Figure 9 to see what happened when the fatalities on weekend nights were singled out of the total body of data. There we see a huge, more than 50%, reduction in fatalities coincident with the implementation of the Road Safety Act in October 1967. We can move from the equivocal results of Figure 2 to the clear certainty of Figure 9, merely by separating the larger body of data.

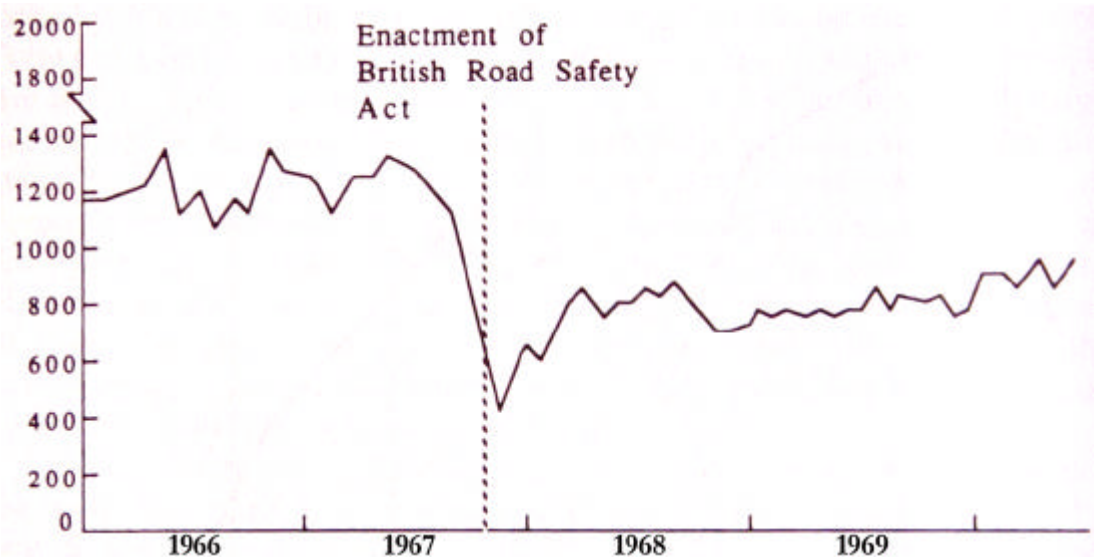


Figure 9. Fatalities for Friday nights, 10 p.m. to midnight; Saturday mornings, midnight to 4 a.m.; Saturday nights, 10 p.m. to midnight; and Sunday mornings, midnight to 4 a.m., corrected for weekend days per month, seasonal variations removed. Broken vertical line represents implementation of the British Road Safety Act. (Source: Ross, Campbell & Glass, 1970).

Why is the effect so apparent in Figure 9 when it was barely discernible (or indeed, not discernible at all) in Figure 2? As was mentioned earlier, an essential feature of the Road Safety Act was a program of roadblocks where drivers were tested for blood alcohol Level. And what time is better for finding drunks on the road than weekend nights? The picture is completed in Figure 10 when one inspects the fatalities curve for the hours commuting to and from work in the morning and late afternoon when there are few drunken drivers. Sorting the data in Figure 2 into two different series in Figures 9 and 10 not only has revealed the intervention effect, but has illuminated the whole question of how the Road Safety Act worked its effect.

To appear in the Handbook of Statistics, Volume 30, Elsevier.

Time Series Analysis with R

A. Ian McLeod, Hao Yu, Esam Mahdi

Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ont., Canada N6A 5B7

The purpose of our article is to provide a summary of a selection of some of the high-quality published computational time series research using R. A more complete overview of time series software available in R for time series analysis is available in the CRAN[1] task views.[2] If you are not already an R user, this article may help you in learning about the R phenomenon and motivate you to learn how to use R. Existing R users may find this selective overview of time series software in R of interest. Books and tutorials for learning R are discussed later in this section. An excellent online introduction from the R Development Core Team is available[3] as well as extensive contributed documentation.[4]

In the area of computational time series analysis, especially for advanced algorithms, R has established itself as the choice of many researchers. R is widely used not only by researchers but also in diverse time series applications and in the teaching of time series courses at all levels. Naturally, there are many other software systems such as *Mathematica* (Wolfram Research, 2011), that have interesting and useful additional capabilities, such as symbolic computation (Smith and Field, 2001; Zhang and McLeod, 2006). For most researchers working with time series, R provides an excellent broad platform.

The history of R has been discussed elsewhere (Gentleman and Ihaka,

Email addresses: aim@stats.uwo.ca (A. Ian McLeod), hyu@stats.uwo.ca (Hao Yu), emahdi@uwo.ca (Esam Mahdi)

[1] Comprehensive R Archive

[2] http://cran.r-project.org/web/views/

[3] http://cran.r-project.org/manuals.html

[4] http://cran.r-project.org/other-docs.html

estimates of the precision using the `boot()` function. These GLM-based time series models are extensively used with longitudinal time series (Li, 1994).

As an illustration, we consider the late night fatality data discussed in Vingilis et al. (2005). The purpose of this analysis was to investigate the effect of the extension of bar closing hours to 2:00 AM that was implemented May 1, 1996. This type of intervention analysis (Box and Tiao, 1975) is known as an interrupted time series design in the social sciences (Shadish et al., 2001). The total fatalities per month for the period starting January 1992 and through to December 1999, corresponding to a time series of length $n = 84$, are shown in Figure 17.

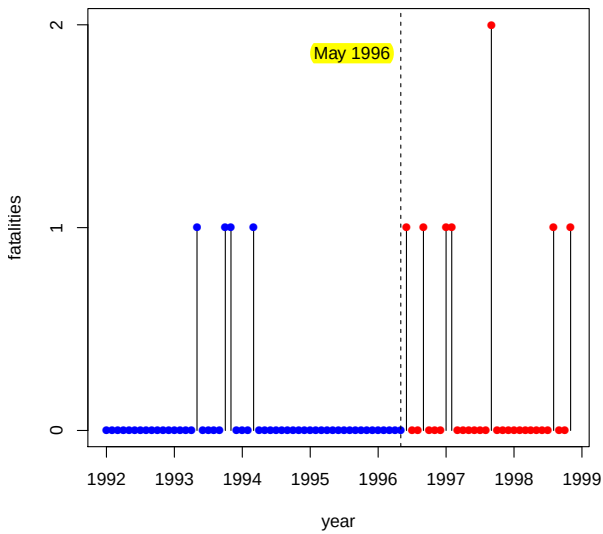


Figure 17: Late night car fatalities in Ontario. Bar closing hours were extended May 1996.

The output from the `glm()` function using `y` as the dependent variable, `y1` as the lagged dependent variable[9], and `x` as the step intervention defined as 0 before May 1, 1996 and 1 after.

```
R >summary(ans)$coefficients
```

[9] `y` and `y1` are the vectors containing the sequence of observed fatalities and its lagged values.

```
              Estimate Std. Error     z value     Pr(>|z|)
(Intercept) -2.53923499  0.5040873 -5.03729193 4.721644e-07
x2           1.16691417  0.6172375  1.89054329 5.868534e-02
y1          -0.06616152  0.6937560 -0.09536712 9.240232e-01
```

The resulting GLM model may be summarized as follows. The total fatalities per month, y_t, are Poisson distributed with mean μ_t, where $\hat{\mu}_t = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\beta}_2 y_{t-1}\}$, $\hat{\beta}_0 \doteq -2.54$, $\hat{\beta}_1 \doteq 1.17$, and $\hat{\beta}_2 \doteq -0.07$. There is no evidence of lagged dependence but the intervention effect, β_2 is significant with $p < 0.10$.

We verified the standard deviation estimates of the parameters by using a non-parametric bootstrap with 1000 bootstrap samples. This computation takes less than 10 seconds on most current PC's. Table 1, produced directly from the R output using the package **xtable**, compares the asymptotic and bootstrap standard deviations. As seen from the table the agreement between the two methods is reasonably good.

	(Intercept)	x2	y1
asymptotic	0.50	0.62	0.69
bootstrap	0.49	0.66	0.75

Table 1: Comparison of asymptotic and bootstrap estimates of the standard deviations in the GLM time series regression

Hidden Markov models provide another time series generalization of Poisson and binomial GLM models (Zucchini and MacDonald, 2009).

```
####### Closing Time, fatalities Interrupted Time Series example
  ## see the Time Series Analysis with R link

> f0 = rpois(50,.5) #prior close fatalities/month
> f1 = rpois(34,1.5) #2 AM close fatalities/month
# additive effect 1 fatality; multiplicative factor 3

> f84 = as.data.frame(c(f0,f1))
> f84$month = 1:84
> names(f84) = c("fatals", "month")
# lag didn't work right for me
> install.packages("DataCombine")
> library(DataCombine)
> f84$fatals_lag = shift(f84$fatals, -1)
Remember to put data in time order before running shift.

> f84$policy = as.numeric(f84$month >50)
> head(f84)
  fatals month fatals_lag policy
1      1     1         NA      0
2      1     2          1      0
3      0     3          1      0
4      0     4          0      0
5      1     5          0      0
6      0     6          1      0

> its = glm(fatals ~ fatals_lag + policy, family = "poisson", data = f84)
> summary(its)
Call: glm(formula = fatals ~ fatals_lag + policy, family = "poisson", data = f84)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9705  -1.1547  -0.3479   0.3798   3.2463

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.29388    0.17878  -1.644    0.100
fatals_lag  -0.11153    0.08264  -1.350    0.177
policy       1.06887    0.22843   4.679 2.88e-06 ***
---
    Null deviance: 122.391  on 82  degrees of freedom
Residual deviance:  99.581  on 80  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 232.17
Number of Fisher Scoring iterations: 5

> exp(coef(its)) # multiplicative effect
(Intercept)  fatals_lag      policy
  0.7453633   0.8944685   2.9120845
> exp(confint(its))
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept) 0.5154239 1.040991
fatals_lag  0.7524371 1.040672
policy      1.8696731 4.591614
> setwd("D:\\drr16\\somgen290\\week8\\")
>
```

Package ‘BayesSingleSub’

February 15, 2013

Type Package

Title Computation of Bayes factors for interrupted time-series designs

Version 0.6.1

Date 2012-11-28

Author Richard D. Morey, Rivka de Vries

Maintainer Richard D. Morey <richarddmorey@gmail.com>

Description The BayesSingleSub package is a suite of functions for computing various Bayes factors for interrupted time-series,based on the models described in de Vries and Morey (submitted).

License GPL-2

Imports coda, mvtnorm, MCMCpack

LazyLoad yes

Repository CRAN

Repository/R-Forge/Project bayesfactorpcl

Repository/R-Forge/Revision 204

Repository/R-Forge/DateTimeStamp 2012-11-29 00:00:43

Date/Publication 2012-11-29 10:04:59

NeedsCompilation yes

R topics documented:

Package ‘Wats’

December 5, 2016

Title Wrap Around Time Series Graphics

Description Wrap-around Time Series (WATS) plots for interrupted time series designs with seasonal patterns.

Version 0.10.3

Date 2015-11-11

Author Will Beasley [aut, cre], Joe Rodgers [aut], Matthew Schuelke [ctb], Ronnie Coleman [ctb], Mark Joseph Lachowicz [ctb]

Maintainer Will Beasley <wibeasley@hotmail.com>

URL https://github.com/OuhscBbmc/Wats

BugReports https://github.com/OuhscBbmc/Wats/issues

Depends R (>= 3.0.0), stats

Imports colorspace, ggplot2, grid, lubridate, plyr, RColorBrewer, testit, zoo

Suggests BayesSingleSub, boot, devtools, knitr, scales, testthat

License MIT + file LICENSE

LazyData TRUE

VignetteBuilder knitr

RoxygenNote 5.0.0

NeedsCompilation no

Repository CRAN

Date/Publication 2016-12-05 18:28:47

R topics documented:

Assesing OKC Fertility with Intercensal Estimates

The MBR manuscript demonstrates WATS plots with data prepared for Rodgers, St. John, & Coleman (2005). In that paper and the MBR manuscript, the denominator of the GFR (General Fertility Rate) is the initial 1990 county population of females ages 15-44.

This vignette uses slightly different Census estimates. The intercensal population estimates (for females ages 15-44) are used for Jan 1990, Jan 1991, Jan 1992, …, Jan 2000. Linear interpolation is then used to complete the remaining 11 months of each year. These monthly estimates become the denominator of each county's monthly GFR.

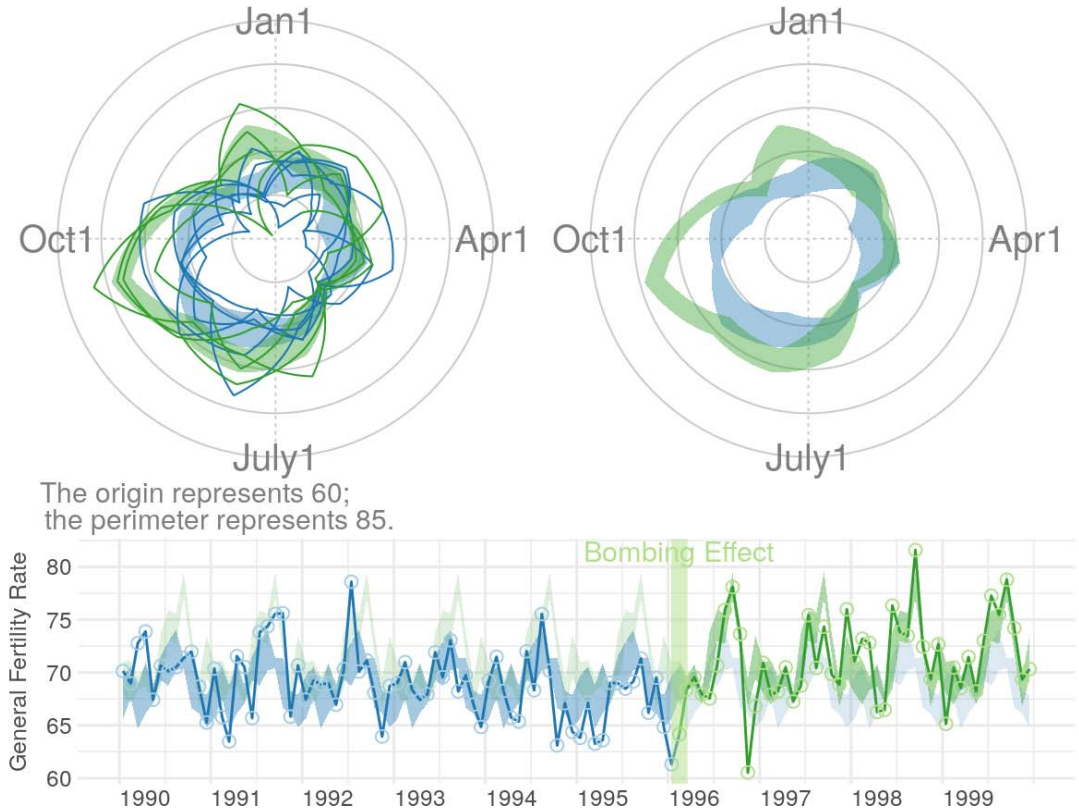
```
changeMonth <- base::as.Date("1996-02-15") #as.Date("1995-04-19") + lubridate::

vpLayout <- function(x, y) { grid::viewport(layout.pos.row=x, layout.pos.col=y)

fullSpread <- function( scores ) {
  return( base::range(scores) ) #A new function isn't necessary.  It's defined
}
hSpread <- function( scores ) {
  return( stats::quantile(x=scores, probs=c(.25, .75)) )
}
seSpread <- function( scores ) {
  return( base::mean(scores) + base::c(-1, 1) * stats::sd(scores) / base::sqrt(
}
bootSpread <- function( scores, conf=.68 ) {
  plugin <- function( d, i ) { base::mean(d[i]) }

  distribution <- boot::boot(data=scores, plugin, R=99) #999 for the publicatio
  ci <- boot::boot.ci(distribution, type = c("bca"), conf=conf)
  return( ci$bca[4:5] ) #The fourth & fifth elements correspond to the lower &
}

darkTheme <- ggplot2::theme(
  axis.title          = ggplot2::element_text(color="gray30", size=9),
  axis.text.x         = ggplot2::element_text(color="gray30", hjust=0),
  axis.text.y         = ggplot2::element_text(color="gray30"),
  axis.ticks.length   = grid::unit(0, "cm"),
  axis.ticks.margin   = grid::unit(.00001, "cm"),
#   panel.grid.minor.y  = element_line(color="gray95", size=.1),
#   panel.grid.major    = element_line(color="gray90", size=.1),
  panel.margin        = grid::unit(c(0, 0, 0, 0), "cm"),
  plot.margin         = grid::unit(c(0, 0, 0, 0), "cm")
)
```



Section 4: Confirmatory Analysis of Interrupted Time Series

The remaining two sections depart from the MBR manuscript analyses. Its goal is to determine if the significant findings of Rodgers, St. John, & Coleman still appear with the modified Census estimates. As shown below, the the post-bombing fertility is still significantly higher than the pre-bombing fertility.

This section uses an approach advocated by McLeod, Yu, & Mahdi (2011), which is consistent other articles, including Rodgers et al. (2005). There are two trends that are de-seasonalized. The first is the 'classic' approach which uses the observed trend line (see `decompose()` on CRAN). The second is a smoothed version, where a loess is passed through the observed data; this smoothed line is then de-seasonalized (see `stl()` on CRAN). Both approaches lead to comparable conclusions. The post-bombing fertility is significantly higher than the pre-bombing fertility (ie, the `step` coefficient is significantly more positive).

tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models

Tobias Liboschik
TU Dortmund University

Konstantinos Fokianos
University of Cyprus

Roland Fried
TU Dortmund University

Abstract

The R package **tscount** provides likelihood-based estimation methods for analysis and modeling of count time series following generalized linear models. This is a flexible class of models which can describe serial correlation in a parsimonious way. The conditional mean of the process is linked to its past values, to past observations and to potential covariate effects. The package allows for models with the identity and with the logarithmic link function. The conditional distribution can be Poisson or Negative Binomial. An important special case of this class is the so-called INGARCH model and its log-linear extension. The package includes methods for model fitting and assessment, prediction and intervention analysis. This paper summarizes the theoretical background of these methods. It gives details on the implementation of the package and provides simulation results for models which have not been studied theoretically before. The usage of the package is illustrated by two data examples. Additionally, we provide a review of R packages which can be used for count time series analysis. This includes a detailed comparison of **tscount** to those packages.

Keywords: aberration detection, autoregressive models, intervention analysis, likelihood, mixed Poisson, model selection, prediction, R, regression model, serial correlation.

This document has been published as a vignette of the R package **tscount**. Last update: May 2016. A stable version has been published as a discussion paper in February 2015 (doi:10.17877/DE290R-7239). Please cite this manuscript as, e.g.:

Tobias Liboschik, Konstantinos Fokianos & Roland Fried (2016). "tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models". Vignette of R package **tscount** version 1.3.0.

1. Introduction

Recently, there has been an increasing interest in regression models for time series of counts and a considerable number of publications on this subject has appeared in the literature. However, most of the proposed methods are not yet available in a statistical software package and hence they cannot be applied easily. We aim at filling this gap and publish a package, named **tscount**, for the popular free and open source software environment R (R Core Team 2016). In fact, our main goal is to develop software for models whose conditional mean depends on previous observations and on its own previous values. These models are quite analogous to the generalized autoregressive conditional heteroscedasticity (GARCH) models (Bollerslev 1986) which were proposed for describing the conditional variance.

Scoring rule	Abbreviation	Definition
logarithmic score	`logarithmic`	$-\log(p_t(y_t))$
quadratic (or Brier) score	`quadratic`	$-2p_t(y_t) + \lVert p_t \rVert^2$
spherical score	`spherical`	$-p_t(y_t)/\lVert p_t \rVert$
ranked probability score	`rankprob`	$\sum_{y=0}^{\infty} (P_t(y) - \mathbb{1}(y_t \leqslant y))^2$
Dawid-Sebastiani score	`dawseb`	$(y_t - \lambda_t)^2/\upsilon_t^2 + 2\log(\upsilon_t)$
normalized squared error score	`normsq`	$(y_t - \lambda_t)^2/\upsilon_t^2$
squared error score	`sqerror`	$(y_t - \lambda_t)^2$

Table 1: Definitions of proper scoring rules $s(P_t, y_t)$ (cf. Czado *et al.* 2009) and their abbreviations in the package; $\lVert p_t \rVert^2 = \sum_{y=0}^{\infty} p_t^2(y)$.

(function `QIC`). We have verified by simulation that in case of a Poisson distribution the QIC approximates the AIC quite satisfactory (not shown here).

6. Intervention analysis

In many applications sudden changes or extraordinary events occur. Box and Tiao (1975) refer to such special events as interventions. This could be for example the outbreak of an epidemic in a time series which counts the weekly number of patients infected with a particular disease. It is of interest to examine the effect of known interventions, for example to judge whether a policy change had the intended impact, or to search for unknown intervention effects and find explanations for them *a posteriori*.

Fokianos and Fried (2010, 2012) model interventions affecting the location by including a deterministic covariate of the form $\delta^{t-\tau}\mathbb{1}(t \geqslant \tau)$, where τ is the time of occurrence and the decay rate δ is a known constant (function `interv_covariate`). This covers various types of interventions for different choices of the constant δ: a singular effect for $\delta = 0$ (spiky outlier), an exponentially decaying change in location for $\delta \in (0, 1)$ (transient shift) and a permanent change of location for $\delta = 1$ (level shift). Similar to the case of covariates, the effect of an intervention is essentially additive for the linear model and multiplicative for the log-linear model. However, the intervention enters the dynamics of the process and therefore its effect on the linear predictor is not purely additive. Our package includes methods to test for such intervention effects developed by Fokianos and Fried (2010, 2012), suitably adapted to the more general model class described in Section 2. The linear predictor of a model with s types of interventions according to parameters $\delta_1, \ldots, \delta_s$ occurring at time points $\tau_1, \ldots, \tau_s$ reads

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \tilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell g(\lambda_{t-j_\ell}) + \boldsymbol{\eta}^\top \boldsymbol{X}_t + \sum_{m=1}^{s} \omega_m \delta_m^{t-\tau_m} \mathbb{1}(t \geqslant \tau_m), \tag{16}$$

where ω_m, $m = 1, \ldots, s$ are the intervention sizes. At the time of its occurrence an intervention changes the level of the time series by adding the magnitude ω_m, for a linear model like (2), or by multiplying the factor $\exp(\omega_m)$, for a log-linear model like (3). In the following paragraphs we briefly outline the proposed intervention detection procedures and refer to the original articles for details.

Our package allows to test whether s interventions of certain types occurring at given time points, according to model (16), have an effect on the observed time series, i.e., to test the hypothesis $H_0 : \omega_1 = \ldots = \omega_s = 0$ against the alternative $H_1 : \omega_\ell \neq 0$ for some $\ell \in \{1, \ldots, s\}$. This is accomplished by employing an approximate score test (function `interv_test`). Under the null hypothesis the score test statistic $T_n(\tau_1, \ldots, \tau_s)$ has asymptotically a χ^2-distribution with s degrees of freedom, assuming some regularity conditions and for a sufficiently large sample size.

For testing whether a single intervention of a certain type occurring at an unknown time point τ has an effect, the package employs the maximum of the score test statistics $T_n(\tau)$ and determines a p value by a parametric bootstrap procedure (function `interv_detect`). If we consider a set D of time points at which the intervention might occur, e.g., $D = \{2, \ldots, n\}$, this test statistic is given by $\widetilde{T}_n = \max_{\tau \in D} T_n(\tau)$. The bootstrap procedure can be computed on multiple cores simultaneously (argument `parallel = TRUE`). The time point of the intervention is estimated to be the value τ which maximizes this test statistic. Our empirical observation is that such an estimator usually has a large variability. It is possible to speed up the computation of the bootstrap test statistics by using the model parameters used for generation of the bootstrap samples instead of estimating them for each bootstrap sample (argument `final.control_bootstrap = NULL`). This results in a conservative procedure, as noted by Fokianos and Fried (2012).

If more than one intervention is suspected in the data, but neither their types nor the time points of its occurrences are known, an iterative detection procedure is used (function `interv_multiple`). Consider the set of possible intervention times D as before and a set of possible intervention types Δ, e.g., $\Delta = \{0, 0.8, 1\}$. In a first step the time series is tested for an intervention of each type $\delta \in \Delta$ as described in the previous paragraph and the p values are corrected to account for the multiple testing by the Bonferroni method. If none of the p values is below a previously specified significance level, the procedure stops and does not identify an intervention effect. Otherwise the procedure detects an intervention of the type corresponding to the lowest p value. In case of equal p values preference is given to interventions with $\delta = 1$, that is level shifts, and then to those with the largest test statistic. In a second step, the effect of the detected intervention is eliminated from the time series and the procedures starts anew and continues until no further intervention effects are detected. Finally, model (16) with all detected intervention effects can be fitted to the data to estimate the intervention sizes and the other parameters jointly. Note that statistical inference for this final model fit has to be done with care.

Liboschik *et al.* (2016) study a model for external intervention effects (modeled by external covariate effects, recall (6) and the related discussion) and compare it to internal intervention effects studied in the two aforementioned publications (argument `external`).

In practical applications, the decay rate δ of a particular intervention effect is often unknown and needs to be estimated. Since the parameter δ is not identifiable when the corresponding intervention size ω is zero, its estimation is nonstandard. As suggested by a reviewer, estimation could be carried out by profiling the likelihood over this parameter. For a single intervention effect this could be done by computing the (quasi) ML estimator of all other parameters for a given decay rate δ. This is repeated for all $\delta \in \Delta$, where Δ is a set of possible decay rates, and the value which results in the maximum value of the log-likelihood is chosen (apply the function `tsglm` repeatedly). Note that this approach affects the validity of the usual statistical inference for the other parameters.

Appendix 3: Using cobalt with Longitudinal Treatments

Noah Greifer

2020-08-31

This is an introduction to the use of `cobalt` with longitudinal treatments. These occur when there are multiple treatment periods spaced over time, with the potential for time-dependent confounding to occur. A common way to estimate treatment effects in these scenarios is to use marginal structural models (MSM), weighted by balancing weights. The goal of applying weights is to simulate a sequential randomization design, where the probability of being assigned to treatment at each time point is independent of each unit's prior covariate and treatment history. For introduction to MSMs in general, see Thoemmes & Ong (2016), VanderWeele, Jackson, & Li (2016), Cole & Hernán (2008), or Robins, Hernán, & Brumback (2000). The key issue addressed by this guide and `cobalt` in general is assessing balance before each treatment period to ensure the removal of confounding.

In preprocessing for MSMs, three types of variables are relevant: baseline covariates, treatments, and intermediate outcomes/time-varying covariates. The goal of balance assessment is to assess whether after preprocessing, the resulting sample is one in which each treatment is independent of baseline covariates, treatment history, and time-varying covariates. The tools in `cobalt` have been developed to satisfy these goals.

The next section describe how to use `cobalt`'s tools to assess balance with longitudinal treatments. First, we'll examine an example data set and identify some tools that can be used to generate weights for MSMs. Next we'll use `bal.tab()`, `bal.plot()`, and `love.plot()` to assess and present balance.

Setup

We're going to use the `iptwExWide` data set in the `twang` package.

```
library("cobalt")
data("iptwExWide", package = "twang")
head(iptwExWide)
```

outcome	gender	age	use0	use1	use2	tx1	tx2	tx3
-0.2782802	0	43	1.1349651	0.4674825	0.3174825	1	1	1
0.5319329	0	50	1.1119318	0.4559659	0.4059659	1	0	1
-0.8173614	1	36	-0.8707776	-0.5353888	-0.5853888	1	0	0
-0.1530853	1	63	0.2107316	0.0053658	-0.1446342	1	1	1
-0.7344267	0	24	0.0693956	-0.0653022	-0.1153022	1	0	1
-0.8519376	1	20	-1.6626489	-0.9313244	-1.0813244	1	1	1

We have the variables `outcome`, which is the outcome, `gender`, `age`, and `use0`, which are the baseline covariates, `use1` and `use2`, which are time-varying covariates measured after treatment periods 1 and 2, and

Note that `CBPS` estimates and assesses balance on MSM weights differently from `twang` and `cobalt`. Its focus is on ensuring balance across all treatment history permutations, whereas `cobalt` focuses on evaluating the similarity to sequential randomization. For this reason, it may appear that `CBMSM` objects have different balance qualities as measured by the two packages.

References

Cole, S. R., & Hernán, M. Á. (2008). Constructing Inverse Probability Weights for Marginal Structural Models. American Journal of Epidemiology, 168(6), 656–664. https://doi.org/10.1093/aje/kwn164

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology, 11(5), 550–560.

Thoemmes, F., & Ong, A. D. (2016). A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models. Emerging Adulthood, 4(1), 40–59. https://doi.org/10.1177/2167696815621645

VanderWeele, T. J., Jackson, J. W., & Li, S. (2016). Causal inference and longitudinal data: a case study of religion and mental health. Social Psychiatry and Psychiatric Epidemiology, 51(11), 1457–1466. https://doi.org/10.1007/s00127-016-1281-9

INSERT RAND TUTORIAL here

Formula uncovers the 'value added'

By CARL CAMPANILE
Last Updated: 6:14 AM, February 25, 2012
Posted: 12:45 AM, February 25, 2012

The teacher rankings released yesterday are based on a sophisticated equation that would require an MIT degree to understand.

Using complex calculations, the formula combines student exam results with 30 variables outside a teacher's control.

Factors affecting the rankings include a poverty index, the number of limited-English-speaking immigrants, kids with disabilities, student suspensions, students who repeated a grade, the number that attended summer school, class sizes, newcomers and demographic information such as race and gender.

As a benchmark, the report looks at how a teacher's students performed on math and reading tests from the prior year.

Based on these controlling factors, the data report projects an achievement score for each student. The prediction is compared with their actual math and reading scores under a teacher.

The difference between the two is called the teacher's "value added" — or contribution to student performance.

Each teacher is then given a ranking from zero to 99.

"The New York model is arguably one of the best in the country in terms of attention to detail. The model is comprehensive, The completeness is a plus," said Rob Meyer,

THIS IS NO WAY TO RATE A TEACHER!

Box 1. A Value-Added Model for a Given Subject, Grade and Year

$$Y_{1i} = \zeta + \lambda Y_{0i} + \beta' X_i + \sum_{l(\text{district})} \sum_{k(\text{school})} \sum_{f(\text{class})} \{\alpha_{jkl}^{\text{class}*} I_{Cijkl} + \alpha_{kl}^{\text{school}*} I_{Sikl} + \alpha_{l}^{\text{district}*} I_{Dil}\} + \varepsilon_i$$

Posttest | Pretest | Student Characteristics | True Total Classroom Effect | True Total School Effect | True Total District Effect | Student Error Term

Classroom Participation Indicator | School Participation Indicator | District Participation Indicator

The New York City Department of Education is making public some inaccurate and misleading information about New York City public school teachers.

The material – contained in what are known as Teacher Data Reports – is supposed to show how well teachers do in their classrooms compared to other teachers.

It doesn't.

In fact, the Teacher Data Reports are compiled using questionable data and employ an unproven and often inaccurate methodology.

1. The reports are based largely on multiple-choice tests, including results for school years where those tests were deemed so flawed by state authorities that they have been abandoned.
2. The methodology, a complex mathematical formula, has a huge margin of error – as much as 54 out of 100 points. This means that teachers identified as top performers could in fact be below average, and teachers identified as low performers could in fact be near the top.
3. The reports themselves are full of bureaucratic errors, including rating teachers for subjects they have not taught and for the progress of students who were never in their classrooms.
4. Even the city's Department of Education admits that these ratings should never be the primary source of judgment about a teacher's performance.

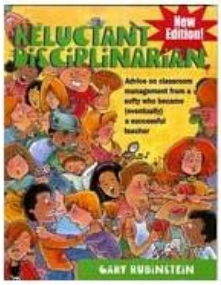
Because this procedure is highly experimental, then-Chancellor Joel Klein promised when it began that the results would be available only to teachers and their supervisors. Then the Department of Education reneged on its pledge and has released them to the public.

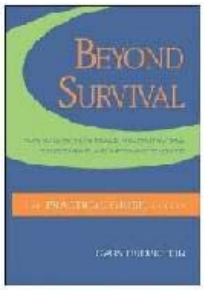
The teachers of this city have dedicated their lives to caring for children. They – and the city's parents – deserve more than judgments based on bad tests, incorrect data and a flawed methodology.

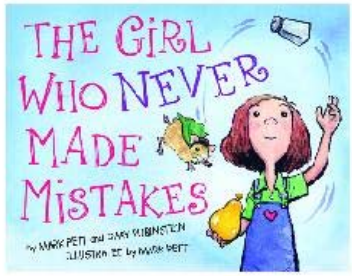
Sincerely,

Michael Mulgrew
President
United Federation of Teachers

UFT
United Federation of Teachers
A Union of Professionals







« IMPACTed Wisdom Truth? Analyzing Released NYC Value-Added Data Part 2 »
Feb 26 2012

Analyzing Released NYC Value-Added Data Part 1

by Gary Rubinstein

Part 1

Part 2

Part 3

Part 4

Part 5

Part 6

Less technical post about VAM: What 'value-added' is and is not
The New York Times, yesterday, released the value-added data on 18,000 New York City teachers collected between 2007 and 2010. Though teachers are irate and various newspapers, The New York Post, in particular, are gleeful, I have mixed feelings.

For sure the 'reformers' have won a battle and have unfairly humiliated thousands of teachers who got inaccurate poor ratings. But I am optimistic that this will be be looked at as one of the turning points in this fight. Up until now, independent researchers like me were unable to support all our claims about how crude a tool value-added metrics still are, though they have been around for nearly 20 years. But with the release of the data, I have been able to test many of my suspicions about value-added. Now I have definitive and indisputable proof which I plan to write about for at least my next five blog posts.

The tricky part about determining the accuracy of these value-added calculations is that there is nothing to compare them to. So a teacher gets an 80 out of 100 on her value added — what does this mean? Does it mean that the teacher would rank 80 out of 100 on some metric that took into account everything that teacher did? As there is no way, at present, to do this, we can't really determine if the 80 was the 'right' score. All we can say is that according to this formula, this teacher got an 80 out of 100. So what we need to 'check' how good of a measure these statistics are some 'objective' truths about teachers — I will describe three which we will see if the value-added measures support.

On The New York Times website they chose to post a limited amount of data. They have the 2010 rating for the teacher and also the career rating for the teacher. These two pieces of data fail to demonstrate the year-to-year variability of these value-added ratings.

I analyzed the data to see if they would agree with three things I think every person would agree upon:

1) A teacher's quality does not change by a huge amount in one year. Maybe they get better or maybe they get worse, but they don't change by that much each year.

2) Teachers generally improve each year. As we tweak our lessons and learn from our mistakes, we improve. Perhaps we slow down when we are very close to retirement, but, in general, we should get better each year.

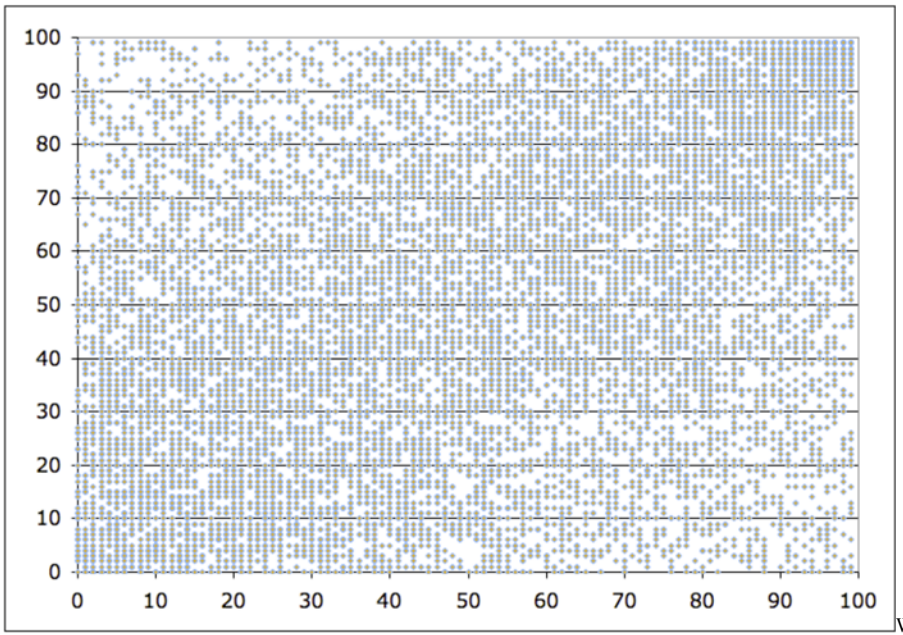
3) A teacher in her second year is way better than that teacher was in her first year. Anyone who taught will admit that they managed to teach way more in their second year. Without expending so much time and energy on classroom management, and also by not having to make all lesson plans from scratch, second year teachers are significantly better than they were in their first year.

Maybe you disagree with my #2. You may even disagree with #1, but you would have to be crazy to disagree with my #3.

Though the Times only showed the data from the 2009-2010 school year, there were actually three files released, 2009-2010, 2008-2009, and 2007-2008. So what I did was 'merge' the 2010 and 2009 files. Of the 18,000 teachers in the 2009-2010 data I found that about 13,000 of them also had ratings from 2008-2009.

Looking over the data, I found that 50% of the teachers had a 21 point 'swing' one way or the other. There were even teachers who had gone up or down as much as 80 points. The average change was 25 points. I also noticed that 49% of the teachers got lower value-added in 2010 than they did in 2009, contrary to my experience that most teachers improve from year to year.

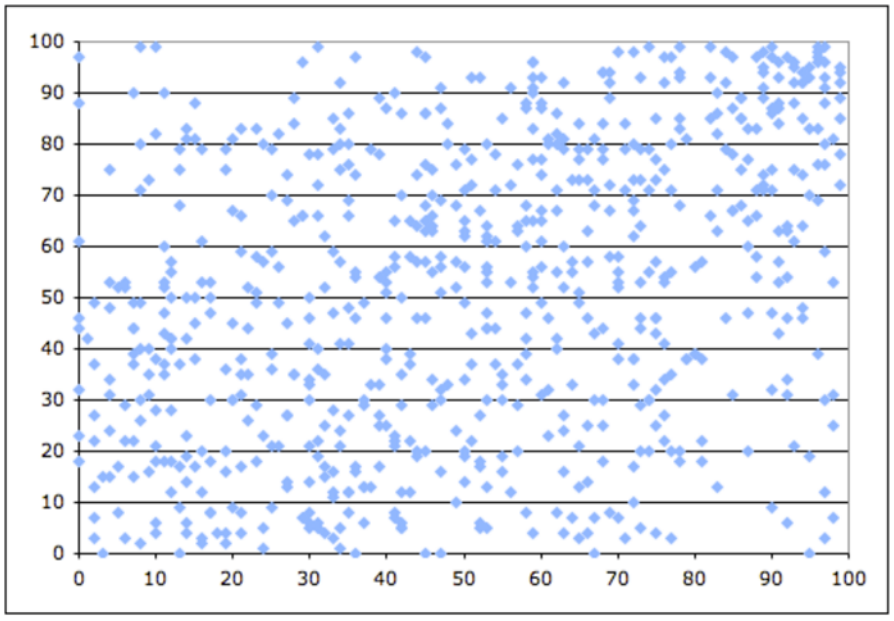
I made a scatter plot with each of these 13,000 teacher's 2008-2009 score on the x-axis and their 2009-2010 score on the y-axis. If the data was consistent, one would expect some kind of correlation with points clustered on an upward sloping line. Instead, I got:



With a correlation coefficient of .35 (and even that is inflated, for reasons I won't get into right now), the scatter plot shows that teachers are not consistent from year to year, contrary to my #1, nor do a good number of them go up, contrary to my #2. (You might argue that 51% go up, which is technically 'most,' but I'd say you'd get about 50% with a random number generator — which is basically what this is.)

But this may not sway you since you do think a teacher's ability can change drastically in one year and also think that teachers get stale with age so you are not surprised that about half went down.

Then I ran the data again. This time, though I used only the 707 teachers who were first year teachers in 2008-2009 and who stayed for a second year in 2009-2010. Just looking at the numbers, I saw that they were similar to the numbers for the whole group. The median amount of change (one way or the other) was still 21 points. The average change was still 25 points. But the amazing thing which definitely proves how inaccurate these measures are, the percent of first year teachers who 'improved' on this metric in their second year was just 52%, contrary to what every teacher in the world knows — that nearly every second year teacher is better in her first year. The scatter plot for teachers who were new teachers in 2008-2009 has the same characteristics of the scatter plot for all 13,000 teachers. Just like the graph above, the x-axis is the value-added score for the first year teacher in 2008-2009 while the y-axis is the value-added score for the same teacher in her second year during 2009-2010.



Reformers beware. I'm just getting started.

EDUCATION | SEPTEMBER 13, 2011

Teachers Are Put to the Test

More States Tie Tenure, Bonuses to New Formulas for Measuring Test Scores

By STEPHANIE BANCHERO And DAVID KESMODEL

As millions of teachers head back to school, many will be facing a new kind of report card that judges them based on how much they help students improve on standardized tests. Stephanie Banchero has details on The News Hub.

MADISON, Wis.—Teacher evaluations for years were based on brief classroom observations by the principal. But now, prodded by President Barack Obama's $4.35 billion Race to the Top program, at least 26 states have agreed to judge teachers based, in part, on results from their students' performance on standardized tests.

So with millions of teachers back in the classroom, many are finding their careers increasingly hinge on obscure formulas like the one that fills a whiteboard in an economist's office here.

Enlarge Image

Narayan Mahon for The Wall Street Journal

Rob Meyer, director of the Value-Added Research Center at the University of Wisconsin, calls his statistical model a 'well-crafted recipe.'

The metric created by Value-Added Research Center, a nonprofit housed at the University of Wisconsin's education department, is a new kind of report card that attempts to gauge how much of students' growth on tests is attributable to the teacher.

For the first time this year, teachers in Rhode Island and Florida will see their evaluations linked to the complex metric. Louisiana and New Jersey will pilot the formulas this year and roll them out next school year. At least a dozen other states and school districts will spend the year finalizing their teacher-rating formulas.

"We have to deliver quality and speed, because [schools] need the data now," said

Rob Meyer, the bowtie-wearing economist who runs the Value-Added Research Center, known as VARC, and calls his statistical model a "well-crafted recipe."

VARC is one of at least eight entities developing such models.

Supporters say the new measuring sticks could improve U.S. educational performance by holding teachers accountable for students' progress. Teachers unions and other critics say the tests' measurements are narrow and that the teachers' scores jump around too much, casting doubt on the validity of the formulas.

Janice Poda, strategic-initiatives director for the Council of Chief State School Officers, said education officials are trying to make sense of the complicated models. "States have to trust the vendor is designing a system that is fair and, right now, a lot of the state officials simply don't have the information they need," she said.

Enlarge Image

Rob Bennett for The Wall Street Journal

Principal Gregory Hodge of New York's Frederick Douglass Academy said data for teachers generally aligns with his classroom observations.

Bill Sanders, who developed the nation's first model to measure teachers' effect on student test scores, advises caution. "People smell the money and there are lots of people rushing out with unsophisticated formulas," said Mr. Sanders, who works as a senior researcher at software firm SAS Institute Inc., which competes with VARC for contracts.

In general, the models use a student's score on, say, a fourth-grade math test to predict how she or he would perform on the fifth-grade test. Some groups, such as VARC, adjust those raw test scores to control for students' outside factors, such as income or race. The actual fifth-grade score is then compared with the expected score, which then translates into the measure of the teacher's added value.

The teacher's overall effectiveness with every student in the classroom is boiled down to one number to rate them from least effective to most effective.

For states and school districts, deciding which vendor to use is critical. The metrics differ in substantial ways and those distinctions can have a significant influence on whether a teacher is rated superior or subpar.

Teaching Moments

1982 Bill Sanders, a professor at the University of Tennessee, begins building value-added models to measure teachers' impact on student achievement. By 1992, Tennessee education officials adopt a refined version of the model to evaluate the state's schools.

2002 President George W. Bush's No Child Left Behind law goes into effect, providing data that can be used to evaluate students' growth.

2005 The University of Wisconsin's Value-Added Research Center, or VARC, is formed by Rob Meyer.

2006 The federal Teacher Incentive Fund begins issuing grants to school systems and states to develop programs to award teachers who raise test scores.

2008 The Houston Independent School District begins issuing bonuses to teachers with high value-added rankings.

2009-2010 New York City starts including value-added data in decisions about whether to grant tenure to teachers.

2010 The $4.35 billion Race to the Top grants create incentives for states to adopt new education policies, including linking test scores to teacher evaluations.

Summer 2010 The Washington, D.C., school district uses value-added data to evaluate and fire teachers.

In August, a New York State Supreme Court judge invalidated a vote by state education officials that would have let districts base 40% of teacher evaluations on state test scores, after the state teachers unions sued saying the law allowed for only 20%. The Los Angeles teachers union has sued to stop the district from launching a pilot program that would grade some teachers using a VARC formula.

Until this year, only a few districts used value-added data. Washington, D.C., used it to fire about 60 teachers; New York City employed it to deny tenure to what it considered underperforming teachers; and Houston relied on it to award bonuses.

Michelle Rhee, who instituted a tough evaluation system when she was schools chancellor in Washington, said she took over a district where many students failed achievement exams, yet virtually every teacher was rated effective.

"While it's not a perfect measure, it was a much fairer, more transparent and consistent way to evaluate teachers," said Ms. Rhee, who now heads StudentsFirst, a nonprofit advocate for education overhauls.

Andy Dewey, an 11th-grade history teacher in Houston, is not a fan. He saw his score bounce

from a positive rating in the 2008-09 school year to a negative rating the following year, decreasing his bonus by about $2,300.

"It's a bunch of garbage," said Mr. Dewey, who is executive vice president of a local teachers union. "These tests are designed to measure students, and they are being used to measure teachers. It's absolutely a misuse of the information."

In New York City, value-added data has been used for the last two years by principals only to make teacher tenure decisions. Last year, 3% of teachers did not receive tenure protection based, in part, on that data. A new state law, passed in an effort to compete for Race the Top, requires the data become an official part of every teacher evaluation.

At Frederick Douglass Academy in Harlem, principal Gregory Hodge uses the value-added results to alter instruction, move teachers to new classroom assignments and pair weak students with the highest performing teachers. Mr. Hodge said the data for teachers generally aligns with his classroom observations. "It's confirming what an experienced principal knows," he said.

—Lisa Fleisher contributed to this article.

Write to David Kesmodel at david.kesmodel@wsj.com

EPI BRIEFING PAPER

ECONOMIC POLICY INSTITUTE • AUGUST 29, 2010 • BRIEFING PAPER #278

PROBLEMS WITH THE USE OF STUDENT TEST SCORES TO EVALUATE TEACHERS

CO-AUTHORED BY SCHOLARS CONVENED BY THE ECONOMIC POLICY INSTITUTE:

EVA L. BAKER, PAUL E. BARTON, LINDA DARLING-HAMMOND, EDWARD HAERTEL, HELEN F. LADD, ROBERT L. LINN, DIANE RAVITCH, RICHARD ROTHSTEIN, RICHARD J. SHAVELSON, AND LORRIE A. SHEPARD

American Statistical Association

Promoting the Practice and Profession of Statistics

ASA Statement on Using Value-Added Models for Educational Assessment

April 8, 2014

Executive Summary

Many states and school districts have adopted Value-Added Models (VAMs) as part of educational accountability systems. The goal of these models, which are also referred to as Value-Added Assessment (VAA) Models, is to estimate effects of individual teachers or schools on student achievement while accounting for differences in student background. VAMs are increasingly promoted or mandated as a component in high-stakes decisions such as determining compensation, evaluating and ranking teachers, hiring or dismissing teachers, awarding tenure, and closing schools.

The American Statistical Association (ASA) makes the following recommendations regarding the use of VAMs:

- The ASA endorses wise use of data, statistical models, and designed experiments for improving the quality of education.
- VAMs are complex statistical models, and high-level statistical expertise is needed to develop the models and interpret their results.
- Estimates from VAMs should always be accompanied by measures of precision and a discussion of the assumptions and possible limitations of the model. These limitations are particularly relevant if VAMs are used for high-stakes purposes.

Research on VAMs has been fairly consistent that aspects of educational effectiveness that are measurable and within teacher control represent a small part of the total variation in student test scores or growth; most estimates in the literature attribute between 1% and 14% of the total variability to teachers. This is not saying that *teachers* have little effect on students, but that *variation* among teachers accounts for a small part of the variation in scores. The majority of the variation in test scores is attributable to factors outside of the teacher's control such as student and family background, poverty, curriculum, and unmeasured influences.

The VAM scores themselves have large standard errors, even when calculated using several years of data. These large standard errors make rankings unstable, even under the best scenarios for modeling. Combining VAMs across multiple years decreases the standard error of VAM scores. Multiple years of data, however, do not help problems caused when a model systematically undervalues teachers who work in specific contexts or with specific types of students, since that systematic undervaluation would be present in every year of data.

A VAM score may provide teachers and administrators with information on their students' performance and identify areas where improvement is needed, but it does not provide information on how to improve the teaching. The models, however, may be used to evaluate effects of policies or teacher training programs by comparing the average VAM scores of teachers from different programs. In these uses, the VAM scores partially adjust for the differing backgrounds of the students, and averaging the results over different teachers improves the stability of the estimates.

Statistical science has an important role to play in raising the quality of education, through developing and refining statistical models for use in education, providing guidance on designing experiments and interpreting statistical results, and applying quality and process improvement expertise to help guide judgments in the presence of uncertainty. The ASA promotes sound use of statistical methodology for improving education.

Using R for Estimating Longitudinal Student Achievement Models

by J.R. Lockwood[1], *Harold Doran and Daniel F. McCaffrey*

Overview

The current environment of test-based accountability in public education has fostered increased interest in analyzing longitudinal data on student achievement. In particular, "value-added models" (VAM) that use longitudinal student achievement data linked to teachers and schools to make inferences about teacher and school effectiveness have burgeoned. Depending on the available data and desired inferences, the models can range from straightforward hierarchical linear models to more complicated and computationally demanding cross-classified models. The purpose of this article is to demonstrate how R, via the `lme` function for linear mixed effects models in the **nlme** package (Pinheiro and Bates, 2000), can be used to estimate all of the most common value-added models used in educational research. After providing background on the substantive problem, we develop notation for the data and model structures that are considered. We then present a sequence of increasingly complex models and demonstrate how to estimate the models in R. We conclude with a discussion of the strengths and limitations of the R facilities for modeling longitudinal student achievement data.

Background

The current education policy environment of test-based accountability has fostered increased interest in collecting and analyzing longitudinal data on student achievement. The key aspect of many such data structures is that students' achievement data are linked to teachers and schools over time. This permits analysts to consider three broad classes of questions: what part of the observed variance in student achievement is attributable to teachers or schools; how effective is an individual teacher or school at producing growth in student achievement; and what characteristics or practices are associated with effective teachers or schools. The models used to make these inferences vary in form and complexity, and collectively are known as "value added models" (VAM, McCaffrey et al., 2004).

However, VAM can be computationally demanding. In order to disentangle the various influences on achievement, models must account simultaneously for the correlations among outcomes within students over time and the correlations among outcomes by students sharing teachers or schools in the current or previous years. The simplest cases have student outcomes fully nested within teachers and schools, in which case standard hierarchical linear models (Raudenbush and Bryk, 2002; Pinheiro and Bates, 2000) are appropriate. When students are linked to changing teachers and/or schools over time, more complicated and computationally challenging cross-classified methods are necessary (Bates and DebRoy, 2003; Raudenbush and Bryk, 2002; Browne et al., 2001). Although the `lme` function of the **nlme** library is designed and optimized for nested structures, its syntax is flexible enough to specify models for more complicated relational structures inherent to educational achievement data.

Data structures

The basic data structures supporting VAM estimation are longitudinal student achievement data $\boldsymbol{Y}_k = (Y_{k0}, \ldots, Y_{kT})$ where k indexes students. Typically the data represent scores on an annual standardized examination for a single subject such as mathematics or reading, with $t = 0, \ldots, T$ indexing years. For clarity, we assume that all students are from the same cohort, so that year is coincident with grade level. More careful consideration of the distinction between grade level and year is necessary when modeling multiple cohorts or students who are held back. We further assume that the scores represent achievement measured on a single developmental scale so that Y_{kt} is expected to increase over time, with the gain scores $(Y_{kt} - Y_{k,t-1})$ representing growth along the scale. The models presented here can be estimated with achievement data that are not scaled as such, but this complicates interpretation (McCaffrey et al., 2004).

As noted, the critical feature of the data underlying VAM inference is that the students are linked, over time, to higher-level units such as teachers and schools. For some data structures and model specifications, these linkages represent a proper nesting relationship, where outcomes are associated with exactly one unit in each hierarchical level (e.g. outcomes nested within students, who are nested within schools). For more complex data, however, the

[1]This material is based on work supported by the National Science Foundation under Grant No. 99-86612. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

A Potential Outcomes View of Value-Added Assessment in Education

Donald B. Rubin, Elizabeth A. Stuart, and Elaine L. Zanutto

Invited discussion to appear in *Journal of Educational and Behavioral Statistics*

November 12, 2003

1 Introduction

1.1 Assessment and Accountability in Education

There has been substantial interest in recent years in the performance and accountability of teachers and schools, partially due to the No Child Left Behind legislation, which requires states to develop a system of sanctions and rewards to hold districts and schools accountable for academic achievement. This focus has lead to an increase in "high-stakes" testing with publicized school rankings and test results. The papers by Ballou et al. (2004), McCaffrey et al. (2004) and Tekwe et al. (2004) approach the estimation of school and teacher effects through a variety of statistical models, known as "value-added" models in the education literature. There are many complex issues involved, and we applaud the authors for addressing this challenging topic.

In this discussion we approach value-added assessment from a "potential outcomes" (Rubin Causal Model, RCM) point of view (Rubin 1974, 1978, 2003; Holland 1986; Little and Rubin 2001), with the goal of clarifying the estimation goals and understanding the limitations of data for the types of comparisons being sought. We discuss the challenges in conceptualizing and obtaining reliable estimates of the causal effects of teachers or schools. We also present an idea for future research that focuses on assessing the effect of implementing reward structures based on value-added models, rather than on assessing the effect of teachers and schools themselves, which we feel is a more relevant policy question, and also one that is more easily addressed.

1.2 Value-added Models–Causal or Descriptive?

The value-added models used to estimate the effectiveness of teachers and schools in Ballou et al. (2004), McCaffrey et al. (2004), and Tekwe et al. (2004) range from a relatively straightforward fixed effects model (Tekwe et al. 2004) to a relatively complex and general multivariate, longitudinal mixed-model (McCaffrey et al. 2004) with either test scores or test score gains as outcomes. These models incorporate parameters for school and teacher effects (including lagged teacher effects), parameters for student-level, classroom level, and school-level covariate effects, and parameters to allow for residual intra-class correlation among outcomes for students in the same class. These models attempt to address problems such as apportioning school effects to more than one school (for students who attended more than one school in the year prior to the test) and the persistence of teacher effects into the future. But none of these articles attempts to define

precisely the quantity that is the target of estimation, except in the rather oblique sense of seeing what the models estimate as samples get larger and larger. Thus, there is a focus on the estimation techniques rather than the definition of the estimand.

The goal of the value-added literature seems to be to estimate the "causal effects" of teachers or schools; that is, to determine how much a particular teacher (or school) has "added value" to their students' test scores. It is implied that the effects being estimated are causal effects: the effect on students of being in school A (or with teacher T) on their test scores, where schools and teachers are to be rewarded or punished "because" of their estimated effects on students. But is it possible to get reliable causal estimates in this setting? The potential outcomes perspective (RCM) provides a framework to clarify this issue concerning whether we are seeking causal or descriptive answers. Before delving into this perspective, it may be helpful to connect this "value-added" problem to one in "hospital profiling", where a relatively substantial literature already exists on a similar problem.

1.3 A Related Problem–Hospital Profiling

The problem of comparing the performance of schools, accounting for the backgrounds of the students they serve, is similar to that addressed in the literature on hospital profiling. In hospital profiling, the aim is to assess the performance of particular hospitals in treating diseases, after accounting for the varying patient populations served by each hospital, so called "case-mix adjustment" (e.g., Goldstein and Spiegelhalter 1996; Christiansen and Morris 1997; Burgess et al. 2001). Additionally, hospital profiling is often done only within subgroups of hospitals, such as by type (teaching, general, psychiatric). Making comparisons within types may also be useful in the school setting, for example comparing public inner-city schools only with other public inner-city schools. This issue will be discussed further in Section 2, which stresses the importance of comparing comparable units.

The school setting appears to be even more complicated than the hospital profiling one, for a variety of reasons. First, there is interest in longitudinal effects with a desire to separate out the effects of last year's and this year's teachers. Thus, we seem to need to have "vertically-linked" test scores that can be compared over time. Second, longitudinal data are not strictly hierarchically nested since students do not remain together as a class over time; not only are students' teachers changing each year, but their classmates are also changing. Finally, there is substantial missing test score data in the school setting and obviously relevant unobserved covariates, such as the motivational levels of the students. These are all issues that appear to be more complex in the school setting than in the hospital profiling setting; nevertheless, workers in the school assessment setting could possibly find relevant ideas in the hospital profiling literature.

The hospital profiling literature also points out other possible problems when such methods for assessing "successful" teachers or schools are implemented. For example, there is the possibility that once any system of assessment is implemented, schools will "game the system" to obtain results that unduly benefit them. In the hospital profiling setting, Green and Wintfeld (1995) describe an increase in reported incidence of risk factors that would increase expected mortality, such as congestive heart failure, after implementation of a system to generate case-mix adjusted physician-specific mortality rates. Presumably, doctors hoped to improve their performance ratings by inflating the entry-level risk statuses of their patients. In the school setting, schools may place more students in special-education or English-as-a-second-language courses so that their student body appears to be more disadvantaged or so that some groups of students are excluded from the overall analysis of test scores.

you have got to be kidding.....

matched districts, pseudo-districts could be formed by matching individual schools within the districts.

With observational data, fully unobserved covariates that may affect both the decision to implement VAA (treatment assignment) and the outcome are a concern. It may be that states in which there is high value placed on education and measurement and thus implement VAA also have higher values of the outcome. Because assignment to implement VAA is not randomized, two states may look identical on observed covariates, but have different political environments or educational values that will affect both whether they implement VAA and their outcomes. In addition to simulations such as in McCaffrey et al. (2004) that assess the effect of omitted variables, sensitivity analyses such as those described in Rosenbaum and Rubin (1983b) could be used to explore the sensitivity of results to an unobserved variable that affects both treatment assignment and outcome, i.e., non-ignorable treatment assignment.

4 Conclusion

Value-added assessment is a complex issue, and we appreciate the efforts of Ballou et al. (2004), McCaffrey et al. (2004) and Tekwe et al. (2004). However, we do not think that their analyses are estimating causal quantities, except under extreme and unrealistic assumptions. We argue that models such as these should not be seen as estimating causal effects of teachers or schools, but rather as providing descriptive measures. It is the reward structures based on such value-added models that should be the objects of assessment, since they can actually be (and are being) implemented. Of course, this focus requires a dramatic shift from current thinking, but a shift towards studying interventions that can be implemented and toward evaluations of them that can be conducted. We look forward to discussion of this approach.

5 References

Abadie, A. and Gardeazabal, J. (2003). "The economic costs of conflict: A case study of the Basque country." *American Economic Review* 93 (1): 113-132.

Ballou, D., Sanders, W., and Wright, P. (2004). "Controlling for student background in value-added assessment of teachers." *Journal of Educational and Behavioral Statistics.*

Barnard, J., Du, J., Hill, J., and Rubin, D.B. (1998) "A Broader Template for Analyzing Broken Randomized Experiments", *Sociological Methods and Research*, 27: 285–317.

Barnard, J., Frangakis, C., Hill, J, and Rubin, D.B. (2003) "A Principal Stratification Approach to Broken Randomized Experiments: A Case Study of Vouchers in New York City" (with discussion and rejoinder) *Journal of the American Statistical Association* 98: 299-323.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York, NY: McGraw Hill.

Burgess et al. (2001). "Medical Profiling: Improving Standards and Risk Adjustments Using Hierarchical Models" *Journal of Health Economics* 19: 291-309.

Christiansen, C.L. and Morris, C.M. (1997). "Improving the Statistical Approach to Health Care Provider Profiling" *Annals of Internal Medicine* 127: 764-768.

Psychological Methods
2007, Vol. 12, No. 3, 247–267

1082-989X/07/$12.00 DOI: 10.1037/1082-989X.12.3.247

Combining Propensity Score Matching and Group-Based Trajectory Analysis in an Observational Study

Amelia Haviland
RAND Corporation

Daniel S. Nagin
Carnegie Mellon University

Paul R. Rosenbaum
University of Pennsylvania

In a nonrandomized or observational study, propensity scores may be used to balance observed covariates and trajectory groups may be used to control baseline or pretreatment measures of outcome. The trajectory groups also aid in characterizing classes of subjects for whom no good matches are available and to define substantively interesting groups between which treatment effects may vary. These and related methods are illustrated using data from a Montreal-based study. The effects on subsequent violence of gang joining at age 14 are studied while controlling for measured characteristics of boys prior to age 14. The boys are divided into trajectory groups based on violence from ages 11 to 13. Within trajectory group, joiners are optimally matched to a variable number of controls using propensity scores, Mahalanobis distances, and a combinatorial optimization algorithm. Use of variable ratio matching results in greater efficiency than pair matching and also greater bias reduction than matching at a fixed ratio. The possible impact of failing to adjust for an important but unmeasured covariate is examined using sensitivity analysis.

Keywords: observational study, propensity score, trajectory group

A key aim of empirical research in developmental psychopathology and life course studies is to measure the effect on a course of development of an intervention or event that occurs at a particular time. Ideally, such effects would be estimated with experimental data, in which the intervention is randomly assigned to some participants and denied to others, but many interventions that affect development cannot be randomized, for ethical or practical reasons. In these situations, inferences must be drawn from observational data. Recalling a suggestion of Dorn (1953), Cochran (1965) argued that the design of an observational study should be organized around the question "How should the study be conducted if it were possible to do it by controlled experimentation?" (p. 236). Certain issues are common to an experiment and an observational study, and these shared issues are brought into focus by thinking about the simpler situation of an experiment. One then tries to reconstruct, to the limited extent possible, the circumstances of the experiment from the observational data. Finally, one tries to address the weaknesses that are present in the observational study but that would have been avoided in an experiment. A similar perspective is discussed in Campbell (1957); Campbell and Stanley (1963); Rubin (1974); Meyer (1995); Shadish, Cook, and Campbell (2002); and Rosenbaum (2002b, 2005a).

A treatment applied at a particular time, say at age 14, may affect subsequent development, but current exposure to the treatment may also affect future exposure to the treatment. In many contexts, the effect of current exposure to the treatment on subsequent exposure to the treatment is quite important. To take an extreme instance, the most conspicuous and immediate effect of exposure to a highly addictive substance may be to seek further and continued exposure to that same treatment. The opposite pattern is also possible.

Amelia Haviland, Statistics Group, RAND Corporation, Pittsburgh, Pennsylvania; Daniel S. Nagin, Heinz School, Carnegie Mellon University; Paul R. Rosenbaum, Statistics Department, Wharton School, University of Pennsylvania.

This work was supported by grants from the Methodology, Measurement and Statistics Program and the Statistics and Probability Program of the U.S. National Science Foundation and Grant RO1 MH65611-01A2 from the National Institute of Mental Health. It also made heavy use of data collected with the support from Québec's Conseil Québécois de la Recherche en Sciences Sociales and Formation des Chercheurs et Aide à la Recherche funding agencies, Canada's National Health Research Development Program and Social Sciences and Humanities Research Council funding agencies, and the Molson Foundation.

Correspondence conerning this article should be addressed to Daniel S. Nagin, H. John Heinz III School of Public Policy and Management, Carnegie Mellon University, 4800 Forbes Avenue, Pittsburgh, PA 15213-3890. E-mail: dn03@andrew.cmu.edu

Developmental Psychology
2008, Vol. 44, No. 2, 422–436

0012-1649/08/$12.00 DOI: 10.1037/0012-1649.44.2.422

Combining Group-Based Trajectory Modeling and Propensity Score Matching for Causal Inferences in Nonexperimental Longitudinal Data

Amelia Haviland
Rand Corporation

Daniel S. Nagin
Carnegie Mellon University

Paul R. Rosenbaum
University of Pennsylvania

Richard E. Tremblay
International Laboratory for Child and Adolescent Mental Health Development, INSERM U669, University College Dublin, and University of Montréal

A central theme of research on human development and psychopathology is whether a therapeutic intervention or a turning-point event, such as a family break-up, alters the trajectory of the behavior under study. This article describes and applies a method for using observational longitudinal data to make more transparent causal inferences about the impact of such events on developmental trajectories. The method combines 2 distinct lines of research: work on the use of finite mixture modeling to analyze developmental trajectories and work on propensity score matching. The propensity scores are used to balance observed covariates and the trajectory groups are used to control pretreatment measures of response. The trajectory groups also aid in characterizing classes of subjects for which no good matches are available. The approach is demonstrated with an analysis of the impact of gang membership on violent delinquency based on data from a large longitudinal study conducted in Montréal, Canada.

Keywords: causal inference, trajectories, propensity scores

A central theme of research on human development and psychopathology is whether a therapeutic intervention or a turning-point event, such as a family break-up, alters the trajectory of the behavior under study. As it is often either unethical or impractical to pursue research on this theme through experimental methods, researchers collect rich nonexperimental longitudinal data and implement methods appropriate for causal analysis of observational data. Although there are always threats to the validity of giving causal interpretation to results from observational data because of selection into the "treatment" group, the threats take a particular form in developmental research based on observational longitudinal data. The problem arises when treatment affects the future direction of the behavior under study, and in addition prior pathways of the behavior predict both entry into treatment and the future direction of the behavior. For instance, early antisocial behavior predicts both later antisocial behavior and school failure, whereas school failure also can affect later antisocial behavior (Maguin, Loeber, & LeMahieu, 1993; Nagin, Pagani, Tremblay, & Vitaro, 2003; Pagani, Boulerice, Vitaro, & Tremblay, 1999).

This article describes and applies a method for using observational longitudinal data to make more transparent causal inferences about the impact of a therapeutic intervention or a turning-point event on developmental trajectories of behavior. This transparency is achieved by modeling our approach after some key attributes of experiments. Although the developmental research scenario we describe has particular goals and challenges, it also has particular benefits. The method we propose, which combines group-based trajectory modeling with propensity score matching, is designed to answer the types of research questions just described and make use of the strengths of the data. It does this in four ways. Propensity score matching is employed to create a control group that is comparable to the treated group with respect to the observed covariates. The group-based trajectory model allows us to take a developmental view of "comparable" by matching treated individuals with individuals who were not treated but who appeared to be on a similar developmental pathway for the behavior under study prior to treatment. In addition, the richness typical of data collected under these circumstances means that participants who are comparable on observed covariates are comparable on an important group of relevant covariates. Second, in addition to the propensity score matching providing more transparent causal inferences, the group-based trajectory groups provide a means of identifying

Amelia Haviland, Rand Corporation, Pittsburgh, PA; Daniel S. Nagin, Heinz School of Public Policy and Management, Carnegie Mellon University; Paul R. Rosenbaum, Department of Statistics, University of Pennsylvania; and Richard E. Tremblay, International Laboratory for Child and Adolescent Mental Health Development, INSERM U669, Paris, France; University College Dublin, Ireland; University of Montréal, Canada.

This research has been supported by National Science Foundation Grants SES-99113700 and SES-0647576, and National Institute of Mental Health Grant RO1 MH65611-01A2. It has also made heavy use of data collected with support from the Fonds du Québec pour la Recherche sur la Société et la Culture and the Fonds de Formation des Chercheurs et d'Aide à la Recherche in Quebec, Canada; the Canadian Institute for Health Research and the Social Sciences and Humanities Research Council of Canada; and the Molson Foundation.

Correspondence concerning this article should be addressed to Daniel S. Nagin, Heinz School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: dn03@andrew.cmu.edu

PSYCHOMETRIKA— VOL. 70, NO. 3, 557–578
SEPTEMBER 2005
DOI: 10.1007/s11336-004-1261-y

CAUSAL INFERENCES WITH GROUP BASED TRAJECTORY MODELS

AMELIA M. HAVILAND

RAND CORPORATION

DANIEL S. NAGIN

CARNEGIE MELLON UNIVERSITY

A central theme of research on human development and psychopathology is whether a therapeutic intervention or a turning-point event, such as a family break-up, alters the trajectory of the behavior under study. This paper lays out and applies a method for using observational longitudinal data to make more confident causal inferences about the impact of such events on developmental trajectories. The method draws upon two distinct lines of research: work on the use of finite mixture modeling to analyze developmental trajectories and work on propensity scores. The essence of the method is to use the posterior probabilities of trajectory group membership from a finite mixture modeling framework, to create balance on lagged outcomes and other covariates established prior to t for the purpose of inferring the impact of first-time treatment at t on the outcome of interest. The approach is demonstrated with an analysis of the impact of gang membership on violent delinquency based on data from a large longitudinal study conducted in Montreal.

Key words: Causal inference, finite mixture models, propensity scores.

1. Introduction

A developmental trajectory describes the course of a behavior over age or time. Two central themes of research on human development and psychopathology are: whether a therapeutic intervention or a turning-point event, such as a family break-up, alters the trajectory of the behavior under study, and whether such "treatment" effects depend upon the prior developmental course of the behavior. A key obstacle to making valid inferences about treatment effects in this problem context is that there may be reciprocal relationships between developmental trajectories and the events that are shaping them. Indeed such reciprocal relationships are at the core of theories of life course development (Elder, 1988; Magnusson, 1988). Specifically, the trajectory itself may affect the likelihood of experiencing these events. For instance, a trajectory of longstanding depression may trigger divorce which, in turn, may deepen depression. Reciprocal relationships of this sort, which Robins, Greenland, and Hu (1999) term "feedback from output to input," can make it very difficult to distinguish cause from effect in observational data.

This paper lays out an approach designed to provide a more confident basis for estimating trajectory contingent treatment effects from the types of observational data commonly collected in longitudinal studies of human development and psychopathology. A hallmark of these studies is repeated measurements of various behaviors and of events that are thought to affect them. The approach is demonstrated with an analysis of the effect of gang membership on trajectories of

The research has been supported by the National Science Foundation (NSF) (SES-99113700) and the National Institute of Mental Health (RO1 MH65611-01A2). It also made heavy use of data collected with the support from Québec's CQRS and FCAR funding agencies, Canada's NHRDP and SSHRC funding agencies, and the Molson Foundation. We thank Stephen Fienberg, Susan Murphy, Paul Rosenbaum, the editor, Paul De Boeck, and two anonymous reviewers for their insightful suggestions.

Requests for reprints should be sent to Amelia M. Haviland, Associate Statistican, Rand Corporation, Pittsburgh, PA 15213, USA. E-mail: Amelia_Haviland@rand.org

Sociological Methods & Research
http://smr.sagepub.com/

Group-based Trajectory Modeling Extended to Account for Nonrandom Participant Attrition

Amelia M. Haviland, Bobby L. Jones and Daniel S. Nagin

Sociological Methods & Research 2011 40: 367 originally published online 21 April 2011
DOI: 10.1177/0049124111400041

The online version of this article can be found at:
http://smr.sagepub.com/content/40/2/367

Published by:
SAGE
http://www.sagepublications.com

Figure 2A simulates the often-reported decrement in adult intelligence as a reflection of marked generational (cohort) differences in the rate of intellectual development and the final level of attainment. All cohorts (1910–1950) are assumed to show an increasing developmental function covering the age period from 10 to 50. However, since any cross-

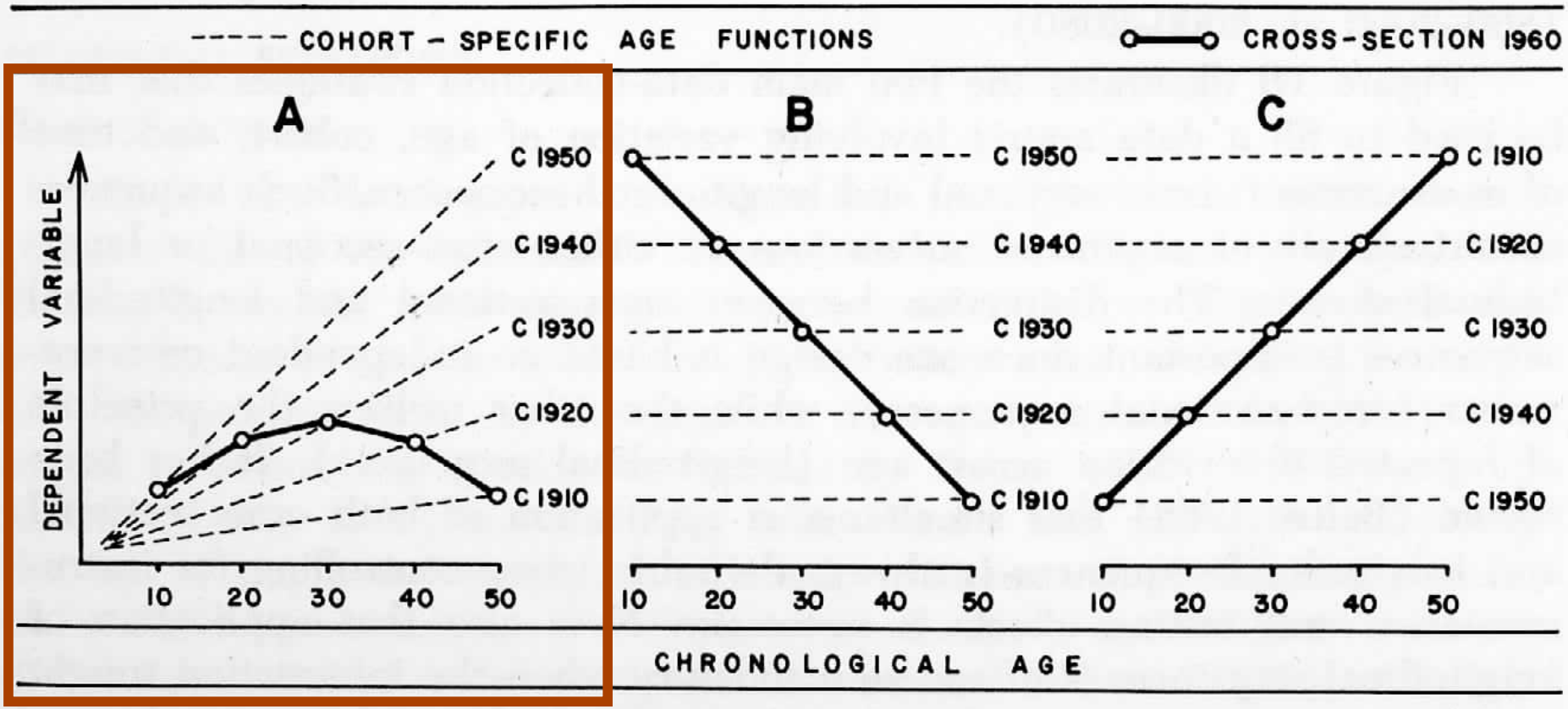


FIGURE 2.—Hypothetical examples illustrating the confounding of age and cohort differences in cross-sectional research.

sectional study is based on a single age-specific observation per cohort only, a cross-sectional study conducted in 1960, for example, results in an inverted U-shape function that suggests decremental changes starting in early adulthood. Obviously, therefore, the conclusions for the outcome illustrated in figure 2*A* must be that the observed cross-sectional *age differences* do not reflect *age changes* but an interaction between age and generation effects. In fact, it is important to note that the resulting cross-sectional gradient corresponds neither to any of the single cohort-specific gradients nor to the average of all of them.

Figure 2*B* and *C* simulates other outcomes involving cohort differences and their effect on cross-sectional gradients, again primarily for didactic purposes. In both cases, the primary objective is to show that cross-sectional gradients of either a decreasing (fig. 2*B*) or increasing (fig. 2*C*) nature can result even if none of the cohorts involved exhibits any developmental change. Hypothesizing, for example, that a behavior class such as conservativism changes in historical "jumps," then figure 2*B* may very well be representative. Similarly, hypothesizing that there are historical trends in such characteristics as body weight (see Damon 1965, e.g.), then figure 2*C* would illustrate how cross-sectional studies could result in increment functions covering the adult age period despite the fact that none of the single cohorts exhibits any ontogenetic change at all for the same age period.

When dealing with sequential data derived from concrete research

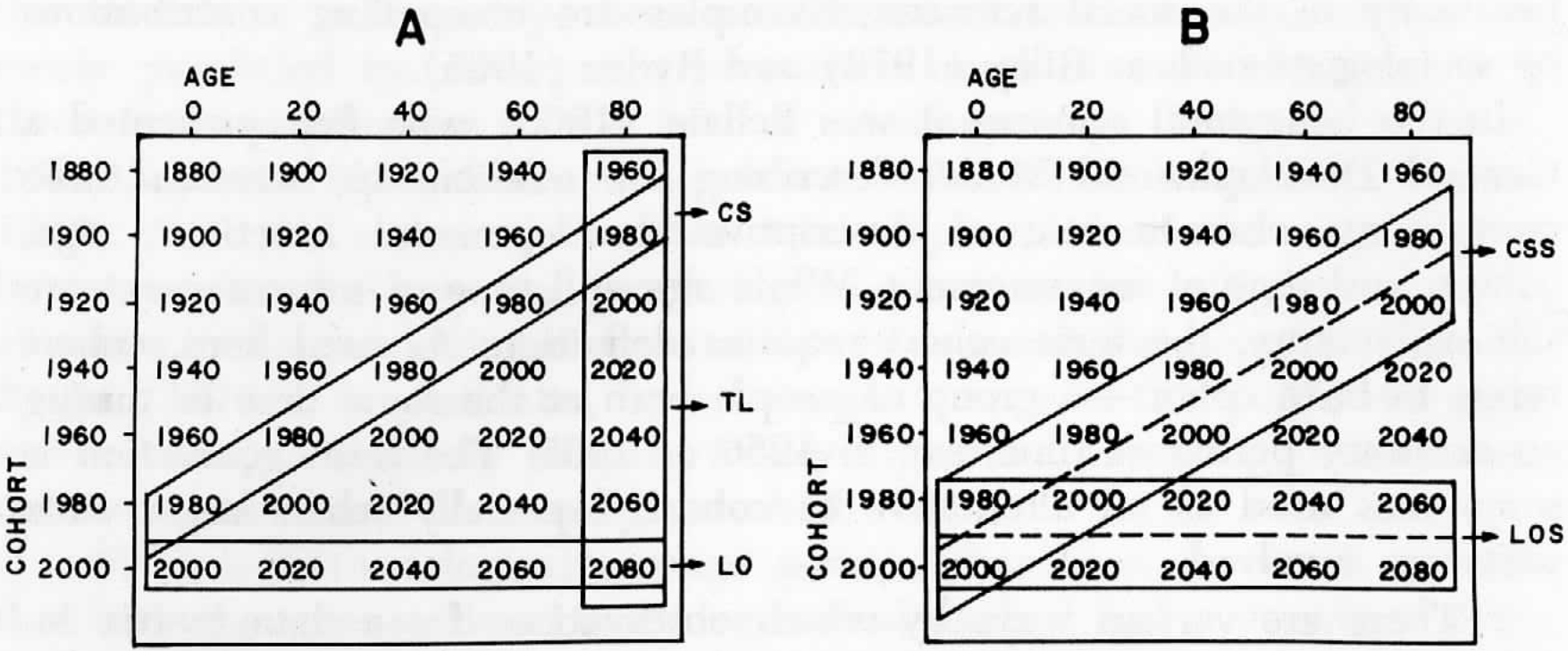


FIGURE 1.—A modified version of Schaie's General Developmental Model illustrating conventional (CS = cross-sectional, LO = longitudinal, TL = time-lag) and sequential (CSS = cross-sectional sequences, LOS = longitudinal sequences) data-collection strategies. Cell entries refer to times of measurement.

TABLE 1

SHORT-TERM LONGITUDINAL SEQUENCES FOR THE STUDY OF ADOLESCENT DEVELOPMENT: DATA COLLECTION AND DESIGN[a]

		Age					
Cohort	Sex	13	14	15	16	17	18
1959	M	1972					
	F						
1958	M	1971	1972				
	F						
1957	M	1970	1971	1972			
	F						
1956	M		1970	1971	1972		
	F						
1955	M			1970	1971	1972	
	F						
1954	M				1970	1971	1972
	F						
1953	M					1970	1971
	F						
1952	M						1970
	F						

NOTE.—Entries represent times of observation (repeated measurement). Mean testing time (range ± 2 months) is January 1 of the year listed. The broken parallelogram indicates the data matrix used for main analyses reported.

[a] To estimate instrumentation and testing effects (internal validity), a set of randomly selected groups of cohorts 1954–1958 were observed for the first and only time in 1972. In addition, to estimate selective dropout effects (external validity), the core longitudinal sample was contrasted with the dropout sample at the first time of measurement (1970).

dropped out with respect to our measurement variables, the external

Nesselroade and Baltes (1974) conducted a study that demonstrated a cohort effect. Longitudinal sequences of cohorts born in 1954, 1955, 1956, and 1957 and tested every year from 1970 to 1972 were used. Over 1,800 subjects were drawn from public schools in West Virginia and given personality and ability tests. Data analyses showed significant main effects of time of measurement on 7 of 10 personality variables and significant main effects of cohort on 2 of 10 personality variables. The main effects of cohort on the personality variables Independence and Achievement can be seen in Figure 1.1. The 14-year-olds tested in 1972 scored much higher in Independence than 14-year-olds tested in 1970 or 1971, while the 14-year-olds tested in 1970 scored higher in Achievement than 14-year-olds tested in 1971 and 1972. The researchers interpreted these findings as suggesting that the social-cultural context is more influential than maturation in adolescent personality development. This study is also important in that it demonstrated retest effects for the mental abilities testing, and attrition influenced the findings, such that those who remained in the study performed better than those who dropped out.

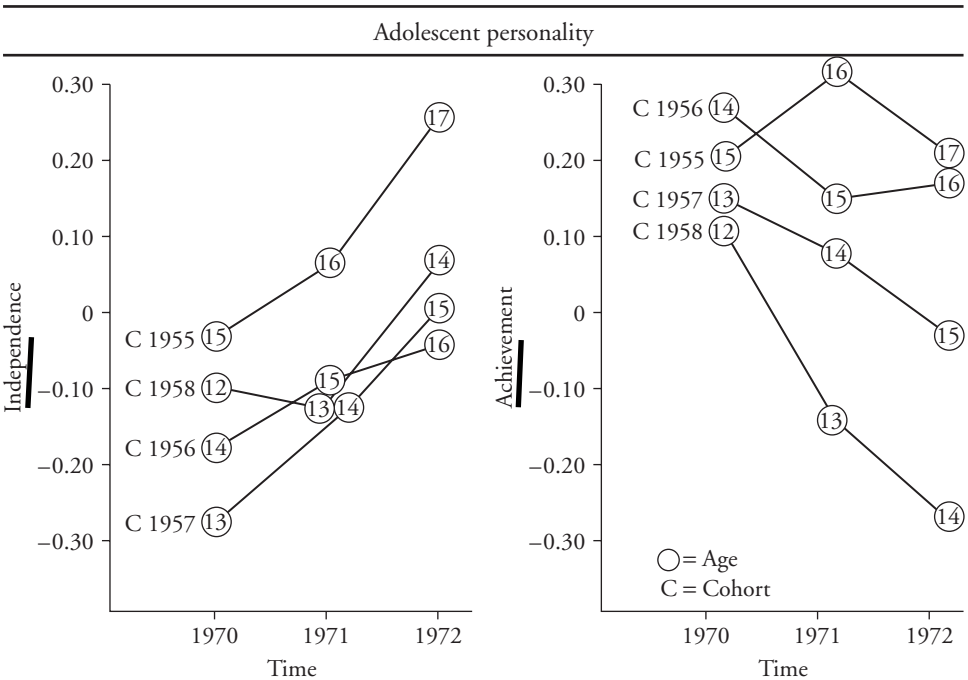


Figure 1.1 Differences in adolescent personality development as a function of the cohort effect. (From Baltes, Cornelius, & Nesselroade (1979). Copyright 1979 by Academic Press. Reprinted by permission.)

Time as a variable

The contribution of time to the developmental function has perhaps been the least understood, and, like the cohort variable, it has tended to be treated as a confound rather than an integral part of development (Schaie, 1984, 1986). Simply acknowledging the influence of time, which can be defined as a marker for historical events (Kosloski, 1986), does not give it explanatory power; it begs the question of what is the underlying psychological process (Caspi & Bem, 1990). Historical time is an essential parameter of developmental research, according to Schaie (1986), because it provides an important context for development.

Schaie (1986) suggests redefining time of measurement in terms of the impact of events on life-span development, which would separate the variable from calendar time. When attempting to determine what influences and processes might be important in terms of historical time, Schaie (1984) suggests looking for "societal changes in technology, customs, and cultural stereotypes that might constrain behavior" (p. 8). The researcher needs to make some conceptualization of historical causation as being distal or proximal as this affects the spacing of observations in longitudinal research (Baltes & Nesselroade, 1979). Finally, it is probably most important for the researcher studying adult/life-span development to hypothesize about the contribution of historical time to the developmental process since it is most likely influential in adulthood (Schaie, 1984, 1986). Indeed, the adult development researcher studying individual differences is really studying cohort and period effects according to Schaie (1986).

Conclusions

Donaldson and Horn (1992, p. 213) noted that "age, cohort, and time constitute a muddle. They are redundant quantities that cannot be independently varied to produce unique contributions to a dependent variable." Psychologists, they argue, tend to ignore cohort and time variables because they are the domains of other disciplines and assert that psychologists alone will not be able to construct a general model of development which takes into account the effects of age, cohort, and period. This is a strong call for interdisciplinary scholarship in developmental research. Indeed, the complexities seen in separating out the influences of age, cohort, and time of measurement on human development mandate the creation of complex models of development which will require the expertise of many disciplines (Baltes, Cornelius, & Nesselroade, 1979; Donaldson & Horn, 1992). At the very least, one undertaking a cross-sectional or simple longitudinal study should recognize that the best solution to the problem of separating age, period, and cohort effects is to measure directly those things that the variables index (Kosloski, 1986).

As indicated above, simple cross-sectional and longitudinal designs have been criticized severely over the years. At the same time, we would argue that, despite the flaws inherent in simple developmental designs, they remain high on the list of design choices

Package ‘plm’

February 15, 2013

Version 1.3-1

Date 2012-12-07

Title Linear Models for Panel Data

Author Yves Croissant <yves.croissant@univ-reunion.fr>, Giovanni Millo <Giovanni_Millo@Generali.com>

Maintainer Yves Croissant <yves.croissant@univ-reunion.fr>

Depends R (>= 2.10), stats, bdsmatrix, nlme, Formula (>= 0.2-0), MASS,sandwich, zoo

Suggests lattice, lmtest, car, AER

Description A set of estimators and tests for panel data.

License GPL (>= 2)

URL http://www.R-project.org

Repository CRAN

Repository/R-Forge/Project plm

Repository/R-Forge/Revision 90

Repository/R-Forge/DateTimeStamp 2012-12-12 08:32:32

Date/Publication 2012-12-13 07:44:56

NeedsCompilation no

R topics documented:

Panel Data Econometrics in R: The plm Package

Yves Croissant
Université Lumière Lyon 2

Giovanni Millo
University of Trieste and Generali SpA

Abstract

This introduction to the **plm** package is a slightly modified version of Croissant and Millo (2008), published in the Journal of Statistical Software.

Panel data econometrics is obviously one of the main fields in the profession, but most of the models used are difficult to estimate with R. **plm** is a package for R which intends to make the estimation of linear panel models straightforward. **plm** provides functions to estimate a wide variety of models and to make (robust) inference.

Keywords:˜panel data, covariance matrix estimators, generalized method of moments, R.

1. Introduction

Panel data econometrics is a continuously developing field. The increasing availability of data observed on cross-sections of units (like households, firms, countries etc.) *and* over time has given rise to a number of estimation approaches exploiting this double dimensionality to cope with some of the typical problems associated with economic data, first of all that of unobserved heterogeneity.

Timewise observation of data from different observational units has long been common in other fields of statistics (where they are often termed *longitudinal* data). In the panel data field as well as in others, the econometric approach is nevertheless peculiar with respect to experimental contexts, as it is emphasizing model specification and testing and tackling a number of issues arising from the particular statistical problems associated with economic data.

Thus, while a very comprehensive software framework for (among many other features) maximum likelihood estimation of linear regression models for longitudinal data, packages **nlme** (Pinheiro, Bates, DebRoy, and the˜R Core˜team 2007) and **lme4** (Bates 2007), is available in the R (R Development Core Team 2008) environment and can be used, e.g., for estimation of random effects panel models, its use is not intuitive for a practicing econometrician, and maximum likelihood estimation is only one of the possible approaches to panel data econometrics. Moreover, economic panel datasets often happen to be *unbalanced* (i.e., they have a different number of observations between groups), which case needs some adaptation to the methods and is not compatible with those in **nlme**. Hence the need for a package doing panel data "from the econometrician's viewpoint" and featuring at a minimum the basic techniques econometricians are used to: random and fixed effects estimation of static linear panel data models, variable coefficients models, generalized method of moments estimation of dynamic models; and the basic toolbox of specification and misspecification diagnostics.

```
R> waldtest(re,update(re,.~.-capital),vcov=function(x) vcovHC(x,method="white2",type="HC3"

Wald test

Model 1: inv ~ value + capital
Model 2: inv ~ value
  Res.Df Df  Chisq Pr(>Chisq)
1    197
2    198 -1 87.828  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Moreover, `linear.hypothesis` from package **car** may be used to test for linear restrictions:

```
R> library("car")
R> linear.hypothesis(re, "2*value=capital", vcov.=vcovHC)

Linear hypothesis test

Hypothesis:
2 value - capital = 0

Model 1: restricted model
Model 2: inv ~ value + capital

Note: Coefficient covariance matrix supplied.

  Res.Df Df  Chisq Pr(>Chisq)
1    198
2    197  1 3.4783    0.06218 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A specific `vcovHC` method for `pgmm` objects is also provided which implements the robust covariance matrix proposed by Windmeijer (2005) for generalized method of moments estimators.

7. plm versus nlme/lme4

The models termed *panel* by the econometricians have counterparts in the statistics literature on *mixed* models (or *hierarchical models*, or *models for longitudinal data*), although there are both differences in jargon and more substantial distinctions. This language inconsistency between the two communities, together with the more complicated general structure of statistical models for longitudinal data and the associated notation in the software, is likely to scare some practicing econometricians away from some potentially useful features of the R environment, so it may be useful to provide here a brief reconciliation between the typical

panel data specifications used in econometrics and the general framework used in statistics for mixed models[20].

R is particularly strong on mixed models' estimation, thanks to the long-standing **nlme** package (see Pinheiro *et˜al.* 2007) and the more recent **lme4** package, based on S4 classes (see Bates 2007)[21]. In the following we will refer to the more established **nlme** to give some examples of "econometric" panel models that can be estimated in a likelihood framework, also including some likelihood ratio tests. Some of them are not feasible in **plm** and make a useful complement to the econometric "toolbox" available in R.

7.1. Fundamental differences between the two approaches

Econometrics deal mostly with non-experimental data. Great emphasis is put on specification procedures and misspecification testing. Model specifications tend therefore to be very simple, while great attention is put on the issues of endogeneity of the regressors, dependence structures in the errors and robustness of the estimators under deviations from normality. The preferred approach is often semi- or non-parametric, and heteroskedasticity-consistent techniques are becoming standard practice both in estimation and testing.

For all these reasons, although the maximum likelihood framework is important in testing[22] and sometimes used in estimation as well, panel model estimation in econometrics is mostly accomplished in the generalized least squares framework based on Aitken's Theorem and, when possible, in its special case OLS, which are free from distributional assumptions (although these kick in at the diagnostic testing stage). On the contrary, longitudinal data models in **nlme** and **lme4** are estimated by (restricted or unrestricted) maximum likelihood. While under normality, homoskedasticity and no serial correlation of the errors OLS are also the maximum likelihood estimator, in all the other cases there are important differences.

The econometric GLS approach has closed-form analytical solutions computable by standard linear algebra and, although the latter can sometimes get computationally heavy on the machine, the expressions for the estimators are usually rather simple. ML estimation of longitudinal models, on the contrary, is based on numerical optimization of nonlinear functions without closed-form solutions and is thus dependent on approximations and convergence criteria. For example, the "GLS" functionality in **nlme** is rather different from its "econometric" counterpart. "Feasible GLS" estimation in **plm** is based on a single two-step procedure, in which an inefficient but consistent estimation method (typically OLS) is employed first in order to get a consistent estimate of the errors' covariance matrix, to be used in GLS at the second step; on the converse, "GLS" estimators in **nlme** are based on iteration until convergence of two-step optimization of the relevant likelihood.

[20] This discussion does not consider GMM models. One of the basic reasons for econometricians not to choose maximum likelihood methods in estimation is that the strict exogeneity of regressors assumption required for consistency of the ML models reported in the following is often inappropriate in economic settings.

[21] The standard reference on the subject of mixed models in S/R is Pinheiro and Bates (2000).

[22] Lagrange Multiplier tests based on the likelihood principle are suitable for testing against more general alternatives on the basis of a maintained model with spherical residuals and find therefore application in testing for departures from the classical hypotheses on the error term. The seminal reference is Breusch and Pagan (1980).

7.2. Some false friends

The *fixed/random effects* terminology in econometrics is often recognized to be misleading, as both are treated as random variates in modern econometrics (see e.g. Wooldridge 2002, 10.2.1). It has been recognized since Mundlak's classic paper (Mundlak 1978) that the fundamental issue is whether the unobserved effects are correlated with the regressors or not. In this last case, they can safely be left in the error term, and the serial correlation they induce is cared for by means of appropriate GLS transformations. On the contrary, in the case of correlation, "fixed effects" methods such as least squares dummy variables or time-demeaning are needed, which explicitly, although inconsistently[23], estimate a group– (or time–) invariant additional parameter for each group (or time period).

Thus, from the point of view of model specification, having *fixed effects* in an econometric model has the meaning of allowing the intercept to vary with group, or time, or both, while the other parameters are generally still assumed to be homogeneous. Having *random effects* means having a group– (or time–, or both) specific component in the error term.

In the mixed models literature, on the contrary, *fixed effect* indicates a parameter that is assumed constant, while *random effects* are parameters that vary randomly around zero according to a joint multivariate Normal distribution.

So, the FE model in econometrics has no counterpart in the mixed models framework, unless reducing it to OLS on a specification with one dummy for each group (often termed *least squares dummy variables*, or LSDV model) which can trivially be estimated by OLS. The RE model is instead a special case of mixed model where only the intercept is specified as a random effect, while the "random" type variable coefficients model can be seen as one that has the same regressors in the fixed and random sets. The unrestricted generalized least squares can in turn be seen, in the **nlme** framework, as a standard linear model with a general error covariance structure within the groups and errors uncorrelated across groups.

7.3. A common taxonomy

To reconcile the two terminologies, in the following we report the specification of the panel models in **plm** according to the general expression of a mixed model in Laird-Ware form (see the web appendix to Fox 2002) and the **nlme** estimation commands for maximum likelihood estimation of an equivalent specification[24].

The Laird-Ware representation for mixed models

A general representation for the linear mixed effects model is given in Laird and Ware (1982).

[23] For fixed effects estimation, as the sample grows (on the dimension on which the fixed effects are specified) so does the number of parameters to be estimated. Estimation of individual fixed effects is T– (but not n–) consistent, and the opposite.

[24] In doing so, we stress that "equivalence" concerns only the specification of the model, and neither the appropriateness nor the relative efficiency of the relevant estimation techniques, which will of course be dependent on the context. Unlike their mixed model counterparts, the specifications in **plm** are, strictly speaking, distribution-free. Nevertheless, for the sake of exposition, in the following we present them in the setting which ensures consistency and efficiency (e.g., we consider the hypothesis of spherical errors part of the specification of pooled OLS and so forth).

$$\begin{array}{rcl} y_{it} & = & \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} \\ & & b_1 z_{1ij} + \ldots + b_p z_{pij} + \epsilon_{ij} \\ b_{ik} & \sim & N(0, \psi_k^2), \;\; Cov(b_k, b_{k'}) = \psi_{kk'} \\ \epsilon_{ij} & \sim & N(0, \sigma^2 \lambda_{ijj}), \;\; Cov(\epsilon_{ij}, \epsilon_{ij'}) = \sigma^2 \lambda_{ijj'} \end{array}$$

where the $x_1, \ldots x_p$ are the fixed effects regressors and the $z_1, \ldots z_p$ are the random effects regressors, assumed to be normally distributed across groups. The covariance of the random effects coefficients $\psi_{kk'}$ is assumed constant across groups and the covariances between the errors in group i, $\sigma^2 \lambda_{ijj'}$, are described by the term $\lambda_{ijj'}$ representing the correlation structure of the errors within each group (e.g., serial correlation over time) scaled by the common error variance σ^2.

Pooling and Within

The *pooling* specification in **plm** is equivalent to a classical linear model (i.e., no random effects regressor and spherical errors: $b_{iq} = 0 \;\; \forall i, q, \;\; \lambda_{ijj} = \sigma^2$ for $j = j'$, 0 else). The *within* one is the same with the regressors' set augmented by $n - 1$ group dummies. There is no point in using **nlme** as parameters can be estimated by OLS which is also ML.

Random effects

In the Laird and Ware notation, the RE specification is a model with only one random effects regressor: the intercept. Formally, $z_{1ij} = 1 \;\; \forall i, j, \;\; z_{qij} = 0 \;\; \forall i, \forall j, \forall q \neq 1 \;\; \lambda_{ij} = 1$ for $i = j$, 0 else). The composite error is therefore $u_{ij} = 1 b_{i1} + \epsilon_{ij}$. Below we report coefficients of Grunfeld's model estimated by GLS and then by ML

```
R> reGLS<-plm(inv~value+capital,data=Grunfeld,model="random")
R> reML<-lme(inv~value+capital,data=Grunfeld,random=~1|firm)
R> coef(reGLS)

(Intercept)       value     capital
-57.8344149   0.1097812   0.3081130

R> summary(reML)$coef$fixed

(Intercept)       value     capital
-57.8644245   0.1097897   0.3081881

R>
```

Variable coefficients, "random"

Swamy's variable coefficients model (Swamy 1970) has coefficients varying randomly (and independently of each other) around a set of fixed values, so the equivalent specification is $z_q = x_q \;\; \forall q$, i.e. the fixed effects and the random effects regressors are the same, and $\psi_{kk'} = \sigma_\mu^2 I_N$, and $\lambda_{ijj} = 1$, $\lambda_{ijj'} = 0$ for $j \neq j'$, that's to say they are not correlated.

Estimation of a mixed model with random coefficients on all regressors is rather demanding from the computational side. Some models from our examples fail to converge. The below example is estimated on the Grunfeld data and model with time effects.

```
R> vcm<-pvcm(inv~value+capital,data=Grunfeld,model="random",effect="time")

[1]  6.318535e-04 -2.453520e-02 -1.410394e+03
attention

R> vcmML<-lme(inv~value+capital,data=Grunfeld,random=~value+capital|year)
R> coef(vcm)

                      y
(Intercept) -18.5538638
value         0.1239595
capital       0.1114579

R> summary(vcmML)$coef$fixed

(Intercept)       value     capital
-26.3558395   0.1241982   0.1381782

R>
```

Variable coefficients, "within"

This specification actually entails separate estimation of T different standard linear models, one for each group in the data, so the estimation approach is the same: OLS. In **nlme** this is done by creating an `lmList` object, so that the two models below are equivalent (output suppressed):

```
R> vcmf<-pvcm(inv~value+capital,data=Grunfeld,model="within",effect="time")
R> vcmfML<-lmList(inv~value+capital|year,data=Grunfeld)
R>
```

Unrestricted fgls

The general, or unrestricted, feasible GLS, `pggls` in the **plm** nomenclature, is equivalent to a model with no random effects regressors ($b_{iq} = 0 \ \forall i, q$) and an error covariance structure which is unrestricted within groups apart from the usual requirements. The function for estimating such models with correlation in the errors but no random effects is `gls()`.

This very general serial correlation and heteroskedasticity structure is not estimable for the original Grunfeld data, which have more time periods than firms, therefore we restrict them to firms 4 to 6.

`Grunfeld` *Grunfeld's Investment Data*

Description

A panel of 10 observations from 1935 to 1954

total number of observations : 200

observation : production units

country : United States

Usage

```
data(Grunfeld)
```

Format

A data frame containing :

firm observation

year date

inv gross Investment

value value of the firm

capital stock of plant and equipment

Source

Online complements to Baltagi (2001).

http://www.wiley.com/legacy/wileychi/baltagi/.

References

Baltagi, Badi H. (2001) *Econometric Analysis of Panel Data*, 2nd ed., John Wiley and Sons.

See Also

For the complete Grunfeld data (11 firms), see `Grunfeld`, in the `AER` package.

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
> install.packages("plm")
> library(plm)  Loading required package: nlme ...
> data(Grunfeld)
> head(Grunfeld)
firm year   inv  value capital |> summary(Grunfeld)
   1 1935 317.6 3078.5     2.8 |      firm          year          inv              value           capital
   1 1936 391.8 4661.7    52.6 | Min.   : 1.0  Min.   :1935  Min.   :   0.93  Min.   :  58.12  Min.   :   0.80
   1 1937 410.6 5387.1   156.9 | 1st Qu.: 3.0  1st Qu.:1940  1st Qu.:  33.56  1st Qu.: 199.97  1st Qu.:  79.17
   1 1938 257.7 2792.2   209.2 | Median : 5.5  Median :1944  Median :  57.48  Median : 517.95  Median : 205.60
   1 1939 330.8 4313.2   203.4 | Mean   : 5.5  Mean   :1944  Mean   : 145.96  Mean   :1081.68  Mean   : 276.02
   1 1940 461.2 4643.9   207.2 | 3rd Qu.: 8.0  3rd Qu.:1949  3rd Qu.: 138.04  3rd Qu.:1679.85  3rd Qu.: 358.10
                               | Max.   :10.0  Max.   :1954  Max.   :1486.70  Max.   :6241.70  Max.   :2226.30

> vcmfML<-lmList(inv~value+capital|year,data=Grunfeld) #Variable coeffcients, "within" plm p.44
> coef(vcmfML)
      (Intercept)       value      capital       value Pr(>|t|)  capital Pr(>|t|)
1935    0.3560335 0.10249786 -0.001994772     2.821848e-03    0.996536081
1936   15.2189424 0.08370736 -0.053641246     1.785969e-04    0.899287732
1937   -3.3864702 0.07651380  0.217722370     4.702279e-05    0.582090879
1938  -17.5819018 0.06801777  0.269114634     4.788130e-02    0.404080267
1939  -21.1542290 0.06552194  0.198664570     6.564605e-03    0.482083718
1940  -27.0470687 0.09539899  0.202290565     2.328544e-05    0.472785960
1941  -16.5194901 0.11476375  0.177465020     1.212807e-06    0.509182484
1942  -17.6182810 0.14282514  0.071024035     9.498889e-06    0.774527355
1943  -22.7637943 0.11860950  0.105411928     6.204563e-06    0.647496743
1944  -15.8281383 0.11816421  0.072207166     1.228336e-06    0.732479953
1945  -10.5196744 0.10847089  0.050220832     9.763134e-07    0.801500261
1946   -5.9906590 0.13794818  0.005413393     1.668821e-09    0.976521034
1947   -3.7324861 0.16392695 -0.003707209     1.640302e-05    0.982440036
1948    8.5388116 0.17866729 -0.042555528     5.598809e-05    0.782630549
1949    5.1782822 0.16159618 -0.036965104     1.496333e-04    0.807913095
1950  -12.1746788 0.17621675 -0.022095639     2.972162e-05    0.876016775
1951   26.1381617 0.18314051 -0.112056960     7.992691e-08    0.404031592
1952    7.2928451 0.19892081 -0.067495013     2.897338e-07    0.616744430
1953  -50.1525452 0.18267386  0.098753342     1.065592e-06    0.456473108
1954 -133.3930999 0.13451161  0.331374645     1.211699e-03    0.003541993
> quantile(coef(vcmfML)[,2])
        0%        25%        50%        75%       100%
0.06552194 0.10072314 0.12656056 0.16699940 0.19892081
> quantile(coef(vcmfML)[,3])
         0%         25%         50%         75%        100%
-0.11205696 -0.02581301  0.06062243  0.18276491  0.33137464

> ?pvcm #pvcm {plm}R Documentation Variable Coefficients Models for Panel Data
> vcmf<-pvcm(inv~value+capital,data=Grunfeld,model="within",effect="time")
> summary(vcmf)
Oneway (time) effect No-pooling model
Call: pvcm(formula = inv ~ value + capital, data = Grunfeld, effect = "time",  model = "within")
Balanced Panel: n=10, T=20, N=200
Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-342.800  -20.100    0.727    0.000   23.650  258.500
Coefficients:
  (Intercept)          value             capital
 Min.   :-133.393  Min.   :0.06552   Min.   :-0.11206
 1st Qu.: -18.502  1st Qu.:0.10072   1st Qu.:-0.02581
 Median : -11.347  Median :0.12656   Median : 0.06062
 Mean   : -14.757  Mean   :0.13060   Mean   : 0.07296
 3rd Qu.:   1.562  3rd Qu.:0.16700   3rd Qu.: 0.18276
 Max.   :  26.138  Max.   :0.19892   Max.   : 0.33137

Total Sum of Squares: 474010000 Residual Sum of Squares: 1205800 Multiple R-Squared: 0.99746

##### Random Effects, plm p.43  ###########
> reGLS<-plm(inv~value+capital,data=Grunfeld,model="random")
>  reML<-lme(inv~value+capital,data=Grunfeld,random=~1|firm)
>  coef(reGLS)
(Intercept)       value     capital
-57.8344149   0.1097812   0.3081130
> summary(reML)$coef$fixed
(Intercept)       value     capital
-57.8644245   0.1097897   0.3081881
```

Review Q, redo as lmer